aws

Bedrock API Reference

Amazon Bedrock



Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon Bedrock: Bedrock API Reference

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Services or capabilities described in AWS documentation might vary by Region. To see the differences applicable to the AWS European Sovereign Cloud Region, see the AWS European Sovereign Cloud User Guide.

Table of Contents

Welcome 1
Amazon Bedrock 1
Agents for Amazon Bedrock 1
Agents for Amazon Bedrock Runtime 1
Data Automation for Amazon Bedrock 1
Runtime for Amazon Bedrock Data Automation 2
Amazon Bedrock Runtime 2
Actions 3
Amazon Bedrock 12
BatchDeleteAdvancedPromptOptimizationJob 17
BatchDeleteEvaluationJob 20
CancelAutomatedReasoningPolicyBuildWorkflow 24
CreateAdvancedPromptOptimizationJob 27
CreateAutomatedReasoningPolicy 33
CreateAutomatedReasoningPolicyTestCase 40
CreateAutomatedReasoningPolicyVersion 46
CreateCustomModel 52
CreateCustomModelDeployment 60
CreateEvaluationJob 66
CreateFoundationModelAgreement 72
CreateGuardrail 76
CreateGuardrailVersion 86
CreateInferenceProfile 90
CreateMarketplaceModelEndpoint 96
CreateModelCopyJob 101
CreateModelCustomizationJob 106
CreateModelImportJob 114
CreateModelInvocationJob 121
CreatePromptRouter 128
CreateProvisionedModelThroughput 133
DeleteAutomatedReasoningPolicy 139
DeleteAutomatedReasoningPolicyBuildWorkflow 142
DeleteAutomatedReasoningPolicyTestCase 145
DeleteCustomModel 148

DeleteCustomModelDeployment 151
DeleteEnforcedGuardrailConfiguration 154
DeleteFoundationModelAgreement 157
DeleteGuardrail 160
DeleteImportedModel 163
DeleteInferenceProfile 166
DeleteMarketplaceModelEndpoint 169
DeleteModelInvocationLoggingConfiguration 172
DeletePromptRouter 174
DeleteProvisionedModelThroughput 177
DeleteResourcePolicy 180
DeregisterMarketplaceModelEndpoint 183
ExportAutomatedReasoningPolicyVersion 186
GetAccountDataRetention 190
GetAdvancedPromptOptimizationJob 193
GetAutomatedReasoningPolicy 198
GetAutomatedReasoningPolicyAnnotations 203
GetAutomatedReasoningPolicyBuildWorkflow 208
GetAutomatedReasoningPolicyBuildWorkflowResultAssets 213
GetAutomatedReasoningPolicyNextScenario 217
GetAutomatedReasoningPolicyTestCase 221
GetAutomatedReasoningPolicyTestResult 225
GetCustomModel 229
GetCustomModelDeployment 236
GetEvaluationJob 241
GetFoundationModel 247
GetFoundationModelAvailability 250
GetGuardrail 254
GetImportedModel 263
GetInferenceProfile 269
GetMarketplaceModelEndpoint 274
GetModelCopyJob 277
GetModelCustomizationJob 283
GetModelImportJob 291
GetModelInvocationJob 297
GetModelInvocationLoggingConfiguration 305

GetPromptRouter 308
GetProvisionedModelThroughput 313
GetResourcePolicy 319
GetUseCaseForModelAccess 322
ListAdvancedPromptOptimizationJobs 325
ListAutomatedReasoningPolicies 328
ListAutomatedReasoningPolicyBuildWorkflows 332
ListAutomatedReasoningPolicyTestCases 336
ListAutomatedReasoningPolicyTestResults 340
ListCustomModelDeployments 344
ListCustomModels 349
ListEnforcedGuardrailsConfiguration 354
ListEvaluationJobs 358
ListFoundationModelAgreementOffers 363
ListFoundationModels 367
ListGuardrails 371
ListImportedModels 375
ListInferenceProfiles 380
ListMarketplaceModelEndpoints 384
ListModelCopyJobs 388
ListModelCustomizationJobs 393
ListModelImportJobs 398
ListModelInvocationJobs 403
ListPromptRouters 409
ListProvisionedModelThroughputs 413
ListTagsForResource 418
PutAccountDataRetention 422
PutEnforcedGuardrailConfiguration 425
PutModelInvocationLoggingConfiguration 429
PutResourcePolicy 432
PutUseCaseForModelAccess 435
RegisterMarketplaceModelEndpoint 438
StartAutomatedReasoningPolicyBuildWorkflow 442
StartAutomatedReasoningPolicyTestWorkflow 448
StopAdvancedPromptOptimizationJob 452
StopEvaluationJob 455

StopModelCustomizationJob ... 458
StopModelInvocationJob ... 461
TagResource ... 464
UntagResource ... 468
UpdateAutomatedReasoningPolicy ... 471
UpdateAutomatedReasoningPolicyAnnotations ... 477
UpdateAutomatedReasoningPolicyTestCase ... 482
UpdateCustomModelDeployment ... 487
UpdateGuardrail ... 491
UpdateMarketplaceModelEndpoint ... 500
UpdateProvisionedModelThroughput ... 504
Agents for Amazon Bedrock ... 507
AssociateAgentCollaborator ... 510
AssociateAgentKnowledgeBase ... 515
CreateAgent ... 519
CreateAgentActionGroup ... 530
CreateAgentAlias ... 539
CreateDataSource ... 545
CreateFlow ... 559
CreateFlowAlias ... 568
CreateFlowVersion ... 575
CreateKnowledgeBase ... 581
CreatePrompt ... 598
CreatePromptVersion ... 606
CreateVpcConfiguration ... 613
DeleteAgent ... 620
DeleteAgentActionGroup ... 624
DeleteAgentAlias ... 627
DeleteAgentVersion ... 631
DeleteDataSource ... 635
DeleteFlow ... 639
DeleteFlowAlias ... 642
DeleteFlowVersion ... 646
DeleteKnowledgeBase ... 650
DeleteKnowledgeBaseDocuments ... 653
DeletePrompt ... 658

DeleteResourcePolicy 662
DeleteVpcConfiguration 666
DisassociateAgentCollaborator 670
DisassociateAgentKnowledgeBase 673
GetAgent 676
GetAgentActionGroup 680
GetAgentAlias 684
GetAgentCollaborator 688
GetAgentKnowledgeBase 692
GetAgentVersion 696
GetDataSource 700
GetFlow 707
GetFlowAlias 713
GetFlowVersion 718
GetIngestionJob 724
GetKnowledgeBase 728
GetKnowledgeBaseDocuments 736
GetPrompt 740
GetResourcePolicy 746
GetVpcConfiguration 750
IngestKnowledgeBaseDocuments 754
ListAgentActionGroups 760
ListAgentAliases 765
ListAgentCollaborators 770
ListAgentKnowledgeBases 774
ListAgents 778
ListAgentVersions 782
ListDataSources 787
ListFlowAliases 791
ListFlows 795
ListFlowVersions 799
ListIngestionJobs 803
ListKnowledgeBaseDocuments 808
ListKnowledgeBases 813
ListPrompts 817
ListTagsForResource 821

ListVpcConfigurations ... 825
PrepareAgent ... 829
PrepareFlow ... 833
PutResourcePolicy ... 837
StartIngestionJob ... 841
StopIngestionJob ... 846
TagResource ... 850
UntagResource ... 854
UpdateAgent ... 857
UpdateAgentActionGroup ... 866
UpdateAgentAlias ... 874
UpdateAgentCollaborator ... 880
UpdateAgentKnowledgeBase ... 885
UpdateDataSource ... 889
UpdateFlow ... 902
UpdateFlowAlias ... 910
UpdateKnowledgeBase ... 916
UpdatePrompt ... 931
ValidateFlowDefinition ... 938
Agents for Amazon Bedrock Runtime ... 941
AgenticRetrieveStream ... 943
CheckIngestedDocumentAcl ... 955
CreateInvocation ... 960
CreateSession ... 965
DeleteAgentMemory ... 971
DeleteSession ... 975
EndSession ... 978
GenerateQuery ... 982
GetAgentMemory ... 987
GetDocumentContent ... 992
GetExecutionFlowSnapshot ... 997
GetFlowExecution ... 1002
GetIngestedDocumentAcl ... 1007
GetInvocationStep ... 1012
GetSession ... 1016
InvokeAgent ... 1021

InvokeFlow 1041
InvokeInlineAgent 1050
ListFlowExecutionEvents 1071
ListFlowExecutions 1076
ListInvocations 1080
ListInvocationSteps 1084
ListSessions 1089
ListTagsForResource 1093
OptimizePrompt 1096
PutInvocationStep 1102
Rerank 1107
Retrieve 1113
RetrieveAndGenerate 1123
RetrieveAndGenerateStream 1136
StartFlowExecution 1151
StopFlowExecution 1157
TagResource 1162
UntagResource 1166
UpdateSession 1169
Data Automation for Amazon Bedrock 1173
CopyBlueprintStage 1175
CreateBlueprint 1178
CreateBlueprintVersion 1183
CreateDataAutomationLibrary 1187
CreateDataAutomationProject 1191
DeleteBlueprint 1201
DeleteDataAutomationLibrary 1204
DeleteDataAutomationProject 1207
GetBlueprint 1210
GetBlueprintOptimizationStatus 1214
GetDataAutomationLibrary 1217
GetDataAutomationLibraryEntity 1220
GetDataAutomationLibraryIngestionJob 1223
GetDataAutomationProject 1226
InvokeBlueprintOptimizationAsync 1233
InvokeDataAutomationLibraryIngestionJob 1238

ListBlueprints ... 1243
ListDataAutomationLibraries ... 1248
ListDataAutomationLibraryEntities ... 1252
ListDataAutomationLibraryIngestionJobs ... 1256
ListDataAutomationProjects ... 1260
ListTagsForResource ... 1265
TagResource ... 1268
UntagResource ... 1271
UpdateBlueprint ... 1274
UpdateDataAutomationLibrary ... 1278
UpdateDataAutomationProject ... 1282
Runtime for Amazon Bedrock Data Automation ... 1290
GetDataAutomationStatus ... 1291
InvokeDataAutomation ... 1294
InvokeDataAutomationAsync ... 1299
ListTagsForResource ... 1304
TagResource ... 1307
UntagResource ... 1310
Amazon Bedrock Runtime ... 1312
ApplyGuardrail ... 1313
Converse ... 1321
ConverseStream ... 1341
CountTokens ... 1360
GetAsyncInvoke ... 1364
InvokeGuardrailChecks ... 1368
InvokeModel ... 1373
InvokeModelWithBidirectionalStream ... 1383
InvokeModelWithResponseStream ... 1390
ListAsyncInvokes ... 1401
StartAsyncInvoke ... 1405
Data Types ... 1410
Amazon Bedrock ... 1445
AccountEnforcedGuardrailInferenceInputConfiguration ... 1455
AccountEnforcedGuardrailOutputConfiguration ... 1457
AdvancedPromptOptimizationInputConfig ... 1460
AdvancedPromptOptimizationJobSummary ... 1461

AdvancedPromptOptimizationOutputConfig 1463
AgreementAvailability 1464
AutomatedEvaluationConfig 1465
AutomatedEvaluationCustomMetricConfig 1467
AutomatedEvaluationCustomMetricSource 1468
AutomatedReasoningCheckFinding 1469
AutomatedReasoningCheckImpossibleFinding 1472
AutomatedReasoningCheckInputTextReference 1473
AutomatedReasoningCheckInvalidFinding 1474
AutomatedReasoningCheckLogicWarning 1475
AutomatedReasoningCheckNoTranslationsFinding 1477
AutomatedReasoningCheckRule 1478
AutomatedReasoningCheckSatisfiableFinding 1479
AutomatedReasoningCheckScenario 1481
AutomatedReasoningCheckTooComplexFinding 1482
AutomatedReasoningCheckTranslation 1483
AutomatedReasoningCheckTranslationAmbiguousFinding 1485
AutomatedReasoningCheckTranslationOption 1486
AutomatedReasoningCheckValidFinding 1487
AutomatedReasoningLogicStatement 1489
AutomatedReasoningPolicyAddRuleAnnotation 1490
AutomatedReasoningPolicyAddRuleFromNaturalLanguageAnnotation 1491
AutomatedReasoningPolicyAddRuleMutation 1492
AutomatedReasoningPolicyAddTypeAnnotation 1493
AutomatedReasoningPolicyAddTypeMutation 1495
AutomatedReasoningPolicyAddTypeValue 1496
AutomatedReasoningPolicyAddVariableAnnotation 1497
AutomatedReasoningPolicyAddVariableMutation 1499
AutomatedReasoningPolicyAnnotatedChunk 1500
AutomatedReasoningPolicyAnnotatedContent 1501
AutomatedReasoningPolicyAnnotatedLine 1502
AutomatedReasoningPolicyAnnotation 1503
AutomatedReasoningPolicyAtomicStatement 1507
AutomatedReasoningPolicyBuildLog 1509
AutomatedReasoningPolicyBuildLogEntry 1510
AutomatedReasoningPolicyBuildResultAssetManifest 1512

AutomatedReasoningPolicyBuildResultAssetManifestEntry 1513
AutomatedReasoningPolicyBuildResultAssets 1515
AutomatedReasoningPolicyBuildStep 1518
AutomatedReasoningPolicyBuildStepContext 1520
AutomatedReasoningPolicyBuildStepMessage 1522
AutomatedReasoningPolicyBuildWorkflowDocument 1523
AutomatedReasoningPolicyBuildWorkflowRepairContent 1525
AutomatedReasoningPolicyBuildWorkflowSource 1526
AutomatedReasoningPolicyBuildWorkflowSummary 1527
AutomatedReasoningPolicyDefinition 1529
AutomatedReasoningPolicyDefinitionElement 1531
AutomatedReasoningPolicyDefinitionQualityReport 1533
AutomatedReasoningPolicyDefinitionRule 1536
AutomatedReasoningPolicyDefinitionType 1538
AutomatedReasoningPolicyDefinitionTypeValue 1540
AutomatedReasoningPolicyDefinitionTypeValuePair 1541
AutomatedReasoningPolicyDefinitionVariable 1542
AutomatedReasoningPolicyDeleteRuleAnnotation 1544
AutomatedReasoningPolicyDeleteRuleMutation 1545
AutomatedReasoningPolicyDeleteTypeAnnotation 1546
AutomatedReasoningPolicyDeleteTypeMutation 1547
AutomatedReasoningPolicyDeleteTypeValue 1548
AutomatedReasoningPolicyDeleteVariableAnnotation 1549
AutomatedReasoningPolicyDeleteVariableMutation 1550
AutomatedReasoningPolicyDisjointRuleSet 1551
AutomatedReasoningPolicyFidelityReport 1553
AutomatedReasoningPolicyGeneratedTestCase 1555
AutomatedReasoningPolicyGeneratedTestCases 1557
AutomatedReasoningPolicyGenerateFidelityReportContent 1558
AutomatedReasoningPolicyIngestContentAnnotation 1559
AutomatedReasoningPolicyIterativeRefinementContent 1560
AutomatedReasoningPolicyMutation 1561
AutomatedReasoningPolicyPlanning 1564
AutomatedReasoningPolicyReportSourceDocument 1565
AutomatedReasoningPolicyRuleReport 1567
AutomatedReasoningPolicyScenario 1569

AutomatedReasoningPolicyScenarios 1571
AutomatedReasoningPolicySourceDocument 1572
AutomatedReasoningPolicyStatementLocation 1574
AutomatedReasoningPolicyStatementReference 1575
AutomatedReasoningPolicySummary 1576
AutomatedReasoningPolicyTestCase 1579
AutomatedReasoningPolicyTestResult 1581
AutomatedReasoningPolicyTypeValueAnnotation 1584
AutomatedReasoningPolicyUpdateFromRuleFeedbackAnnotation 1586
AutomatedReasoningPolicyUpdateFromScenarioFeedbackAnnotation 1588
AutomatedReasoningPolicyUpdateRuleAnnotation 1590
AutomatedReasoningPolicyUpdateRuleMutation 1591
AutomatedReasoningPolicyUpdateTypeAnnotation 1592
AutomatedReasoningPolicyUpdateTypeMutation 1594
AutomatedReasoningPolicyUpdateTypeValue 1595
AutomatedReasoningPolicyUpdateVariableAnnotation 1597
AutomatedReasoningPolicyUpdateVariableMutation 1599
AutomatedReasoningPolicyVariableReport 1600
AutomatedReasoningPolicyWorkflowTypeContent 1602
BatchDeleteAdvancedPromptOptimizationJobError 1604
BatchDeleteAdvancedPromptOptimizationJobItem 1606
BatchDeleteEvaluationJobError 1607
BatchDeleteEvaluationJobItem 1609
BedrockEvaluatorModel 1610
ByteContentDoc 1611
CloudWatchConfig 1613
CustomizationConfig 1615
CustomMetricBedrockEvaluatorModel 1616
CustomMetricDefinition 1617
CustomMetricEvaluatorModelConfig 1619
CustomModelDataSource 1620
CustomModelDeploymentSummary 1621
CustomModelDeploymentUpdateDetails 1624
CustomModelSummary 1626
CustomModelUnits 1629
DataProcessingDetails 1631

DimensionalPriceRate 1633
DistillationConfig 1635
EndpointConfig 1636
EvaluationBedrockModel 1637
EvaluationConfig 1639
EvaluationDataset 1640
EvaluationDatasetLocation 1642
EvaluationDatasetMetricConfig 1643
EvaluationInferenceConfig 1645
EvaluationInferenceConfigSummary 1647
EvaluationModelConfig 1648
EvaluationModelConfigSummary 1649
EvaluationOutputDataConfig 1651
EvaluationPrecomputedInferenceSource 1652
EvaluationPrecomputedRagSourceConfig 1653
EvaluationPrecomputedRetrieveAndGenerateSourceConfig 1655
EvaluationPrecomputedRetrieveSourceConfig 1656
EvaluationRagConfigSummary 1657
EvaluationSummary 1659
EvaluatorModelConfig 1664
ExternalSource 1665
ExternalSourcesGenerationConfiguration 1666
ExternalSourcesRetrieveAndGenerateConfiguration 1668
FieldForReranking 1670
FilterAttribute 1671
FoundationModelDetails 1672
FoundationModelLifecycle 1675
FoundationModelSummary 1677
GenerationConfiguration 1680
GraderConfig 1682
GuardrailAutomatedReasoningPolicy 1683
GuardrailAutomatedReasoningPolicyConfig 1685
GuardrailConfiguration 1687
GuardrailContentFilter 1688
GuardrailContentFilterConfig 1692
GuardrailContentFiltersTier 1696

GuardrailContentFiltersTierConfig 1697
GuardrailContentPolicy 1698
GuardrailContentPolicyConfig 1699
GuardrailContextualGroundingFilter 1700
GuardrailContextualGroundingFilterConfig 1702
GuardrailContextualGroundingPolicy 1704
GuardrailContextualGroundingPolicyConfig 1705
GuardrailCrossRegionConfig 1706
GuardrailCrossRegionDetails 1707
GuardrailManagedWords 1709
GuardrailManagedWordsConfig 1711
GuardrailPiiEntity 1713
GuardrailPiiEntityConfig 1716
GuardrailRegex 1723
GuardrailRegexConfig 1726
GuardrailSensitiveInformationPolicy 1729
GuardrailSensitiveInformationPolicyConfig 1730
GuardrailSummary 1731
GuardrailTopic 1734
GuardrailTopicConfig 1737
GuardrailTopicPolicy 1740
GuardrailTopicPolicyConfig 1741
GuardrailTopicsTier 1742
GuardrailTopicsTierConfig 1743
GuardrailWord 1744
GuardrailWordConfig 1746
GuardrailWordPolicy 1748
GuardrailWordPolicyConfig 1749
HumanEvaluationConfig 1750
HumanEvaluationCustomMetric 1752
HumanWorkflowConfig 1754
ImplicitFilterConfiguration 1756
ImportedModelSummary 1758
InferenceConfiguration 1760
InferenceProfileModel 1762
InferenceProfileModelSource 1763

InferenceProfileSummary 1764
InvocationLogsConfig 1767
InvocationLogSource 1769
KbInferenceConfig 1770
KnowledgeBaseConfig 1771
KnowledgeBaseRetrievalConfiguration 1772
KnowledgeBaseRetrieveAndGenerateConfiguration 1773
KnowledgeBaseVectorSearchConfiguration 1775
LambdaGraderConfig 1777
LegalTerm 1778
LoggingConfig 1779
MarketplaceModelEndpoint 1781
MarketplaceModelEndpointSummary 1784
MetadataAttributeSchema 1786
MetadataConfigurationForReranking 1788
ModelConfiguration 1789
ModelCopyJobSummary 1791
ModelCustomizationJobSummary 1795
ModelDataSource 1799
ModelEnforcement 1800
ModelImportJobSummary 1802
ModelInvocationJobInputDataConfig 1805
ModelInvocationJobOutputDataConfig 1806
ModelInvocationJobS3InputDataConfig 1807
ModelInvocationJobS3OutputDataConfig 1809
ModelInvocationJobSummary 1811
ModelPackageArnDataSource 1817
Offer 1818
OrchestrationConfiguration 1819
OutputDataConfig 1820
PerformanceConfiguration 1821
PricingTerm 1822
PromptRouterSummary 1823
PromptRouterTargetModel 1826
PromptTemplate 1827
ProvisionedModelSummary 1828

QueryTransformationConfiguration 1832
RAGConfig 1833
RatingScaleItem 1835
RatingScaleItemValue 1836
RequestMetadataBaseFilters 1837
RequestMetadataFilters 1839
RerankingMetadataSelectiveModeConfiguration 1841
RetrievalFilter 1843
RetrieveAndGenerateConfiguration 1848
RetrieveConfig 1850
RFTConfig 1851
RFTHyperParameters 1852
RoutingCriteria 1855
S3Config 1856
S3DataSource 1857
S3ObjectDoc 1858
SageMakerEndpoint 1859
SelectiveContentGuarding 1861
StatusDetails 1862
SupportTerm 1864
Tag 1865
TeacherModelConfig 1866
TermDetails 1867
TextInferenceConfig 1869
TrainingDataConfig 1871
TrainingDetails 1872
TrainingMetrics 1874
ValidationDataConfig 1875
ValidationDetails 1876
Validator 1878
ValidatorMetric 1879
ValidityTerm 1880
VectorSearchBedrockRerankingConfiguration 1881
VectorSearchBedrockRerankingModelConfiguration 1883
VectorSearchRerankingConfiguration 1884
VpcConfig 1885

Agents for Amazon Bedrock 1886
ActionGroupExecutor 1895
ActionGroupSummary 1897
Agent 1899
AgentActionGroup 1905
AgentAlias 1910
AgentAliasHistoryEvent 1914
AgentAliasRoutingConfigurationListItem 1916
AgentAliasSummary 1917
AgentCollaborator 1920
AgentCollaboratorSummary 1923
AgentDescriptor 1926
AgentFlowNodeConfiguration 1927
AgentKnowledgeBase 1928
AgentKnowledgeBaseSummary 1930
AgentSummary 1932
AgentVersion 1934
AgentVersionSummary 1939
AnyToolChoice 1941
APISchema 1942
AudioConfiguration 1944
AudioExtractionConfiguration 1945
AudioSegmentationConfiguration 1946
AutoToolChoice 1947
BedrockDataAutomationConfiguration 1948
BedrockEmbeddingModelConfiguration 1949
BedrockFoundationModelConfiguration 1952
BedrockFoundationModelContextEnrichmentConfiguration 1954
ByteContentDoc 1956
CachePointBlock 1958
ChatPromptTemplateConfiguration 1959
ChunkingConfiguration 1961
CollectorFlowNodeConfiguration 1963
ConditionFlowNodeConfiguration 1964
ConfluenceCrawlerConfiguration 1965
ConfluenceDataSourceConfiguration 1966

ConfluenceSourceConfiguration 1967
ContentBlock 1969
ContextEnrichmentConfiguration 1970
CrawlFilterConfiguration 1971
CuratedQuery 1972
CustomContent 1973
CustomDocumentIdentifier 1975
CustomOrchestration 1976
CustomS3Location 1977
CustomTransformationConfiguration 1978
CyclicConnectionFlowValidationDetails 1980
DailySchedule 1981
DataSource 1982
DataSourceConfiguration 1985
DataSourceSummary 1988
DayOfMonth 1990
DeletionProtectionConfiguration 1992
DocumentAccessControlEntry 1993
DocumentContent 1995
DocumentIdentifier 1997
DocumentMetadata 1998
DuplicateConditionExpressionFlowValidationDetails 2000
DuplicateConnectionsFlowValidationDetails 2001
EmbeddingModelConfiguration 2002
EnrichmentStrategyConfiguration 2003
FieldForReranking 2004
FixedSizeChunkingConfiguration 2005
FlowAliasConcurrencyConfiguration 2006
FlowAliasRoutingConfigurationListItem 2008
FlowAliasSummary 2009
FlowCondition 2012
FlowConditionalConnectionConfiguration 2013
FlowConnection 2014
FlowConnectionConfiguration 2016
FlowDataConnectionConfiguration 2017
FlowDefinition 2018

FlowNode 2019
FlowNodeConfiguration 2021
FlowNodeInput 2025
FlowNodeOutput 2027
FlowSummary 2028
FlowValidation 2031
FlowValidationDetails 2033
FlowVersionSummary 2040
Function 2042
FunctionSchema 2044
GuardrailConfiguration 2046
HierarchicalChunkingConfiguration 2047
HierarchicalChunkingLevelConfiguration 2049
ImageExtractionConfiguration 2050
IncompatibleConnectionDataTypeFlowValidationDetails 2051
InferenceConfiguration 2052
IngestionJob 2054
IngestionJobFilter 2057
IngestionJobSortBy 2059
IngestionJobStatistics 2060
IngestionJobSummary 2062
InlineCodeFlowNodeConfiguration 2065
InlineContent 2067
InputFlowNodeConfiguration 2069
IntermediateStorage 2070
InvalidLoopBoundaryFlowValidationDetails 2071
IteratorFlowNodeConfiguration 2073
KendraKnowledgeBaseConfiguration 2074
KnowledgeBase 2075
KnowledgeBaseConfiguration 2078
KnowledgeBaseDocument 2080
KnowledgeBaseDocumentDetail 2081
KnowledgeBaseFlowNodeConfiguration 2084
KnowledgeBaseOrchestrationConfiguration 2087
KnowledgeBasePromptTemplate 2089
KnowledgeBaseSummary 2090

LambdaFunctionFlowNodeConfiguration ... 2092
LastDayOfMonth ... 2093
LexFlowNodeConfiguration ... 2094
LoopControllerFlowNodeConfiguration ... 2095
LoopFlowNodeConfiguration ... 2096
LoopIncompatibleNodeTypeFlowValidationDetails ... 2098
LoopInputFlowNodeConfiguration ... 2100
MalformedConditionExpressionFlowValidationDetails ... 2101
MalformedNodeInputExpressionFlowValidationDetails ... 2103
ManagedKnowledgeBaseConfiguration ... 2105
ManagedKnowledgeBaseConnectorConfiguration ... 2107
MediaExtractionConfiguration ... 2109
MemoryConfiguration ... 2110
Message ... 2112
MetadataAttribute ... 2113
MetadataAttributeValue ... 2114
MetadataConfigurationForReranking ... 2116
MismatchedNodeInputTypeFlowValidationDetails ... 2117
MismatchedNodeOutputTypeFlowValidationDetails ... 2119
MissingConnectionConfigurationFlowValidationDetails ... 2121
MissingDefaultConditionFlowValidationDetails ... 2122
MissingEndingNodesFlowValidationDetails ... 2123
MissingLoopControllerNodeFlowValidationDetails ... 2124
MissingLoopInputNodeFlowValidationDetails ... 2125
MissingNodeConfigurationFlowValidationDetails ... 2126
MissingNodeInputFlowValidationDetails ... 2127
MissingNodeOutputFlowValidationDetails ... 2128
MissingStartingNodesFlowValidationDetails ... 2129
MongoDbAtlasConfiguration ... 2130
MongoDbAtlasFieldMapping ... 2133
MonthlySchedule ... 2135
MultipleLoopControllerNodesFlowValidationDetails ... 2136
MultipleLoopInputNodesFlowValidationDetails ... 2137
MultipleNodeInputConnectionsFlowValidationDetails ... 2138
NeptuneAnalyticsConfiguration ... 2139
NeptuneAnalyticsFieldMapping ... 2140

OpenSearchManagedClusterConfiguration 2141
OpenSearchManagedClusterFieldMapping 2143
OpenSearchServerlessConfiguration 2145
OpenSearchServerlessFieldMapping 2147
OrchestrationExecutor 2149
OutputFlowNodeConfiguration 2150
ParameterDetail 2151
ParsingConfiguration 2153
ParsingPrompt 2155
PatternObjectFilter 2156
PatternObjectFilterConfiguration 2158
PerformanceConfiguration 2159
PineconeConfiguration 2160
PineconeFieldMapping 2162
PromptAgentResource 2163
PromptConfiguration 2164
PromptFlowNodeConfiguration 2167
PromptFlowNodeInlineConfiguration 2168
PromptFlowNodeResourceConfiguration 2170
PromptFlowNodeSourceConfiguration 2171
PromptGenAiResource 2172
PromptInferenceConfiguration 2173
PromptInputVariable 2174
PromptMetadataEntry 2175
PromptModelInferenceConfiguration 2176
PromptOverrideConfiguration 2178
PromptSummary 2180
PromptTemplateConfiguration 2183
PromptVariant 2184
QueryGenerationColumn 2187
QueryGenerationConfiguration 2189
QueryGenerationContext 2190
QueryGenerationTable 2191
RdsConfiguration 2193
RdsFieldMapping 2195
RedisEnterpriseCloudConfiguration 2197

RedisEnterpriseCloudFieldMapping 2199
RedshiftConfiguration 2201
RedshiftProvisionedAuthConfiguration 2203
RedshiftProvisionedConfiguration 2205
RedshiftQueryEngineAwsDataCatalogStorageConfiguration 2206
RedshiftQueryEngineConfiguration 2207
RedshiftQueryEngineRedshiftStorageConfiguration 2209
RedshiftQueryEngineStorageConfiguration 2210
RedshiftServerlessAuthConfiguration 2212
RedshiftServerlessConfiguration 2213
RerankingMetadataSelectiveModeConfiguration 2214
RetrievalFlowNodeConfiguration 2216
RetrievalFlowNodeS3Configuration 2217
RetrievalFlowNodeServiceConfiguration 2218
S3Content 2219
S3DataSourceConfiguration 2220
S3Identifier 2222
S3Location 2223
S3VectorsConfiguration 2224
SalesforceCrawlerConfiguration 2226
SalesforceDataSourceConfiguration 2227
SalesforceSourceConfiguration 2228
SeedUrl 2230
SemanticChunkingConfiguration 2231
ServerSideEncryptionConfiguration 2233
SessionSummaryConfiguration 2234
SharePointCrawlerConfiguration 2235
SharePointDataSourceConfiguration 2236
SharePointSourceConfiguration 2237
SpecificToolChoice 2239
SqlKnowledgeBaseConfiguration 2240
StorageConfiguration 2241
StorageFlowNodeConfiguration 2244
StorageFlowNodeS3Configuration 2245
StorageFlowNodeServiceConfiguration 2246
SupplementalDataStorageConfiguration 2247

SupplementalDataStorageLocation 2248
SyncSchedule 2249
SystemContentBlock 2251
TextContentDoc 2253
TextPromptTemplateConfiguration 2254
Tool 2256
ToolChoice 2257
ToolConfiguration 2259
ToolInputSchema 2260
ToolSpecification 2261
Transformation 2263
TransformationFunction 2264
TransformationLambdaConfiguration 2265
UnfulfilledNodeInputFlowValidationDetails 2266
UnknownConnectionConditionFlowValidationDetails 2267
UnknownConnectionSourceFlowValidationDetails 2268
UnknownConnectionSourceOutputFlowValidationDetails 2269
UnknownConnectionTargetFlowValidationDetails 2270
UnknownConnectionTargetInputFlowValidationDetails 2271
UnknownNodeInputFlowValidationDetails 2272
UnknownNodeOutputFlowValidationDetails 2273
UnreachableNodeFlowValidationDetails 2274
UnsatisfiedConnectionConditionsFlowValidationDetails 2275
UnspecifiedFlowValidationDetails 2276
UrlConfiguration 2277
ValidationExceptionField 2278
VectorIngestionConfiguration 2279
VectorKnowledgeBaseConfiguration 2281
VectorSearchBedrockRerankingConfiguration 2283
VectorSearchBedrockRerankingModelConfiguration 2284
VectorSearchRerankingConfiguration 2285
VideoConfiguration 2286
VideoExtractionConfiguration 2287
VideoSegmentationConfiguration 2288
VpcConfiguration 2289
VpcConfigurationSummary 2293

WebCrawlerConfiguration ... 2297
WebCrawlerLimits ... 2300
WebDataSourceConfiguration ... 2301
WebSourceConfiguration ... 2302
WeeklySchedule ... 2303
Agents for Amazon Bedrock Runtime ... 2303
ActionGroupExecutor ... 2314
ActionGroupInvocationInput ... 2316
ActionGroupInvocationOutput ... 2318
AgentActionGroup ... 2319
AgentCollaboratorInputPayload ... 2322
AgentCollaboratorInvocationInput ... 2323
AgentCollaboratorInvocationOutput ... 2325
AgentCollaboratorOutputPayload ... 2327
AgenticRetrieveAction ... 2328
AgenticRetrieveActionDetails ... 2329
AgenticRetrieveBedrockGuardrailConfiguration ... 2330
AgenticRetrieveBedrockRerankingConfiguration ... 2331
AgenticRetrieveBedrockRerankingModelConfiguration ... 2332
AgenticRetrieveCitation ... 2333
AgenticRetrieveCitationReference ... 2334
AgenticRetrieveConfiguration ... 2335
AgenticRetrieveFailure ... 2337
AgenticRetrieveFullDocExpansionDetails ... 2338
AgenticRetrieveGeneratedResponse ... 2339
AgenticRetrieveGuardrailWarning ... 2340
AgenticRetrieveMemoryConfiguration ... 2342
AgenticRetrieveMemoryMetadataFilter ... 2344
AgenticRetrieveMemoryMetadataFilterLeft ... 2346
AgenticRetrieveMemoryMetadataFilterRight ... 2347
AgenticRetrieveMemoryMetadataValue ... 2348
AgenticRetrieveMemoryRetrievalConfig ... 2350
AgenticRetrieveMemoryRetrieveDetails ... 2352
AgenticRetrieveMemorySessionBinding ... 2354
AgenticRetrieveMessage ... 2356
AgenticRetrieveMessageContent ... 2357

AgenticRetrievePolicyConfiguration ... 2358
AgenticRetriever ... 2359
AgenticRetrieveRerankingConfiguration ... 2360
AgenticRetrieveResponseEvent ... 2361
AgenticRetrieveResultEvent ... 2362
AgenticRetrieveResultItem ... 2364
AgenticRetrieveSourceMetadata ... 2365
AgenticRetrieveSourceRetriever ... 2366
AgenticRetrieveStreamResponseOutput ... 2367
AgenticRetrieveTraceEvent ... 2370
AgenticRetrieveTraceEventAttributes ... 2371
AgenticRetrieveTraceResultItem ... 2374
AgenticRetrieveWarning ... 2375
AgenticRetrieveWarningMessage ... 2376
AnalyzePromptEvent ... 2377
ApiInvocationInput ... 2378
ApiParameter ... 2380
ApiRequestBody ... 2382
ApiResult ... 2383
APISchema ... 2386
Attribution ... 2387
AudioSegment ... 2388
BedrockFoundationModelConfiguration ... 2389
BedrockFoundationModelModelConfiguration ... 2390
BedrockModelConfigurations ... 2391
BedrockRerankingConfiguration ... 2392
BedrockRerankingModelConfiguration ... 2393
BedrockSessionContentBlock ... 2394
ByteContentDoc ... 2396
ByteContentFile ... 2398
Caller ... 2399
Citation ... 2400
CitationEvent ... 2401
CodeInterpreterInvocationInput ... 2402
CodeInterpreterInvocationOutput ... 2403
Collaborator ... 2405

CollaboratorConfiguration 2409
ConditionResultEvent 2411
ContentBlock 2413
ContentBody 2414
ConversationHistory 2415
CustomOrchestration 2416
CustomOrchestrationTrace 2417
CustomOrchestrationTraceEvent 2418
DocumentAcl 2419
DocumentAclCondition 2420
DocumentAclGroup 2422
DocumentAclMembership 2423
DocumentAclUser 2424
ExternalSource 2425
ExternalSourcesGenerationConfiguration 2426
ExternalSourcesRetrieveAndGenerateConfiguration 2428
FailureTrace 2430
FieldForReranking 2432
FilePart 2433
FileSource 2434
FilterAttribute 2435
FinalResponse 2436
FlowCompletionEvent 2437
FlowExecutionContent 2438
FlowExecutionError 2439
FlowExecutionEvent 2441
FlowExecutionInputEvent 2444
FlowExecutionOutputEvent 2446
FlowExecutionSummary 2448
FlowFailureEvent 2451
FlowInput 2453
FlowInputContent 2455
FlowInputField 2456
FlowMultiTurnInputContent 2458
FlowMultiTurnInputRequestEvent 2459
FlowOutputContent 2461

FlowOutputEvent 2462
FlowOutputField 2464
FlowResponseStream 2466
FlowTrace 2470
FlowTraceCondition 2472
FlowTraceConditionNodeResultEvent 2473
FlowTraceDependencyEvent 2475
FlowTraceEvent 2477
FlowTraceNodeActionEvent 2478
FlowTraceNodeInputContent 2480
FlowTraceNodeInputEvent 2481
FlowTraceNodeInputExecutionChainItem 2483
FlowTraceNodeInputField 2485
FlowTraceNodeInputSource 2487
FlowTraceNodeOutputContent 2489
FlowTraceNodeOutputEvent 2490
FlowTraceNodeOutputField 2492
FlowTraceNodeOutputNext 2494
FoundationModelConfiguration 2495
FunctionDefinition 2496
FunctionInvocationInput 2498
FunctionParameter 2500
FunctionResult 2502
FunctionSchema 2504
GeneratedQuery 2505
GeneratedResponsePart 2506
GenerationConfiguration 2507
GuardrailAssessment 2509
GuardrailConfiguration 2511
GuardrailConfigurationWithArn 2512
GuardrailContentFilter 2513
GuardrailContentPolicyAssessment 2515
GuardrailCustomWord 2516
GuardrailEvent 2517
GuardrailManagedWord 2518
GuardrailPiiEntityFilter 2520

GuardrailRegexFilter 2522
GuardrailSensitiveInformationPolicyAssessment 2524
GuardrailTopic 2525
GuardrailTopicPolicyAssessment 2527
GuardrailTrace 2528
GuardrailWordPolicyAssessment 2530
ImageBlock 2531
ImageInput 2532
ImageInputSource 2533
ImageSource 2534
ImplicitFilterConfiguration 2536
InferenceConfig 2537
InferenceConfiguration 2538
InlineAgentFilePart 2540
InlineAgentPayloadPart 2541
InlineAgentResponseStream 2542
InlineAgentReturnControlPayload 2546
InlineAgentTracePart 2547
InlineBedrockModelConfigurations 2549
InlineSessionState 2550
InputFile 2552
InputImage 2553
InputPrompt 2554
InvocationInput 2555
InvocationInputMember 2557
InvocationResultMember 2559
InvocationStep 2561
InvocationStepPayload 2563
InvocationStepSummary 2564
InvocationSummary 2566
KnowledgeBase 2568
KnowledgeBaseConfiguration 2570
KnowledgeBaseLookupInput 2571
KnowledgeBaseLookupOutput 2572
KnowledgeBaseQuery 2573
KnowledgeBaseRetrievalConfiguration 2575

KnowledgeBaseRetrievalResult 2576
KnowledgeBaseRetrieveAndGenerateConfiguration 2578
KnowledgeBaseRetrieverConfiguration 2580
KnowledgeBaseVectorSearchConfiguration 2581
ManagedSearchBedrockRerankingConfiguration 2583
ManagedSearchBedrockRerankingModelConfiguration 2584
ManagedSearchConfiguration 2585
ManagedSearchRerankingConfiguration 2587
Memory 2588
MemorySessionSummary 2589
Message 2591
Metadata 2592
MetadataAttributeSchema 2594
MetadataConfigurationForReranking 2596
ModelInvocationInput 2597
ModelPerformanceConfiguration 2600
NodeActionEvent 2601
NodeDependencyEvent 2603
NodeExecutionContent 2605
NodeFailureEvent 2606
NodeInputEvent 2608
NodeInputExecutionChainItem 2610
NodeInputField 2612
NodeInputSource 2614
NodeOutputEvent 2616
NodeOutputField 2618
NodeOutputNext 2620
NodeTraceElements 2621
Observation 2622
OptimizedPrompt 2625
OptimizedPromptEvent 2626
OptimizedPromptStream 2627
OrchestrationConfiguration 2630
OrchestrationExecutor 2632
OrchestrationModelInvocationOutput 2633
OrchestrationTrace 2635

OutputFile ... 2637
Parameter ... 2638
ParameterDetail ... 2639
PayloadPart ... 2641
PerformanceConfiguration ... 2642
PostProcessingModelInvocationOutput ... 2643
PostProcessingParsedResponse ... 2645
PostProcessingTrace ... 2646
PreProcessingModelInvocationOutput ... 2648
PreProcessingParsedResponse ... 2650
PreProcessingTrace ... 2651
PromptConfiguration ... 2653
PromptCreationConfigurations ... 2656
PromptOverrideConfiguration ... 2658
PromptTemplate ... 2660
PropertyParameters ... 2661
QueryGenerationInput ... 2662
QueryTransformationConfiguration ... 2663
Rationale ... 2664
RawResponse ... 2665
ReasoningContentBlock ... 2666
ReasoningTextBlock ... 2668
RepromptResponse ... 2669
RequestBody ... 2670
RerankDocument ... 2671
RerankingConfiguration ... 2673
RerankingMetadataSelectiveModeConfiguration ... 2674
RerankQuery ... 2676
RerankResult ... 2677
RerankSource ... 2678
RerankTextDocument ... 2679
ResponseStream ... 2680
RetrievalContent ... 2684
RetrievalFilter ... 2685
RetrievalOverrides ... 2690
RetrievalResultConfluenceLocation ... 2691

RetrievalResultContent 2692
RetrievalResultContentColumn 2694
RetrievalResultCustomDocumentLocation 2695
RetrievalResultGoogleDriveLocation 2696
RetrievalResultKendraDocumentLocation 2697
RetrievalResultLocation 2698
RetrievalResultOneDriveLocation 2701
RetrievalResultS3Location 2702
RetrievalResultSalesforceLocation 2703
RetrievalResultSharePointLocation 2704
RetrievalResultSqlLocation 2705
RetrievalResultWebLocation 2706
RetrieveAndGenerateConfiguration 2707
RetrieveAndGenerateInput 2709
RetrieveAndGenerateOutput 2710
RetrieveAndGenerateOutputEvent 2711
RetrieveAndGenerateSessionConfiguration 2712
RetrieveAndGenerateStreamResponseOutput 2713
RetrievedReference 2717
RetrieverConfiguration 2719
ReturnControlPayload 2720
ReturnControlResults 2721
RoutingClassifierModelInvocationOutput 2722
RoutingClassifierTrace 2723
S3Identifier 2725
S3Location 2726
S3ObjectDoc 2727
S3ObjectFile 2728
SatisfiedCondition 2729
SessionState 2730
SessionSummary 2733
Span 2735
StreamingConfigurations 2737
TextInferenceConfig 2739
TextPrompt 2741
TextResponsePart 2742

TextToSqlConfiguration ... 2743
TextToSqlKnowledgeBaseConfiguration ... 2744
Trace ... 2745
TraceElements ... 2748
TracePart ... 2749
TransformationConfiguration ... 2752
Usage ... 2753
UserContext ... 2754
VectorSearchBedrockRerankingConfiguration ... 2755
VectorSearchBedrockRerankingModelConfiguration ... 2756
VectorSearchRerankingConfiguration ... 2757
VideoSegment ... 2758
Data Automation for Amazon Bedrock ... 2759
AudioExtractionCategory ... 2762
AudioExtractionCategoryTypeConfiguration ... 2764
AudioLanguageConfiguration ... 2765
AudioOverrideConfiguration ... 2767
AudioStandardExtraction ... 2768
AudioStandardGenerativeField ... 2769
AudioStandardOutputConfiguration ... 2770
Blueprint ... 2771
BlueprintFilter ... 2775
BlueprintItem ... 2777
BlueprintOptimizationObject ... 2779
BlueprintOptimizationOutputConfiguration ... 2780
BlueprintOptimizationSample ... 2781
BlueprintSummary ... 2782
ChannelLabelingConfiguration ... 2784
CustomOutputConfiguration ... 2785
DataAutomationLibrary ... 2786
DataAutomationLibraryConfiguration ... 2789
DataAutomationLibraryEntitySummary ... 2790
DataAutomationLibraryFilter ... 2791
DataAutomationLibraryIngestionJob ... 2792
DataAutomationLibraryIngestionJobSummary ... 2795
DataAutomationLibraryItem ... 2797

DataAutomationLibrarySummary 2798
DataAutomationProject 2800
DataAutomationProjectFilter 2804
DataAutomationProjectSummary 2805
DeleteEntitiesInfo 2807
DocumentBoundingBox 2808
DocumentCustomOutputConfiguration 2809
DocumentExtractionGranularity 2810
DocumentOutputAdditionalFileFormat 2811
DocumentOutputFormat 2812
DocumentOutputTextFormat 2813
DocumentOverrideConfiguration 2814
DocumentStandardExtraction 2815
DocumentStandardGenerativeField 2816
DocumentStandardOutputConfiguration 2817
EncryptionConfiguration 2818
EntityDetails 2820
EntityTypeInfo 2821
EventBridgeConfiguration 2822
ImageBoundingBox 2823
ImageExtractionCategory 2824
ImageOverrideConfiguration 2825
ImageStandardExtraction 2826
ImageStandardGenerativeField 2827
ImageStandardOutputConfiguration 2828
InlinePayload 2829
InputConfiguration 2830
ModalityProcessingConfiguration 2831
ModalityRoutingConfiguration 2832
NotificationConfiguration 2834
OutputConfiguration 2835
OverrideConfiguration 2836
Phrase 2838
PIIEntitiesConfiguration 2839
S3Object 2841
SensitiveDataConfiguration 2842

SpeakerLabelingConfiguration ... 2844
SplitterConfiguration ... 2845
StandardOutputConfiguration ... 2846
Tag ... 2848
TranscriptConfiguration ... 2849
UpsertEntityInfo ... 2850
ValidationExceptionField ... 2851
VideoBoundingBox ... 2852
VideoExtractionCategory ... 2853
VideoOverrideConfiguration ... 2854
VideoStandardExtraction ... 2855
VideoStandardGenerativeField ... 2856
VideoStandardOutputConfiguration ... 2857
VocabularyEntity ... 2858
VocabularyEntityInfo ... 2860
VocabularyEntitySummary ... 2862
Runtime for Amazon Bedrock Data Automation ... 2863
AssetProcessingConfiguration ... 2865
Blueprint ... 2866
DataAutomationConfiguration ... 2868
EncryptionConfiguration ... 2869
EventBridgeConfiguration ... 2871
InputConfiguration ... 2872
NotificationConfiguration ... 2873
OutputConfiguration ... 2874
OutputSegment ... 2875
SyncInputConfiguration ... 2876
Tag ... 2877
TimestampSegment ... 2878
VideoAssetProcessingConfiguration ... 2879
VideoSegmentConfiguration ... 2880
Amazon Bedrock Runtime ... 2880
AnyToolChoice ... 2887
AppliedGuardrailDetails ... 2888
AsyncInvokeOutputDataConfig ... 2890
AsyncInvokeS3OutputDataConfig ... 2891

AsyncInvokeSummary ... 2893
AudioBlock ... 2896
AudioSource ... 2898
AutoToolChoice ... 2900
BidirectionalInputPayloadPart ... 2901
BidirectionalOutputPayloadPart ... 2902
CacheDetail ... 2903
CachePointBlock ... 2904
Citation ... 2905
CitationGeneratedContent ... 2907
CitationLocation ... 2908
CitationsConfig ... 2910
CitationsContentBlock ... 2911
CitationsDelta ... 2912
CitationSourceContent ... 2914
CitationSourceContentDelta ... 2915
ContentBlock ... 2916
ContentBlockDelta ... 2919
ContentBlockDeltaEvent ... 2921
ContentBlockStart ... 2922
ContentBlockStartEvent ... 2924
ContentBlockStopEvent ... 2925
ConverseMetrics ... 2926
ConverseOutput ... 2927
ConverseStreamMetadataEvent ... 2928
ConverseStreamMetrics ... 2930
ConverseStreamOutput ... 2931
ConverseStreamTrace ... 2934
ConverseTokensRequest ... 2935
ConverseTrace ... 2937
CountTokensInput ... 2938
DocumentBlock ... 2940
DocumentCharLocation ... 2942
DocumentChunkLocation ... 2944
DocumentContentBlock ... 2946
DocumentPageLocation ... 2947

DocumentSource 2949
ErrorBlock 2951
GuardrailAssessment 2952
GuardrailAutomatedReasoningFinding 2954
GuardrailAutomatedReasoningImpossibleFinding 2957
GuardrailAutomatedReasoningInputTextReference 2958
GuardrailAutomatedReasoningInvalidFinding 2959
GuardrailAutomatedReasoningLogicWarning 2960
GuardrailAutomatedReasoningNoTranslationsFinding 2962
GuardrailAutomatedReasoningPolicyAssessment 2963
GuardrailAutomatedReasoningRule 2964
GuardrailAutomatedReasoningSatisfiableFinding 2965
GuardrailAutomatedReasoningScenario 2967
GuardrailAutomatedReasoningStatement 2968
GuardrailAutomatedReasoningTooComplexFinding 2969
GuardrailAutomatedReasoningTranslation 2970
GuardrailAutomatedReasoningTranslationAmbiguousFinding 2972
GuardrailAutomatedReasoningTranslationOption 2973
GuardrailAutomatedReasoningValidFinding 2974
GuardrailChecksConfig 2976
GuardrailChecksContentBlock 2977
GuardrailChecksContentFilterCategoryConfig 2978
GuardrailChecksContentFilterConfig 2979
GuardrailChecksContentFilterResult 2980
GuardrailChecksContentFilterResultEntry 2981
GuardrailChecksContentFilterUsage 2982
GuardrailChecksMessage 2983
GuardrailChecksPromptAttackCategoryConfig 2984
GuardrailChecksPromptAttackConfig 2985
GuardrailChecksPromptAttackResult 2986
GuardrailChecksPromptAttackResultEntry 2987
GuardrailChecksPromptAttackUsage 2988
GuardrailChecksResults 2989
GuardrailChecksSensitiveInformationConfig 2990
GuardrailChecksSensitiveInformationEntityConfig 2991
GuardrailChecksSensitiveInformationResult 2992

GuardrailChecksSensitiveInformationResultEntry ... 2993
GuardrailChecksSensitiveInformationUsage ... 2995
GuardrailChecksUsageResults ... 2996
GuardrailConfiguration ... 2997
GuardrailContentBlock ... 2999
GuardrailContentFilter ... 3000
GuardrailContentPolicyAssessment ... 3002
GuardrailContextualGroundingFilter ... 3003
GuardrailContextualGroundingPolicyAssessment ... 3005
GuardrailConverseContentBlock ... 3006
GuardrailConverseImageBlock ... 3007
GuardrailConverseImageSource ... 3008
GuardrailConverseTextBlock ... 3009
GuardrailCoverage ... 3010
GuardrailCustomWord ... 3011
GuardrailImageBlock ... 3012
GuardrailImageCoverage ... 3013
GuardrailImageSource ... 3014
GuardrailInvocationMetrics ... 3015
GuardrailManagedWord ... 3016
GuardrailOutputContent ... 3018
GuardrailPiiEntityFilter ... 3019
GuardrailRegexFilter ... 3021
GuardrailSensitiveInformationPolicyAssessment ... 3023
GuardrailStreamConfiguration ... 3024
GuardrailTextBlock ... 3026
GuardrailTextCharactersCoverage ... 3027
GuardrailTopic ... 3028
GuardrailTopicPolicyAssessment ... 3030
GuardrailTraceAssessment ... 3031
GuardrailUsage ... 3033
GuardrailWordPolicyAssessment ... 3035
ImageBlock ... 3036
ImageBlockDelta ... 3038
ImageBlockStart ... 3039
ImageSource ... 3040

InferenceConfiguration 3042
InvokeModelTokensRequest 3044
InvokeModelWithBidirectionalStreamInput 3045
InvokeModelWithBidirectionalStreamOutput 3046
JsonSchemaDefinition 3048
Message 3049
MessageStartEvent 3051
MessageStopEvent 3052
OutputConfig 3053
OutputFormat 3054
OutputFormatStructure 3055
PayloadPart 3056
PerformanceConfiguration 3057
PromptRouterTrace 3058
PromptVariableValues 3059
ReasoningContentBlock 3060
ReasoningContentBlockDelta 3062
ReasoningTextBlock 3064
ResponseStream 3065
S3Location 3067
SearchResultBlock 3068
SearchResultContentBlock 3070
SearchResultLocation 3071
ServiceTier 3073
SpecificToolChoice 3074
SystemContentBlock 3075
SystemTool 3077
Tag 3078
TokenUsage 3079
Tool 3081
ToolChoice 3083
ToolConfiguration 3085
ToolInputSchema 3086
ToolResultBlock 3087
ToolResultBlockDelta 3089
ToolResultBlockStart 3090

ToolResultContentBlock 3092
ToolSpecification 3094
ToolUseBlock 3096
ToolUseBlockDelta 3098
ToolUseBlockStart 3099
VideoBlock 3101
VideoSource 3102
WebLocation 3104
Common Parameters 3105
Common Error Types 3108

Welcome

Amazon Bedrock

Describes the API operations for creating, managing, fine-turning, and evaluating Amazon Bedrock models.

Agents for Amazon Bedrock

> **Note**
>
> Amazon Bedrock Agents (now Amazon Bedrock Agents Classic) is no longer open to new customers. For capabilities similar to Bedrock Agents Classic, explore Amazon Bedrock AgentCore. Existing customers can continue to use the service as normal. For more information, see Amazon Bedrock Agents Classic availability change.

Describes the API operations for creating and managing Amazon Bedrock agents.

Agents for Amazon Bedrock Runtime

> **Note**
>
> Amazon Bedrock Agents (now Amazon Bedrock Agents Classic) is no longer open to new customers. For capabilities similar to Bedrock Agents Classic, explore Amazon Bedrock AgentCore. Existing customers can continue to use the service as normal. For more information, see Amazon Bedrock Agents Classic availability change.

Contains APIs related to model invocation and querying of knowledge bases.

Data Automation for Amazon Bedrock

Describes the API operations for creating and managing Data Automation resources and generate insights.

Runtime for Amazon Bedrock Data Automation

Process data with a data automation project.

Amazon Bedrock Runtime

Describes the API operations for running inference using Amazon Bedrock models.

Actions

The following actions are supported by Amazon Bedrock:

- BatchDeleteAdvancedPromptOptimizationJob
- BatchDeleteEvaluationJob
- CancelAutomatedReasoningPolicyBuildWorkflow
- CreateAdvancedPromptOptimizationJob
- CreateAutomatedReasoningPolicy
- CreateAutomatedReasoningPolicyTestCase
- CreateAutomatedReasoningPolicyVersion
- CreateCustomModel
- CreateCustomModelDeployment
- CreateEvaluationJob
- CreateFoundationModelAgreement
- CreateGuardrail
- CreateGuardrailVersion
- CreateInferenceProfile
- CreateMarketplaceModelEndpoint
- CreateModelCopyJob
- CreateModelCustomizationJob
- CreateModelImportJob
- CreateModelInvocationJob
- CreatePromptRouter
- CreateProvisionedModelThroughput
- DeleteAutomatedReasoningPolicy
- DeleteAutomatedReasoningPolicyBuildWorkflow
- DeleteAutomatedReasoningPolicyTestCase
- DeleteCustomModel
- DeleteCustomModelDeployment
- DeleteEnforcedGuardrailConfiguration

- DeleteFoundationModelAgreement
- DeleteGuardrail
- DeleteImportedModel
- DeleteInferenceProfile
- DeleteMarketplaceModelEndpoint
- DeleteModelInvocationLoggingConfiguration
- DeletePromptRouter
- DeleteProvisionedModelThroughput
- DeleteResourcePolicy
- DeregisterMarketplaceModelEndpoint
- ExportAutomatedReasoningPolicyVersion
- GetAccountDataRetention
- GetAdvancedPromptOptimizationJob
- GetAutomatedReasoningPolicy
- GetAutomatedReasoningPolicyAnnotations
- GetAutomatedReasoningPolicyBuildWorkflow
- GetAutomatedReasoningPolicyBuildWorkflowResultAssets
- GetAutomatedReasoningPolicyNextScenario
- GetAutomatedReasoningPolicyTestCase
- GetAutomatedReasoningPolicyTestResult
- GetCustomModel
- GetCustomModelDeployment
- GetEvaluationJob
- GetFoundationModel
- GetFoundationModelAvailability
- GetGuardrail
- GetImportedModel
- GetInferenceProfile
- GetMarketplaceModelEndpoint
- GetModelCopyJob

- GetModelCustomizationJob
- GetModelImportJob
- GetModelInvocationJob
- GetModelInvocationLoggingConfiguration
- GetPromptRouter
- GetProvisionedModelThroughput
- GetResourcePolicy
- GetUseCaseForModelAccess
- ListAdvancedPromptOptimizationJobs
- ListAutomatedReasoningPolicies
- ListAutomatedReasoningPolicyBuildWorkflows
- ListAutomatedReasoningPolicyTestCases
- ListAutomatedReasoningPolicyTestResults
- ListCustomModelDeployments
- ListCustomModels
- ListEnforcedGuardrailsConfiguration
- ListEvaluationJobs
- ListFoundationModelAgreementOffers
- ListFoundationModels
- ListGuardrails
- ListImportedModels
- ListInferenceProfiles
- ListMarketplaceModelEndpoints
- ListModelCopyJobs
- ListModelCustomizationJobs
- ListModelImportJobs
- ListModelInvocationJobs
- ListPromptRouters
- ListProvisionedModelThroughputs
- ListTagsForResource

- PutAccountDataRetention
- PutEnforcedGuardrailConfiguration
- PutModelInvocationLoggingConfiguration
- PutResourcePolicy
- PutUseCaseForModelAccess
- RegisterMarketplaceModelEndpoint
- StartAutomatedReasoningPolicyBuildWorkflow
- StartAutomatedReasoningPolicyTestWorkflow
- StopAdvancedPromptOptimizationJob
- StopEvaluationJob
- StopModelCustomizationJob
- StopModelInvocationJob
- TagResource
- UntagResource
- UpdateAutomatedReasoningPolicy
- UpdateAutomatedReasoningPolicyAnnotations
- UpdateAutomatedReasoningPolicyTestCase
- UpdateCustomModelDeployment
- UpdateGuardrail
- UpdateMarketplaceModelEndpoint
- UpdateProvisionedModelThroughput

The following actions are supported by Agents for Amazon Bedrock:

- AssociateAgentCollaborator
- AssociateAgentKnowledgeBase
- CreateAgent
- CreateAgentActionGroup
- CreateAgentAlias
- CreateDataSource
- CreateFlow

- CreateFlowAlias
- CreateFlowVersion
- CreateKnowledgeBase
- CreatePrompt
- CreatePromptVersion
- CreateVpcConfiguration
- DeleteAgent
- DeleteAgentActionGroup
- DeleteAgentAlias
- DeleteAgentVersion
- DeleteDataSource
- DeleteFlow
- DeleteFlowAlias
- DeleteFlowVersion
- DeleteKnowledgeBase
- DeleteKnowledgeBaseDocuments
- DeletePrompt
- DeleteResourcePolicy
- DeleteVpcConfiguration
- DisassociateAgentCollaborator
- DisassociateAgentKnowledgeBase
- GetAgent
- GetAgentActionGroup
- GetAgentAlias
- GetAgentCollaborator
- GetAgentKnowledgeBase
- GetAgentVersion
- GetDataSource
- GetFlow
- GetFlowAlias

- GetFlowVersion
- GetIngestionJob
- GetKnowledgeBase
- GetKnowledgeBaseDocuments
- GetPrompt
- GetResourcePolicy
- GetVpcConfiguration
- IngestKnowledgeBaseDocuments
- ListAgentActionGroups
- ListAgentAliases
- ListAgentCollaborators
- ListAgentKnowledgeBases
- ListAgents
- ListAgentVersions
- ListDataSources
- ListFlowAliases
- ListFlows
- ListFlowVersions
- ListIngestionJobs
- ListKnowledgeBaseDocuments
- ListKnowledgeBases
- ListPrompts
- ListTagsForResource
- ListVpcConfigurations
- PrepareAgent
- PrepareFlow
- PutResourcePolicy
- StartIngestionJob
- StopIngestionJob
- TagResource

- UntagResource
- UpdateAgent
- UpdateAgentActionGroup
- UpdateAgentAlias
- UpdateAgentCollaborator
- UpdateAgentKnowledgeBase
- UpdateDataSource
- UpdateFlow
- UpdateFlowAlias
- UpdateKnowledgeBase
- UpdatePrompt
- ValidateFlowDefinition

The following actions are supported by Agents for Amazon Bedrock Runtime:

- AgenticRetrieveStream
- CheckIngestedDocumentAcl
- CreateInvocation
- CreateSession
- DeleteAgentMemory
- DeleteSession
- EndSession
- GenerateQuery
- GetAgentMemory
- GetDocumentContent
- GetExecutionFlowSnapshot
- GetFlowExecution
- GetIngestedDocumentAcl
- GetInvocationStep
- GetSession
- InvokeAgent

- InvokeFlow
- InvokeInlineAgent
- ListFlowExecutionEvents
- ListFlowExecutions
- ListInvocations
- ListInvocationSteps
- ListSessions
- ListTagsForResource
- OptimizePrompt
- PutInvocationStep
- Rerank
- Retrieve
- RetrieveAndGenerate
- RetrieveAndGenerateStream
- StartFlowExecution
- StopFlowExecution
- TagResource
- UntagResource
- UpdateSession

The following actions are supported by Data Automation for Amazon Bedrock:

- CopyBlueprintStage
- CreateBlueprint
- CreateBlueprintVersion
- CreateDataAutomationLibrary
- CreateDataAutomationProject
- DeleteBlueprint
- DeleteDataAutomationLibrary
- DeleteDataAutomationProject
- GetBlueprint

- GetBlueprintOptimizationStatus
- GetDataAutomationLibrary
- GetDataAutomationLibraryEntity
- GetDataAutomationLibraryIngestionJob
- GetDataAutomationProject
- InvokeBlueprintOptimizationAsync
- InvokeDataAutomationLibraryIngestionJob
- ListBlueprints
- ListDataAutomationLibraries
- ListDataAutomationLibraryEntities
- ListDataAutomationLibraryIngestionJobs
- ListDataAutomationProjects
- ListTagsForResource
- TagResource
- UntagResource
- UpdateBlueprint
- UpdateDataAutomationLibrary
- UpdateDataAutomationProject

The following actions are supported by Runtime for Amazon Bedrock Data Automation:

- GetDataAutomationStatus
- InvokeDataAutomation
- InvokeDataAutomationAsync
- ListTagsForResource
- TagResource
- UntagResource

The following actions are supported by Amazon Bedrock Runtime:

- ApplyGuardrail
- Converse

- ConverseStream
- CountTokens
- GetAsyncInvoke
- InvokeGuardrailChecks
- InvokeModel
- InvokeModelWithBidirectionalStream
- InvokeModelWithResponseStream
- ListAsyncInvokes
- StartAsyncInvoke

Amazon Bedrock

The following actions are supported by Amazon Bedrock:

- BatchDeleteAdvancedPromptOptimizationJob
- BatchDeleteEvaluationJob
- CancelAutomatedReasoningPolicyBuildWorkflow
- CreateAdvancedPromptOptimizationJob
- CreateAutomatedReasoningPolicy
- CreateAutomatedReasoningPolicyTestCase
- CreateAutomatedReasoningPolicyVersion
- CreateCustomModel
- CreateCustomModelDeployment
- CreateEvaluationJob
- CreateFoundationModelAgreement
- CreateGuardrail
- CreateGuardrailVersion
- CreateInferenceProfile
- CreateMarketplaceModelEndpoint
- CreateModelCopyJob
- CreateModelCustomizationJob

- CreateModelImportJob
- CreateModelInvocationJob
- CreatePromptRouter
- CreateProvisionedModelThroughput
- DeleteAutomatedReasoningPolicy
- DeleteAutomatedReasoningPolicyBuildWorkflow
- DeleteAutomatedReasoningPolicyTestCase
- DeleteCustomModel
- DeleteCustomModelDeployment
- DeleteEnforcedGuardrailConfiguration
- DeleteFoundationModelAgreement
- DeleteGuardrail
- DeleteImportedModel
- DeleteInferenceProfile
- DeleteMarketplaceModelEndpoint
- DeleteModelInvocationLoggingConfiguration
- DeletePromptRouter
- DeleteProvisionedModelThroughput
- DeleteResourcePolicy
- DeregisterMarketplaceModelEndpoint
- ExportAutomatedReasoningPolicyVersion
- GetAccountDataRetention
- GetAdvancedPromptOptimizationJob
- GetAutomatedReasoningPolicy
- GetAutomatedReasoningPolicyAnnotations
- GetAutomatedReasoningPolicyBuildWorkflow
- GetAutomatedReasoningPolicyBuildWorkflowResultAssets
- GetAutomatedReasoningPolicyNextScenario
- GetAutomatedReasoningPolicyTestCase
- GetAutomatedReasoningPolicyTestResult

- GetCustomModel
- GetCustomModelDeployment
- GetEvaluationJob
- GetFoundationModel
- GetFoundationModelAvailability
- GetGuardrail
- GetImportedModel
- GetInferenceProfile
- GetMarketplaceModelEndpoint
- GetModelCopyJob
- GetModelCustomizationJob
- GetModelImportJob
- GetModelInvocationJob
- GetModelInvocationLoggingConfiguration
- GetPromptRouter
- GetProvisionedModelThroughput
- GetResourcePolicy
- GetUseCaseForModelAccess
- ListAdvancedPromptOptimizationJobs
- ListAutomatedReasoningPolicies
- ListAutomatedReasoningPolicyBuildWorkflows
- ListAutomatedReasoningPolicyTestCases
- ListAutomatedReasoningPolicyTestResults
- ListCustomModelDeployments
- ListCustomModels
- ListEnforcedGuardrailsConfiguration
- ListEvaluationJobs
- ListFoundationModelAgreementOffers
- ListFoundationModels
- ListGuardrails

- ListImportedModels
- ListInferenceProfiles
- ListMarketplaceModelEndpoints
- ListModelCopyJobs
- ListModelCustomizationJobs
- ListModelImportJobs
- ListModelInvocationJobs
- ListPromptRouters
- ListProvisionedModelThroughputs
- ListTagsForResource
- PutAccountDataRetention
- PutEnforcedGuardrailConfiguration
- PutModelInvocationLoggingConfiguration
- PutResourcePolicy
- PutUseCaseForModelAccess
- RegisterMarketplaceModelEndpoint
- StartAutomatedReasoningPolicyBuildWorkflow
- StartAutomatedReasoningPolicyTestWorkflow
- StopAdvancedPromptOptimizationJob
- StopEvaluationJob
- StopModelCustomizationJob
- StopModelInvocationJob
- TagResource
- UntagResource
- UpdateAutomatedReasoningPolicy
- UpdateAutomatedReasoningPolicyAnnotations
- UpdateAutomatedReasoningPolicyTestCase
- UpdateCustomModelDeployment
- UpdateGuardrail
- UpdateMarketplaceModelEndpoint

- UpdateProvisionedModelThroughput

BatchDeleteAdvancedPromptOptimizationJob

Service: Amazon Bedrock

Deletes one or more advanced prompt optimization jobs.

Request Syntax

```
POST /advanced-prompt-optimization-job/batch-delete HTTP/1.1
Content-type: application/json

{
   "jobIdentifiers": [ "string" ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

jobIdentifiers

A list of ARNs or IDs of the advanced prompt optimization jobs to delete.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/)?[a-z0-9]{12})`

Required: Yes

Response Syntax

```
HTTP/1.1 202
```

```
Content-type: application/json

{
   "advancedPromptOptimizationJobs": [
      {
         "jobIdentifier": "string",
         "jobStatus": "string"
      }
   ],
   "errors": [
      {
         "code": "string",
         "jobIdentifier": "string",
         "message": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

advancedPromptOptimizationJobs

A list of jobs that were successfully deleted.

Type: Array of BatchDeleteAdvancedPromptOptimizationJobItem objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

errors

A list of errors for jobs that could not be deleted.

Type: Array of BatchDeleteAdvancedPromptOptimizationJobError objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

BatchDeleteEvaluationJob

Service: Amazon Bedrock

Deletes a batch of evaluation jobs. An evaluation job can only be deleted if it has following status `FAILED`, `COMPLETED`, and `STOPPED`. You can request up to 25 model evaluation jobs be deleted in a single request.

Request Syntax

```
POST /evaluation-jobs/batch-delete HTTP/1.1
Content-type: application/json

{
   "jobIdentifiers": [ "string" ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

jobIdentifiers

A list of one or more evaluation job Amazon Resource Names (ARNs) you want to delete.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})`

Required: Yes

Response Syntax

```
HTTP/1.1 202
```

```
Content-type: application/json

{
   "errors": [
      {
         "code": "string",
         "jobIdentifier": "string",
         "message": "string"
      }
   ],
   "evaluationJobs": [
      {
         "jobIdentifier": "string",
         "jobStatus": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

errors

A JSON object containing the HTTP status codes and the ARNs of evaluation jobs that failed to be deleted.

Type: Array of BatchDeleteEvaluationJobError objects

Array Members: Minimum number of 0 items. Maximum number of 25 items.

evaluationJobs

The list of evaluation jobs for deletion.

Type: Array of BatchDeleteEvaluationJobItem objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CancelAutomatedReasoningPolicyBuildWorkflow

Service: Amazon Bedrock

Cancels a running Automated Reasoning policy build workflow. This stops the policy generation process and prevents further processing of the source documents.

Request Syntax

```
POST /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/cancel
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow to cancel. You can get this ID from the StartAutomatedReasoningPolicyBuildWorkflow response or by listing build workflows.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose build workflow you want to cancel.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAdvancedPromptOptimizationJob

Service: Amazon Bedrock

Creates an advanced prompt optimization job.

Request Syntax

```
POST /advanced-prompt-optimization-jobs HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "encryptionKeyArn": "string",
   "inputConfig": {
      "s3Uri": "string"
   },
   "jobDescription": "string",
   "jobName": "string",
   "modelConfigurations": [
      {
         "additionalModelRequestFields": {
            "string" : JSON value
         },
         "inferenceConfig": {
            "maxTokens": number,
            "stopSequences": [ "string" ],
            "temperature": number,
            "topP": number
         },
         "modelId": "string"
      }
   ],
   "outputConfig": {
      "s3Uri": "string"
   },
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure idempotency of the request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

encryptionKeyArn

The ARN of the KMS key to use for encrypting the output data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

inputConfig

The input data configuration for the optimization job.

Type: AdvancedPromptOptimizationInputConfig object

Required: Yes

jobDescription

A description of the advanced prompt optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: No

jobName

A name for the advanced prompt optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9.+-]*`

Required: Yes

modelConfigurations

A list of model configurations for the optimization job.

Type: Array of ModelConfiguration objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

outputConfig

The output data configuration for the optimization job.

Type: AdvancedPromptOptimizationOutputConfig object

Required: Yes

tags

Tags to associate with the advanced prompt optimization job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

jobArn

The ARN of the created advanced prompt optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAutomatedReasoningPolicy

Service: Amazon Bedrock

Creates an Automated Reasoning policy for Amazon Bedrock Guardrails. Automated Reasoning policies use mathematical techniques to detect hallucinations, suggest corrections, and highlight unstated assumptions in the responses of your GenAI application.

To create a policy, you upload a source document that describes the rules that you're encoding. Automated Reasoning extracts important concepts from the source document that will become variables in the policy and infers policy rules.

Request Syntax

```
POST /automated-reasoning-policies HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "description": "string",
   "kmsKeyId": "string",
   "name": "string",
   "policyDefinition": {
      "rules": [
         {
            "alternateExpression": "string",
            "expression": "string",
            "id": "string"
         }
      ],
      "types": [
         {
            "description": "string",
            "name": "string",
            "values": [
               {
                  "description": "string",
                  "value": "string"
               }
            ]
         }
      ],
      "variables": [
```

```
            {
               "description": "string",
               "name": "string",
               "type": "string"
            }
         ],
         "version": "string"
      },
      "tags": [
         {
            "key": "string",
            "value": "string"
         }
      ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than once. If this token matches a previous request, Amazon Bedrock ignores the request but doesn't return an error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

description

A description of the Automated Reasoning policy. Use this to provide context about the policy's purpose and the types of validations it performs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

kmsKeyId

The identifier of the AWS KMS key to use for encrypting the automated reasoning policy and its associated artifacts. If you don't specify a AWS KMS key, Amazon Bedrock uses an AWS KMS managed key for encryption. For enhanced security and control, you can specify a customer managed AWS KMS key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

name

A unique name for the Automated Reasoning policy. The name must be between 1 and 256 characters and can contain letters, numbers, hyphens, and underscores.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

Required: Yes

policyDefinition

The policy definition that contains the formal logic rules, variables, and custom variable types used to validate foundation model responses in your application.

Type: AutomatedReasoningPolicyDefinition object

Required: No

tags

A list of tags to associate with the Automated Reasoning policy. Tags help you organize and manage your policies.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "definitionHash": "string",
   "description": "string",
   "name": "string",
   "policyArn": "string",
   "updatedAt": "string",
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp when the policy was created.

Type: Timestamp

definitionHash

The hash of the policy definition. This is used as a concurrency token for creating policy versions that you can use in your application.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

description

The description of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

name

The name of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy that you created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

updatedAt

The timestamp when the policy was last updated.

Type: Timestamp

version

The version number of the newly created Automated Reasoning policy. The initial version is always DRAFT.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `([1-9][0-9]{0,11})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAutomatedReasoningPolicyTestCase

Service: Amazon Bedrock

Creates a test for an Automated Reasoning policy. Tests validate that your policy works as expected by providing sample inputs and expected outcomes. Use tests to verify policy behavior before deploying to production.

Request Syntax

```
POST /automated-reasoning-policies/policyArn/test-cases HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "confidenceThreshold": number,
   "expectedAggregatedFindingsResult": "string",
   "guardContent": "string",
   "queryContent": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy for which to create the test.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

confidenceThreshold

The minimum confidence level for logic validation. Content that meets the threshold is considered a high-confidence finding that can be validated.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

expectedAggregatedFindingsResult

The expected result of the Automated Reasoning check. Valid values include: , TOO_COMPLEX, and NO_TRANSLATIONS.

- `VALID` - The claims are true. The claims are implied by the premises and the Automated Reasoning policy. Given the Automated Reasoning policy and premises, it is not possible for these claims to be false. In other words, there are no alternative answers that are true that contradict the claims.
- `INVALID` - The claims are false. The claims are not implied by the premises and Automated Reasoning policy. Furthermore, there exists different claims that are consistent with the premises and Automated Reasoning policy.
- `SATISFIABLE` - The claims can be true or false. It depends on what assumptions are made for the claim to be implied from the premises and Automated Reasoning policy rules. In this situation, different assumptions can make input claims false and alternative claims true.
- `IMPOSSIBLE` - Automated Reasoning can't make a statement about the claims. This can happen if the premises are logically incorrect, or if there is a conflict within the Automated Reasoning policy itself.

- TRANSLATION_AMBIGUOUS - Detected an ambiguity in the translation meant it would be unsound to continue with validity checking. Additional context or follow-up questions might be needed to get translation to succeed.
- TOO_COMPLEX - The input contains too much information for Automated Reasoning to process within its latency limits.
- NO_TRANSLATIONS - Identifies that some or all of the input prompt wasn't translated into logic. This can happen if the input isn't relevant to the Automated Reasoning policy, or if the policy doesn't have variables to model relevant input. If Automated Reasoning can't translate anything, you get a single NO_TRANSLATIONS finding. You might also see a NO_TRANSLATIONS (along with other findings) if some part of the validation isn't translated.

Type: String

Valid Values: VALID | INVALID | SATISFIABLE | IMPOSSIBLE | TRANSLATION_AMBIGUOUS | TOO_COMPLEX | NO_TRANSLATION

Required: Yes

guardContent

The output content that's validated by the Automated Reasoning policy. This represents the foundation model response that will be checked for accuracy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

queryContent

The input query or prompt that generated the content. This provides context for the validation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "policyArn": "string",
   "testCaseId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

policyArn

The Amazon Resource Name (ARN) of the policy for which the test was created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

testCaseId

The unique identifier of the created test.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAutomatedReasoningPolicyVersion

Service: Amazon Bedrock

Creates a new version of an existing Automated Reasoning policy. This allows you to iterate on your policy rules while maintaining previous versions for rollback or comparison purposes.

Request Syntax

```
POST /automated-reasoning-policies/policyArn/versions HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "lastUpdatedDefinitionHash": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy for which to create a version.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

lastUpdatedDefinitionHash

The hash of the current policy definition used as a concurrency token to ensure the policy hasn't been modified since you last retrieved it.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

Required: Yes

tags

A list of tags to associate with the policy version.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
```

```
    "createdAt": "string",
    "definitionHash": "string",
    "description": "string",
    "name": "string",
    "policyArn": "string",
    "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp when the policy version was created.

Type: Timestamp

definitionHash

The hash of the policy definition for this version.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

description

The description of the policy version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

name

The name of the policy version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

policyArn

The versioned Amazon Resource Name (ARN) of the policy version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

version

The version number of the policy version.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `([1-9][0-9]{0,11})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateCustomModel

Service: Amazon Bedrock

Creates a new custom model in Amazon Bedrock. After the model is active, you can use it for inference.

You can provide the model data source in one of the following ways:

- `customModelDataSource` — Specify a SageMaker AI model package ARN. Amazon Bedrock resolves the model package to retrieve the model artifacts. This is the preferred method for new SageMaker AI training outputs.
- `modelSourceConfig` — Specify an Amazon S3 URI pointing to the Amazon-managed Amazon S3 bucket containing your model artifacts.

To use the model for inference, you must purchase Provisioned Throughput for it. You can't use On-demand inference with these custom models. For more information about Provisioned Throughput, see Provisioned Throughput.

The model appears in `ListCustomModels` with a `customizationType` of `imported`. To track the status of the new model, you use the `GetCustomModel` API operation. The model can be in the following states:

- `Creating` - Initial state during validation and registration
- `Active` - Model is ready for use in inference
- `Failed` - Creation process encountered an error

Related APIs

- GetCustomModel
- ListCustomModels
- DeleteCustomModel

Request Syntax

```
POST /custom-models/create-custom-model HTTP/1.1
Content-type: application/json
```

```
{
   "clientRequestToken": "string",
   "customModelDataSource": { ... },
   "modelKmsKeyArn": "string",
   "modelName": "string",
   "modelSourceConfig": { ... },
   "modelTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "roleArn": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

customModelDataSource

The data source for the custom model. Use this field to specify a SageMaker AI model package ARN as the source for your custom model. Amazon Bedrock resolves the model package to retrieve the model artifacts.

You can specify either `customModelDataSource` or `modelSourceConfig`, but not both.

Type: CustomModelDataSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

modelKmsKeyArn

The Amazon Resource Name (ARN) of the customer managed AWS KMS key to encrypt the custom model. If you don't provide a AWS KMS key, Amazon Bedrock uses an AWS-managed AWS KMS key to encrypt the model.

If you provide a customer managed AWS KMS key, your Amazon Bedrock service role must have permissions to use it. For more information see Encryption of imported models.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

modelName

A unique name for the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

modelSourceConfig

The data source for the model. The Amazon S3 URI in the model source must be for the Amazon-managed Amazon S3 bucket containing your model artifacts.

Type: ModelDataSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

modelTags

A list of key-value pairs to associate with the custom model resource. You can use these tags to organize and identify your resources.

For more information, see Tagging resources in the Amazon Bedrock User Guide.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

roleArn

The Amazon Resource Name (ARN) of an IAM service role that Amazon Bedrock assumes to perform tasks on your behalf. This role must have permissions to access the Amazon S3 bucket containing your model artifacts and the AWS KMS key (if specified). For more information, see Setting up an IAM service role for importing models in the Amazon Bedrock User Guide.

This field is required when you use `modelSourceConfig` with an Amazon S3 data source. It is not required when you use `customModelDataSource` with a model package ARN, because Amazon Bedrock uses its own credentials to access the model artifacts.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json
```

```
{
   "modelArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

modelArn

The Amazon Resource Name (ARN) of the new custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request using a model package ARN

This example illustrates one usage of CreateCustomModel.

```
POST /custom-models/create-custom-model HTTP/1.1
Content-type: application/json

{
    "modelName": "my-custom-nova-model",
    "customModelDataSource": {
        "modelPackageArnDataSource": {
            "modelPackageArn": "arn:aws:sagemaker:us-east-1:123456789012:model-package/
my-model-package/1"
        }
    },
    "modelKmsKeyArn": "arn:aws:kms:us-
east-1:123456789012:key/12345678-1234-1234-1234-123456789012",
    "clientRequestToken": "unique-request-token-123",
    "modelTags": [
        {
            "key": "Environment",
            "value": "Production"
        },
        {
            "key": "Project",
            "value": "CustomNova"
        }
    ]
}
```

Example request using an S3 data source

This example illustrates one usage of CreateCustomModel.

```
POST /custom-models/create-custom-model HTTP/1.1
Content-type: application/json

{
    "modelName": "my-custom-nova-model",
    "modelSourceConfig": {
        "s3Uri": "s3://amzn-s3-demo-bucket/"
    },
    "roleArn": "arn:aws:iam::123456789012:role/BedrockCustomModelRole",
    "modelKmsKeyArn": "arn:aws:kms:us-
east-1:123456789012:key/12345678-1234-1234-1234-123456789012",
    "clientRequestToken": "unique-request-token-123",
    "modelTags": [
```

```
        {
            "key": "Environment",
            "value": "Production"
        },
        {
            "key": "Project",
            "value": "CustomNova"
        }
    ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateCustomModelDeployment

Service: Amazon Bedrock

Deploys a custom model for on-demand inference in Amazon Bedrock. After you deploy your custom model, you use the deployment's Amazon Resource Name (ARN) as the `modelId` parameter when you submit prompts and generate responses with model inference.

For more information about setting up on-demand inference for custom models, see Set up inference for a custom model.

The following actions are related to the `CreateCustomModelDeployment` operation:

- GetCustomModelDeployment
- ListCustomModelDeployments
- DeleteCustomModelDeployment

Request Syntax

```
POST /model-customization/custom-model-deployments HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "description": "string",
   "modelArn": "string",
   "modelDeploymentName": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

description

A description for the custom model deployment to help you identify its purpose.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

Required: No

modelArn

The Amazon Resource Name (ARN) of the custom model to deploy for on-demand inference. The custom model must be in the `Active` state.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

Required: Yes

modelDeploymentName

The name for the custom model deployment. The name must be unique within your AWS account and Region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

tags

Tags to assign to the custom model deployment. You can use tags to organize and track your AWS resources for cost allocation and management purposes.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "customModelDeploymentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

customModelDeploymentArn

The Amazon Resource Name (ARN) of the custom model deployment. Use this ARN as the `modelId` parameter when invoking the model with the `InvokeModel` or `Converse` operations.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of CreateCustomModelDeployment.

```
POST /model-customization/custom-model-deployments HTTP/1.1
Content-type: application/json

{
    "clientRequestToken": "unique-deployment-token-456",
    "description": "Production deployment of my custom model for customer support
 chatbot",
    "modelArn": "arn:aws:bedrock:us-west-2:123456789012:custom-model-deployment/
abc123def456",
    "modelDeploymentName": "customer-support-model-deployment",
    "tags": [
        {
            "key": "Environment",
            "value": "Production"
        },
        {
            "key": "Application",
            "value": "CustomerSupport"
        },
        {
            "key": "CostCenter",
            "value": "Engineering"
        }
    ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateEvaluationJob

Service: Amazon Bedrock

Creates an evaluation job.

For code examples that demonstrate how to create evaluation jobs, see Create a model evaluation job in Amazon Bedrock.

Request Syntax

```
POST /evaluation-jobs HTTP/1.1
Content-type: application/json

{
   "applicationType": "string",
   "clientRequestToken": "string",
   "customerEncryptionKeyId": "string",
   "evaluationConfig": { ... },
   "inferenceConfig": { ... },
   "jobDescription": "string",
   "jobName": "string",
   "jobTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "outputDataConfig": {
      "s3Uri": "string"
   },
   "roleArn": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

applicationType

Specifies whether the evaluation job is for evaluating a model or evaluating a knowledge base (retrieval and response generation).

Type: String

Valid Values: `ModelEvaluation | RagEvaluation`

Required: No

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

customerEncryptionKeyId

Specify your customer managed encryption key Amazon Resource Name (ARN) that will be used to encrypt your evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

evaluationConfig

Contains the configuration details of either an automated or human-based evaluation job.

Type: EvaluationConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

inferenceConfig

Contains the configuration details of the inference model for the evaluation job.

For model evaluation jobs, automated jobs support a single model or inference profile, and jobs that use human workers support two models or inference profiles.

Type: EvaluationInferenceConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

jobDescription

A description of the evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

Required: No

jobName

A name for the evaluation job. Names must unique with your AWS account, and your account's AWS region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

Required: Yes

jobTags

Tags to attach to the model evaluation job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

outputDataConfig

Contains the configuration details of the Amazon S3 bucket for storing the results of the evaluation job.

Type: EvaluationOutputDataConfig object

Required: Yes

roleArn

The Amazon Resource Name (ARN) of an IAM service role that Amazon Bedrock can assume to perform tasks on your behalf. To learn more about the required permissions, see Required permissions for model evaluations.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

jobArn

The Amazon Resource Name (ARN) of the evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateFoundationModelAgreement

Service: Amazon Bedrock

Request a model access agreement for the specified model.

Request Syntax

```
POST /create-foundation-model-agreement HTTP/1.1
Content-type: application/json

{
   "modelId": "string",
   "offerToken": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

modelId

Model Id of the model for the access request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

offerToken

An offer token encapsulates the information for an offer.

Type: String

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "modelId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

modelId

Model Id of the model for the access request.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateGuardrail

Service: Amazon Bedrock

Creates a guardrail to detect and filter harmful content in your generative AI application.

Amazon Bedrock Guardrails provides the following safeguards (also known as policies) to detect and filter harmful content:

- **Content filters** - Detect and filter harmful text or image content in input prompts or model responses. Filtering is done based on detection of certain predefined harmful content categories: Hate, Insults, Sexual, Violence, Misconduct and Prompt Attack. You also can adjust the filter strength for each of these categories.
- **Denied topics** - Define a set of topics that are undesirable in the context of your application. The filter will help block them if detected in user queries or model responses.
- **Word filters** - Configure filters to help block undesirable words, phrases, and profanity (exact match). Such words can include offensive terms, competitor names, etc.
- **Sensitive information filters** - Configure filters to help block or mask sensitive information, such as personally identifiable information (PII), or custom regex in user inputs and model responses. Blocking or masking is done based on probabilistic detection of sensitive information in standard formats in entities such as SSN number, Date of Birth, address, etc. This also allows configuring regular expression based detection of patterns for identifiers.
- **Contextual grounding check** - Help detect and filter hallucinations in model responses based on grounding in a source and relevance to the user query.

For more information, see How Amazon Bedrock Guardrails works.

Request Syntax

```
POST /guardrails HTTP/1.1
Content-type: application/json

{
   "automatedReasoningPolicyConfig": {
      "confidenceThreshold": number,
      "policies": [ "string" ]
   },
   "blockedInputMessaging": "string",
   "blockedOutputsMessaging": "string",
```

```
   "clientRequestToken": "string",
   "contentPolicyConfig": {
      "filtersConfig": [
         {
            "inputAction": "string",
            "inputEnabled": boolean,
            "inputModalities": [ "string" ],
            "inputStrength": "string",
            "outputAction": "string",
            "outputEnabled": boolean,
            "outputModalities": [ "string" ],
            "outputStrength": "string",
            "type": "string"
         }
      ],
      "tierConfig": {
         "tierName": "string"
      }
   },
   "contextualGroundingPolicyConfig": {
      "filtersConfig": [
         {
            "action": "string",
            "enabled": boolean,
            "threshold": number,
            "type": "string"
         }
      ]
   },
   "crossRegionConfig": {
      "guardrailProfileIdentifier": "string"
   },
   "description": "string",
   "kmsKeyId": "string",
   "name": "string",
   "sensitiveInformationPolicyConfig": {
      "piiEntitiesConfig": [
         {
            "action": "string",
            "inputAction": "string",
            "inputEnabled": boolean,
            "outputAction": "string",
            "outputEnabled": boolean,
            "type": "string"
```

```
            }
        ],
        "regexesConfig": [
            {
                "action": "string",
                "description": "string",
                "inputAction": "string",
                "inputEnabled": boolean,
                "name": "string",
                "outputAction": "string",
                "outputEnabled": boolean,
                "pattern": "string"
            }
        ]
    },
    "tags": [
        {
            "key": "string",
            "value": "string"
        }
    ],
    "topicPolicyConfig": {
        "tierConfig": {
            "tierName": "string"
        },
        "topicsConfig": [
            {
                "definition": "string",
                "examples": [ "string" ],
                "inputAction": "string",
                "inputEnabled": boolean,
                "name": "string",
                "outputAction": "string",
                "outputEnabled": boolean,
                "type": "string"
            }
        ]
    },
    "wordPolicyConfig": {
        "managedWordListsConfig": [
            {
                "inputAction": "string",
                "inputEnabled": boolean,
                "outputAction": "string",
```

```
            "outputEnabled": boolean,
            "type": "string"
         }
      ],
      "wordsConfig": [
         {
            "inputAction": "string",
            "inputEnabled": boolean,
            "outputAction": "string",
            "outputEnabled": boolean,
            "text": "string"
         }
      ]
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

automatedReasoningPolicyConfig

Optional configuration for integrating Automated Reasoning policies with the new guardrail.

Type: GuardrailAutomatedReasoningPolicyConfig object

Required: No

blockedInputMessaging

The message to return when the guardrail blocks a prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

blockedOutputsMessaging

The message to return when the guardrail blocks a model response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than once. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency in the *Amazon S3 User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

contentPolicyConfig

The content filter policies to configure for the guardrail.

Type: GuardrailContentPolicyConfig object

Required: No

contextualGroundingPolicyConfig

The contextual grounding policy configuration used to create a guardrail.

Type: GuardrailContextualGroundingPolicyConfig object

Required: No

crossRegionConfig

The system-defined guardrail profile that you're using with your guardrail. Guardrail profiles define the destination AWS Regions where guardrail inference requests can be automatically routed.

For more information, see the Amazon Bedrock User Guide.

Type: GuardrailCrossRegionConfig object

Required: No

description

A description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

kmsKeyId

The ARN of the AWS KMS key that you use to encrypt the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

name

The name to give the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `[0-9a-zA-Z-_]+`

Required: Yes

sensitiveInformationPolicyConfig

The sensitive information policy to configure for the guardrail.

Type: GuardrailSensitiveInformationPolicyConfig object

Required: No

tags

The tags that you want to attach to the guardrail.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

topicPolicyConfig

The topic policies to configure for the guardrail.

Type: GuardrailTopicPolicyConfig object

Required: No

wordPolicyConfig

The word policy you configure for the guardrail.

Type: GuardrailWordPolicyConfig object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "createdAt": "string",
   "guardrailArn": "string",
   "guardrailId": "string",
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

createdAt

The time at which the guardrail was created.

Type: Timestamp

guardrailArn

The ARN of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+`

guardrailId

The unique identifier of the guardrail that was created.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

version

The version of the guardrail that was created. This value will always be `DRAFT`.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateGuardrailVersion

Service: Amazon Bedrock

Creates a version of the guardrail. Use this API to create a snapshot of the guardrail when you are satisfied with a configuration, or to compare the configuration with another version.

Request Syntax

```
POST /guardrails/guardrailIdentifier HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "description": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail. This can be an ID or the ARN.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than once. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency in the *Amazon S3 User Guide*.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

description

A description of the guardrail version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "guardrailId": "string",
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

guardrailId

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

version

The number of the version of the guardrail.

Type: String

Pattern: `[1-9][0-9]{0,7}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateInferenceProfile

Service: Amazon Bedrock

Creates an application inference profile to track metrics and costs when invoking a model. To create an application inference profile for a foundation model in one region, specify the ARN of the model in that region. To create an application inference profile for a foundation model across multiple regions, specify the ARN of the system-defined inference profile that contains the regions that you want to route requests to. For more information, see Increase throughput and resilience with cross-region inference in Amazon Bedrock. in the Amazon Bedrock User Guide.

Request Syntax

```
POST /inference-profiles HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "description": "string",
   "inferenceProfileName": "string",
   "modelSource": { ... },
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

description

A description for the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `([0-9a-zA-Z:.][ _-]?)+`

Required: No

inferenceProfileName

A name for the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `([0-9a-zA-Z][ _-]?)+`

Required: Yes

modelSource

The foundation model or system-defined inference profile that the inference profile will track metrics and costs for.

Type: InferenceProfileModelSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

tags

An array of objects, each of which contains a tag and its value. For more information, see Tagging resources in the Amazon Bedrock User Guide.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "inferenceProfileArn": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

inferenceProfileArn

The ARN of the inference profile that you created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+`

status

The status of the inference profile. `ACTIVE` means that the inference profile is ready to be used.

Type: String

Valid Values: `ACTIVE`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Create an application inference profile from a foundation model

Run the following example to create an application inference profile from the Anthropic Claude 3 Sonnet model in the `us-west-2` region, replacing `${AccountId}` with your AWS account ID:

Sample Request

```
POST /inference-profiles HTTP/1.1
Content-type: application/json

{
   "inferenceProfileName": "USClaudeSonnetApplicationIP",
   "modelSource": {
      "copyFrom": "arn:aws:bedrock:us-west-2::foundation-model/anthropic.claude-3-
sonnet-20240229-v1:0"
   },
   "tags": [
      {
         "key": "projectId",
         "value": "abcdef123456"
      }
   ]
}
```

Create an application inference profile from a cross-region (system-defined) inference profile

Run the following example to create an application inference profile from the US Anthropic Claude 3 Sonnet inference profile in the `us-west-2` region, replacing `${AccountId}` with your AWS account ID:

Sample Request

```
POST /inference-profiles HTTP/1.1
Content-type: application/json

{
   "inferenceProfileName": "USClaudeSonnetApplicationIP",
   "modelSource": {
      "copyFrom": "arn:aws:bedrock:us-west-2:${AccountId}:inference-profile/
us.anthropic.claude-3-sonnet-20240229-v1:0"
   },
   "tags": [
      {
         "key": "projectId",
         "value": "abcdef123456"
      }
   ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateMarketplaceModelEndpoint

Service: Amazon Bedrock

Creates an endpoint for a model from Amazon Bedrock Marketplace. The endpoint is hosted by Amazon SageMaker.

Request Syntax

```
POST /marketplace-model/endpoints HTTP/1.1
Content-type: application/json

{
   "acceptEula": boolean,
   "clientRequestToken": "string",
   "endpointConfig": { ... },
   "endpointName": "string",
   "modelSourceIdentifier": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

acceptEula

Indicates whether you accept the end-user license agreement (EULA) for the model. Set to `true` to accept the EULA.

Type: Boolean

Required: No

clientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. This token is listed as not required because AWS SDKs automatically generate it for you and set this parameter. If you're not using the AWS SDK or the AWS CLI, you must provide this token or the action will fail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

endpointConfig

The configuration for the endpoint, including the number and type of instances to use.

Type: EndpointConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

endpointName

The name of the endpoint. This name must be unique within your AWS account and region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 30.

Required: Yes

modelSourceIdentifier

The ARN of the model from Amazon Bedrock Marketplace that you want to deploy to the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*arn:aws:sagemaker:.*:hub-content/SageMakerPublicHub/Model/.*`

Required: Yes

tags

An array of key-value pairs to apply to the underlying Amazon SageMaker endpoint. You can use these tags to organize and identify your AWS resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "marketplaceModelEndpoint": {
      "createdAt": "string",
      "endpointArn": "string",
      "endpointConfig": { ... },
      "endpointStatus": "string",
      "endpointStatusMessage": "string",
      "modelSourceIdentifier": "string",
      "status": "string",
      "statusMessage": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

marketplaceModelEndpoint

Details about the created endpoint.

Type: MarketplaceModelEndpoint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateModelCopyJob

Service: Amazon Bedrock

Copies a model to another region so that it can be used there. For more information, see Copy models to be used in other regions in the Amazon Bedrock User Guide.

Request Syntax

```
POST /model-copy-jobs HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "modelKmsKeyId": "string",
   "sourceModelArn": "string",
   "targetModelName": "string",
   "targetModelTags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

modelKmsKeyId

The ARN of the AWS KMS key that you use to encrypt the model copy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

sourceModelArn

The Amazon Resource Name (ARN) of the model to be copied.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

targetModelName

A name for the copied model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

targetModelTags

Tags to associate with the target model. For more information, see Tag resources in the Amazon Bedrock User Guide.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

jobArn

The Amazon Resource Name (ARN) of the model copy job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-copy-job/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

Examples

Create a model copy job (CLI)

The following example shows how to copy a custom model from `us-west-2` into the region from which the request is made, using the AWS CLI.

```
aws bedrock create-model-copy-job --source-model-arn arn:aws:bedrock:us-
west-2:123456789012:custom-model/amazon.titan-text-lite-v1:0:4k/MyCustomModel --target-
model-name MyCustomModelCopy
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateModelCustomizationJob

Service: Amazon Bedrock

Creates a fine-tuning job to customize a base model.

You specify the base foundation model and the location of the training data. After the model-customization job completes successfully, your custom model resource will be ready to use. Amazon Bedrock returns validation loss metrics and output generations after the job completes.

For information on the format of training and validation data, see Prepare the datasets.

Model-customization jobs are asynchronous and the completion time depends on the base model and the training/validation data size. To monitor a job, use the `GetModelCustomizationJob` operation to retrieve the job status.

For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
POST /model-customization-jobs HTTP/1.1
Content-type: application/json

{
   "baseModelIdentifier": "string",
   "clientRequestToken": "string",
   "customizationConfig": { ... },
   "customizationType": "string",
   "customModelKmsKeyId": "string",
   "customModelName": "string",
   "customModelTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "hyperParameters": {
      "string" : "string"
   },
   "jobName": "string",
   "jobTags": [
      {
         "key": "string",
```

```
            "value": "string"
         }
      ],
      "outputDataConfig": {
         "s3Uri": "string"
      },
      "roleArn": "string",
      "trainingDataConfig": {
         "invocationLogsConfig": {
            "invocationLogSource": { ... },
            "requestMetadataFilters": { ... },
            "usePromptResponse": boolean
         },
         "s3Uri": "string"
      },
      "validationDataConfig": {
         "validators": [
            {
               "s3Uri": "string"
            }
         ]
      },
      "vpcConfig": {
         "securityGroupIds": [ "string" ],
         "subnetIds": [ "string" ]
      }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

baseModelIdentifier

Name of the base model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

customizationConfig

The customization configuration for the model customization job.

Type: CustomizationConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

customizationType

The customization type.

Type: String

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION`

Required: No

customModelKmsKeyId

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

customModelName

A name for the resulting custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

customModelTags

Tags to attach to the resulting custom model.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

hyperParameters

Parameters related to tuning the model. For details on the format for different models, see Custom model hyperparameters.

Type: String to string map

Required: No

jobName

A name for the fine-tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

Required: Yes

jobTags

Tags to attach to the job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

outputDataConfig

S3 location for the output data.

Type: OutputDataConfig object

Required: Yes

roleArn

The Amazon Resource Name (ARN) of an IAM service role that Amazon Bedrock can assume to perform tasks on your behalf. For example, during model training, Amazon Bedrock needs your permission to read input data from an S3 bucket, write model artifacts to an S3 bucket. To pass this role to Amazon Bedrock, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

trainingDataConfig

Information about the training dataset.

Type: TrainingDataConfig object

Required: Yes

validationDataConfig

Information about the validation dataset.

Type: ValidationDataConfig object

Required: No

vpcConfig

The configuration of the Virtual Private Cloud (VPC) that contains the resources that you're using for this job. For more information, see Protect your model customization jobs using a VPC.

Type: VpcConfig object

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

jobArn

Amazon Resource Name (ARN) of the fine tuning job

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateModelImportJob

Service: Amazon Bedrock

Creates a model import job to import model that you have customized in other environments, such as Amazon SageMaker. For more information, see Import a customized model

Request Syntax

```
POST /model-import-jobs HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "importedModelKmsKeyId": "string",
   "importedModelName": "string",
   "importedModelTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "jobName": "string",
   "jobTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "modelDataSource": { ... },
   "roleArn": "string",
   "vpcConfig": {
      "securityGroupIds": [ "string" ],
      "subnetIds": [ "string" ]
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

importedModelKmsKeyId

The imported model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

importedModelName

The name of the imported model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Required: Yes

importedModelTags

Tags to attach to the imported model.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

jobName

The name of the import job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

Required: Yes

jobTags

Tags to attach to this import job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

modelDataSource

The data source for the imported model.

Type: ModelDataSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the service role with the proper custom model import permissions. You can use the console to automatically create a service role with the proper permissions or create a custom service role. For more information about creating a custom service role, see Create a service role for model import.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

vpcConfig

VPC configuration parameters for the private Virtual Private Cloud (VPC) that contains the resources you are using for the import job.

Type: VpcConfig object

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

jobArn

The Amazon Resource Name (ARN) of the model import job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-import-job/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Create a model import job

Creates a model import job.

```
POST /model-import-jobs/ HTTP/1.1
Content-type: application/json
{
    "importedModelName": "ImportedModelName",
    "jobName": "MyImportedModelJobName",
    "modelDataSource": {
        "s3DataSource": {
            "s3Uri": "S3://amzn-s3-demo-bucket/key-name"
        }
    },
    "roleArn": "arn:aws:iam::111122223333:role/Role_Name"
}
```

Example response

Response for the above request.

```
HTTP/1.1 200
Content-type: application/json

{
    "jobArn": "arn:aws:bedrock:us-east-1:111122223333:model-import-job/yggb47n4xlml"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateModelInvocationJob

Service: Amazon Bedrock

Creates a batch inference job to invoke a model on multiple prompts. Format your data according to Format your inference data and upload it to an Amazon S3 bucket. For more information, see Process multiple prompts with batch inference.

The response returns a `jobArn` that you can use to stop or get details about the job.

Request Syntax

```
POST /model-invocation-job HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "inputDataConfig": { ... },
   "jobName": "string",
   "modelId": "string",
   "modelInvocationType": "string",
   "outputDataConfig": { ... },
   "roleArn": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "timeoutDurationInHours": number,
   "vpcConfig": {
      "securityGroupIds": [ "string" ],
      "subnetIds": [ "string" ]
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]{1,256}(-*[a-zA-Z0-9]){0,256}`

Required: No

inputDataConfig

Details about the location of the input to the batch inference job.

Type: ModelInvocationJobInputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

jobName

A name to give the batch inference job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]{1,63}(-*[a-zA-Z0-9\+\-\.]){0,63}`

Required: Yes

modelId

The unique identifier of the foundation model to use for the batch inference job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-:]{1,63}/[a-z0-9]{12}$)|(:foundation-`

```
model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:]
[a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$)|([0-9]{12}:(inference-
profile|application-inference-profile)/[a-zA-Z0-9-:.]+$)))|([a-z0-9-]
{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63})
{0,2})|(([0-9a-zA-Z][_-]?)+)
```

Required: Yes

modelInvocationType

The invocation endpoint for ModelInvocationJob

Type: String

Valid Values: `InvokeModel | Converse`

Required: No

outputDataConfig

Details about the location of the output of the batch inference job.

Type: ModelInvocationJobOutputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the service role with permissions to carry out and manage batch inference. You can use the console to create a default service role or follow the steps at Create a service role for batch inference.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

tags

Any tags to associate with the batch inference job. For more information, see Tagging Amazon Bedrock resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

timeoutDurationInHours

The number of hours after which to force the batch inference job to time out.

Type: Integer

Valid Range: Minimum value of 24. Maximum value of 168.

Required: No

vpcConfig

The configuration of the Virtual Private Cloud (VPC) for the data in the batch inference job. For more information, see Protect batch inference jobs using a VPC.

Type: VpcConfig object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

jobArn

The Amazon Resource Name (ARN) of the batch inference job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-invocation-job/[a-z0-9]{12})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Create a batch inference job

This example illustrates one usage of CreateModelInvocationJob.

```
POST /model-invocation-job HTTP/1.1
Content-type: application/json

{
    "clientRequestToken": "string",
    "inputDataConfig": {
        "s3InputDataConfig": {
            "s3Uri": "s3://input-bucket/abc.jsonl"
        }
    },
    "jobName": "my-batch-job",
    "modelId": "anthropic.claude-3-haiku-20240307-v1:0",
    "outputDataConfig": {
        "s3OutputDataConfig": {
            "s3Uri": "s3://output-bucket/"
        }
    },
    "roleArn": "arn:aws:iam::123456789012:role/MyBatchInferenceRole"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreatePromptRouter

Service: Amazon Bedrock

Creates a prompt router that manages the routing of requests between multiple foundation models based on the routing criteria.

Request Syntax

```
POST /prompt-routers HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "description": "string",
   "fallbackModel": {
      "modelArn": "string"
   },
   "models": [
      {
         "modelArn": "string"
      }
   ],
   "promptRouterName": "string",
   "routingCriteria": {
      "responseQualityDifference": number
   },
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier that you provide to ensure idempotency of your requests. If not specified, the AWS SDK automatically generates one for you.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

description

An optional description of the prompt router to help identify its purpose.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `([0-9a-zA-Z:.][ _-]?)+`

Required: No

fallbackModel

The default model to use when the routing criteria is not met.

Type: PromptRouterTargetModel object

Required: Yes

models

A list of foundation models that the prompt router can route requests to. At least one model must be specified.

Type: Array of PromptRouterTargetModel objects

Required: Yes

promptRouterName

The name of the prompt router. The name must be unique within your AWS account in the current region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `([0-9a-zA-Z][ _-]?)+`

Required: Yes

routingCriteria

The criteria, which is the response quality difference, used to determine how incoming requests are routed to different models.

Type: RoutingCriteria object

Required: Yes

tags

An array of key-value pairs to apply to this resource as tags. You can use tags to categorize and manage your AWS resources.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "promptRouterArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

promptRouterArn

The Amazon Resource Name (ARN) that uniquely identifies the prompt router.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:default-prompt-router/[a-zA-Z0-9-:.]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateProvisionedModelThroughput

Service: Amazon Bedrock

Creates dedicated throughput for a base or custom model with the model units and for the duration that you specify. For pricing details, see Amazon Bedrock Pricing. For more information, see Provisioned Throughput in the Amazon Bedrock User Guide.

Request Syntax

```
POST /provisioned-model-throughput HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "commitmentDuration": "string",
   "modelId": "string",
   "modelUnits": number,
   "provisionedModelName": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency in the Amazon S3 User Guide.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

commitmentDuration

The commitment duration requested for the Provisioned Throughput. Billing occurs hourly and is discounted for longer commitment terms. To request a no-commit Provisioned Throughput, omit this field.

Custom models support all levels of commitment. To see which base models support no commitment, see Supported regions and models for Provisioned Throughput in the Amazon Bedrock User Guide

Type: String

Valid Values: `OneMonth | SixMonths`

Required: No

modelId

The model ID (for a foundation model), name (for a custom model), or Amazon Resource Name (ARN) of the model to associate with this Provisioned Throughput. For a list of models for which you can purchase Provisioned Throughput, see Amazon Bedrock model IDs for purchasing Provisioned Throughput in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

modelUnits

Number of model units to allocate. A model unit delivers a specific throughput level for the specified model. The throughput level of a model unit specifies the total number of input and output tokens that it can process and generate within a span of one minute. By default, your account has no model units for purchasing Provisioned Throughputs with commitment. You must first visit the AWS support center to request MUs.

For model unit quotas, see Provisioned Throughput quotas in the Amazon Bedrock User Guide.

For more information about what an MU specifies, contact your AWS account manager.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

provisionedModelName

The name for this Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Required: Yes

tags

Tags to associate with this Provisioned Throughput.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
```

```
Content-type: application/json

{
   "provisionedModelArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

provisionedModelArn

The Amazon Resource Name (ARN) for this Provisioned Throughput.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAutomatedReasoningPolicy

Service: Amazon Bedrock

Deletes an Automated Reasoning policy or policy version. This operation is idempotent. If you delete a policy more than once, each call succeeds. Deleting a policy removes it permanently and cannot be undone.

Request Syntax

```
DELETE /automated-reasoning-policies/policyArn?force=force HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

force

Specifies whether to force delete the automated reasoning policy even if it has active resources. When `false`, Amazon Bedrock validates if all artifacts have been deleted (e.g. policy version, test case, test result) for a policy before deletion. When `true`, Amazon Bedrock will delete the policy and all its artifacts without validation. Default is `false`.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAutomatedReasoningPolicyBuildWorkflow

Service: Amazon Bedrock

Deletes an Automated Reasoning policy build workflow and its associated artifacts. This permanently removes the workflow history and any generated assets.

Request Syntax

```
DELETE /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId?
updatedAt=lastUpdatedAt HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow to delete.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

lastUpdatedAt

The timestamp when the build workflow was last updated. This is used for optimistic concurrency control to prevent accidental deletion of workflows that have been modified.

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose build workflow you want to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAutomatedReasoningPolicyTestCase

Service: Amazon Bedrock

Deletes an Automated Reasoning policy test. This operation is idempotent; if you delete a test more than once, each call succeeds.

Request Syntax

```
DELETE /automated-reasoning-policies/policyArn/test-cases/testCaseId?
updatedAt=lastUpdatedAt HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

lastUpdatedAt

The timestamp when the test was last updated. This is used as a concurrency token to prevent conflicting modifications.

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy that contains the test.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

testCaseId

The unique identifier of the test to delete.

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteCustomModel

Service: Amazon Bedrock

Deletes a custom model that you created earlier. For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /custom-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

Name of the model to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteCustomModelDeployment

Service: Amazon Bedrock

Deletes a custom model deployment. This operation stops the deployment and removes it from your account. After deletion, the deployment ARN can no longer be used for inference requests.

The following actions are related to the `DeleteCustomModelDeployment` operation:

- CreateCustomModelDeployment
- GetCustomModelDeployment
- ListCustomModelDeployments

Request Syntax

```
DELETE /model-customization/custom-model-deployments/customModelDeploymentIdentifier
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

customModelDeploymentIdentifier

The Amazon Resource Name (ARN) or name of the custom model deployment to delete.

Length Constraints: Minimum length of 1. Maximum length of 93.

Pattern: `(arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12})|^([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteEnforcedGuardrailConfiguration

Service: Amazon Bedrock

Deletes the account-level enforced guardrail configuration.

Request Syntax

```
DELETE /enforcedGuardrailsConfiguration/configId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

configId

Unique ID for the account enforced configuration.

Pattern: `[a-z0-9]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

DeleteFoundationModelAgreement

Service: Amazon Bedrock

Delete the model access agreement for the specified model.

Request Syntax

```
POST /delete-foundation-model-agreement HTTP/1.1
Content-type: application/json

{
   "modelId": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

modelId

Model Id of the model access to delete.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteGuardrail

Service: Amazon Bedrock

Deletes a guardrail.

- To delete a guardrail, only specify the ARN of the guardrail in the `guardrailIdentifier` field. If you delete a guardrail, all of its versions will be deleted.
- To delete a version of a guardrail, specify the ARN of the guardrail in the `guardrailIdentifier` field and the version in the `guardrailVersion` field.

Request Syntax

```
DELETE /guardrails/guardrailIdentifier?guardrailVersion=guardrailVersion HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail. This can be an ID or the ARN.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

guardrailVersion

The version of the guardrail.

Pattern: `[1-9][0-9]{0,7}`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteImportedModel

Service: Amazon Bedrock

Deletes a custom model that you imported earlier. For more information, see Import a customized model in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /imported-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

Name of the imported model to delete.

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:imported-model/[a-z0-9]{12})|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Delete an imported model

Deletes a model that you have previously imported.

```
DELETE /imported-models/{modelIdentifier} HTTP/1.1
Content-type: application/json
```

Example response

Response for the above request.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteInferenceProfile

Service: Amazon Bedrock

Deletes an application inference profile. For more information, see Increase throughput and resilience with cross-region inference in Amazon Bedrock. in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /inference-profiles/inferenceProfileIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

inferenceProfileIdentifier

The Amazon Resource Name (ARN) or ID of the application inference profile to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|application-inference-profile)/)?[a-zA-Z0-9-:.]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Delete an application inference profile

Assuming you've created an application inference profile called `USClaudeSonnetApplicationIP`, run the following example to delete it:

Sample Request

```
DELETE /inference-profiles/USClaudeSonnetApplicationIP HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteMarketplaceModelEndpoint

Service: Amazon Bedrock

Deletes an endpoint for a model from Amazon Bedrock Marketplace.

Request Syntax

```
DELETE /marketplace-model/endpoints/endpointArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

endpointArn

The Amazon Resource Name (ARN) of the endpoint you want to delete.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

DeleteModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Delete the invocation logging.

Request Syntax

```
DELETE /logging/modelinvocations HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeletePromptRouter

Service: Amazon Bedrock

Deletes a specified prompt router. This action cannot be undone.

Request Syntax

```
DELETE /prompt-routers/promptRouterArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

promptRouterArn

The Amazon Resource Name (ARN) of the prompt router to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:default-prompt-router/[a-zA-Z0-9-:.]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteProvisionedModelThroughput

Service: Amazon Bedrock

Deletes a Provisioned Throughput. You can't delete a Provisioned Throughput before the commitment term is over. For more information, see Provisioned Throughput in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /provisioned-model-throughput/provisionedModelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId

The Amazon Resource Name (ARN) or ID of the Provisioned Throughput.

Pattern: `(([0-9a-zA-Z][_-]?)+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteResourcePolicy

Service: Amazon Bedrock

Deletes a previously created Bedrock resource policy.

Request Syntax

```
DELETE /resource-policy/resourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The ARN of the Bedrock resource to which this resource policy applies.

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

DeregisterMarketplaceModelEndpoint

Service: Amazon Bedrock

Deregisters an endpoint for a model from Amazon Bedrock Marketplace. This operation removes the endpoint's association with Amazon Bedrock but does not delete the underlying Amazon SageMaker endpoint.

Request Syntax

```
DELETE /marketplace-model/endpoints/endpointArn/registration HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

endpointArn

The Amazon Resource Name (ARN) of the endpoint you want to deregister.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceUnavailableException

Returned if the service cannot complete the request.

HTTP Status Code: 503

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ExportAutomatedReasoningPolicyVersion

Service: Amazon Bedrock

Exports the policy definition for an Automated Reasoning policy version. Returns the complete policy definition including rules, variables, and custom variable types in a structured format.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/export HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy to export. Can be either the unversioned ARN for the draft policy or a versioned ARN for a specific policy version.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "rules": [
      {
         "alternateExpression": "string",
         "expression": "string",
```

```
            "id": "string"
         }
      ],
      "types": [
         {
            "description": "string",
            "name": "string",
            "values": [
               {
                  "description": "string",
                  "value": "string"
               }
            ]
         }
      ],
      "variables": [
         {
            "description": "string",
            "name": "string",
            "type": "string"
         }
      ],
      "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

rules

The formal logic rules extracted from the source document. Rules define the logical constraints that determine whether model responses are valid, invalid, or satisfiable.

Type: Array of AutomatedReasoningPolicyDefinitionRule objects

Array Members: Minimum number of 0 items. Maximum number of 1500 items.

types

The custom user-defined vairable types used in the policy. Types are enum-based variable types that provide additional context beyond the predefined variable types.

Type: Array of AutomatedReasoningPolicyDefinitionType objects

Array Members: Minimum number of 0 items. Maximum number of 150 items.

variables

The variables that represent concepts in the policy. Variables can have values assigned when translating natural language into formal logic. Their descriptions are crucial for accurate translation.

Type: Array of AutomatedReasoningPolicyDefinitionVariable objects

Array Members: Minimum number of 0 items. Maximum number of 600 items.

version

The version of the policy definition format.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAccountDataRetention

Service: Amazon Bedrock

Returns the account-wide data retention mode for Amazon Bedrock.

Request Syntax

```
GET /data-retention HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "mode": "string",
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

mode

The data retention mode configured for the account.

Type: String

Valid Values: `default | none | aws_review | provider_data_share | inherit`

updatedAt

The time at which the data retention mode was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAdvancedPromptOptimizationJob

Service: Amazon Bedrock

Gets information about an advanced prompt optimization job.

Request Syntax

```
GET /advanced-prompt-optimization-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The ARN or ID of the advanced prompt optimization job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/)?[a-z0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "creationTime": "string",
   "encryptionKeyArn": "string",
   "failureMessage": "string",
   "inputConfig": {
      "s3Uri": "string"
   },
   "jobArn": "string",
```

```
   "jobDescription": "string",
   "jobName": "string",
   "jobStatus": "string",
   "lastModifiedTime": "string",
   "modelConfigurations": [
      {
         "additionalModelRequestFields": {
            "string" : JSON value
         },
         "inferenceConfig": {
            "maxTokens": number,
            "stopSequences": [ "string" ],
            "temperature": number,
            "topP": number
         },
         "modelId": "string"
      }
   ],
   "outputConfig": {
      "s3Uri": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

creationTime

The time at which the job was created.

Type: Timestamp

encryptionKeyArn

The ARN of the KMS key used to encrypt the output data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

failureMessage

If the job failed, a message describing the failure.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

inputConfig

The input data configuration for the optimization job.

Type: AdvancedPromptOptimizationInputConfig object

jobArn

The ARN of the advanced prompt optimization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/[a-z0-9]{12}`

jobDescription

The description of the advanced prompt optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

jobName

The name of the advanced prompt optimization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9.+-]*`

jobStatus

The status of the advanced prompt optimization job.

Type: String

Valid Values: `InProgress | Completed | Failed | PartiallyCompleted | Stopping | Stopped | Deleting`

lastModifiedTime

The time at which the job was last modified.

Type: Timestamp

modelConfigurations

The model configurations used in the optimization job.

Type: Array of ModelConfiguration objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

outputConfig

The output data configuration for the optimization job.

Type: AdvancedPromptOptimizationOutputConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAutomatedReasoningPolicy

Service: Amazon Bedrock

Retrieves details about an Automated Reasoning policy or policy version. Returns information including the policy definition, metadata, and timestamps.

Request Syntax

```
GET /automated-reasoning-policies/policyArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy to retrieve. Can be either the unversioned ARN for the draft policy or an ARN for a specific policy version.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "definitionHash": "string",
   "description": "string",
   "kmsKeyArn": "string",
```

```
    "name": "string",
    "policyArn": "string",
    "policyId": "string",
    "updatedAt": "string",
    "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp when the policy was created.

Type: Timestamp

definitionHash

The hash of the policy definition used as a concurrency token.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

description

The description of the policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

kmsKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key used to encrypt the automated reasoning policy and its associated artifacts. If a AWS KMS key is not provided during the initial CreateAutomatedReasoningPolicyRequest, the kmsKeyArn won't be included in the GetAutomatedReasoningPolicyResponse.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

name

The name of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

policyArn

The Amazon Resource Name (ARN) of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

policyId

The unique identifier of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[a-z0-9]{12}`

updatedAt

The timestamp when the policy was last updated.

Type: Timestamp

version

The version of the policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `([1-9][0-9]{0,11})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAutomatedReasoningPolicyAnnotations

Service: Amazon Bedrock

Retrieves the current annotations for an Automated Reasoning policy build workflow. Annotations contain corrections to the rules, variables and types to be applied to the policy.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/annotations
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow whose annotations you want to retrieve.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose annotations you want to retrieve.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "annotations": [
      { ... }
   ],
   "annotationSetHash": "string",
   "buildWorkflowId": "string",
   "name": "string",
   "policyArn": "string",
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

annotations

The current set of annotations containing rules, variables, and types extracted from the source documents. These can be modified before finalizing the policy.

Type: Array of AutomatedReasoningPolicyAnnotation objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

annotationSetHash

A hash value representing the current state of the annotations. This is used for optimistic concurrency control when updating annotations.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

buildWorkflowId

The unique identifier of the build workflow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

name

The name of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

updatedAt

The timestamp when the annotations were last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAutomatedReasoningPolicyBuildWorkflow

Service: Amazon Bedrock

Retrieves detailed information about an Automated Reasoning policy build workflow, including its status, configuration, and metadata.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow to retrieve.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose build workflow you want to retrieve.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "buildWorkflowId": "string",
   "buildWorkflowType": "string",
   "createdAt": "string",
   "documentContentType": "string",
   "documentDescription": "string",
   "documentName": "string",
   "policyArn": "string",
   "status": "string",
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

buildWorkflowId

The unique identifier of the build workflow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

buildWorkflowType

The type of build workflow being executed (e.g., DOCUMENT_INGESTION, POLICY_REPAIR).

Type: String

Valid Values: `INGEST_CONTENT | REFINE_POLICY | IMPORT_POLICY | GENERATE_FIDELITY_REPORT | GENERATE_POLICY_SCENARIOS | RESOLVE_POLICY_AMBIGUITIES | ITERATIVELY_REFINE_POLICY`

createdAt

The timestamp when the build workflow was created.

Type: Timestamp

documentContentType

The content type of the source document (e.g., text/plain, application/pdf).

Type: String

Valid Values: `pdf | txt`

documentDescription

A detailed description of the document's content and how it should be used in the policy generation process.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4000.

documentName

The name of the source document used in the build workflow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

status

The current status of the build workflow (e.g., RUNNING, COMPLETED, FAILED, CANCELLED).

Type: String

Valid Values: `SCHEDULED | CANCEL_REQUESTED | PREPROCESSING | BUILDING | TESTING | COMPLETED | FAILED | CANCELLED`

updatedAt

The timestamp when the build workflow was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAutomatedReasoningPolicyBuildWorkflowResultAssets

Service: Amazon Bedrock

Retrieves the resulting assets from a completed Automated Reasoning policy build workflow, including build logs, quality reports, and generated policy artifacts.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/result-assets?assetId=assetId&assetType=assetType HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

assetId

The unique identifier of the specific asset to retrieve when multiple assets of the same type exist. This is required when retrieving SOURCE_DOCUMENT assets, as multiple source documents may have been used in the workflow. The asset ID can be obtained from the asset manifest.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[0-9a-fA-F\-]+`

assetType

The type of asset to retrieve (e.g., BUILD_LOG, QUALITY_REPORT, POLICY_DEFINITION, GENERATED_TEST_CASES, POLICY_SCENARIOS, FIDELITY_REPORT, ASSET_MANIFEST, SOURCE_DOCUMENT).

Valid Values: `BUILD_LOG | QUALITY_REPORT | POLICY_DEFINITION | GENERATED_TEST_CASES | POLICY_SCENARIOS | FIDELITY_REPORT | ASSET_MANIFEST | SOURCE_DOCUMENT`

Required: Yes

buildWorkflowId

The unique identifier of the build workflow whose result assets you want to retrieve.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose build workflow assets you want to retrieve.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "buildWorkflowAssets": { ... },
   "buildWorkflowId": "string",
   "policyArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

buildWorkflowAssets

The requested build workflow asset. This is a union type that returns only one of the available asset types (logs, reports, or generated artifacts) based on the specific asset type requested in the API call.

Type: AutomatedReasoningPolicyBuildResultAssets object

Note: This object is a Union. Only one member of this object can be specified or returned.

buildWorkflowId

The unique identifier of the build workflow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAutomatedReasoningPolicyNextScenario

Service: Amazon Bedrock

Retrieves the next test scenario for validating an Automated Reasoning policy. This is used during the interactive policy refinement process to test policy behavior.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/scenarios
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow associated with the test scenarios.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy for which you want to get the next test scenario.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "policyArn": "string",
   "scenario": {
      "alternateExpression": "string",
      "expectedResult": "string",
      "expression": "string",
      "ruleIds": [ "string" ]
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

scenario

The next test scenario to validate, including the test expression and expected results.

Type: AutomatedReasoningPolicyScenario object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAutomatedReasoningPolicyTestCase

Service: Amazon Bedrock

Retrieves details about a specific Automated Reasoning policy test.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/test-cases/testCaseId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy that contains the test.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

testCaseId

The unique identifier of the test to retrieve.

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```

```
{
   "policyArn": "string",
   "testCase": {
      "confidenceThreshold": number,
      "createdAt": "string",
      "expectedAggregatedFindingsResult": "string",
      "guardContent": "string",
      "queryContent": "string",
      "testCaseId": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

policyArn

The Amazon Resource Name (ARN) of the policy that contains the test.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

testCase

The test details including the content, query, expected result, and metadata.

Type: AutomatedReasoningPolicyTestCase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

GetAutomatedReasoningPolicyTestResult

Service: Amazon Bedrock

Retrieves the test result for a specific Automated Reasoning policy test. Returns detailed validation findings and execution status.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/test-cases/testCaseId/test-results HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The build workflow identifier. The build workflow must display a COMPLETED status to get results.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

testCaseId

The unique identifier of the test for which to retrieve results.

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "testResult": {
      "aggregatedTestFindingsResult": "string",
      "policyArn": "string",
      "testCase": {
         "confidenceThreshold": number,
         "createdAt": "string",
         "expectedAggregatedFindingsResult": "string",
         "guardContent": "string",
         "queryContent": "string",
         "testCaseId": "string",
         "updatedAt": "string"
      },
      "testFindings": [
         { ... }
      ],
      "testRunResult": "string",
      "testRunStatus": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

testResult

The test result containing validation findings, execution status, and detailed analysis.

Type: AutomatedReasoningPolicyTestResult object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetCustomModel

Service: Amazon Bedrock

Get the properties associated with a Amazon Bedrock custom model that you have created. For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
GET /custom-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

Name or Amazon Resource Name (ARN) of the custom model.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "baseModelArn": "string",
   "creationTime": "string",
```

```
    "customizationConfig": { ... },
    "customizationType": "string",
    "failureMessage": "string",
    "hyperParameters": {
       "string" : "string"
    },
    "jobArn": "string",
    "jobName": "string",
    "modelArn": "string",
    "modelKmsKeyArn": "string",
    "modelName": "string",
    "modelStatus": "string",
    "outputDataConfig": {
       "s3Uri": "string"
    },
    "trainingDataConfig": {
       "invocationLogsConfig": {
          "invocationLogSource": { ... },
          "requestMetadataFilters": { ... },
          "usePromptResponse": boolean
       },
       "s3Uri": "string"
    },
    "trainingMetrics": {
       "trainingLoss": number
    },
    "validationDataConfig": {
       "validators": [
          {
             "s3Uri": "string"
          }
       ]
    },
    "validationMetrics": [
       {
          "validationLoss": number
       }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

baseModelArn

Amazon Resource Name (ARN) of the base model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

creationTime

Creation time of the model.

Type: Timestamp

customizationConfig

The customization configuration for the custom model.

Type: CustomizationConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

customizationType

The type of model customization.

Type: String

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION | REINFORCEMENT_FINE_TUNING | IMPORTED`

failureMessage

A failure message for any issues that occurred when creating the custom model. This is included for only a failed CreateCustomModel operation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

hyperParameters

Hyperparameter values associated with this model. For details on the format for different models, see Custom model hyperparameters.

Type: String to string map

jobArn

Job Amazon Resource Name (ARN) associated with this model. For models that you create with the CreateCustomModel API operation, this is NULL.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}`

jobName

Job name associated with this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

modelArn

Amazon Resource Name (ARN) associated with this model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

modelKmsKeyArn

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

modelName

Model name associated with this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

modelStatus

The current status of the custom model. Possible values include:

- `Creating` - The model is being created and validated.
- `Active` - The model has been successfully created and is ready for use.
- `Failed` - The model creation process failed. Check the `failureMessage` field for details.

Type: String

Valid Values: `Active | Creating | Failed`

outputDataConfig

Output data configuration associated with this custom model.

Type: OutputDataConfig object

trainingDataConfig

Contains information about the training dataset.

Type: TrainingDataConfig object

trainingMetrics

Contains training metrics from the job creation.

Type: TrainingMetrics object

validationDataConfig

Contains information about the validation dataset.

Type: ValidationDataConfig object

validationMetrics

The validation metrics from the job creation.

Type: Array of ValidatorMetric objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetCustomModelDeployment

Service: Amazon Bedrock

Retrieves information about a custom model deployment, including its status, configuration, and metadata. Use this operation to monitor the deployment status and retrieve details needed for inference requests.

The following actions are related to the `GetCustomModelDeployment` operation:

- CreateCustomModelDeployment
- ListCustomModelDeployments
- DeleteCustomModelDeployment

Request Syntax

```
GET /model-customization/custom-model-deployments/customModelDeploymentIdentifier
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

customModelDeploymentIdentifier

The Amazon Resource Name (ARN) or name of the custom model deployment to retrieve information about.

Length Constraints: Minimum length of 1. Maximum length of 93.

Pattern: `(arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12})|^([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "customModelDeploymentArn": "string",
   "description": "string",
   "failureMessage": "string",
   "lastUpdatedAt": "string",
   "modelArn": "string",
   "modelDeploymentName": "string",
   "status": "string",
   "updateDetails": {
      "modelArn": "string",
      "updateStatus": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The date and time when the custom model deployment was created.

Type: Timestamp

customModelDeploymentArn

The Amazon Resource Name (ARN) of the custom model deployment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12}`

description

The description of the custom model deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*`

failureMessage

If the deployment status is `FAILED`, this field contains a message describing the failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

lastUpdatedAt

The date and time when the custom model deployment was last updated.

Type: Timestamp

modelArn

The Amazon Resource Name (ARN) of the custom model associated with this deployment.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

modelDeploymentName

The name of the custom model deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

status

The status of the custom model deployment. Possible values are:

- `CREATING` - The deployment is being set up and prepared for inference.
- `ACTIVE` - The deployment is ready and available for inference requests.
- `FAILED` - The deployment failed to be created or became unavailable.

Type: String

Valid Values: `Creating | Active | Failed`

updateDetails

Details about any pending or completed updates to the custom model deployment, including the new model ARN and update status.

Type: CustomModelDeploymentUpdateDetails object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetEvaluationJob

Service: Amazon Bedrock

Gets information about an evaluation job, such as the status of the job.

Request Syntax

```
GET /evaluation-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The Amazon Resource Name (ARN) of the evaluation job you want get information on.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "applicationType": "string",
   "creationTime": "string",
   "customerEncryptionKeyId": "string",
   "evaluationConfig": { ... },
   "failureMessages": [ "string" ],
   "inferenceConfig": { ... },
   "jobArn": "string",
```

```
    "jobDescription": "string",
    "jobName": "string",
    "jobType": "string",
    "lastModifiedTime": "string",
    "outputDataConfig": {
       "s3Uri": "string"
    },
    "roleArn": "string",
    "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

applicationType

Specifies whether the evaluation job is for evaluating a model or evaluating a knowledge base (retrieval and response generation).

Type: String

Valid Values: `ModelEvaluation | RagEvaluation`

creationTime

The time the evaluation job was created.

Type: Timestamp

customerEncryptionKeyId

The Amazon Resource Name (ARN) of the customer managed encryption key specified when the evaluation job was created.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

evaluationConfig

Contains the configuration details of either an automated or human-based evaluation job.

Type: EvaluationConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

failureMessages

A list of strings that specify why the evaluation job failed to create.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

inferenceConfig

Contains the configuration details of the inference model used for the evaluation job.

Type: EvaluationInferenceConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

jobArn

The Amazon Resource Name (ARN) of the evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12}`

jobDescription

The description of the evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.+`

jobName

The name for the evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

jobType

Specifies whether the evaluation job is automated or human-based.

Type: String

Valid Values: `Human | Automated`

lastModifiedTime

The time the evaluation job was last modified.

Type: Timestamp

outputDataConfig

Contains the configuration details of the Amazon S3 bucket for storing the results of the evaluation job.

Type: EvaluationOutputDataConfig object

roleArn

The Amazon Resource Name (ARN) of the IAM service role used in the evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

status

The current status of the evaluation job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetFoundationModel

Service: Amazon Bedrock

Get details about a Amazon Bedrock foundation model.

Request Syntax

```
GET /foundation-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

The model identifier.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelDetails": {
      "customizationsSupported": [ "string" ],
      "inferenceTypesSupported": [ "string" ],
      "inputModalities": [ "string" ],
```

```
      "modelArn": "string",
      "modelId": "string",
      "modelLifecycle": {
         "endOfLifeTime": "string",
         "legacyTime": "string",
         "publicExtendedAccessTime": "string",
         "startOfLifeTime": "string",
         "status": "string"
      },
      "modelName": "string",
      "outputModalities": [ "string" ],
      "providerName": "string",
      "responseStreamingSupported": boolean
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelDetails

Information about the foundation model.

Type: FoundationModelDetails object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetFoundationModelAvailability

Service: Amazon Bedrock

Get information about the Foundation model availability.

Request Syntax

```
GET /foundation-model-availability/modelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelId

The model Id of the foundation model.

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agreementAvailability": {
      "errorMessage": "string",
      "status": "string"
   },
   "authorizationStatus": "string",
   "entitlementAvailability": "string",
   "modelId": "string",
```

```
   "regionAvailability": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agreementAvailability

Agreement availability.

Type: AgreementAvailability object

authorizationStatus

Authorization status.

Type: String

Valid Values: `AUTHORIZED | NOT_AUTHORIZED`

entitlementAvailability

Entitlement availability.

Type: String

Valid Values: `AVAILABLE | NOT_AVAILABLE`

modelId

The model Id of the foundation model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

regionAvailability

Region availability.

Type: String

Valid Values: `AVAILABLE | NOT_AVAILABLE`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetGuardrail

Service: Amazon Bedrock

Gets details about a guardrail. If you don't specify a version, the response returns details for the DRAFT version.

Request Syntax

```
GET /guardrails/guardrailIdentifier?guardrailVersion=guardrailVersion HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail for which to get details. This can be an ID or the ARN.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

guardrailVersion

The version of the guardrail for which to get details. If you don't specify a version, the response returns details for the DRAFT version.

Pattern: `((1-9][0-9]{0,7})|(DRAFT))`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "automatedReasoningPolicy": {
```

```
      "confidenceThreshold": number,
      "policies": [ "string" ]
   },
   "blockedInputMessaging": "string",
   "blockedOutputsMessaging": "string",
   "contentPolicy": {
      "filters": [
         {
            "inputAction": "string",
            "inputEnabled": boolean,
            "inputModalities": [ "string" ],
            "inputStrength": "string",
            "outputAction": "string",
            "outputEnabled": boolean,
            "outputModalities": [ "string" ],
            "outputStrength": "string",
            "type": "string"
         }
      ],
      "tier": {
         "tierName": "string"
      }
   },
   "contextualGroundingPolicy": {
      "filters": [
         {
            "action": "string",
            "enabled": boolean,
            "threshold": number,
            "type": "string"
         }
      ]
   },
   "createdAt": "string",
   "crossRegionDetails": {
      "guardrailProfileArn": "string",
      "guardrailProfileId": "string"
   },
   "description": "string",
   "failureRecommendations": [ "string" ],
   "guardrailArn": "string",
   "guardrailId": "string",
   "kmsKeyArn": "string",
   "name": "string",
```

```
   "sensitiveInformationPolicy": {
      "piiEntities": [
         {
            "action": "string",
            "inputAction": "string",
            "inputEnabled": boolean,
            "outputAction": "string",
            "outputEnabled": boolean,
            "type": "string"
         }
      ],
      "regexes": [
         {
            "action": "string",
            "description": "string",
            "inputAction": "string",
            "inputEnabled": boolean,
            "name": "string",
            "outputAction": "string",
            "outputEnabled": boolean,
            "pattern": "string"
         }
      ]
   },
   "status": "string",
   "statusReasons": [ "string" ],
   "topicPolicy": {
      "tier": {
         "tierName": "string"
      },
      "topics": [
         {
            "definition": "string",
            "examples": [ "string" ],
            "inputAction": "string",
            "inputEnabled": boolean,
            "name": "string",
            "outputAction": "string",
            "outputEnabled": boolean,
            "type": "string"
         }
      ]
   },
   "updatedAt": "string",
```

```
    "version": "string",
    "wordPolicy": {
       "managedWordLists": [
          {
             "inputAction": "string",
             "inputEnabled": boolean,
             "outputAction": "string",
             "outputEnabled": boolean,
             "type": "string"
          }
       ],
       "words": [
          {
             "inputAction": "string",
             "inputEnabled": boolean,
             "outputAction": "string",
             "outputEnabled": boolean,
             "text": "string"
          }
       ]
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

automatedReasoningPolicy

The current Automated Reasoning policy configuration for the guardrail, if any is configured.

Type: GuardrailAutomatedReasoningPolicy object

blockedInputMessaging

The message that the guardrail returns when it blocks a prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

blockedOutputsMessaging

The message that the guardrail returns when it blocks a model response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

contentPolicy

The content policy that was configured for the guardrail.

Type: GuardrailContentPolicy object

contextualGroundingPolicy

The contextual grounding policy used in the guardrail.

Type: GuardrailContextualGroundingPolicy object

createdAt

The date and time at which the guardrail was created.

Type: Timestamp

crossRegionDetails

Details about the system-defined guardrail profile that you're using with your guardrail, including the guardrail profile ID and Amazon Resource Name (ARN).

Type: GuardrailCrossRegionDetails object

description

The description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

failureRecommendations

Appears if the `status` of the guardrail is `FAILED`. A list of recommendations to carry out before retrying the request.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Minimum length of 1. Maximum length of 200.

guardrailArn

The ARN of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+`

guardrailId

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

kmsKeyArn

The ARN of the AWS KMS key that encrypts the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

name

The name of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `[0-9a-zA-Z-_]+`

sensitiveInformationPolicy

The sensitive information policy that was configured for the guardrail.

Type: GuardrailSensitiveInformationPolicy object

status

The status of the guardrail.

Type: String

Valid Values: `CREATING | UPDATING | VERSIONING | READY | FAILED | DELETING`

statusReasons

Appears if the `status` is `FAILED`. A list of reasons for why the guardrail failed to be created, updated, versioned, or deleted.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Minimum length of 1. Maximum length of 200.

topicPolicy

The topic policy that was configured for the guardrail.

Type: GuardrailTopicPolicy object

updatedAt

The date and time at which the guardrail was updated.

Type: Timestamp

version

The version of the guardrail.

Type: String

Pattern: `(([1-9][0-9]{0,7})|(DRAFT))`

wordPolicy

The word policy that was configured for the guardrail.

Type: GuardrailWordPolicy object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetImportedModel

Service: Amazon Bedrock

Gets properties associated with a customized model you imported.

Request Syntax

```
GET /imported-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier

Name or Amazon Resource Name (ARN) of the imported model.

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:imported-model/[a-z0-9]{12})|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "creationTime": "string",
   "customModelUnits": {
      "customModelUnitsPerModelCopy": number,
      "customModelUnitsVersion": "string"
   },
   "instructSupported": boolean,
   "jobArn": "string",
```

```
    "jobName": "string",
    "modelArchitecture": "string",
    "modelArn": "string",
    "modelDataSource": { ... },
    "modelKmsKeyArn": "string",
    "modelName": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

creationTime

Creation time of the imported model.

Type: Timestamp

customModelUnits

Information about the hardware utilization for a single copy of the model.

Type: CustomModelUnits object

instructSupported

Specifies if the imported model supports converse.

Type: Boolean

jobArn

Job Amazon Resource Name (ARN) associated with the imported model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-import-job/[a-z0-9]{12}`

jobName

Job name associated with the imported model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

modelArchitecture

The architecture of the imported model.

Type: String

modelArn

The Amazon Resource Name (ARN) associated with this imported model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:imported-model/[a-z0-9]{12}`

modelDataSource

The data source for this imported model.

Type: ModelDataSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

modelKmsKeyArn

The imported model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

modelName

The name of the imported model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Get model properties

Get the properties associated with a customized model that you imported.

```
GET /imported-models/{modelIdentifier} HTTP/1.1
Content-type: application/json
```

Example response

Response for the above request.

```
HTTP/1.1 200
Content-type: application/json

{
    "modelArn": "arn:aws:bedrock:us-east-1:111122223333:imported-model/s4dt0wly5gud",
    "modelName": "SomeImportedModelName",
    "jobName": "importJob-20240713T121942",
    "jobArn": "arn:aws:bedrock:us-east-1:111122223333:model-import-job/dchh9ny8e0dv",
    "modelDataSource": {
        "s3DataSource": {
            "s3Uri": "S3://amzn-s3-demo-bucket/key-name"
        }
    },
    "creationTime": "2024-08-13T19:20:14.058Z",
    "modelArchitecture": "mistral",
     "customModelUnits": {
         "customModelUnitsPerModelCopy": 8,
         "customModelUnitsVersion": "v1.0"
     }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetInferenceProfile

Service: Amazon Bedrock

Gets information about an inference profile. For more information, see Increase throughput and resilience with cross-region inference in Amazon Bedrock. in the Amazon Bedrock User Guide.

Request Syntax

```
GET /inference-profiles/inferenceProfileIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

inferenceProfileIdentifier

The ID or Amazon Resource Name (ARN) of the inference profile.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|application-inference-profile)/)?[a-zA-Z0-9-:.]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "description": "string",
   "inferenceProfileArn": "string",
   "inferenceProfileId": "string",
```

```
    "inferenceProfileName": "string",
    "models": [
       {
          "modelArn": "string"
       }
    ],
    "status": "string",
    "type": "string",
    "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The time at which the inference profile was created.

Type: Timestamp

description

The description of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `([0-9a-zA-Z:.][ _-]?)+`

inferenceProfileArn

The Amazon Resource Name (ARN) of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+`

inferenceProfileId

The unique identifier of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9-:.]+`

inferenceProfileName

The name of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `([0-9a-zA-Z][ _-]?)+`

models

A list of information about each model in the inference profile.

Type: Array of InferenceProfileModel objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

status

The status of the inference profile. `ACTIVE` means that the inference profile is ready to be used.

Type: String

Valid Values: `ACTIVE`

type

The type of the inference profile. The following types are possible:

- `SYSTEM_DEFINED` – The inference profile is defined by Amazon Bedrock. You can route inference requests across regions with these inference profiles.
- `APPLICATION` – The inference profile was created by a user. This type of inference profile can track metrics and costs when invoking the model in it. The inference profile may route requests to one or multiple regions.

Type: String

Valid Values: `SYSTEM_DEFINED | APPLICATION`

updatedAt

The time at which the inference profile was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Get information about an inference profile

Run the following example to get information about the US Anthropic Claude 3 Sonnet inference profile:

Sample Request

```
GET /inference-profiles/us.anthropic.claude-3-sonnet-20240229-v1:0 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetMarketplaceModelEndpoint

Service: Amazon Bedrock

Retrieves details about a specific endpoint for a model from Amazon Bedrock Marketplace.

Request Syntax

```
GET /marketplace-model/endpoints/endpointArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

endpointArn

The Amazon Resource Name (ARN) of the endpoint you want to get information about.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "marketplaceModelEndpoint": {
      "createdAt": "string",
      "endpointArn": "string",
      "endpointConfig": { ... },
      "endpointStatus": "string",
      "endpointStatusMessage": "string",
      "modelSourceIdentifier": "string",
      "status": "string",
      "statusMessage": "string",
      "updatedAt": "string"
```

```
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

marketplaceModelEndpoint

Details about the requested endpoint.

Type: MarketplaceModelEndpoint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetModelCopyJob

Service: Amazon Bedrock

Retrieves information about a model copy job. For more information, see Copy models to be used in other regions in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-copy-jobs/jobArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobArn

The Amazon Resource Name (ARN) of the model copy job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-copy-job/[a-z0-9]{12}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "creationTime": "string",
   "failureMessage": "string",
   "jobArn": "string",
   "sourceAccountId": "string",
   "sourceModelArn": "string",
   "sourceModelName": "string",
```

```
    "status": "string",
    "targetModelArn": "string",
    "targetModelKmsKeyArn": "string",
    "targetModelName": "string",
    "targetModelTags": [
       {
          "key": "string",
          "value": "string"
       }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

creationTime

The time at which the model copy job was created.

Type: Timestamp

failureMessage

An error message for why the model copy job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

jobArn

The Amazon Resource Name (ARN) of the model copy job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-copy-job/[a-z0-9]{12}`

sourceAccountId

The unique identifier of the account that the model being copied originated from.

Type: String

Pattern: `[0-9]{12}`

sourceModelArn

The Amazon Resource Name (ARN) of the original model being copied.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

sourceModelName

The name of the original model being copied.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

status

The status of the model copy job.

Type: String

Valid Values: `InProgress | Completed | Failed`

targetModelArn

The Amazon Resource Name (ARN) of the copied model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-`

`z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

targetModelKmsKeyArn

The Amazon Resource Name (ARN) of the KMS key encrypting the copied model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

targetModelName

The name of the copied model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

targetModelTags

The tags associated with the copied model.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Get a model copy job (CLI)

The following example shows how to get information about a model copy job using the AWS CLI.

```
aws bedrock get-model-copy-job --job-arn arn:aws:bedrock:us-east-1:123456789012:model-copy-job/amazon.titan-text-lite-v1:0:4k/abcdef123456
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetModelCustomizationJob

Service: Amazon Bedrock

Retrieves the properties associated with a model-customization job, including the status of the job. For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-customization-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

Identifier for the customization job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "baseModelArn": "string",
   "clientRequestToken": "string",
   "creationTime": "string",
   "customizationConfig": { ... },
```

```
   "customizationType": "string",
   "endTime": "string",
   "failureMessage": "string",
   "hyperParameters": {
      "string" : "string"
   },
   "jobArn": "string",
   "jobName": "string",
   "lastModifiedTime": "string",
   "outputDataConfig": {
      "s3Uri": "string"
   },
   "outputModelArn": "string",
   "outputModelKmsKeyArn": "string",
   "outputModelName": "string",
   "roleArn": "string",
   "status": "string",
   "statusDetails": {
      "dataProcessingDetails": {
         "creationTime": "string",
         "lastModifiedTime": "string",
         "status": "string"
      },
      "trainingDetails": {
         "creationTime": "string",
         "lastModifiedTime": "string",
         "status": "string"
      },
      "validationDetails": {
         "creationTime": "string",
         "lastModifiedTime": "string",
         "status": "string"
      }
   },
   "trainingDataConfig": {
      "invocationLogsConfig": {
         "invocationLogSource": { ... },
         "requestMetadataFilters": { ... },
         "usePromptResponse": boolean
      },
      "s3Uri": "string"
   },
   "trainingMetrics": {
      "trainingLoss": number
```

```
    },
    "validationDataConfig": {
       "validators": [
          {
             "s3Uri": "string"
          }
       ]
    },
    "validationMetrics": [
       {
          "validationLoss": number
       }
    ],
    "vpcConfig": {
       "securityGroupIds": [ "string" ],
       "subnetIds": [ "string" ]
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

baseModelArn

Amazon Resource Name (ARN) of the base model.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

clientRequestToken

The token that you specified in the `CreateCustomizationJob` request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

creationTime

Time that the resource was created.

Type: Timestamp

customizationConfig

The customization configuration for the model customization job.

Type: CustomizationConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

customizationType

The type of model customization.

Type: String

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION | REINFORCEMENT_FINE_TUNING | IMPORTED`

endTime

Time that the resource transitioned to terminal state.

Type: Timestamp

failureMessage

Information about why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

hyperParameters

The hyperparameter values for the job. For details on the format for different models, see Custom model hyperparameters.

Type: String to string map

jobArn

The Amazon Resource Name (ARN) of the customization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}`

jobName

The name of the customization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

lastModifiedTime

Time that the resource was last modified.

Type: Timestamp

outputDataConfig

Output data configuration

Type: OutputDataConfig object

outputModelArn

The Amazon Resource Name (ARN) of the output model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

outputModelKmsKeyArn

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

outputModelName

The name of the output model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

roleArn

The Amazon Resource Name (ARN) of the IAM role.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

status

The status of the job. A successful job transitions from in-progress to completed when the output model is ready to use. If the job failed, the failure message contains information about why the job failed.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

statusDetails

For a Distillation job, the details about the statuses of the sub-tasks of the customization job.

Type: StatusDetails object

trainingDataConfig

Contains information about the training dataset.

Type: TrainingDataConfig object

trainingMetrics

Contains training metrics from the job creation.

Type: TrainingMetrics object

validationDataConfig

Contains information about the validation dataset.

Type: ValidationDataConfig object

validationMetrics

The loss metric for each validator that you provided in the createjob request.

Type: Array of ValidatorMetric objects

vpcConfig

VPC configuration for the custom model job.

Type: VpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetModelImportJob

Service: Amazon Bedrock

Retrieves the properties associated with import model job, including the status of the job. For more information, see Import a customized model in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-import-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The identifier of the import job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-import-job/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "creationTime": "string",
   "endTime": "string",
   "failureMessage": "string",
   "importedModelArn": "string",
   "importedModelKmsKeyArn": "string",
   "importedModelName": "string",
```

```
    "jobArn": "string",
    "jobName": "string",
    "lastModifiedTime": "string",
    "modelDataSource": { ... },
    "roleArn": "string",
    "status": "string",
    "vpcConfig": {
       "securityGroupIds": [ "string" ],
       "subnetIds": [ "string" ]
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

creationTime

The time the resource was created.

Type: Timestamp

endTime

Time that the resource transitioned to terminal state.

Type: Timestamp

failureMessage

Information about why the import job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

importedModelArn

The Amazon Resource Name (ARN) of the imported model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:imported-model/[a-z0-9]{12}`

importedModelKmsKeyArn

The imported model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

importedModelName

The name of the imported model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

jobArn

The Amazon Resource Name (ARN) of the import job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-import-job/[a-z0-9]{12}`

jobName

The name of the import job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

lastModifiedTime

Time the resource was last modified.

Type: Timestamp

modelDataSource

The data source for the imported model.

Type: ModelDataSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

roleArn

The Amazon Resource Name (ARN) of the IAM role associated with this job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

status

The status of the job. A successful job transitions from in-progress to completed when the imported model is ready to use. If the job failed, the failure message contains information about why the job failed.

Type: String

Valid Values: `InProgress | Completed | Failed`

vpcConfig

The Virtual Private Cloud (VPC) configuration of the import model job.

Type: VpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Get model import job

Gets the properties of a model import job.

```
GET /model-import-jobs/{jobIdentifier} HTTP/1.1
Content-type: application/json
```

Example response

Response for the above request.

```
HTTP/1.1 200
Content-type: application/json

{
    "jobArn": "arn:aws:bedrock:us-east-1:111122223333:model-import-job/yggb47n4xlml",
    "jobName": "MyImportedModelJobName",
```

```
    "importedModelName": "ImportedModelName",
    "roleArn": "arn:aws:iam::111122223333:role/Role_Name",
    "modelDataSource": {
        "s3DataSource": {
            "s3Uri": "S3://amzn-s3-demo-bucket/key-name"
        }
    },
    "status": "InProgress",
    "creationTime": "2024-08-13T23:38:42.457Z",
    "lastModifiedTime": "2024-08-13T23:39:25.158Z"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetModelInvocationJob

Service: Amazon Bedrock

Gets details about a batch inference job. For more information, see Monitor batch inference jobs

Request Syntax

```
GET /model-invocation-job/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The Amazon Resource Name (ARN) of the batch inference job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-invocation-job/)?[a-z0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "clientRequestToken": "string",
   "endTime": "string",
   "errorRecordCount": number,
   "inputDataConfig": { ... },
   "jobArn": "string",
   "jobExpirationTime": "string",
   "jobName": "string",
```

```
    "lastModifiedTime": "string",
    "message": "string",
    "modelId": "string",
    "modelInvocationType": "string",
    "outputDataConfig": { ... },
    "processedRecordCount": number,
    "roleArn": "string",
    "status": "string",
    "submitTime": "string",
    "successRecordCount": number,
    "timeoutDurationInHours": number,
    "totalRecordCount": number,
    "vpcConfig": {
       "securityGroupIds": [ "string" ],
       "subnetIds": [ "string" ]
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]{1,256}(-*[a-zA-Z0-9]){0,256}`

endTime

The time at which the batch inference job ended.

Type: Timestamp

errorRecordCount

The number of records that failed to process in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

inputDataConfig

Details about the location of the input to the batch inference job.

Type: ModelInvocationJobInputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

jobArn

The Amazon Resource Name (ARN) of the batch inference job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-invocation-job/[a-z0-9]{12})`

jobExpirationTime

The time at which the batch inference job times or timed out.

Type: Timestamp

jobName

The name of the batch inference job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]{1,63}(-*[a-zA-Z0-9\+\-\.]){0,63}`

lastModifiedTime

The time at which the batch inference job was last modified.

Type: Timestamp

message

If the batch inference job failed, this field contains a message describing why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

modelId

The unique identifier of the foundation model used for model inference.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-:]{1,63}/[a-z0-9]{12}$)|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$)|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+$)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)`

modelInvocationType

The invocation endpoint for ModelInvocationJob

Type: String

Valid Values: `InvokeModel | Converse`

outputDataConfig

Details about the location of the output of the batch inference job.

Type: ModelInvocationJobOutputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

processedRecordCount

The number of records that have been processed in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

roleArn

The Amazon Resource Name (ARN) of the service role with permissions to carry out and manage batch inference. You can use the console to create a default service role or follow the steps at Create a service role for batch inference.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

status

The status of the batch inference job.

The following statuses are possible:

- Submitted – This job has been submitted to a queue for validation.
- Validating – This job is being validated for the requirements described in Format and upload your batch inference data. The criteria include the following:
  - Your IAM service role has access to the Amazon S3 buckets containing your files.
  - Your files are .jsonl files and each individual record is a JSON object in the correct format. Note that validation doesn't check if the `modelInput` value matches the request body for the model.
  - Your files fulfill the requirements for file size and number of records. For more information, see Quotas for Amazon Bedrock.
- Scheduled – This job has been validated and is now in a queue. The job will automatically start when it reaches its turn.
- Expired – This job timed out because it was scheduled but didn't begin before the set timeout duration. Submit a new job request.
- InProgress – This job has begun. You can start viewing the results in the output S3 location.
- Completed – This job has successfully completed. View the output files in the output S3 location.
- PartiallyCompleted – This job has partially completed. Not all of your records could be processed in time. View the output files in the output S3 location.

- Failed – This job has failed. Check the failure message for any further details. For further assistance, reach out to the Support Center.
- Stopped – This job was stopped by a user.
- Stopping – This job is being stopped by a user.

Type: String

Valid Values: `Submitted | InProgress | Completed | Failed | Stopping | Stopped | PartiallyCompleted | Expired | Validating | Scheduled`

submitTime

The time at which the batch inference job was submitted.

Type: Timestamp

successRecordCount

The number of records that were successfully processed in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

timeoutDurationInHours

The number of hours after which batch inference job was set to time out.

Type: Integer

Valid Range: Minimum value of 24. Maximum value of 168.

totalRecordCount

The total number of records in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

vpcConfig

The configuration of the Virtual Private Cloud (VPC) for the data in the batch inference job. For more information, see Protect batch inference jobs using a VPC.

Type: VpcConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Get a batch inference job

This example illustrates one usage of GetModelInvocationJob.

```
GET /model-invocation-job/BATCHJOB1234 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Get the current configuration values for model invocation logging.

Request Syntax

```
GET /logging/modelinvocations HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "loggingConfig": {
      "audioDataDeliveryEnabled": boolean,
      "cloudWatchConfig": {
         "largeDataDeliveryS3Config": {
            "bucketName": "string",
            "keyPrefix": "string"
         },
         "logGroupName": "string",
         "roleArn": "string"
      },
      "embeddingDataDeliveryEnabled": boolean,
      "imageDataDeliveryEnabled": boolean,
      "s3Config": {
         "bucketName": "string",
         "keyPrefix": "string"
      },
      "textDataDeliveryEnabled": boolean,
      "videoDataDeliveryEnabled": boolean
   }
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

loggingConfig

The current configuration values.

Type: LoggingConfig object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetPromptRouter

Service: Amazon Bedrock

Retrieves details about a prompt router.

Request Syntax

```
GET /prompt-routers/promptRouterArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

promptRouterArn

The prompt router's ARN

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:default-prompt-router/[a-zA-Z0-9-:.]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "description": "string",
   "fallbackModel": {
      "modelArn": "string"
   },
   "models": [
      {
```

```
            "modelArn": "string"
        }
    ],
    "promptRouterArn": "string",
    "promptRouterName": "string",
    "routingCriteria": {
        "responseQualityDifference": number
    },
    "status": "string",
    "type": "string",
    "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

When the router was created.

Type: Timestamp

description

The router's description.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `([0-9a-zA-Z:.][ _-]?)+`

fallbackModel

The router's fallback model.

Type: PromptRouterTargetModel object

models

The router's models.

Type: Array of PromptRouterTargetModel objects

promptRouterArn

The prompt router's ARN

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:default-prompt-router/[a-zA-Z0-9-:.]+`

promptRouterName

The router's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `([0-9a-zA-Z][ _-]?)+`

routingCriteria

The router's routing criteria.

Type: RoutingCriteria object

status

The router's status.

Type: String

Valid Values: `AVAILABLE`

type

The router's type.

Type: String

Valid Values: `custom | default`

updatedAt

When the router was updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetProvisionedModelThroughput

Service: Amazon Bedrock

Returns details for a Provisioned Throughput. For more information, see Provisioned Throughput in the Amazon Bedrock User Guide.

Request Syntax

```
GET /provisioned-model-throughput/provisionedModelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId

The Amazon Resource Name (ARN) or ID of the Provisioned Throughput.

Pattern: `((([0-9a-zA-Z][_-]?)+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "commitmentDuration": "string",
   "commitmentExpirationTime": "string",
   "creationTime": "string",
   "desiredModelArn": "string",
   "desiredModelUnits": number,
   "failureMessage": "string",
   "foundationModelArn": "string",
```

```
    "lastModifiedTime": "string",
    "modelArn": "string",
    "modelUnits": number,
    "provisionedModelArn": "string",
    "provisionedModelName": "string",
    "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

commitmentDuration

Commitment duration of the Provisioned Throughput.

Type: String

Valid Values: `OneMonth | SixMonths`

commitmentExpirationTime

The timestamp for when the commitment term for the Provisioned Throughput expires.

Type: Timestamp

creationTime

The timestamp of the creation time for this Provisioned Throughput.

Type: Timestamp

desiredModelArn

The Amazon Resource Name (ARN) of the model requested to be associated to this Provisioned Throughput. This value differs from the `modelArn` if updating hasn't completed.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63})`

`{0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

desiredModelUnits

The number of model units that was requested for this Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

failureMessage

A failure message for any issues that occurred during creation, updating, or deletion of the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

foundationModelArn

The Amazon Resource Name (ARN) of the base model for which the Provisioned Throughput was created, or of the base model that the custom model for which the Provisioned Throughput was created was customized.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

lastModifiedTime

The timestamp of the last time that this Provisioned Throughput was modified.

Type: Timestamp

modelArn

The Amazon Resource Name (ARN) of the model associated with this Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

modelUnits

The number of model units allocated to this Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

provisionedModelArn

The Amazon Resource Name (ARN) of the Provisioned Throughput.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}`

provisionedModelName

The name of the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

status

The status of the Provisioned Throughput.

Type: String

Valid Values: `Creating | InService | Updating | Failed`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetResourcePolicy

Service: Amazon Bedrock

Gets the resource policy document for a Bedrock resource

Request Syntax

```
GET /resource-policy/resourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The ARN of the Bedrock resource to which this resource policy applies.

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "resourcePolicy": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

resourcePolicy

The JSON string representing the Bedrock resource policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20480.

Pattern: `[ -ÿ]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetUseCaseForModelAccess

Service: Amazon Bedrock

Returns the use case for requesting access to Anthropic models.

Request Syntax

```
GET /use-case-for-model-access HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "formData": blob
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

formData

Returns form data from the Anthropic first time user request.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 10. Maximum length of 16384.

Errors

For information about the errors that are common to all actions, see Common Error Types.

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAdvancedPromptOptimizationJobs

Service: Amazon Bedrock

Lists the advanced prompt optimization jobs in your account.

Request Syntax

```
GET /advanced-prompt-optimization-jobs?
maxResults=maxResults&nextToken=nextToken&sortBy=sortBy&sortOrder=sortOrder HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

A token to get the next page of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

The field to sort the results by.

Valid Values: `CreationTime`

sortOrder

The sort order for the results.

Valid Values: `Ascending | Descending`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "jobSummaries": [
      {
         "creationTime": "string",
         "jobArn": "string",
         "jobName": "string",
         "jobStatus": "string",
         "lastModifiedTime": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

jobSummaries

A list of advanced prompt optimization job summaries.

Type: Array of AdvancedPromptOptimizationJobSummary objects

nextToken

A token to get the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAutomatedReasoningPolicies

Service: Amazon Bedrock

Lists all Automated Reasoning policies in your account, with optional filtering by policy ARN. This helps you manage and discover existing policies.

Request Syntax

```
GET /automated-reasoning-policies?
maxResults=maxResults&nextToken=nextToken&policyArn=policyArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of policies to return in a single call.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

The pagination token from a previous request to retrieve the next page of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

policyArn

Optional filter to list only the policy versions with the specified Amazon Resource Name (ARN). If not provided, the DRAFT versions for all policies are listed.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "automatedReasoningPolicySummaries": [
      {
         "createdAt": "string",
         "description": "string",
         "name": "string",
         "policyArn": "string",
         "policyId": "string",
         "updatedAt": "string",
         "version": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

automatedReasoningPolicySummaries

A list of Automated Reasoning policy summaries.

Type: Array of AutomatedReasoningPolicySummary objects

Array Members: Minimum number of 0 items. Maximum number of 1000 items.

nextToken

The pagination token to use in a subsequent request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAutomatedReasoningPolicyBuildWorkflows

Service: Amazon Bedrock

Lists all build workflows for an Automated Reasoning policy, showing the history of policy creation and modification attempts.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows?
maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of build workflows to return in a single response. Valid range is 1-100.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

A pagination token from a previous request to continue listing build workflows from where the previous request left off.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose build workflows you want to list.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "automatedReasoningPolicyBuildWorkflowSummaries": [
      {
         "buildWorkflowId": "string",
         "buildWorkflowType": "string",
         "createdAt": "string",
         "policyArn": "string",
         "status": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

automatedReasoningPolicyBuildWorkflowSummaries

A list of build workflow summaries, each containing key information about a build workflow including its status and timestamps.

Type: Array of AutomatedReasoningPolicyBuildWorkflowSummary objects

Array Members: Minimum number of 0 items. Maximum number of 1000 items.

nextToken

A pagination token to use in subsequent requests to retrieve additional build workflows.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAutomatedReasoningPolicyTestCases

Service: Amazon Bedrock

Lists tests for an Automated Reasoning policy. We recommend using pagination to ensure that the operation returns quickly and successfully.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/test-cases?
maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of tests to return in a single call.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

The pagination token from a previous request to retrieve the next page of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy for which to list tests.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "nextToken": "string",
   "testCases": [
      {
         "confidenceThreshold": number,
         "createdAt": "string",
         "expectedAggregatedFindingsResult": "string",
         "guardContent": "string",
         "queryContent": "string",
         "testCaseId": "string",
         "updatedAt": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

The pagination token to use in a subsequent request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

testCases

A list of tests for the specified policy.

Type: Array of AutomatedReasoningPolicyTestCase objects

Array Members: Minimum number of 0 items. Maximum number of 1000 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAutomatedReasoningPolicyTestResults

Service: Amazon Bedrock

Lists test results for an Automated Reasoning policy, showing how the policy performed against various test scenarios and validation checks.

Request Syntax

```
GET /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/test-results?maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow whose test results you want to list.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

maxResults

The maximum number of test results to return in a single response. Valid range is 1-100.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

A pagination token from a previous request to continue listing test results from where the previous request left off.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose test results you want to list.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "nextToken": "string",
   "testResults": [
      {
         "aggregatedTestFindingsResult": "string",
         "policyArn": "string",
         "testCase": {
            "confidenceThreshold": number,
            "createdAt": "string",
            "expectedAggregatedFindingsResult": "string",
            "guardContent": "string",
            "queryContent": "string",
            "testCaseId": "string",
            "updatedAt": "string"
         },
         "testFindings": [
            { ... }
         ],
         "testRunResult": "string",
         "testRunStatus": "string",
         "updatedAt": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

A pagination token to use in subsequent requests to retrieve additional test results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

testResults

A list of test results, each containing information about how the policy performed on specific test scenarios.

Type: Array of AutomatedReasoningPolicyTestResult objects

Array Members: Minimum number of 0 items. Maximum number of 5000 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListCustomModelDeployments

Service: Amazon Bedrock

Lists custom model deployments in your account. You can filter the results by creation time, name, status, and associated model. Use this operation to manage and monitor your custom model deployments.

We recommend using pagination to ensure that the operation returns quickly and successfully.

The following actions are related to the `ListCustomModelDeployments` operation:

- CreateCustomModelDeployment
- GetCustomModelDeployment
- DeleteCustomModelDeployment

Request Syntax

```
GET /model-customization/custom-model-deployments?
createdAfter=createdAfter&createdBefore=createdBefore&maxResults=maxResults&modelArnEquals=mode
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

createdAfter

Filters deployments created after the specified date and time.

createdBefore

Filters deployments created before the specified date and time.

maxResults

The maximum number of results to return in a single call.

Valid Range: Minimum value of 1. Maximum value of 1000.

modelArnEquals

Filters deployments by the Amazon Resource Name (ARN) of the associated custom model.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

nameContains

Filters deployments whose names contain the specified string.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

nextToken

The token for the next set of results. Use this token to retrieve additional results when the response is truncated.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

The field to sort the results by. The only supported value is `CreationTime`.

Valid Values: `CreationTime`

sortOrder

The sort order for the results. Valid values are `Ascending` and `Descending`. Default is `Descending`.

Valid Values: `Ascending | Descending`

statusEquals

Filters deployments by status. Valid values are `CREATING`, `ACTIVE`, and `FAILED`.

Valid Values: `Creating | Active | Failed`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelDeploymentSummaries": [
      {
         "createdAt": "string",
         "customModelDeploymentArn": "string",
         "customModelDeploymentName": "string",
         "failureMessage": "string",
         "lastUpdatedAt": "string",
         "modelArn": "string",
         "status": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelDeploymentSummaries

A list of custom model deployment summaries.

Type: Array of CustomModelDeploymentSummary objects

nextToken

The token for the next set of results. This value is null when there are no more results to return.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListCustomModels

Service: Amazon Bedrock

Returns a list of the custom models that you have created with the `CreateModelCustomizationJob` operation.

For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
GET /custom-models?
baseModelArnEquals=baseModelArnEquals&creationTimeAfter=creationTimeAfter&creationTimeBefore=cr
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

baseModelArnEquals

Return custom models only if the base model Amazon Resource Name (ARN) matches this parameter.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

creationTimeAfter

Return custom models created after the specified time.

creationTimeBefore

Return custom models created before the specified time.

foundationModelArnEquals

Return custom models only if the foundation model Amazon Resource Name (ARN) matches this parameter.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

isOwned

Return custom models depending on if the current account owns them (`true`) or if they were shared with the current account (`false`).

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

modelStatus

The status of them model to filter results by. Possible values include:

- `Creating` - Include only models that are currently being created and validated.
- `Active` - Include only models that have been successfully created and are ready for use.
- `Failed` - Include only models where the creation process failed.

If you don't specify a status, the API returns models in all states.

Valid Values: `Active | Creating | Failed`

nameContains

Return custom models only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

The field to sort by in the returned list of models.

Valid Values: `CreationTime`

sortOrder

The sort order of the results.

Valid Values: `Ascending | Descending`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelSummaries": [
      {
         "baseModelArn": "string",
         "baseModelName": "string",
         "creationTime": "string",
         "customizationType": "string",
         "modelArn": "string",
         "modelName": "string",
         "modelStatus": "string",
         "ownerAccountId": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries

Model summaries.

Type: Array of CustomModelSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListEnforcedGuardrailsConfiguration

Service: Amazon Bedrock

Lists the account-level enforced guardrail configurations.

Request Syntax

```
GET /enforcedGuardrailsConfiguration?nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

nextToken

Opaque continuation token of previous paginated response.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "guardrailsConfig": [
      {
         "configId": "string",
         "createdAt": "string",
         "createdBy": "string",
         "guardrailArn": "string",
         "guardrailId": "string",
         "guardrailVersion": "string",
         "inputTags": "string",
```

```
            "modelEnforcement": {
               "excludedModels": [ "string" ],
               "includedModels": [ "string" ]
            },
            "owner": "string",
            "selectiveContentGuarding": {
               "messages": "string",
               "system": "string"
            },
            "updatedAt": "string",
            "updatedBy": "string"
         }
      ],
      "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

guardrailsConfig

Array of AccountEnforcedGuardrailOutputConfiguration objects.

Type: Array of AccountEnforcedGuardrailOutputConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

nextToken

Opaque continuation token of previous paginated response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListEvaluationJobs

Service: Amazon Bedrock

Lists all existing evaluation jobs.

Request Syntax

```
GET /evaluation-jobs?
applicationTypeEquals=applicationTypeEquals&creationTimeAfter=creationTimeAfter&creationTimeBef
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

applicationTypeEquals

A filter to only list evaluation jobs that are either model evaluations or knowledge base evaluations.

Valid Values: `ModelEvaluation | RagEvaluation`

creationTimeAfter

A filter to only list evaluation jobs created after a specified time.

creationTimeBefore

A filter to only list evaluation jobs created before a specified time.

maxResults

The maximum number of results to return.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains

A filter to only list evaluation jobs that contain a specified string in the job name.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

nextToken

Continuation token from the previous response, for Amazon Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

sortBy

Specifies a creation time to sort the list of evaluation jobs by when they were created.

Valid Values: `CreationTime`

sortOrder

Specifies whether to sort the list of evaluation jobs by either ascending or descending order.

Valid Values: `Ascending | Descending`

statusEquals

A filter to only list evaluation jobs that are of a certain status.

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "jobSummaries": [
      {
         "applicationType": "string",
         "creationTime": "string",
         "customMetricsEvaluatorModelIdentifiers": [ "string" ],
         "evaluationTaskTypes": [ "string" ],
```

```
         "evaluatorModelIdentifiers": [ "string" ],
         "inferenceConfigSummary": {
            "modelConfigSummary": {
               "bedrockModelIdentifiers": [ "string" ],
               "precomputedInferenceSourceIdentifiers": [ "string" ]
            },
            "ragConfigSummary": {
               "bedrockKnowledgeBaseIdentifiers": [ "string" ],
               "precomputedRagSourceIdentifiers": [ "string" ]
            }
         },
         "jobArn": "string",
         "jobName": "string",
         "jobType": "string",
         "modelIdentifiers": [ "string" ],
         "ragIdentifiers": [ "string" ],
         "status": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

jobSummaries

A list of summaries of the evaluation jobs.

Type: Array of EvaluationSummary objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

nextToken

Continuation token from the previous response, for Amazon Bedrock to list the next set of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFoundationModelAgreementOffers

Service: Amazon Bedrock

Get the offers associated with the specified model.

Request Syntax

```
GET /list-foundation-model-agreement-offers/modelId?offerType=offerType HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelId

Model Id of the foundation model.

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

offerType

Type of offer associated with the model.

Valid Values: `ALL | PUBLIC`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelId": "string",
   "offers": [
```

```
      {
         "offerId": "string",
         "offerToken": "string",
         "termDetails": {
            "legalTerm": {
               "url": "string"
            },
            "supportTerm": {
               "refundPolicyDescription": "string"
            },
            "usageBasedPricingTerm": {
               "rateCard": [
                  {
                     "description": "string",
                     "dimension": "string",
                     "price": "string",
                     "unit": "string"
                  }
               ]
            },
            "validityTerm": {
               "agreementDuration": "string"
            }
         }
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelId

Model Id of the foundation model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

offers

List of the offers associated with the specified model.

Type: Array of Offer objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFoundationModels

Service: Amazon Bedrock

Lists Amazon Bedrock foundation models that you can use. You can filter the results with the request parameters. For more information, see Foundation models in the Amazon Bedrock User Guide.

Request Syntax

```
GET /foundation-models?
byCustomizationType=byCustomizationType&byInferenceType=byInferenceType&byOutputModality=byOutp
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

byCustomizationType

Return models that support the customization type that you specify. For more information, see Custom models in the Amazon Bedrock User Guide.

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION`

byInferenceType

Return models that support the inference type that you specify. For more information, see Provisioned Throughput in the Amazon Bedrock User Guide.

Valid Values: `ON_DEMAND | PROVISIONED`

byOutputModality

Return models that support the output modality that you specify.

Valid Values: `TEXT | IMAGE | EMBEDDING`

byProvider

Return models belonging to the model provider that you specify.

Pattern: `[A-Za-z0-9- ]{1,63}`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelSummaries": [
      {
         "customizationsSupported": [ "string" ],
         "inferenceTypesSupported": [ "string" ],
         "inputModalities": [ "string" ],
         "modelArn": "string",
         "modelId": "string",
         "modelLifecycle": {
            "endOfLifeTime": "string",
            "legacyTime": "string",
            "publicExtendedAccessTime": "string",
            "startOfLifeTime": "string",
            "status": "string"
         },
         "modelName": "string",
         "outputModalities": [ "string" ],
         "providerName": "string",
         "responseStreamingSupported": boolean
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries

A list of Amazon Bedrock foundation models.

Type: Array of FoundationModelSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListGuardrails

Service: Amazon Bedrock

Lists details about all the guardrails in an account. To list the DRAFT version of all your guardrails, don't specify the `guardrailIdentifier` field. To list all versions of a guardrail, specify the ARN of the guardrail in the `guardrailIdentifier` field.

You can set the maximum number of results to return in a response in the `maxResults` field. If there are more results than the number you set, the response returns a `nextToken` that you can send in another `ListGuardrails` request to see the next batch of results.

Request Syntax

```
GET /guardrails?
guardrailIdentifier=guardrailIdentifier&maxResults=maxResults&nextToken=nextToken
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail. This can be an ID or the ARN.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

maxResults

The maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If there are more results than were returned in the response, the response returns a `nextToken` that you can send in another `ListGuardrails` request to see the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "guardrails": [
      {
         "arn": "string",
         "createdAt": "string",
         "crossRegionDetails": {
            "guardrailProfileArn": "string",
            "guardrailProfileId": "string"
         },
         "description": "string",
         "id": "string",
         "name": "string",
         "status": "string",
         "updatedAt": "string",
         "version": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

guardrails

A list of objects, each of which contains details about a guardrail.

Type: Array of GuardrailSummary objects

Array Members: Minimum number of 0 items. Maximum number of 1000 items.

nextToken

If there are more results than were returned in the response, the response returns a `nextToken` that you can send in another `ListGuardrails` request to see the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListImportedModels

Service: Amazon Bedrock

Returns a list of models you've imported. You can filter the results to return based on one or more criteria. For more information, see Import a customized model in the Amazon Bedrock User Guide.

Request Syntax

```
GET /imported-models?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter

Return imported models that were created after the specified time.

creationTimeBefore

Return imported models that created before the specified time.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains

Return imported models only if the model name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

sortBy

The field to sort by in the returned list of imported models.

Valid Values: CreationTime

sortOrder

Specifies whetehr to sort the results in ascending or descending order.

Valid Values: Ascending | Descending

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelSummaries": [
      {
         "creationTime": "string",
         "instructSupported": boolean,
         "modelArchitecture": "string",
         "modelArn": "string",
         "modelName": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries

Model summaries.

Type: Array of ImportedModelSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

List the imported models

Lists the models that you have imported.

```
GET /imported-models/ HTTP/1.1
Content-type: application/json
```

Example response

Response for the above request.

```
HTTP/1.1 200
Content-type: application/json

{
    "modelSummaries": [
        {
            "modelArn": "arn:aws:bedrock:us-east-1:111122223333:imported-model/
s4dt0wly5gud",
            "modelName": "SomeImportedModelName",
            "creationTime": "2024-08-13T19:20:14.058Z"
        }
    ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListInferenceProfiles

Service: Amazon Bedrock

Returns a list of inference profiles that you can use. For more information, see Increase throughput and resilience with cross-region inference in Amazon Bedrock. in the Amazon Bedrock User Guide.

Request Syntax

```
GET /inference-profiles?maxResults=maxResults&nextToken=nextToken&type=typeEquals
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

typeEquals

Filters for inference profiles that match the type you specify.

- `SYSTEM_DEFINED` – The inference profile is defined by Amazon Bedrock. You can route inference requests across regions with these inference profiles.
- `APPLICATION` – The inference profile was created by a user. This type of inference profile can track metrics and costs when invoking the model in it. The inference profile may route requests to one or multiple regions.

Valid Values: `SYSTEM_DEFINED | APPLICATION`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "inferenceProfileSummaries": [
      {
         "createdAt": "string",
         "description": "string",
         "inferenceProfileArn": "string",
         "inferenceProfileId": "string",
         "inferenceProfileName": "string",
         "models": [
            {
               "modelArn": "string"
            }
         ],
         "status": "string",
         "type": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

inferenceProfileSummaries

A list of information about each inference profile that you can use.

Type: Array of InferenceProfileSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

List information about inference profiles in your Region

Run the following example to list information for up to 5 inference profiles in your region:

Sample Request

```
GET /inference-profiles?maxResults=5 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListMarketplaceModelEndpoints

Service: Amazon Bedrock

Lists the endpoints for models from Amazon Bedrock Marketplace in your AWS account.

Request Syntax

```
GET /marketplace-model/endpoints?
maxResults=maxResults&modelSourceIdentifier=modelSourceEquals&nextToken=nextToken
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in a single call. If more results are available, the operation returns a `NextToken` value.

Valid Range: Minimum value of 1. Maximum value of 1000.

modelSourceEquals

If specified, only endpoints for the given model source identifier are returned.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*arn:aws:sagemaker:.*:hub-content/SageMakerPublicHub/Model/.*`

nextToken

The token for the next set of results. You receive this token from a previous `ListMarketplaceModelEndpoints` call.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "marketplaceModelEndpoints": [
      {
         "createdAt": "string",
         "endpointArn": "string",
         "modelSourceIdentifier": "string",
         "status": "string",
         "statusMessage": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

marketplaceModelEndpoints

An array of endpoint summaries.

Type: Array of MarketplaceModelEndpointSummary objects

Array Members: Minimum number of 0 items. Maximum number of 1000 items.

nextToken

The token for the next set of results. Use this token to get the next set of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListModelCopyJobs

Service: Amazon Bedrock

Returns a list of model copy jobs that you have submitted. You can filter the jobs to return based on one or more criteria. For more information, see Copy models to be used in other regions in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-copy-jobs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter

Filters for model copy jobs created after the specified time.

creationTimeBefore

Filters for model copy jobs created before the specified time.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

sortBy

The field to sort by in the returned list of model copy jobs.

Valid Values: `CreationTime`

sortOrder

Specifies whether to sort the results in ascending or descending order.

Valid Values: `Ascending | Descending`

sourceAccountEquals

Filters for model copy jobs in which the account that the source model belongs to is equal to the value that you specify.

Pattern: `[0-9]{12}`

sourceModelArnEquals

Filters for model copy jobs in which the Amazon Resource Name (ARN) of the source model to is equal to the value that you specify.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

statusEquals

Filters for model copy jobs whose status matches the value that you specify.

Valid Values: `InProgress | Completed | Failed`

targetModelNameContains

Filters for model copy jobs in which the name of the copied model contains the string that you specify.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelCopyJobSummaries": [
      {
         "creationTime": "string",
         "failureMessage": "string",
         "jobArn": "string",
         "sourceAccountId": "string",
         "sourceModelArn": "string",
         "sourceModelName": "string",
         "status": "string",
         "targetModelArn": "string",
         "targetModelKmsKeyArn": "string",
         "targetModelName": "string",
         "targetModelTags": [
            {
               "key": "string",
               "value": "string"
            }
         ]
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelCopyJobSummaries

A list of information about each model copy job.

Type: Array of ModelCopyJobSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

List model copy jobs (CLI)

The following example shows how to return information about ten model copy jobs in an account, using the AWS CLI.

```
aws bedrock list-model-copy-jobs --max-results 10
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListModelCustomizationJobs

Service: Amazon Bedrock

Returns a list of model customization jobs that you have submitted. You can filter the jobs to return based on one or more criteria.

For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-customization-jobs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter

Return customization jobs created after the specified time.

creationTimeBefore

Return customization jobs created before the specified time.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains

Return customization jobs only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

The field to sort by in the returned list of jobs.

Valid Values: `CreationTime`

sortOrder

The sort order of the results.

Valid Values: `Ascending | Descending`

statusEquals

Return customization jobs with the specified status.

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelCustomizationJobSummaries": [
      {
         "baseModelArn": "string",
         "creationTime": "string",
         "customizationType": "string",
         "customModelArn": "string",
         "customModelName": "string",
```

```
            "endTime": "string",
            "jobArn": "string",
            "jobName": "string",
            "lastModifiedTime": "string",
            "status": "string",
            "statusDetails": {
               "dataProcessingDetails": {
                  "creationTime": "string",
                  "lastModifiedTime": "string",
                  "status": "string"
               },
               "trainingDetails": {
                  "creationTime": "string",
                  "lastModifiedTime": "string",
                  "status": "string"
               },
               "validationDetails": {
                  "creationTime": "string",
                  "lastModifiedTime": "string",
                  "status": "string"
               }
            }
         }
      ],
      "nextToken": "string"
   }
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelCustomizationJobSummaries

Job summaries.

Type: Array of ModelCustomizationJobSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListModelImportJobs

Service: Amazon Bedrock

Returns a list of import jobs you've submitted. You can filter the results to return based on one or more criteria. For more information, see Import a customized model in the Amazon Bedrock User Guide.

Request Syntax

```
GET /model-import-jobs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter

Return import jobs that were created after the specified time.

creationTimeBefore

Return import jobs that were created before the specified time.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains

Return imported jobs only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

The field to sort by in the returned list of imported jobs.

Valid Values: `CreationTime`

sortOrder

Specifies whether to sort the results in ascending or descending order.

Valid Values: `Ascending | Descending`

statusEquals

Return imported jobs with the specified status.

Valid Values: `InProgress | Completed | Failed`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelImportJobSummaries": [
      {
         "creationTime": "string",
         "endTime": "string",
         "importedModelArn": "string",
         "importedModelName": "string",
```

```
            "jobArn": "string",
            "jobName": "string",
            "lastModifiedTime": "string",
            "status": "string"
        }
    ],
    "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelImportJobSummaries

Import job summaries.

Type: Array of ModelImportJobSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

List model import jobs

Gets a list of the import jobs that you have submitted.

```
GET /model-import-jobs/ HTTP/1.1
Content-type: application/json
```

Example response

Response for the above request.

```
HTTP/1.1 200
Content-type: application/json

{
    "modelImportJobSummaries": [
        {
            "jobArn": "arn:aws:bedrock:us-east-1:111122223333:model-import-job/
yggb47n4xlml",
            "jobName": "MyImportedModelJobName",
            "status": "InProgress",
            "lastModifiedTime": "2024-08-13T23:40:47.517Z",
            "creationTime": "2024-08-13T23:38:42.457Z",
```

```
            "importedModelName": "ImportedModelName"
        },
        {
            "jobArn": "arn:aws:bedrock:us-east-1:111122223333:model-import-job/
dchh9ny8e0dv",
            "jobName": "SomeJobName",
            "status": "Completed",
            "lastModifiedTime": "2024-08-13T19:36:33.674Z",
            "creationTime": "2024-08-13T19:20:14.058Z",
            "endTime": "2024-08-13T19:36:33.492Z",
            "importedModelArn": "arn:aws:bedrock:us-east-1:111122223333:imported-model/
s4dt0wly5gud",
            "importedModelName": "SomeImportedModelName"
        }
    ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListModelInvocationJobs

Service: Amazon Bedrock

Lists all batch inference jobs in the account. For more information, see View details about a batch inference job.

Request Syntax

```
GET /model-invocation-jobs?
maxResults=maxResults&nameContains=nameContains&nextToken=nextToken&sortBy=sortBy&sortOrder=sor
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return. If there are more results than the number that you specify, a `nextToken` value is returned. Use the `nextToken` in a request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains

Specify a string to filter for batch inference jobs whose names contain the string.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]{1,63}(-*[a-zA-Z0-9\+\-\.]){0,63}`

nextToken

If there were more results than the value you specified in the `maxResults` field in a previous `ListModelInvocationJobs` request, the response would have returned a `nextToken` value. To see the next batch of results, send the `nextToken` value in another request.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

An attribute by which to sort the results.

Valid Values: `CreationTime`

sortOrder

Specifies whether to sort the results by ascending or descending order.

Valid Values: `Ascending | Descending`

statusEquals

Specify a status to filter for batch inference jobs whose statuses match the string you specify.

The following statuses are possible:

- Submitted – This job has been submitted to a queue for validation.
- Validating – This job is being validated for the requirements described in Format and upload your batch inference data. The criteria include the following:
  - Your IAM service role has access to the Amazon S3 buckets containing your files.
  - Your files are .jsonl files and each individual record is a JSON object in the correct format. Note that validation doesn't check if the `modelInput` value matches the request body for the model.
  - Your files fulfill the requirements for file size and number of records. For more information, see Quotas for Amazon Bedrock.
- Scheduled – This job has been validated and is now in a queue. The job will automatically start when it reaches its turn.
- Expired – This job timed out because it was scheduled but didn't begin before the set timeout duration. Submit a new job request.
- InProgress – This job has begun. You can start viewing the results in the output S3 location.
- Completed – This job has successfully completed. View the output files in the output S3 location.
- PartiallyCompleted – This job has partially completed. Not all of your records could be processed in time. View the output files in the output S3 location.
- Failed – This job has failed. Check the failure message for any further details. For further assistance, reach out to the Support Center.
- Stopped – This job was stopped by a user.

- Stopping – This job is being stopped by a user.

Valid Values: `Submitted | InProgress | Completed | Failed | Stopping | Stopped | PartiallyCompleted | Expired | Validating | Scheduled`

submitTimeAfter

Specify a time to filter for batch inference jobs that were submitted after the time you specify.

submitTimeBefore

Specify a time to filter for batch inference jobs that were submitted before the time you specify.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "invocationJobSummaries": [
      {
         "clientRequestToken": "string",
         "endTime": "string",
         "errorRecordCount": number,
         "inputDataConfig": { ... },
         "jobArn": "string",
         "jobExpirationTime": "string",
         "jobName": "string",
         "lastModifiedTime": "string",
         "message": "string",
         "modelId": "string",
         "modelInvocationType": "string",
         "outputDataConfig": { ... },
         "processedRecordCount": number,
         "roleArn": "string",
         "status": "string",
         "submitTime": "string",
         "successRecordCount": number,
         "timeoutDurationInHours": number,
```

```
            "totalRecordCount": number,
            "vpcConfig": {
               "securityGroupIds": [ "string" ],
               "subnetIds": [ "string" ]
            }
         }
      ],
      "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationJobSummaries

A list of items, each of which contains a summary about a batch inference job.

Type: Array of ModelInvocationJobSummary objects

nextToken

If there are more results than can fit in the response, a `nextToken` is returned. Use the `nextToken` in a request to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

List up to ten model invocation jobs

This example illustrates one usage of ListModelInvocationJobs.

```
GET /model-invocation-jobs?maxResults=10&sortBy=CreationTime&sortOrder=Descending
 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListPromptRouters

Service: Amazon Bedrock

Retrieves a list of prompt routers.

Request Syntax

```
GET /prompt-routers?maxResults=maxResults&nextToken=nextToken&type=type HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of prompt routers to return in one page of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

type

The type of the prompt routers, such as whether it's default or custom.

Valid Values: `custom | default`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
```

```
   "nextToken": "string",
   "promptRouterSummaries": [
      {
         "createdAt": "string",
         "description": "string",
         "fallbackModel": {
            "modelArn": "string"
         },
         "models": [
            {
               "modelArn": "string"
            }
         ],
         "promptRouterArn": "string",
         "promptRouterName": "string",
         "routingCriteria": {
            "responseQualityDifference": number
         },
         "status": "string",
         "type": "string",
         "updatedAt": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

promptRouterSummaries

A list of prompt router summaries.

Type: Array of PromptRouterSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListProvisionedModelThroughputs

Service: Amazon Bedrock

Lists the Provisioned Throughputs in the account. For more information, see Provisioned Throughput in the Amazon Bedrock User Guide.

Request Syntax

```
GET /provisioned-model-throughputs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter

A filter that returns Provisioned Throughputs created after the specified time.

creationTimeBefore

A filter that returns Provisioned Throughputs created before the specified time.

maxResults

THe maximum number of results to return in the response. If there are more results than the number you specified, the response returns a `nextToken` value. To see the next batch of results, send the `nextToken` value in another list request.

Valid Range: Minimum value of 1. Maximum value of 1000.

modelArnEquals

A filter that returns Provisioned Throughputs whose model Amazon Resource Name (ARN) is equal to the value that you specify.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

nameContains

A filter that returns Provisioned Throughputs if their name contains the expression that you specify.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

nextToken

If there are more results than the number you specified in the `maxResults` field, the response returns a `nextToken` value. To see the next batch of results, specify the `nextToken` value in this field.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

The field by which to sort the returned list of Provisioned Throughputs.

Valid Values: `CreationTime`

sortOrder

The sort order of the results.

Valid Values: `Ascending | Descending`

statusEquals

A filter that returns Provisioned Throughputs if their statuses matches the value that you specify.

Valid Values: `Creating | InService | Updating | Failed`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "nextToken": "string",
   "provisionedModelSummaries": [
      {
         "commitmentDuration": "string",
         "commitmentExpirationTime": "string",
         "creationTime": "string",
         "desiredModelArn": "string",
         "desiredModelUnits": number,
         "foundationModelArn": "string",
         "lastModifiedTime": "string",
         "modelArn": "string",
         "modelUnits": number,
         "provisionedModelArn": "string",
         "provisionedModelName": "string",
         "status": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

If there are more results than the number you specified in the `maxResults` field, this value is returned. To see the next batch of results, include this value in the `nextToken` field in another list request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

provisionedModelSummaries

A list of summaries, one for each Provisioned Throughput in the response.

Type: Array of ProvisionedModelSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListTagsForResource

Service: Amazon Bedrock

List the tags associated with the specified resource.

For more information, see Tagging resources in the Amazon Bedrock User Guide.

Request Syntax

```
POST /listTagsForResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The Amazon Resource Name (ARN) of the resource.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^[a-zA-Z0-9][a-zA-Z0-9\-]*$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported)/[a-z0-9]{12}$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})$)|(:guardrail/[a-z0-9]+$)|(:automated-reasoning-policy/[a-zA-Z0-9]+(:[a-zA-Z0-9]+)?$)|(:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+$)|(:(provisioned-model|`

`model-invocation-job|model-evaluation-job|evaluation-job|model-import-job|imported-model|async-invoke|provisioned-model-v2|provisioned-model-reservation|prompt-router|custom-model-deployment)/[a-z0-9]{12}$))).*`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

An array of the tags associated with this resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

PutAccountDataRetention

Service: Amazon Bedrock

Sets the account-wide data retention mode for Amazon Bedrock.

Request Syntax

```
PUT /data-retention HTTP/1.1
Content-type: application/json

{
   "mode": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

mode

The data retention mode to set for the account.

Type: String

Valid Values: `default | none | aws_review | provider_data_share | inherit`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "mode": "string",
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

mode

The data retention mode set for the account.

Type: String

Valid Values: `default | none | aws_review | provider_data_share | inherit`

updatedAt

The time at which the data retention mode was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PutEnforcedGuardrailConfiguration

Service: Amazon Bedrock

Sets the account-level enforced guardrail configuration.

Request Syntax

```
PUT /enforcedGuardrailsConfiguration HTTP/1.1
Content-type: application/json

{
   "configId": "string",
   "guardrailInferenceConfig": {
      "guardrailIdentifier": "string",
      "guardrailVersion": "string",
      "modelEnforcement": {
         "excludedModels": [ "string" ],
         "includedModels": [ "string" ]
      },
      "selectiveContentGuarding": {
         "messages": "string",
         "system": "string"
      }
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

configId

Unique ID for the account enforced configuration.

Type: String

Pattern: `[a-z0-9]+`

Required: No

guardrailInferenceConfig

Account-level enforced guardrail input configuration.

Type: AccountEnforcedGuardrailInferenceInputConfiguration object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "configId": "string",
   "updatedAt": "string",
   "updatedBy": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

configId

Unique ID for the account enforced configuration.

Type: String

Pattern: `[a-z0-9]+`

updatedAt

Timestamp.

Type: Timestamp

updatedBy

The ARN of the role used to update the configuration.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PutModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Set the configuration values for model invocation logging.

Request Syntax

```
PUT /logging/modelinvocations HTTP/1.1
Content-type: application/json

{
   "loggingConfig": {
      "audioDataDeliveryEnabled": boolean,
      "cloudWatchConfig": {
         "largeDataDeliveryS3Config": {
            "bucketName": "string",
            "keyPrefix": "string"
         },
         "logGroupName": "string",
         "roleArn": "string"
      },
      "embeddingDataDeliveryEnabled": boolean,
      "imageDataDeliveryEnabled": boolean,
      "s3Config": {
         "bucketName": "string",
         "keyPrefix": "string"
      },
      "textDataDeliveryEnabled": boolean,
      "videoDataDeliveryEnabled": boolean
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

loggingConfig

The logging configuration values to set.

Type: LoggingConfig object

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PutResourcePolicy

Service: Amazon Bedrock

Adds a resource policy for a Bedrock resource.

Request Syntax

```
POST /resource-policy HTTP/1.1
Content-type: application/json

{
   "resourceArn": "string",
   "resourcePolicy": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceArn

The ARN of the Bedrock resource to which this resource policy applies.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: Yes

resourcePolicy

The JSON string representing the Bedrock resource policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20480.

Pattern: `[ -ÿ]+`

Required: Yes

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "resourceArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

resourceArn

The ARN of the Bedrock resource to which this resource policy applies.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PutUseCaseForModelAccess

Service: Amazon Bedrock

Returns the use case for model access. The example shown below under *Examples* shows the form data json configuration.

Request Syntax

```
POST /use-case-for-model-access HTTP/1.1
Content-type: application/json

{
   "formData": blob
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

formData

Returns the form data from the Anthropic first time user request.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 10. Maximum length of 16384.

Required: Yes

Response Syntax

```
HTTP/1.1 201
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example

This example illustrates one usage of PutUseCaseForModelAccess.

```
form_data = {
    "companyName": COMPANY_NAME,
    "companyWebsite": COMPANY_WEBSITE,
    "intendedUsers": INTENDED_USERS,
    "industryOption": INDUSTRY_OPTION,
    "otherIndustryOption": OTHER_INDUSTRY_OPTION,
    "useCases": USE_CASES
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

RegisterMarketplaceModelEndpoint

Service: Amazon Bedrock

Registers an existing Amazon SageMaker endpoint with Amazon Bedrock Marketplace, allowing it to be used with Amazon Bedrock APIs.

Request Syntax

```
POST /marketplace-model/endpoints/endpointIdentifier/registration HTTP/1.1
Content-type: application/json

{
   "modelSourceIdentifier": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

endpointIdentifier

The ARN of the Amazon SageMaker endpoint you want to register with Amazon Bedrock Marketplace.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

Request Body

The request accepts the following data in JSON format.

modelSourceIdentifier

The ARN of the model from Amazon Bedrock Marketplace that is deployed on the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*arn:aws:sagemaker:.*:hub-content/SageMakerPublicHub/Model/.*`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "marketplaceModelEndpoint": {
      "createdAt": "string",
      "endpointArn": "string",
      "endpointConfig": { ... },
      "endpointStatus": "string",
      "endpointStatusMessage": "string",
      "modelSourceIdentifier": "string",
      "status": "string",
      "statusMessage": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

marketplaceModelEndpoint

Details about the registered endpoint.

Type: MarketplaceModelEndpoint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceUnavailableException

Returned if the service cannot complete the request.

HTTP Status Code: 503

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StartAutomatedReasoningPolicyBuildWorkflow

Service: Amazon Bedrock

Starts a new build workflow for an Automated Reasoning policy. This initiates the process of analyzing source documents and generating policy rules, variables, and types.

Request Syntax

```
POST /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowType/start
 HTTP/1.1
x-amz-client-token: clientRequestToken
Content-type: application/json

{
   "policyDefinition": {
      "rules": [
         {
            "alternateExpression": "string",
            "expression": "string",
            "id": "string"
         }
      ],
      "types": [
         {
            "description": "string",
            "name": "string",
            "values": [
               {
                  "description": "string",
                  "value": "string"
               }
            ]
         }
      ],
      "variables": [
         {
            "description": "string",
            "name": "string",
            "type": "string"
         }
      ],
      "version": "string"
   },
```

```
    "workflowContent": { ... }
}
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowType

The type of build workflow to start (e.g., DOCUMENT_INGESTION for processing new documents, POLICY_REPAIR for fixing existing policies).

Valid Values: `INGEST_CONTENT | REFINE_POLICY | IMPORT_POLICY | GENERATE_FIDELITY_REPORT | GENERATE_POLICY_SCENARIOS | RESOLVE_POLICY_AMBIGUITIES | ITERATIVELY_REFINE_POLICY`

Required: Yes

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than once. If this token matches a previous request, Amazon Bedrock ignores the request but doesn't return an error.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy for which to start the build workflow.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

policyDefinition

An existing policy definition that serves as the starting point for the build workflow, typically used in policy repair or update scenarios.

Type: AutomatedReasoningPolicyDefinition object

Required: No

workflowContent

The actual content to be processed in the build workflow, such as documents to analyze or repair instructions to apply.

Type: AutomatedReasoningPolicyWorkflowTypeContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "buildWorkflowId": "string",
   "policyArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

buildWorkflowId

The unique identifier of the newly started build workflow. Use this ID to track the workflow's progress and retrieve its results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StartAutomatedReasoningPolicyTestWorkflow

Service: Amazon Bedrock

Initiates a test workflow to validate Automated Reasoning policy tests. The workflow executes the specified tests against the policy and generates validation results.

Request Syntax

```
POST /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/test-workflows HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "testCaseIds": [ "string" ]
}
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The build workflow identifier. The build workflow must show a COMPLETED status before running tests.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy to test.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request but doesn't return an error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

testCaseIds

The list of test identifiers to run. If not provided, all tests for the policy are run.

Type: Array of strings

Array Members: Fixed number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "policyArn": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

policyArn

The Amazon Resource Name (ARN) of the policy for which the test workflow was started.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StopAdvancedPromptOptimizationJob

Service: Amazon Bedrock

Stops an advanced prompt optimization job that is in progress.

Request Syntax

```
POST /advanced-prompt-optimization-jobs/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The ARN or ID of the advanced prompt optimization job to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/)?[a-z0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StopEvaluationJob

Service: Amazon Bedrock

Stops an evaluation job that is current being created or running.

Request Syntax

```
POST /evaluation-job/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The Amazon Resource Name (ARN) of the evaluation job you want to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StopModelCustomizationJob

Service: Amazon Bedrock

Stops an active model customization job. For more information, see Custom models in the Amazon Bedrock User Guide.

Request Syntax

```
POST /model-customization-jobs/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

Job identifier of the job to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*)`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StopModelInvocationJob

Service: Amazon Bedrock

Stops a batch inference job. You're only charged for tokens that were already processed. For more information, see Stop a batch inference job.

Request Syntax

```
POST /model-invocation-job/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier

The Amazon Resource Name (ARN) of the batch inference job to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-invocation-job/)?[a-z0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Stop a model invocation job

This example illustrates one usage of StopModelInvocationJob.

```
POST /model-invocation-job/BATCHJOB1234/stop HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

TagResource

Service: Amazon Bedrock

Associate tags with a resource. For more information, see Tagging resources in the Amazon Bedrock User Guide.

Request Syntax

```
POST /tagResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The Amazon Resource Name (ARN) of the resource to tag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^[a-zA-Z0-9][a-zA-Z0-9\-]*$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported)/[a-z0-9]{12}$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-`

```
z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]
{1,63}){0,2}(/[a-z0-9]{12})$)|(:guardrail/[a-z0-9]+$)|(:automated-
reasoning-policy/[a-zA-Z0-9]+(:[a-zA-Z0-9]+)?$)|(:(inference-profile|
application-inference-profile)/[a-zA-Z0-9-:.]+$)|(:(provisioned-model|
model-invocation-job|model-evaluation-job|evaluation-job|model-import-
job|imported-model|async-invoke|provisioned-model-v2|provisioned-model-
reservation|prompt-router|custom-model-deployment)/[a-z0-9]{12}$))).*
```

Required: Yes

tags

Tags to associate with the resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UntagResource

Service: Amazon Bedrock

Remove one or more tags from a resource. For more information, see Tagging resources in the Amazon Bedrock User Guide.

Request Syntax

```
POST /untagResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string",
   "tagKeys": [ "string" ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The Amazon Resource Name (ARN) of the resource to untag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^[a-zA-Z0-9][a-zA-Z0-9\-]*$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported)/[a-z0-9]{12}$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})$)|(:guardrail/[a-z0-9]+$)|(:automated-reasoning-policy/[a-zA-Z0-9]+(:[a-zA-Z0-9]+)?$)|(:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+$)|(:(provisioned-model|`

`model-invocation-job|model-evaluation-job|evaluation-job|model-import-job|imported-model|async-invoke|provisioned-model-v2|provisioned-model-reservation|prompt-router|custom-model-deployment)/[a-z0-9]{12}$))).*`

Required: Yes

tagKeys

Tag keys of the tags to remove from the resource.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAutomatedReasoningPolicy

Service: Amazon Bedrock

Updates an existing Automated Reasoning policy with new rules, variables, or configuration. This creates a new version of the policy while preserving the previous version.

Request Syntax

```
PATCH /automated-reasoning-policies/policyArn HTTP/1.1
Content-type: application/json

{
   "description": "string",
   "name": "string",
   "policyDefinition": {
      "rules": [
         {
            "alternateExpression": "string",
            "expression": "string",
            "id": "string"
         }
      ],
      "types": [
         {
            "description": "string",
            "name": "string",
            "values": [
               {
                  "description": "string",
                  "value": "string"
               }
            ]
         }
      ],
      "variables": [
         {
            "description": "string",
            "name": "string",
            "type": "string"
         }
      ],
      "version": "string"
   }
```

```
}
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy to update. This must be the ARN of a draft policy.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

The updated description for the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

name

The updated name for the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

Required: No

policyDefinition

The updated policy definition containing the formal logic rules, variables, and types.

Type: AutomatedReasoningPolicyDefinition object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "definitionHash": "string",
   "name": "string",
   "policyArn": "string",
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

definitionHash

The hash of the updated policy definition.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

name

The updated name of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

policyArn

The Amazon Resource Name (ARN) of the updated policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

updatedAt

The timestamp when the policy was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

resourceName

The name of the resource with too many tags.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAutomatedReasoningPolicyAnnotations

Service: Amazon Bedrock

Updates the annotations for an Automated Reasoning policy build workflow. This allows you to modify extracted rules, variables, and types before finalizing the policy.

Request Syntax

```
PATCH /automated-reasoning-policies/policyArn/build-workflows/buildWorkflowId/
annotations HTTP/1.1
Content-type: application/json

{
   "annotations": [
      { ... }
   ],
   "lastUpdatedAnnotationSetHash": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

buildWorkflowId

The unique identifier of the build workflow whose annotations you want to update.

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy whose annotations you want to update.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

annotations

The updated annotations containing modified rules, variables, and types for the policy.

Type: Array of AutomatedReasoningPolicyAnnotation objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

lastUpdatedAnnotationSetHash

The hash value of the annotation set that you're updating. This is used for optimistic concurrency control to prevent conflicting updates.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "annotationSetHash": "string",
   "buildWorkflowId": "string",
   "policyArn": "string",
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

annotationSetHash

The new hash value representing the updated state of the annotations.

Type: String

Length Constraints: Fixed length of 128.

Pattern: `[0-9a-z]{128}`

buildWorkflowId

The unique identifier of the build workflow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

updatedAt

The timestamp when the annotations were updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAutomatedReasoningPolicyTestCase

Service: Amazon Bedrock

Updates an existing Automated Reasoning policy test. You can modify the content, query, expected result, and confidence threshold.

Request Syntax

```
PATCH /automated-reasoning-policies/policyArn/test-cases/testCaseId HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "confidenceThreshold": number,
   "expectedAggregatedFindingsResult": "string",
   "guardContent": "string",
   "lastUpdatedAt": "string",
   "queryContent": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy that contains the test.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

testCaseId

The unique identifier of the test to update.

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier to ensure that the operation completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

confidenceThreshold

The updated minimum confidence level for logic validation. If null is provided, the threshold will be removed.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

expectedAggregatedFindingsResult

The updated expected result of the Automated Reasoning check.

Type: String

Valid Values: `VALID | INVALID | SATISFIABLE | IMPOSSIBLE | TRANSLATION_AMBIGUOUS | TOO_COMPLEX | NO_TRANSLATION`

Required: Yes

guardContent

The updated content to be validated by the Automated Reasoning policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

lastUpdatedAt

The timestamp when the test was last updated. This is used as a concurrency token to prevent conflicting modifications.

Type: Timestamp

Required: Yes

queryContent

The updated input query or prompt that generated the content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "policyArn": "string",
   "testCaseId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

policyArn

The Amazon Resource Name (ARN) of the policy that contains the updated test.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

testCaseId

The unique identifier of the updated test.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceInUseException

Thrown when attempting to delete or modify a resource that is currently being used by other resources or operations. For example, trying to delete an Automated Reasoning policy that is referenced by an active guardrail.

HTTP Status Code: 400

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateCustomModelDeployment

Service: Amazon Bedrock

Updates a custom model deployment with a new custom model. This allows you to deploy updated models without creating new deployment endpoints.

Request Syntax

```
PATCH /model-customization/custom-model-deployments/customModelDeploymentIdentifier
 HTTP/1.1
Content-type: application/json

{
   "modelArn": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

customModelDeploymentIdentifier

Identifier of the custom model deployment to update with the new custom model.

Length Constraints: Minimum length of 1. Maximum length of 93.

Pattern: `(arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12})|^([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

modelArn

ARN of the new custom model to deploy. This replaces the currently deployed model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "customModelDeploymentArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

customModelDeploymentArn

ARN of the custom model deployment being updated.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

UpdateGuardrail

Service: Amazon Bedrock

Updates a guardrail with the values you specify.

- Specify a `name` and optional `description`.
- Specify messages for when the guardrail successfully blocks a prompt or a model response in the `blockedInputMessaging` and `blockedOutputsMessaging` fields.
- Specify topics for the guardrail to deny in the `topicPolicyConfig` object. Each GuardrailTopicConfig object in the `topicsConfig` list pertains to one topic.
  - Give a `name` and `description` so that the guardrail can properly identify the topic.
  - Specify `DENY` in the `type` field.
  - (Optional) Provide up to five prompts that you would categorize as belonging to the topic in the `examples` list.
- Specify filter strengths for the harmful categories defined in Amazon Bedrock in the `contentPolicyConfig` object. Each GuardrailContentFilterConfig object in the `filtersConfig` list pertains to a harmful category. For more information, see Content filters. For more information about the fields in a content filter, see GuardrailContentFilterConfig.
  - Specify the category in the `type` field.
  - Specify the strength of the filter for prompts in the `inputStrength` field and for model responses in the `strength` field of the GuardrailContentFilterConfig.
- (Optional) For security, include the ARN of a AWS KMS key in the `kmsKeyId` field.

Request Syntax

```
PUT /guardrails/guardrailIdentifier HTTP/1.1
Content-type: application/json

{
   "automatedReasoningPolicyConfig": {
      "confidenceThreshold": number,
      "policies": [ "string" ]
   },
   "blockedInputMessaging": "string",
   "blockedOutputsMessaging": "string",
   "contentPolicyConfig": {
      "filtersConfig": [
```

```
            {
               "inputAction": "string",
               "inputEnabled": boolean,
               "inputModalities": [ "string" ],
               "inputStrength": "string",
               "outputAction": "string",
               "outputEnabled": boolean,
               "outputModalities": [ "string" ],
               "outputStrength": "string",
               "type": "string"
            }
         ],
         "tierConfig": {
            "tierName": "string"
         }
      },
      "contextualGroundingPolicyConfig": {
         "filtersConfig": [
            {
               "action": "string",
               "enabled": boolean,
               "threshold": number,
               "type": "string"
            }
         ]
      },
      "crossRegionConfig": {
         "guardrailProfileIdentifier": "string"
      },
      "description": "string",
      "kmsKeyId": "string",
      "name": "string",
      "sensitiveInformationPolicyConfig": {
         "piiEntitiesConfig": [
            {
               "action": "string",
               "inputAction": "string",
               "inputEnabled": boolean,
               "outputAction": "string",
               "outputEnabled": boolean,
               "type": "string"
            }
         ],
         "regexesConfig": [
```

```
            {
               "action": "string",
               "description": "string",
               "inputAction": "string",
               "inputEnabled": boolean,
               "name": "string",
               "outputAction": "string",
               "outputEnabled": boolean,
               "pattern": "string"
            }
         ]
      },
      "topicPolicyConfig": {
         "tierConfig": {
            "tierName": "string"
         },
         "topicsConfig": [
            {
               "definition": "string",
               "examples": [ "string" ],
               "inputAction": "string",
               "inputEnabled": boolean,
               "name": "string",
               "outputAction": "string",
               "outputEnabled": boolean,
               "type": "string"
            }
         ]
      },
      "wordPolicyConfig": {
         "managedWordListsConfig": [
            {
               "inputAction": "string",
               "inputEnabled": boolean,
               "outputAction": "string",
               "outputEnabled": boolean,
               "type": "string"
            }
         ],
         "wordsConfig": [
            {
               "inputAction": "string",
               "inputEnabled": boolean,
               "outputAction": "string",
```

```
            "outputEnabled": boolean,
            "text": "string"
         }
      ]
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The unique identifier of the guardrail. This can be an ID or the ARN.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

Request Body

The request accepts the following data in JSON format.

automatedReasoningPolicyConfig

Updated configuration for Automated Reasoning policies associated with the guardrail.

Type: GuardrailAutomatedReasoningPolicyConfig object

Required: No

blockedInputMessaging

The message to return when the guardrail blocks a prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

blockedOutputsMessaging

The message to return when the guardrail blocks a model response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

contentPolicyConfig

The content policy to configure for the guardrail.

Type: GuardrailContentPolicyConfig object

Required: No

contextualGroundingPolicyConfig

The contextual grounding policy configuration used to update a guardrail.

Type: GuardrailContextualGroundingPolicyConfig object

Required: No

crossRegionConfig

The system-defined guardrail profile that you're using with your guardrail. Guardrail profiles define the destination AWS Regions where guardrail inference requests can be automatically routed.

For more information, see the Amazon Bedrock User Guide.

Type: GuardrailCrossRegionConfig object

Required: No

description

A description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

kmsKeyId

The ARN of the AWS KMS key with which to encrypt the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

name

A name for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `[0-9a-zA-Z-_]+`

Required: Yes

sensitiveInformationPolicyConfig

The sensitive information policy to configure for the guardrail.

Type: GuardrailSensitiveInformationPolicyConfig object

Required: No

topicPolicyConfig

The topic policy to configure for the guardrail.

Type: GuardrailTopicPolicyConfig object

Required: No

wordPolicyConfig

The word policy to configure for the guardrail.

Type: GuardrailWordPolicyConfig object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "guardrailArn": "string",
   "guardrailId": "string",
   "updatedAt": "string",
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

guardrailArn

The ARN of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+`

guardrailId

The unique identifier of the guardrail

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

updatedAt

The date and time at which the guardrail was updated.

Type: Timestamp

version

The version of the guardrail.

Type: String

Length Constraints: Fixed length of 5.

Pattern: DRAFT

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateMarketplaceModelEndpoint

Service: Amazon Bedrock

Updates the configuration of an existing endpoint for a model from Amazon Bedrock Marketplace.

Request Syntax

```
PATCH /marketplace-model/endpoints/endpointArn HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "endpointConfig": { ... }
}
```

URI Request Parameters

The request uses the following URI parameters.

endpointArn

The Amazon Resource Name (ARN) of the endpoint you want to update.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientRequestToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. This token is listed as not required because AWS SDKs automatically generate it for you and set this parameter. If you're not using the AWS SDK or the AWS CLI, you must provide this token or the action will fail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([-a-zA-Z0-9]{0,254}[a-zA-Z0-9])?`

Required: No

endpointConfig

The new configuration for the endpoint, including the number and type of instances to use.

Type: EndpointConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "marketplaceModelEndpoint": {
      "createdAt": "string",
      "endpointArn": "string",
      "endpointConfig": { ... },
      "endpointStatus": "string",
      "endpointStatusMessage": "string",
      "modelSourceIdentifier": "string",
      "status": "string",
      "statusMessage": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

marketplaceModelEndpoint

Details about the updated endpoint.

Type: MarketplaceModelEndpoint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateProvisionedModelThroughput

Service: Amazon Bedrock

Updates the name or associated model for a Provisioned Throughput. For more information, see Provisioned Throughput in the Amazon Bedrock User Guide.

Request Syntax

```
PATCH /provisioned-model-throughput/provisionedModelId HTTP/1.1
Content-type: application/json

{
   "desiredModelId": "string",
   "desiredProvisionedModelName": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId

The Amazon Resource Name (ARN) or ID of the Provisioned Throughput to update.

Pattern: `(([0-9a-zA-Z][_-]?)+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))`

Required: Yes

Request Body

The request accepts the following data in JSON format.

desiredModelId

The model ID (for a foundation model), name (for a custom model), or Amazon Resource Name (ARN) of the new model to associate with this Provisioned Throughput. You can't specify this field if this Provisioned Throughput is associated with a base model.

If this Provisioned Throughput is associated with a custom model, you can specify one of the following options:

- The base model from which the custom model was customized.
- Another custom model that was customized from the same base model as the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)`

Required: No

desiredProvisionedModelName

The new name for this Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Required: No

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Agents for Amazon Bedrock

The following actions are supported by Agents for Amazon Bedrock:

- AssociateAgentCollaborator
- AssociateAgentKnowledgeBase
- CreateAgent
- CreateAgentActionGroup
- CreateAgentAlias
- CreateDataSource
- CreateFlow
- CreateFlowAlias
- CreateFlowVersion
- CreateKnowledgeBase
- CreatePrompt
- CreatePromptVersion
- CreateVpcConfiguration
- DeleteAgent
- DeleteAgentActionGroup
- DeleteAgentAlias
- DeleteAgentVersion
- DeleteDataSource
- DeleteFlow
- DeleteFlowAlias
- DeleteFlowVersion
- DeleteKnowledgeBase

- DeleteKnowledgeBaseDocuments
- DeletePrompt
- DeleteResourcePolicy
- DeleteVpcConfiguration
- DisassociateAgentCollaborator
- DisassociateAgentKnowledgeBase
- GetAgent
- GetAgentActionGroup
- GetAgentAlias
- GetAgentCollaborator
- GetAgentKnowledgeBase
- GetAgentVersion
- GetDataSource
- GetFlow
- GetFlowAlias
- GetFlowVersion
- GetIngestionJob
- GetKnowledgeBase
- GetKnowledgeBaseDocuments
- GetPrompt
- GetResourcePolicy
- GetVpcConfiguration
- IngestKnowledgeBaseDocuments
- ListAgentActionGroups
- ListAgentAliases
- ListAgentCollaborators
- ListAgentKnowledgeBases
- ListAgents
- ListAgentVersions
- ListDataSources

- ListFlowAliases
- ListFlows
- ListFlowVersions
- ListIngestionJobs
- ListKnowledgeBaseDocuments
- ListKnowledgeBases
- ListPrompts
- ListTagsForResource
- ListVpcConfigurations
- PrepareAgent
- PrepareFlow
- PutResourcePolicy
- StartIngestionJob
- StopIngestionJob
- TagResource
- UntagResource
- UpdateAgent
- UpdateAgentActionGroup
- UpdateAgentAlias
- UpdateAgentCollaborator
- UpdateAgentKnowledgeBase
- UpdateDataSource
- UpdateFlow
- UpdateFlowAlias
- UpdateKnowledgeBase
- UpdatePrompt
- ValidateFlowDefinition

AssociateAgentCollaborator

Service: Agents for Amazon Bedrock

Makes an agent a collaborator for another agent.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/agentcollaborators/ HTTP/1.1
Content-type: application/json

{
   "agentDescriptor": {
      "aliasArn": "string"
   },
   "clientToken": "string",
   "collaborationInstruction": "string",
   "collaboratorName": "string",
   "relayConversationHistory": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The agent's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

An agent version.

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentDescriptor

The alias of the collaborator agent.

Type: AgentDescriptor object

Required: Yes

clientToken

A client token.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

collaborationInstruction

Instruction for the collaborator.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: Yes

collaboratorName

A name for the collaborator.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

relayConversationHistory

A relay conversation history for the collaborator.

Type: String

Valid Values: `TO_COLLABORATOR | DISABLED`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentCollaborator": {
      "agentDescriptor": {
         "aliasArn": "string"
      },
      "agentId": "string",
      "agentVersion": "string",
      "clientToken": "string",
      "collaborationInstruction": "string",
      "collaboratorId": "string",
      "collaboratorName": "string",
      "createdAt": "string",
      "lastUpdatedAt": "string",
      "relayConversationHistory": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentCollaborator

Details about the collaborator.

Type: AgentCollaborator object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

AssociateAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Associates a knowledge base with an agent. If a knowledge base is associated and its `indexState` is set to `Enabled`, the agent queries the knowledge base for information to augment its response to the user.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/knowledgebases/ HTTP/1.1
Content-type: application/json

{
   "description": "string",
   "knowledgeBaseId": "string",
   "knowledgeBaseState": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent with which you want to associate the knowledge base.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent with which you want to associate the knowledge base.

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

A description of what the agent should use the knowledge base for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to associate with the agent.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseState

Specifies whether to use the knowledge base or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentKnowledgeBase": {
      "agentId": "string",
      "agentVersion": "string",
      "createdAt": "string",
      "description": "string",
      "knowledgeBaseId": "string",
      "knowledgeBaseState": "string",
      "updatedAt": "string"
```

```
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentKnowledgeBase

Contains details about the knowledge base that has been associated with the agent.

Type: AgentKnowledgeBase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAgent

Service: Agents for Amazon Bedrock

> **Note**
>
> Amazon Bedrock Agents (now Amazon Bedrock Agents Classic) is no longer open to new customers. For capabilities similar to Bedrock Agents Classic, explore Amazon Bedrock AgentCore. Existing customers can continue to use the service as normal. For more information, see Amazon Bedrock Agents Classic availability change.

Creates an agent that orchestrates interactions between foundation models, data sources, software applications, user conversations, and APIs to carry out tasks to help customers.

- Specify the following fields for security purposes.
  - `agentResourceRoleArn` – The Amazon Resource Name (ARN) of the role with permissions to invoke API operations on an agent.
  - (Optional) `customerEncryptionKeyArn` – The Amazon Resource Name (ARN) of a AWS KMS key to encrypt the creation of the agent.
  - (Optional) `idleSessionTTLinSeconds` – Specify the number of seconds for which the agent should maintain session information. After this time expires, the subsequent `InvokeAgent` request begins a new session.
- To enable your agent to retain conversational context across multiple sessions, include a `memoryConfiguration` object. For more information, see Configure memory.
- To override the default prompt behavior for agent orchestration and to use advanced prompts, include a `promptOverrideConfiguration` object. For more information, see Advanced prompts.
- If your agent fails to be created, the response returns a list of `failureReasons` alongside a list of `recommendedActions` for you to troubleshoot.
- The agent instructions will not be honored if your agent has only one knowledge base, uses default prompts, has no action group, and user input is disabled.

Request Syntax

```
PUT /agents/ HTTP/1.1
Content-type: application/json
```

```
{
   "agentCollaboration": "string",
   "agentName": "string",
   "agentResourceRoleArn": "string",
   "clientToken": "string",
   "customerEncryptionKeyArn": "string",
   "customOrchestration": {
      "executor": { ... }
   },
   "description": "string",
   "foundationModel": "string",
   "guardrailConfiguration": {
      "guardrailIdentifier": "string",
      "guardrailVersion": "string"
   },
   "idleSessionTTLInSeconds": number,
   "instruction": "string",
   "memoryConfiguration": {
      "enabledMemoryTypes": [ "string" ],
      "sessionSummaryConfiguration": {
         "maxRecentSessions": number
      },
      "storageDays": number
   },
   "orchestrationType": "string",
   "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
         {
            "additionalModelRequestFields": JSON value,
            "basePromptTemplate": "string",
            "foundationModel": "string",
            "inferenceConfiguration": {
               "maximumLength": number,
               "stopSequences": [ "string" ],
               "temperature": number,
               "topK": number,
               "topP": number
            },
            "parserMode": "string",
            "promptCreationMode": "string",
            "promptState": "string",
            "promptType": "string"
```

```
            }
         ]
      },
      "tags": {
         "string" : "string"
      }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

agentCollaboration

The agent's collaboration role.

Type: String

Valid Values: `SUPERVISOR | SUPERVISOR_ROUTER | DISABLED`

Required: No

agentName

A name for the agent that you create.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key with which to encrypt the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

customOrchestration

Contains details of the custom orchestration configured for the agent.

Type: CustomOrchestration object

Required: No

description

A description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

foundationModel

The identifier for the model that you want to be used for orchestration by the agent you create.

The `modelId` to provide depends on the type of model or throughput that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see Amazon Bedrock base model IDs (on-demand throughput) in the Amazon Bedrock User Guide.
- If you use an inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for cross-region inference in the Amazon Bedrock User Guide.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see Run inference using a Provisioned Throughput in the Amazon Bedrock User Guide.
- If you use a custom model, first purchase Provisioned Throughput for it. Then specify the ARN of the resulting provisioned model. For more information, see Use a custom model in Amazon Bedrock in the Amazon Bedrock User Guide.
- If you use an imported model, specify the ARN of the imported model. You can get the model ARN from a successful call to CreateModelImportJob or from the Imported models page in the Amazon Bedrock console.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

guardrailConfiguration

The unique Guardrail configuration assigned to the agent when it is created.

Type: GuardrailConfiguration object

Required: No

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 5400.

Required: No

instruction

Instructions that tell the agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 4000.

Required: No

memoryConfiguration

Contains the details of the memory configured for the agent.

Type: MemoryConfiguration object

Required: No

orchestrationType

Specifies the type of orchestration strategy for the agent. This is set to `DEFAULT` orchestration type, by default.

Type: String

Valid Values: `DEFAULT | CUSTOM_ORCHESTRATION`

Required: No

promptOverrideConfiguration

Contains configurations to override prompts in different parts of an agent sequence. For more information, see Advanced prompts.

Type: PromptOverrideConfiguration object

Required: No

tags

Any tags that you want to attach to the agent.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agent": {
      "agentArn": "string",
      "agentCollaboration": "string",
      "agentId": "string",
      "agentName": "string",
      "agentResourceRoleArn": "string",
      "agentStatus": "string",
      "agentVersion": "string",
      "clientToken": "string",
      "createdAt": "string",
      "customerEncryptionKeyArn": "string",
      "customOrchestration": {
         "executor": { ... }
      },
      "description": "string",
      "failureReasons": [ "string" ],
      "foundationModel": "string",
      "guardrailConfiguration": {
```

```
            "guardrailIdentifier": "string",
            "guardrailVersion": "string"
         },
         "idleSessionTTLInSeconds": number,
         "instruction": "string",
         "memoryConfiguration": {
            "enabledMemoryTypes": [ "string" ],
            "sessionSummaryConfiguration": {
               "maxRecentSessions": number
            },
            "storageDays": number
         },
         "orchestrationType": "string",
         "preparedAt": "string",
         "promptOverrideConfiguration": {
            "overrideLambda": "string",
            "promptConfigurations": [
               {
                  "additionalModelRequestFields": JSON value,
                  "basePromptTemplate": "string",
                  "foundationModel": "string",
                  "inferenceConfiguration": {
                     "maximumLength": number,
                     "stopSequences": [ "string" ],
                     "temperature": number,
                     "topK": number,
                     "topP": number
                  },
                  "parserMode": "string",
                  "promptCreationMode": "string",
                  "promptState": "string",
                  "promptType": "string"
               }
            ]
         },
         "recommendedActions": [ "string" ],
         "updatedAt": "string"
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agent

Contains details about the agent created.

Type: Agent object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of CreateAgent.

```
PUT /agents/ HTTP/1.1
Content-type: application/json

{
  "agentName": "o1nvve1",
  "agentResourceRoleArn": "arn:aws:iam::123456789012:role/
AmazonBedrockExecutionRoleForAgents_user",
  "instruction": "You are an IT agent who solves customer's problems",
  "description": "Description is here",
  "idleSessionTTLInSeconds": 900,
  "foundationModel": "anthropic.claude-v2"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAgentActionGroup

Service: Agents for Amazon Bedrock

Creates an action group for an agent. An action group represents the actions that an agent can carry out for the customer by defining the APIs that an agent can call and the logic for calling them.

To allow your agent to request the user for additional information when trying to complete a task, add an action group with the `parentActionGroupSignature` field set to `AMAZON.UserInput`.

To allow your agent to generate, run, and troubleshoot code when trying to complete a task, add an action group with the `parentActionGroupSignature` field set to `AMAZON.CodeInterpreter`.

You must leave the `description`, `apiSchema`, and `actionGroupExecutor` fields blank for this action group. During orchestration, if your agent determines that it needs to invoke an API in an action group, but doesn't have enough information to complete the API request, it will invoke this action group instead and return an Observation reprompting the user for more information.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/actiongroups/ HTTP/1.1
Content-type: application/json

{
   "actionGroupExecutor": { ... },
   "actionGroupName": "string",
   "actionGroupState": "string",
   "apiSchema": { ... },
   "clientToken": "string",
   "description": "string",
   "functionSchema": { ... },
   "parentActionGroupSignature": "string",
   "parentActionGroupSignatureParams": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent for which to create the action group.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent for which to create the action group.

Length Constraints: Fixed length of 5.

Pattern: DRAFT

Required: Yes

Request Body

The request accepts the following data in JSON format.

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Type: ActionGroupExecutor object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

actionGroupName

The name to give the action group.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see Action group OpenAPI schemas.

Type: APISchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

A description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Contains details about the function schema for the action group or the JSON or YAML-formatted payload defining the schema.

Type: FunctionSchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionGroupSignature

Specify a built-in or computer use action for this action group. If you specify a value, you must leave the `description`, `apiSchema`, and `actionGroupExecutor` fields empty for this action group.

- To allow your agent to request the user for additional information when trying to complete a task, set this field to `AMAZON.UserInput`.
- To allow your agent to generate, run, and troubleshoot code when trying to complete a task, set this field to `AMAZON.CodeInterpreter`.
- To allow your agent to use an Anthropic computer use tool, specify one of the following values.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Anthropic Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. When operating computer use functionality, we recommend taking additional security precautions, such as executing computer actions in virtual environments with restricted data access and limited internet connectivity. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

- `ANTHROPIC.Computer` - Gives the agent permission to use the mouse and keyboard and take screenshots.
- `ANTHROPIC.TextEditor` - Gives the agent permission to view, create and edit files.
- `ANTHROPIC.Bash` - Gives the agent permission to run commands in a bash shell.

Type: String

Valid Values: `AMAZON.UserInput | AMAZON.CodeInterpreter | ANTHROPIC.Computer | ANTHROPIC.Bash | ANTHROPIC.TextEditor`

Required: No

parentActionGroupSignatureParams

The configuration settings for a computer use action.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Anthropic Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Type: String to string map

Key Length Constraints: Minimum length of 0. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 100.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentActionGroup": {
      "actionGroupExecutor": { ... },
      "actionGroupId": "string",
      "actionGroupName": "string",
      "actionGroupState": "string",
      "agentId": "string",
      "agentVersion": "string",
      "apiSchema": { ... },
      "clientToken": "string",
      "createdAt": "string",
```

```
      "description": "string",
      "functionSchema": { ... },
      "parentActionGroupSignatureParams": {
         "string" : "string"
      },
      "parentActionSignature": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentActionGroup

Contains details about the action group that was created.

Type: AgentActionGroup object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Create an action group using an OpenAPI schema and a Lambda function

The following example creates an action group using an OpenAPI schema uploaded to an Amazon S3 bucket and sends the information elicited from the user to a Lambda function.

```
PUT /agents/AGENT12345/agentversions/DRAFT/actiongroups/ HTTP/1.1
Content-type: application/json

{
   "actionGroupName": "Test Action",
   "actionGroupState": "ENABLED",
   "apiSchema": {
```

```
        "s3": {
            "s3BucketName": "apischema-s3",
            "s3ObjectKey": "it_agent_openapi.json"
        }
    },
   "description": "Testing latest IT Management action",
   "actionGroupExecutor": {
        "lambda": "arn:aws:lambda:us-west-2:123456789012:function:ItAgentLambda"
    }
}
```

Create an action group using an OpenAPI schema and return control

The following example creates an action group using an OpenAPI schema uploaded to an Amazon S3 bucket and returns control by sending the information in the `InvokeAgent` response.

```
{
    "actionGroupName": "WeatherAPIs",
    "description": "Actions to get current weather and historical trends for a
 location",
    "actionGroupState": "ENABLED",
    "apiSchema": {
        "s3": {
            "s3BucketName": "openapi-spec-iad",
            "s3ObjectKey": "get_weather_openapi.yaml"
        }
    },
    "actionGroupExecutor": {
        "customControl": "RETURN_CONTROL"
    }
}
```

Create an action group using function details and return control

The following example creates an action group using function details and returns control by sending the information in the `InvokeAgent` response

```
PUT /agents/AGENT12345/agentversions/DRAFT/actiongroups/ HTTP/1.1
Content-type: application/json

{
    "actionGroupName": "OrderManagementAction",
```

```
    "description": "Action to get the order history, product details, product
 availability and to update the order",
    "actionGroupState": "ENABLED",
    "actionGroupExecutor": {
        "customControl": "RETURN_CONTROL"
    },
    "functionSchema": {
        "functions": [{
            "name": "GetOrderDetails",
            "description": "Retrieves the order history for a given OrderId and returns
 productId, color, productName, size, productType, quantity, and status."
            "parameters": {
                "orderId": {
                    "type": "string",
                    "required": true
                }
            }
        }]
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateAgentAlias

Service: Agents for Amazon Bedrock

Creates an alias of an agent that can be used to deploy the agent.

Request Syntax

```
PUT /agents/agentId/agentaliases/ HTTP/1.1
Content-type: application/json

{
   "agentAliasName": "string",
   "clientToken": "string",
   "description": "string",
   "routingConfiguration": [
      {
         "agentVersion": "string",
         "provisionedThroughput": "string"
      }
   ],
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentAliasName

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

A description of the alias of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

routingConfiguration

Contains details about the routing configuration of the alias.

Type: Array of AgentAliasRoutingConfigurationListItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

tags

Any tags that you want to attach to the alias of the agent.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agentAlias": {
      "agentAliasArn": "string",
      "agentAliasHistoryEvents": [
         {
            "endDate": "string",
            "routingConfiguration": [
               {
                  "agentVersion": "string",
                  "provisionedThroughput": "string"
               }
            ],
            "startDate": "string"
         }
      ],
      "agentAliasId": "string",
      "agentAliasName": "string",
      "agentAliasStatus": "string",
      "agentId": "string",
      "aliasInvocationState": "string",
      "clientToken": "string",
      "createdAt": "string",
```

```
      "description": "string",
      "failureReasons": [ "string" ],
      "routingConfiguration": [
         {
            "agentVersion": "string",
            "provisionedThroughput": "string"
         }
      ],
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentAlias

Contains details about the alias that was created.

Type: AgentAlias object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of CreateAgentAlias.

```
PUT /agents/ABCDEFGHIJ/agentaliases/ HTTP/1.1
Content-type: application/json

{
 "agentAliasName": "TestName",
 "description": "Alias is test"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateDataSource

Service: Agents for Amazon Bedrock

Connects a knowledge base to a data source. You specify the configuration for the specific data source service in the `dataSourceConfiguration` field.

> **Important**
>
> You can't change the `chunkingConfiguration` after you create the data source connector.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "dataDeletionPolicy": "string",
   "dataSourceConfiguration": {
      "confluenceConfiguration": {
         "crawlerConfiguration": {
            "filterConfiguration": {
               "patternObjectFilter": {
                  "filters": [
                     {
                        "exclusionFilters": [ "string" ],
                        "inclusionFilters": [ "string" ],
                        "objectType": "string"
                     }
                  ]
               },
               "type": "string"
            }
         },
         "sourceConfiguration": {
            "authType": "string",
            "credentialsSecretArn": "string",
            "hostType": "string",
            "hostUrl": "string"
         }
```

```
      },
      "managedKnowledgeBaseConnectorConfiguration": {
         "connectorParameters": JSON value,
         "deletionProtectionConfiguration": {
            "deletionProtectionStatus": "string",
            "deletionProtectionThreshold": number
         },
         "mediaExtractionConfiguration": {
            "audioExtractionConfiguration": {
               "audioExtractionStatus": "string"
            },
            "imageExtractionConfiguration": {
               "imageExtractionStatus": "string"
            },
            "videoExtractionConfiguration": {
               "videoExtractionStatus": "string"
            }
         },
         "syncSchedule": { ... }
      },
      "s3Configuration": {
         "bucketArn": "string",
         "bucketOwnerAccountId": "string",
         "inclusionPrefixes": [ "string" ]
      },
      "salesforceConfiguration": {
         "crawlerConfiguration": {
            "filterConfiguration": {
               "patternObjectFilter": {
                  "filters": [
                     {
                        "exclusionFilters": [ "string" ],
                        "inclusionFilters": [ "string" ],
                        "objectType": "string"
                     }
                  ]
               },
               "type": "string"
            }
         },
         "sourceConfiguration": {
            "authType": "string",
            "credentialsSecretArn": "string",
            "hostUrl": "string"
```

```
            }
         },
         "sharePointConfiguration": {
            "crawlerConfiguration": {
               "filterConfiguration": {
                  "patternObjectFilter": {
                     "filters": [
                        {
                           "exclusionFilters": [ "string" ],
                           "inclusionFilters": [ "string" ],
                           "objectType": "string"
                        }
                     ]
                  },
                  "type": "string"
               }
            },
            "sourceConfiguration": {
               "authType": "string",
               "credentialsSecretArn": "string",
               "domain": "string",
               "hostType": "string",
               "siteUrls": [ "string" ],
               "tenantId": "string"
            }
         },
         "type": "string",
         "webConfiguration": {
            "crawlerConfiguration": {
               "crawlerLimits": {
                  "maxPages": number,
                  "rateLimit": number
               },
               "exclusionFilters": [ "string" ],
               "inclusionFilters": [ "string" ],
               "scope": "string",
               "userAgent": "string",
               "userAgentHeader": "string"
            },
            "sourceConfiguration": {
               "urlConfiguration": {
                  "seedUrls": [
                     {
                        "url": "string"
```

```
                }
             ]
          }
       }
    }
 },
 "description": "string",
 "name": "string",
 "serverSideEncryptionConfiguration": {
    "kmsKeyArn": "string"
 },
 "vectorIngestionConfiguration": {
    "chunkingConfiguration": {
       "chunkingStrategy": "string",
       "fixedSizeChunkingConfiguration": {
          "maxTokens": number,
          "overlapPercentage": number
       },
       "hierarchicalChunkingConfiguration": {
          "levelConfigurations": [
             {
                "maxTokens": number
             }
          ],
          "overlapTokens": number
       },
       "semanticChunkingConfiguration": {
          "breakpointPercentileThreshold": number,
          "bufferSize": number,
          "maxTokens": number
       }
    },
    "contextEnrichmentConfiguration": {
       "bedrockFoundationModelConfiguration": {
          "enrichmentStrategyConfiguration": {
             "method": "string"
          },
          "modelArn": "string"
       },
       "type": "string"
    },
    "customTransformationConfiguration": {
       "intermediateStorage": {
          "s3Location": {
```

```
                "uri": "string"
             }
          },
          "transformations": [
             {
                "stepToApply": "string",
                "transformationFunction": {
                   "transformationLambdaConfiguration": {
                      "lambdaArn": "string"
                   }
                }
             }
          ]
       },
       "parsingConfiguration": {
          "bedrockDataAutomationConfiguration": {
             "parsingModality": "string"
          },
          "bedrockFoundationModelConfiguration": {
             "modelArn": "string",
             "parsingModality": "string",
             "parsingPrompt": {
                "parsingPromptText": "string"
             }
          },
          "parsingStrategy": "string"
       }
    }
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to which to add the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

dataDeletionPolicy

The data deletion policy for the data source.

You can set the data deletion policy to:

- DELETE: Deletes all data from your data source that's converted into vector embeddings upon deletion of a knowledge base or data source resource. Note that the **vector store itself is not deleted**, only the data. This flag is ignored if an AWS account is deleted.
- RETAIN: Retains all data from your data source that's converted into vector embeddings upon deletion of a knowledge base or data source resource. Note that the **vector store itself is not deleted** if you delete a knowledge base or data source resource.

> **Note**
>
> For managed knowledge bases, the only supported option is `DELETE`, which is also the default.

Type: String

Valid Values: `RETAIN | DELETE`

Required: No

dataSourceConfiguration

The connection configuration for the data source.

Type: DataSourceConfiguration object

Required: Yes

description

A description of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

The name of the data source.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

serverSideEncryptionConfiguration

Contains details about the server-side encryption for the data source.

Type: ServerSideEncryptionConfiguration object

Required: No

vectorIngestionConfiguration

Contains details about how to ingest the documents in the data source.

Type: VectorIngestionConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "dataSource": {
      "createdAt": "string",
      "dataDeletionPolicy": "string",
      "dataSourceConfiguration": {
         "confluenceConfiguration": {
            "crawlerConfiguration": {
               "filterConfiguration": {
                  "patternObjectFilter": {
                     "filters": [
                        {
                           "exclusionFilters": [ "string" ],
                           "inclusionFilters": [ "string" ],
                           "objectType": "string"
                        }
                     ]
                  },
                  "type": "string"
               }
            },
            "sourceConfiguration": {
               "authType": "string",
               "credentialsSecretArn": "string",
               "hostType": "string",
               "hostUrl": "string"
            }
         },
         "managedKnowledgeBaseConnectorConfiguration": {
            "connectorParameters": JSON value,
            "deletionProtectionConfiguration": {
               "deletionProtectionStatus": "string",
               "deletionProtectionThreshold": number
            },
            "mediaExtractionConfiguration": {
               "audioExtractionConfiguration": {
                  "audioExtractionStatus": "string"
               },
               "imageExtractionConfiguration": {
```

```
                "imageExtractionStatus": "string"
            },
            "videoExtractionConfiguration": {
                "videoExtractionStatus": "string"
            }
        },
        "syncSchedule": { ... }
    },
    "s3Configuration": {
        "bucketArn": "string",
        "bucketOwnerAccountId": "string",
        "inclusionPrefixes": [ "string" ]
    },
    "salesforceConfiguration": {
        "crawlerConfiguration": {
            "filterConfiguration": {
                "patternObjectFilter": {
                    "filters": [
                        {
                            "exclusionFilters": [ "string" ],
                            "inclusionFilters": [ "string" ],
                            "objectType": "string"
                        }
                    ]
                },
                "type": "string"
            }
        },
        "sourceConfiguration": {
            "authType": "string",
            "credentialsSecretArn": "string",
            "hostUrl": "string"
        }
    },
    "sharePointConfiguration": {
        "crawlerConfiguration": {
            "filterConfiguration": {
                "patternObjectFilter": {
                    "filters": [
                        {
                            "exclusionFilters": [ "string" ],
                            "inclusionFilters": [ "string" ],
                            "objectType": "string"
                        }
```

```
                    ]
                },
                "type": "string"
            }
        },
        "sourceConfiguration": {
            "authType": "string",
            "credentialsSecretArn": "string",
            "domain": "string",
            "hostType": "string",
            "siteUrls": [ "string" ],
            "tenantId": "string"
        }
    },
    "type": "string",
    "webConfiguration": {
        "crawlerConfiguration": {
            "crawlerLimits": {
                "maxPages": number,
                "rateLimit": number
            },
            "exclusionFilters": [ "string" ],
            "inclusionFilters": [ "string" ],
            "scope": "string",
            "userAgent": "string",
            "userAgentHeader": "string"
        },
        "sourceConfiguration": {
            "urlConfiguration": {
                "seedUrls": [
                    {
                        "url": "string"
                    }
                ]
            }
        }
    }
},
"dataSourceId": "string",
"description": "string",
"failureReasons": [ "string" ],
"knowledgeBaseId": "string",
"name": "string",
"serverSideEncryptionConfiguration": {
```

```
            "kmsKeyArn": "string"
        },
        "status": "string",
        "updatedAt": "string",
        "vectorIngestionConfiguration": {
            "chunkingConfiguration": {
                "chunkingStrategy": "string",
                "fixedSizeChunkingConfiguration": {
                    "maxTokens": number,
                    "overlapPercentage": number
                },
                "hierarchicalChunkingConfiguration": {
                    "levelConfigurations": [
                        {
                            "maxTokens": number
                        }
                    ],
                    "overlapTokens": number
                },
                "semanticChunkingConfiguration": {
                    "breakpointPercentileThreshold": number,
                    "bufferSize": number,
                    "maxTokens": number
                }
            },
            "contextEnrichmentConfiguration": {
                "bedrockFoundationModelConfiguration": {
                    "enrichmentStrategyConfiguration": {
                        "method": "string"
                    },
                    "modelArn": "string"
                },
                "type": "string"
            },
            "customTransformationConfiguration": {
                "intermediateStorage": {
                    "s3Location": {
                        "uri": "string"
                    }
                },
                "transformations": [
                    {
                        "stepToApply": "string",
                        "transformationFunction": {
```

```
                    "transformationLambdaConfiguration": {
                       "lambdaArn": "string"
                    }
                 }
              }
           ]
        },
        "parsingConfiguration": {
           "bedrockDataAutomationConfiguration": {
              "parsingModality": "string"
           },
           "bedrockFoundationModelConfiguration": {
              "modelArn": "string",
              "parsingModality": "string",
              "parsingPrompt": {
                 "parsingPromptText": "string"
              }
           },
           "parsingStrategy": "string"
        }
     }
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSource

Contains details about the data source.

Type: DataSource object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateFlow

Service: Agents for Amazon Bedrock

Creates a prompt flow that you can use to send an input through various steps to yield an output. Configure nodes, each of which corresponds to a step of the flow, and create connections between the nodes to create paths to different outputs. For more information, see How it works and Create a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
POST /flows/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "customerEncryptionKeyArn": "string",
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
                  "type": "string"
               }
            ],
            "name": "string",
            "outputs": [
               {
                  "name": "string",
                  "type": "string"
```

```
                }
            ],
            "type": "string"
        }
    ]
  },
  "description": "string",
  "executionRoleArn": "string",
  "name": "string",
  "tags": {
     "string" : "string"
  }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to encrypt the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

definition

A definition of the nodes and connections between nodes in the flow.

Type: FlowDefinition object

Required: No

description

A description for the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create and manage a flow. For more information, see Create a service role for flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

Required: Yes

name

A name for the flow.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

tags

Any tags that you want to attach to the flow. For more information, see Tagging resources in Amazon Bedrock.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
```

```
                    "expression": "string",
                    "name": "string",
                    "type": "string"
                 }
              ],
              "name": "string",
              "outputs": [
                 {
                    "name": "string",
                    "type": "string"
                 }
              ],
              "type": "string"
           }
        ]
    },
    "description": "string",
    "executionRoleArn": "string",
    "id": "string",
    "name": "string",
    "status": "string",
    "updatedAt": "string",
    "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

createdAt

The time at which the flow was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key that you encrypted the flow with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

definition

A definition of the nodes and connections between nodes in the flow.

Type: FlowDefinition object

description

The description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create a flow. For more information, see Create a service role for flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the flow.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

status

The status of the flow. When you submit this request, the status will be `NotPrepared`. If creation fails, the status becomes `Failed`.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

updatedAt

The time at which the flow was last updated.

Type: Timestamp

version

The version of the flow. When you create a flow, the version created is the `DRAFT` version.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateFlowAlias

Service: Agents for Amazon Bedrock

Creates an alias of a flow for deployment. For more information, see Deploy a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
POST /flows/flowIdentifier/aliases HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "concurrencyConfiguration": {
      "maxConcurrency": number,
      "type": "string"
   },
   "description": "string",
   "name": "string",
   "routingConfiguration": [
      {
         "flowVersion": "string"
      }
   ],
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow for which to create an alias.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

concurrencyConfiguration

The configuration that specifies how nodes in the flow are executed in parallel.

Type: FlowAliasConcurrencyConfiguration object

Required: No

description

A description for the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

A name for the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

routingConfiguration

Contains information about the version to which to map the alias.

Type: Array of FlowAliasRoutingConfigurationListItem objects

Array Members: Fixed number of 1 item.

Required: Yes

tags

Any tags that you want to attach to the alias of the flow. For more information, see Tagging resources in Amazon Bedrock.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "arn": "string",
   "concurrencyConfiguration": {
      "maxConcurrency": number,
      "type": "string"
   },
   "createdAt": "string",
   "description": "string",
   "flowId": "string",
   "id": "string",
   "name": "string",
```

```
   "routingConfiguration": [
      {
         "flowVersion": "string"
      }
   ],
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the alias.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/(TSTALIASID|[0-9a-zA-Z]{10})`

concurrencyConfiguration

The configuration that specifies how nodes in the flow are executed in parallel.

Type: FlowAliasConcurrencyConfiguration object

createdAt

The time at which the alias was created.

Type: Timestamp

description

The description of the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

flowId

The unique identifier of the flow that the alias belongs to.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

id

The unique identifier of the alias.

Type: String

Pattern: `(TSTALIASID|[0-9a-zA-Z]{10})`

name

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

routingConfiguration

Contains information about the version that the alias is mapped to.

Type: Array of FlowAliasRoutingConfigurationListItem objects

Array Members: Fixed number of 1 item.

updatedAt

The time at which the alias of the flow was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateFlowVersion

Service: Agents for Amazon Bedrock

Creates a version of the flow that you can deploy. For more information, see Deploy a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
POST /flows/flowIdentifier/versions HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "description": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow that you want to create a version of.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

A description of the version of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
```

```
                    "type": "string"
                }
            ],
            "name": "string",
            "outputs": [
                {
                    "name": "string",
                    "type": "string"
                }
            ],
            "type": "string"
        }
    ]
  },
  "description": "string",
  "executionRoleArn": "string",
  "id": "string",
  "name": "string",
  "status": "string",
  "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

createdAt

The time at which the flow was created.

Type: Timestamp

customerEncryptionKeyArn

The KMS key that the flow is encrypted with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

definition

A definition of the nodes and connections in the flow.

Type: FlowDefinition object

description

The description of the version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create a flow. For more information, see Create a service role for flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the version.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

status

The status of the flow.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

version

The version of the flow that was created. Versions are numbered incrementally, starting from 1.

Type: String

Pattern: `[0-9]{1,5}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateKnowledgeBase

Service: Agents for Amazon Bedrock

Creates a knowledge base. A knowledge base contains your data sources so that Large Language Models (LLMs) can use your data. To create a knowledge base, you must first set up your data sources and configure a supported vector store. For more information, see Set up a knowledge base.

> **Note**
>
> To create a managed knowledge base, provide a `managedKnowledgeBaseConfiguration` during creation. For more information, see Build a managed knowledge base.

- Provide the `name` and an optional `description`.
- Provide the Amazon Resource Name (ARN) with permissions to create a knowledge base in the `roleArn` field.
- For managed knowledge bases, set `embeddingModelType` to `MANAGED` to use the service-managed embedding model, or `CUSTOM` with an `embeddingModelArn` to use your own. To use your own KMS key for encryption, provide the ARN in `serverSideEncryptionConfiguration`. No vector store configuration is required for managed knowledge bases.
- For self-managed knowledge bases, provide the embedding model to use in the `embeddingModelArn` field in the `knowledgeBaseConfiguration` object.
- For self-managed knowledge bases, provide the configuration for your vector store in the `storageConfiguration` object.
  - For an Amazon OpenSearch Service database, use the `opensearchServerlessConfiguration` object. For more information, see Create a vector store in Amazon OpenSearch Service.
  - For an Amazon Aurora database, use the `RdsConfiguration` object. For more information, see Create a vector store in Amazon Aurora.
  - For a Pinecone database, use the `pineconeConfiguration` object. For more information, see Create a vector store in Pinecone.

- For a Redis Enterprise Cloud database, use the `redisEnterpriseCloudConfiguration` object. For more information, see Create a vector store in Redis Enterprise Cloud.

Request Syntax

```
PUT /knowledgebases/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "description": "string",
   "knowledgeBaseConfiguration": {
      "kendraKnowledgeBaseConfiguration": {
         "kendraIndexArn": "string"
      },
      "managedKnowledgeBaseConfiguration": {
         "embeddingModelArn": "string",
         "embeddingModelConfiguration": {
            "bedrockEmbeddingModelConfiguration": {
               "audio": [
                  {
                     "segmentationConfiguration": {
                        "fixedLengthDuration": number
                     }
                  }
               ],
               "dimensions": number,
               "embeddingDataType": "string",
               "modelConfiguration": JSON value,
               "video": [
                  {
                     "segmentationConfiguration": {
                        "fixedLengthDuration": number
                     }
                  }
               ]
            }
         },
         "embeddingModelType": "string",
         "serverSideEncryptionConfiguration": {
            "kmsKeyArn": "string"
         },
```

```
      "supplementalDataStorageConfiguration": {
         "storageLocations": [
            {
               "s3Location": {
                  "uri": "string"
               },
               "type": "string"
            }
         ]
      }
   },
   "sqlKnowledgeBaseConfiguration": {
      "redshiftConfiguration": {
         "queryEngineConfiguration": {
            "provisionedConfiguration": {
               "authConfiguration": {
                  "databaseUser": "string",
                  "type": "string",
                  "usernamePasswordSecretArn": "string"
               },
               "clusterIdentifier": "string"
            },
            "serverlessConfiguration": {
               "authConfiguration": {
                  "type": "string",
                  "usernamePasswordSecretArn": "string"
               },
               "workgroupArn": "string"
            },
            "type": "string"
         },
         "queryGenerationConfiguration": {
            "executionTimeoutSeconds": number,
            "generationContext": {
               "curatedQueries": [
                  {
                     "naturalLanguage": "string",
                     "sql": "string"
                  }
               ],
               "tables": [
                  {
                     "columns": [
                        {
```

```
                        "description": "string",
                        "inclusion": "string",
                        "name": "string"
                     }
                  ],
                  "description": "string",
                  "inclusion": "string",
                  "name": "string"
               }
            ]
         }
      },
      "storageConfigurations": [
         {
            "awsDataCatalogConfiguration": {
               "tableNames": [ "string" ]
            },
            "redshiftConfiguration": {
               "databaseName": "string"
            },
            "type": "string"
         }
      ]
   },
   "type": "string"
},
"type": "string",
"vectorKnowledgeBaseConfiguration": {
   "embeddingModelArn": "string",
   "embeddingModelConfiguration": {
      "bedrockEmbeddingModelConfiguration": {
         "audio": [
            {
               "segmentationConfiguration": {
                  "fixedLengthDuration": number
               }
            }
         ],
         "dimensions": number,
         "embeddingDataType": "string",
         "modelConfiguration": JSON value,
         "video": [
            {
               "segmentationConfiguration": {
```

```
                        "fixedLengthDuration": number
                     }
                  }
               ]
            }
         },
         "supplementalDataStorageConfiguration": {
            "storageLocations": [
               {
                  "s3Location": {
                     "uri": "string"
                  },
                  "type": "string"
               }
            ]
         }
      }
   },
   "name": "string",
   "roleArn": "string",
   "storageConfiguration": {
      "mongoDbAtlasConfiguration": {
         "collectionName": "string",
         "credentialsSecretArn": "string",
         "databaseName": "string",
         "endpoint": "string",
         "endpointServiceName": "string",
         "fieldMapping": {
            "metadataField": "string",
            "textField": "string",
            "vectorField": "string"
         },
         "textIndexName": "string",
         "vectorIndexName": "string"
      },
      "neptuneAnalyticsConfiguration": {
         "fieldMapping": {
            "metadataField": "string",
            "textField": "string"
         },
         "graphArn": "string"
      },
      "opensearchManagedClusterConfiguration": {
         "domainArn": "string",
```

```
            "domainEndpoint": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "opensearchServerlessConfiguration": {
            "collectionArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "pineconeConfiguration": {
            "connectionString": "string",
            "credentialsSecretArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "namespace": "string"
         },
         "rdsConfiguration": {
            "credentialsSecretArn": "string",
            "databaseName": "string",
            "fieldMapping": {
               "customMetadataField": "string",
               "metadataField": "string",
               "primaryKeyField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "resourceArn": "string",
            "tableName": "string"
         },
         "redisEnterpriseCloudConfiguration": {
            "credentialsSecretArn": "string",
            "endpoint": "string",
            "fieldMapping": {
               "metadataField": "string",
```

```
            "textField": "string",
            "vectorField": "string"
         },
         "vectorIndexName": "string"
      },
      "s3VectorsConfiguration": {
         "indexArn": "string",
         "indexName": "string",
         "vectorBucketArn": "string"
      },
      "type": "string"
   },
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

A description of the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

knowledgeBaseConfiguration

Contains details about the embeddings model used for the knowledge base.

Type: KnowledgeBaseConfiguration object

Required: Yes

name

A name for the knowledge base.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

storageConfiguration

Contains details about the configuration of the vector database used for the knowledge base.

Type: StorageConfiguration object

Required: No

tags

Specify the key-value pairs for the tags that you want to attach to your knowledge base in this object.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "knowledgeBase": {
      "createdAt": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "knowledgeBaseArn": "string",
      "knowledgeBaseConfiguration": {
         "kendraKnowledgeBaseConfiguration": {
            "kendraIndexArn": "string"
         },
         "managedKnowledgeBaseConfiguration": {
            "embeddingModelArn": "string",
            "embeddingModelConfiguration": {
               "bedrockEmbeddingModelConfiguration": {
                  "audio": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ],
                  "dimensions": number,
                  "embeddingDataType": "string",
                  "modelConfiguration": JSON value,
                  "video": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
```

```
                        }
                    }
                ]
            }
        },
        "embeddingModelType": "string",
        "serverSideEncryptionConfiguration": {
            "kmsKeyArn": "string"
        },
        "supplementalDataStorageConfiguration": {
            "storageLocations": [
                {
                    "s3Location": {
                        "uri": "string"
                    },
                    "type": "string"
                }
            ]
        }
    },
    "sqlKnowledgeBaseConfiguration": {
        "redshiftConfiguration": {
            "queryEngineConfiguration": {
                "provisionedConfiguration": {
                    "authConfiguration": {
                        "databaseUser": "string",
                        "type": "string",
                        "usernamePasswordSecretArn": "string"
                    },
                    "clusterIdentifier": "string"
                },
                "serverlessConfiguration": {
                    "authConfiguration": {
                        "type": "string",
                        "usernamePasswordSecretArn": "string"
                    },
                    "workgroupArn": "string"
                },
                "type": "string"
            },
            "queryGenerationConfiguration": {
                "executionTimeoutSeconds": number,
                "generationContext": {
                    "curatedQueries": [
```

```
                        {
                           "naturalLanguage": "string",
                           "sql": "string"
                        }
                     ],
                     "tables": [
                        {
                           "columns": [
                              {
                                 "description": "string",
                                 "inclusion": "string",
                                 "name": "string"
                              }
                           ],
                           "description": "string",
                           "inclusion": "string",
                           "name": "string"
                        }
                     ]
                  }
               },
               "storageConfigurations": [
                  {
                     "awsDataCatalogConfiguration": {
                        "tableNames": [ "string" ]
                     },
                     "redshiftConfiguration": {
                        "databaseName": "string"
                     },
                     "type": "string"
                  }
               ]
            },
            "type": "string"
         },
         "type": "string",
         "vectorKnowledgeBaseConfiguration": {
            "embeddingModelArn": "string",
            "embeddingModelConfiguration": {
               "bedrockEmbeddingModelConfiguration": {
                  "audio": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
```

```
                        }
                     }
                  ],
                  "dimensions": number,
                  "embeddingDataType": "string",
                  "modelConfiguration": JSON value,
                  "video": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ]
               }
            },
            "supplementalDataStorageConfiguration": {
               "storageLocations": [
                  {
                     "s3Location": {
                        "uri": "string"
                     },
                     "type": "string"
                  }
               ]
            }
         }
      },
      "knowledgeBaseId": "string",
      "name": "string",
      "roleArn": "string",
      "status": "string",
      "storageConfiguration": {
         "mongoDbAtlasConfiguration": {
            "collectionName": "string",
            "credentialsSecretArn": "string",
            "databaseName": "string",
            "endpoint": "string",
            "endpointServiceName": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "textIndexName": "string",
```

```
            "vectorIndexName": "string"
         },
         "neptuneAnalyticsConfiguration": {
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "graphArn": "string"
         },
         "opensearchManagedClusterConfiguration": {
            "domainArn": "string",
            "domainEndpoint": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "opensearchServerlessConfiguration": {
            "collectionArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "pineconeConfiguration": {
            "connectionString": "string",
            "credentialsSecretArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "namespace": "string"
         },
         "rdsConfiguration": {
            "credentialsSecretArn": "string",
            "databaseName": "string",
            "fieldMapping": {
               "customMetadataField": "string",
               "metadataField": "string",
               "primaryKeyField": "string",
```

```
                "textField": "string",
                "vectorField": "string"
             },
             "resourceArn": "string",
             "tableName": "string"
          },
          "redisEnterpriseCloudConfiguration": {
             "credentialsSecretArn": "string",
             "endpoint": "string",
             "fieldMapping": {
                "metadataField": "string",
                "textField": "string",
                "vectorField": "string"
             },
             "vectorIndexName": "string"
          },
          "s3VectorsConfiguration": {
             "indexArn": "string",
             "indexName": "string",
             "vectorBucketArn": "string"
          },
          "type": "string"
       },
       "updatedAt": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

knowledgeBase

Contains details about the knowledge base.

Type: KnowledgeBase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Create a managed knowledge base with a multimodal embedding model configuration

The following example creates a managed knowledge base that uses a multimodal embedding model. The `modelConfiguration` field configures dynamic segmentation for audio and fixed-length segmentation for video.

```
PUT /knowledgebases/ HTTP/1.1
Content-type: application/json

{
    "name": "my-multimodal-knowledge-base",
    "roleArn": "arn:aws:iam::123456789012:role/
AmazonBedrockExecutionRoleForKnowledgeBase",
    "knowledgeBaseConfiguration": {
        "type": "MANAGED",
        "managedKnowledgeBaseConfiguration": {
            "embeddingModelType": "CUSTOM",
            "embeddingModelArn": "arn:aws:bedrock:us-east-1::foundation-model/
twelvelabs.marengo-embed-3-0-v1:0",
            "embeddingModelConfiguration": {
                "bedrockEmbeddingModelConfiguration": {
                    "embeddingDataType": "FLOAT32",
                    "modelConfiguration": {
                        "version": "1",
                        "audio": {
                            "segmentation": {
                                "method": "dynamic",
                                "dynamic": {
                                    "minDurationSec": 4
                                }
                            }
                        },
                        "video": {
                            "segmentation": {
                                "method": "fixed",
                                "fixed": {
                                    "durationSec": 6
                                }
                            }
                        }
                    }
                }
            },
            "supplementalDataStorageConfiguration": {
                "storageLocations": [
                    {
                        "type": "S3",
                        "s3Location": {
                            "uri": "s3://amzn-s3-demo-bucket"
```

```
                        }
                    }
                ]
            }
        }
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreatePrompt

Service: Agents for Amazon Bedrock

Creates a prompt in your prompt library that you can add to a flow. For more information, see Prompt management in Amazon Bedrock, Create a prompt using Prompt management and Prompt flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
POST /prompts/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "customerEncryptionKeyArn": "string",
   "defaultVariant": "string",
   "description": "string",
   "name": "string",
   "tags": {
      "string" : "string"
   },
   "variants": [
      {
         "additionalModelRequestFields": JSON value,
         "genAiResource": { ... },
         "inferenceConfiguration": { ... },
         "metadata": [
            {
               "key": "string",
               "value": "string"
            }
         ],
         "modelId": "string",
         "name": "string",
         "templateConfiguration": { ... },
         "templateType": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to encrypt the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

defaultVariant

The name of the default variant for the prompt. This value must match the `name` field in the relevant PromptVariant object.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

description

A description for the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

A name for the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

tags

Any tags that you want to attach to the prompt. For more information, see Tagging resources in Amazon Bedrock.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

variants

A list of objects, each containing details about a variant of the prompt.

Type: Array of PromptVariant objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "defaultVariant": "string",
   "description": "string",
   "id": "string",
   "name": "string",
   "updatedAt": "string",
   "variants": [
      {
         "additionalModelRequestFields": JSON value,
         "genAiResource": { ... },
         "inferenceConfiguration": { ... },
         "metadata": [
            {
               "key": "string",
               "value": "string"
            }
         ],
         "modelId": "string",
         "name": "string",
         "templateConfiguration": { ... },
         "templateType": "string"
      }
   ],
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the prompt.

Type: String

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?)`

createdAt

The time at which the prompt was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key that you encrypted the prompt with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

defaultVariant

The name of the default variant for your prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

description

The description of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

id

The unique identifier of the prompt.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

updatedAt

The time at which the prompt was last updated.

Type: Timestamp

variants

A list of objects, each containing details about a variant of the prompt.

Type: Array of PromptVariant objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

version

The version of the prompt. When you create a prompt, the version created is the DRAFT version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreatePromptVersion

Service: Agents for Amazon Bedrock

Creates a static snapshot of your prompt that can be deployed to production. For more information, see Deploy prompts using Prompt management by creating versions in the Amazon Bedrock User Guide.

Request Syntax

```
POST /prompts/promptIdentifier/versions HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "description": "string",
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

promptIdentifier

The unique identifier of the prompt that you want to create a version of.

Pattern: `([0-9a-zA-Z]{10})|(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10})(?::[0-9]{1,5})?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

A description for the version of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

tags

Any tags that you want to attach to the version of the prompt. For more information, see Tagging resources in Amazon Bedrock.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json
```

```
{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "defaultVariant": "string",
   "description": "string",
   "id": "string",
   "name": "string",
   "updatedAt": "string",
   "variants": [
      {
         "additionalModelRequestFields": JSON value,
         "genAiResource": { ... },
         "inferenceConfiguration": { ... },
         "metadata": [
            {
               "key": "string",
               "value": "string"
            }
         ],
         "modelId": "string",
         "name": "string",
         "templateConfiguration": { ... },
         "templateType": "string"
      }
   ],
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the version of the prompt.

Type: String

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?)`

createdAt

The time at which the prompt was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to encrypt the version of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

defaultVariant

The name of the default variant for the prompt. This value must match the `name` field in the relevant PromptVariant object.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

description

A description for the version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

id

The unique identifier of the prompt.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

updatedAt

The time at which the prompt was last updated.

Type: Timestamp

variants

A list of objects, each containing details about a variant of the prompt.

Type: Array of PromptVariant objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

version

The version of the prompt that was created. Versions are numbered incrementally, starting from 1.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateVpcConfiguration

Service: Agents for Amazon Bedrock

Creates a VPC configuration that lets a knowledge base connect to a resource in your private VPC. This operation is asynchronous: it returns a `vpcConfigurationId` with status `CREATING`. Poll `GetVpcConfiguration` until the status becomes `CREATED` or `CREATE_FAILED`.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/vpcconfigurations/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "description": "string",
   "hostHeader": "string",
   "name": "string",
   "port": number,
   "protocol": "string",
   "resolutionMode": "string",
   "resourceTarget": "string",
   "subnetIds": [ "string" ],
   "tlsServerName": "string",
   "vpcId": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to associate this VPC configuration with.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the operation completes no more than one time. If this token matches a previous request, the service ignores the request but does not return an error.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

An optional description of the VPC configuration. If you don't specify a description, the VPC configuration has no description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `[^\p{C}]*`

Required: No

hostHeader

An optional HTTP `Host` header value to send when invoking the resource. Set this only if your resource (or an upstream router or ingress) routes by the `Host` header and that host differs from the target. This setting is independent of `tlsServerName`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[A-Za-z0-9._:\[\]-]{1,255}`

Required: No

name

An optional human-readable name for the VPC configuration. If you don't specify a name, the VPC configuration has no name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9 _-]*[a-zA-Z0-9])?`

Required: No

port

The port on which to reach the resource.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 65535.

Required: Yes

protocol

The protocol used to connect to the resource. Specify `HTTP` for plaintext or `HTTPS` for TLS. When you specify `HTTPS`, you must also provide `tlsServerName`.

Type: String

Valid Values: `HTTP | HTTPS`

Required: Yes

resolutionMode

Controls how a domain-name `resourceTarget` is resolved. This applies only when the target is a domain name; it has no effect for IP-address targets, which have no name to resolve. In all cases the resolved address must be reachable from inside your VPC. Valid values:

- `IN_VPC` (default, recommended) – The target domain name is resolved privately, using the DNS resolvers of the VPC, such as private Route 53 hosted zones or on-premises DNS reachable from the VPC. Use this for targets that are private to your VPC, such as internal load balancers, private hosted-zone names, or on-premises hosts.
- `PUBLIC` – The target domain name is resolved against public DNS resolvers. Select this only when the target's domain name must be resolved through public DNS and the resulting address is still reachable from the VPC, an uncommon split-horizon configuration. If you are unsure, use `IN_VPC`.

Type: String

Valid Values: `PUBLIC | IN_VPC`

Required: Yes

resourceTarget

The private IPv4 address or DNS name of the resource you want the knowledge base to reach. The target must be privately reachable from inside your VPC, such as an internal load balancer or a private IP. The following are not supported:

- Internet-facing endpoints
- Loopback addresses
- Link-local addresses
- Wildcard addresses
- Multicast addresses
- IPv6 literals

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: Yes

subnetIds

The subnets, in the VPC identified by `vpcId`, that the knowledge base uses to connect to the resource.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `subnet-[a-zA-Z0-9]+`

Required: Yes

tlsServerName

The expected TLS server name. The service matches this value against the Subject Alternative Names on your resource's TLS certificate during invocation. This field is required when `protocol` is `HTTPS`. Set it to a hostname on your certificate, such

as `app.internal.example.com`. You can use a single leftmost wildcard, such as `*.example.com`. The value must be a hostname without a port.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 253.

Pattern: `(\*\.)?([A-Za-z0-9]([A-Za-z0-9-]{0,61}[A-Za-z0-9])?\.)*[A-Za-z0-9]([A-Za-z0-9-]{0,61}[A-Za-z0-9])?`

Required: No

vpcId

The identifier of the VPC that the knowledge base connects through to reach the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `vpc-[a-zA-Z0-9]+`

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "status": "string",
   "vpcConfigurationId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

status

The current status of the VPC configuration. Immediately after creation this is `CREATING`.

Type: String

Valid Values: `CREATING | CREATED | DELETING | CREATE_FAILED | DELETE_FAILED`

vpcConfigurationId

The unique identifier of the VPC configuration that was created.

Type: String

Length Constraints: Fixed length of 32.

Pattern: `[a-z0-9](?:[a-z0-9-]{30}[a-z0-9])`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAgent

Service: Agents for Amazon Bedrock

Deletes an agent.

Request Syntax

```
DELETE /agents/agentId/?skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent to delete.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agentId": "string",
   "agentStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentId

The unique identifier of the agent that was deleted.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

agentStatus

The status of the agent.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteAgent.

```
DELETE /agents/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAgentActionGroup

Service: Agents for Amazon Bedrock

Deletes an action group in an agent.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion/actiongroups/actionGroupId/?
skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

actionGroupId

The unique identifier of the action group to delete.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentId

The unique identifier of the agent that the action group belongs to.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent that the action group belongs to.

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 204
```

Response Elements

If the action is successful, the service sends back an HTTP 204 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteAgentActionGroup.

```
DELETE /agents/ABCDEFGHIJ/agentversions/1/actiongroups/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAgentAlias

Service: Agents for Amazon Bedrock

Deletes an alias of an agent.

Request Syntax

```
DELETE /agents/agentId/agentaliases/agentAliasId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of the alias to delete.

Length Constraints: Fixed length of 10.

Pattern: `(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

agentId

The unique identifier of the agent that the alias belongs to.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agentAliasId": "string",
```

```
    "agentAliasStatus": "string",
    "agentId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentAliasId

The unique identifier of the alias that was deleted.

Type: String

Length Constraints: Fixed length of 10.

Pattern: `(\bTSTALIASID\b|[0-9a-zA-Z]+)`

agentAliasStatus

The status of the alias.

Type: String

Valid Values: `CREATING | PREPARED | FAILED | UPDATING | DELETING | DISSOCIATED`

agentId

The unique identifier of the agent that the alias belongs to.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteAgentAlias.

```
DELETE /agents/ABCDEFGHIJ/agentaliases/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAgentVersion

Service: Agents for Amazon Bedrock

Deletes a version of an agent.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion/?
skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent that the version belongs to.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent to delete.

Pattern: `[0-9]{1,5}`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

```
Content-type: application/json

{
   "agentId": "string",
   "agentStatus": "string",
   "agentVersion": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentId

The unique identifier of the agent that the version belongs to.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

agentStatus

The status of the agent version.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

agentVersion

The version that was deleted.

Type: String

Pattern: `[0-9]{1,5}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteAgentVersion.

```
DELETE /agents/ABCDEFGHIJ/agentversions/1/?skipResourceInUseCheck=true HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteDataSource

Service: Agents for Amazon Bedrock

Deletes a data source from a knowledge base.

Request Syntax

```
DELETE /knowledgebases/knowledgeBaseId/datasources/dataSourceId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source to delete.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base from which to delete the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "dataSourceId": "string",
   "knowledgeBaseId": "string",
   "status": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

dataSourceId

The unique identifier of the data source that was deleted.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

knowledgeBaseId

The unique identifier of the knowledge base to which the data source that was deleted belonged.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

status

The status of the data source.

Type: String

Valid Values: `AVAILABLE | DELETING | DELETE_UNSUCCESSFUL | CREATING | UPDATING | FAILED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteFlow

Service: Agents for Amazon Bedrock

Deletes a flow.

Request Syntax

```
DELETE /flows/flowIdentifier/?skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "id": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteFlowAlias

Service: Agents for Amazon Bedrock

Deletes an alias of a flow.

Request Syntax

```
DELETE /flows/flowIdentifier/aliases/aliasIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

aliasIdentifier

The unique identifier of the alias to be deleted.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(TSTALIASID|[0-9a-zA-Z]{10})`

Required: Yes

flowIdentifier

The unique identifier of the flow that the alias belongs to.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
```

```
    "flowId": "string",
    "id": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

flowId

The unique identifier of the flow that the alias belongs to.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

id

The unique identifier of the flow.

Type: String

Pattern: `(TSTALIASID|[0-9a-zA-Z]{10})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteFlowVersion

Service: Agents for Amazon Bedrock

Deletes a version of a flow.

Request Syntax

```
DELETE /flows/flowIdentifier/versions/flowVersion/?
skipResourceInUseCheck=skipResourceInUseCheck HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow whose version that you want to delete

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

flowVersion

The version of the flow that you want to delete.

Pattern: `[0-9]{1,5}`

Required: Yes

skipResourceInUseCheck

By default, this value is `false` and deletion is stopped if the resource is in use. If you set it to `true`, the resource will be deleted even if the resource is in use.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "id": "string",
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

version

The version of the flow being deleted.

Type: String

Pattern: `[0-9]{1,5}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteKnowledgeBase

Service: Agents for Amazon Bedrock

Deletes a knowledge base. Before deleting a knowledge base, you should disassociate the knowledge base from any agents that it is associated with by making a DisassociateAgentKnowledgeBase request.

Request Syntax

```
DELETE /knowledgebases/knowledgeBaseId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to delete.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "knowledgeBaseId": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

knowledgeBaseId

The unique identifier of the knowledge base that was deleted.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

status

The status of the knowledge base and whether it has been successfully deleted.

Type: String

Valid Values: `CREATING | ACTIVE | DELETING | UPDATING | FAILED | DELETE_UNSUCCESSFUL | UPDATE_UNSUCCESSFUL`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteKnowledgeBaseDocuments

Service: Agents for Amazon Bedrock

Deletes documents from a data source and syncs the changes to the knowledge base that is connected to it. For more information, see Ingest changes directly into a knowledge base in the Amazon Bedrock User Guide.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/documents/deleteDocuments
 HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "documentIdentifiers": [
      {
         "custom": {
            "id": "string"
         },
         "dataSourceType": "string",
         "s3": {
            "uri": "string"
         }
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source that contains the documents.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that is connected to the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

documentIdentifiers

A list of objects, each of which contains information to identify a document to delete.

Type: Array of DocumentIdentifier objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "documentDetails": [
      {
         "dataSourceId": "string",
         "identifier": {
```

```
            "custom": {
               "id": "string"
            },
            "dataSourceType": "string",
            "s3": {
               "uri": "string"
            }
         },
         "knowledgeBaseId": "string",
         "status": "string",
         "statusReason": "string",
         "updatedAt": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

documentDetails

A list of objects, each of which contains information about the documents that were deleted.

Type: Array of KnowledgeBaseDocumentDetail objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeletePrompt

Service: Agents for Amazon Bedrock

Deletes a prompt or a version of it, depending on whether you include the `promptVersion` field or not. For more information, see Delete prompts from the Prompt management tool and Delete a version of a prompt from the Prompt management tool in the Amazon Bedrock User Guide.

Request Syntax

```
DELETE /prompts/promptIdentifier/?promptVersion=promptVersion HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

promptIdentifier

The unique identifier of the prompt.

Pattern: `([0-9a-zA-Z]{10})|(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10})(?::[0-9]{1,5})?`

Required: Yes

promptVersion

The version of the prompt to delete. To delete the prompt, omit this field.

Pattern: `[0-9]{1,5}`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "id": "string",
   "version": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

id

The unique identifier of the prompt that was deleted.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

version

The version of the prompt that was deleted.

Type: String

Pattern: `[0-9]{1,5}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Delete a prompt

The following request deletes the prompt in the account with the identifier PROMPT12345:

```
DELETE /prompts/PROMPT12345/ HTTP/1.1
```

Delete a version of a prompt

The following request deletes version 1 of the prompt in the account with the identifier PROMPT12345:

```
GET /prompts/PROMPT12345?promptVersion=1 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteResourcePolicy

Service: Agents for Amazon Bedrock

Removes the resource policy associated with a knowledge base. After deletion, other AWS accounts can no longer access the knowledge base using cross-account permissions.

Request Syntax

```
DELETE /resourcepolicy/resourceArn?expectedRevisionId=expectedRevisionId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

expectedRevisionId

The expected revision identifier of the resource policy. Use this to prevent conflicts when multiple users update the same policy concurrently.

Length Constraints: Minimum length of 1. Maximum length of 255.

resourceArn

The Amazon Resource Name (ARN) of the knowledge base to remove the resource policy from.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```

```
{
   "resourceArn": "string",
   "revisionId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

resourceArn

The ARN of the knowledge base that the resource policy was removed from.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

revisionId

The revision identifier after the resource policy was deleted.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteVpcConfiguration

Service: Agents for Amazon Bedrock

Deletes a VPC configuration. This operation is asynchronous: it returns status `DELETING`. Poll `GetVpcConfiguration` until it returns a `ResourceNotFoundException`, indicating the configuration is deleted. Delete requests are idempotent and safe to retry.

Request Syntax

```
DELETE /knowledgebases/knowledgeBaseId/vpcconfigurations/vpcConfigurationId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base that owns the VPC configuration.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

vpcConfigurationId

The unique identifier of the VPC configuration to delete.

Length Constraints: Fixed length of 32.

Pattern: `[a-z0-9](?:[a-z0-9-]{30}[a-z0-9])`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json
```

```
{
   "status": "string",
   "vpcConfigurationId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

status

The current status of the VPC configuration. Immediately after a delete request this is `DELETING`.

Type: String

Valid Values: `CREATING | CREATED | DELETING | CREATE_FAILED | DELETE_FAILED`

vpcConfigurationId

The unique identifier of the VPC configuration being deleted.

Type: String

Length Constraints: Fixed length of 32.

Pattern: `[a-z0-9](?:[a-z0-9-]{30}[a-z0-9])`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DisassociateAgentCollaborator

Service: Agents for Amazon Bedrock

Disassociates an agent collaborator.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion/agentcollaborators/collaboratorId/
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

An agent ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The agent's version.

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

collaboratorId

The collaborator's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 204
```

Response Elements

If the action is successful, the service sends back an HTTP 204 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DisassociateAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Disassociates a knowledge base from an agent.

Request Syntax

```
DELETE /agents/agentId/agentversions/agentVersion/knowledgebases/knowledgeBaseId/
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent from which to disassociate the knowledge base.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent from which to disassociate the knowledge base.

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to disassociate.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 204
```

Response Elements

If the action is successful, the service sends back an HTTP 204 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgent

Service: Agents for Amazon Bedrock

Gets information about an agent.

Request Syntax

```
GET /agents/agentId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agent": {
      "agentArn": "string",
      "agentCollaboration": "string",
      "agentId": "string",
      "agentName": "string",
      "agentResourceRoleArn": "string",
      "agentStatus": "string",
      "agentVersion": "string",
      "clientToken": "string",
      "createdAt": "string",
      "customerEncryptionKeyArn": "string",
```

```
      "customOrchestration": {
         "executor": { ... }
      },
      "description": "string",
      "failureReasons": [ "string" ],
      "foundationModel": "string",
      "guardrailConfiguration": {
         "guardrailIdentifier": "string",
         "guardrailVersion": "string"
      },
      "idleSessionTTLInSeconds": number,
      "instruction": "string",
      "memoryConfiguration": {
         "enabledMemoryTypes": [ "string" ],
         "sessionSummaryConfiguration": {
            "maxRecentSessions": number
         },
         "storageDays": number
      },
      "orchestrationType": "string",
      "preparedAt": "string",
      "promptOverrideConfiguration": {
         "overrideLambda": "string",
         "promptConfigurations": [
            {
               "additionalModelRequestFields": JSON value,
               "basePromptTemplate": "string",
               "foundationModel": "string",
               "inferenceConfiguration": {
                  "maximumLength": number,
                  "stopSequences": [ "string" ],
                  "temperature": number,
                  "topK": number,
                  "topP": number
               },
               "parserMode": "string",
               "promptCreationMode": "string",
               "promptState": "string",
               "promptType": "string"
            }
         ]
      },
      "recommendedActions": [ "string" ],
      "updatedAt": "string"
```

```
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agent

Contains details about the agent.

Type: Agent object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetAgent.

```
GET /agents/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgentActionGroup

Service: Agents for Amazon Bedrock

Gets information about an action group for an agent.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/actiongroups/actionGroupId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

actionGroupId

The unique identifier of the action group for which to get information.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentId

The unique identifier of the agent that the action group belongs to.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent that the action group belongs to.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentActionGroup": {
      "actionGroupExecutor": { ... },
      "actionGroupId": "string",
      "actionGroupName": "string",
      "actionGroupState": "string",
      "agentId": "string",
      "agentVersion": "string",
      "apiSchema": { ... },
      "clientToken": "string",
      "createdAt": "string",
      "description": "string",
      "functionSchema": { ... },
      "parentActionGroupSignatureParams": {
         "string" : "string"
      },
      "parentActionSignature": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentActionGroup

Contains details about the action group.

Type: AgentActionGroup object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetAgentActionGroup.

```
GET /agents/AGENT12345/agentversions/1/actiongroups/ACTION1234/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgentAlias

Service: Agents for Amazon Bedrock

Gets information about an alias of an agent.

Request Syntax

```
GET /agents/agentId/agentaliases/agentAliasId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of the alias for which to get information.

Length Constraints: Fixed length of 10.

Pattern: `(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

agentId

The unique identifier of the agent to which the alias to get information belongs.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
```

```
   "agentAlias": {
      "agentAliasArn": "string",
      "agentAliasHistoryEvents": [
         {
            "endDate": "string",
            "routingConfiguration": [
               {
                  "agentVersion": "string",
                  "provisionedThroughput": "string"
               }
            ],
            "startDate": "string"
         }
      ],
      "agentAliasId": "string",
      "agentAliasName": "string",
      "agentAliasStatus": "string",
      "agentId": "string",
      "aliasInvocationState": "string",
      "clientToken": "string",
      "createdAt": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "routingConfiguration": [
         {
            "agentVersion": "string",
            "provisionedThroughput": "string"
         }
      ],
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentAlias

Contains information about the alias.

Type: AgentAlias object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetAgentAlias.

```
GET /agents/ABCDEFGHIJ/agentaliases/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgentCollaborator

Service: Agents for Amazon Bedrock

Retrieves information about an agent's collaborator.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/agentcollaborators/collaboratorId/
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The agent's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The agent's version.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

collaboratorId

The collaborator's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentCollaborator": {
      "agentDescriptor": {
         "aliasArn": "string"
      },
      "agentId": "string",
      "agentVersion": "string",
      "clientToken": "string",
      "collaborationInstruction": "string",
      "collaboratorId": "string",
      "collaboratorName": "string",
      "createdAt": "string",
      "lastUpdatedAt": "string",
      "relayConversationHistory": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentCollaborator

Details about the collaborator.

Type: AgentCollaborator object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Gets information about a knowledge base associated with an agent.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/knowledgebases/knowledgeBaseId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent with which the knowledge base is associated.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent with which the knowledge base is associated.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base associated with the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentKnowledgeBase": {
      "agentId": "string",
      "agentVersion": "string",
      "createdAt": "string",
      "description": "string",
      "knowledgeBaseId": "string",
      "knowledgeBaseState": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentKnowledgeBase

Contains details about a knowledge base attached to an agent.

Type: AgentKnowledgeBase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgentVersion

Service: Agents for Amazon Bedrock

Gets details about a version of an agent.

Request Syntax

```
GET /agents/agentId/agentversions/agentVersion/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent.

Pattern: `[0-9]{1,5}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentVersion": {
      "agentArn": "string",
      "agentCollaboration": "string",
      "agentId": "string",
```

```
    "agentName": "string",
    "agentResourceRoleArn": "string",
    "agentStatus": "string",
    "createdAt": "string",
    "customerEncryptionKeyArn": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "foundationModel": "string",
    "guardrailConfiguration": {
       "guardrailIdentifier": "string",
       "guardrailVersion": "string"
    },
    "idleSessionTTLInSeconds": number,
    "instruction": "string",
    "memoryConfiguration": {
       "enabledMemoryTypes": [ "string" ],
       "sessionSummaryConfiguration": {
          "maxRecentSessions": number
       },
       "storageDays": number
    },
    "promptOverrideConfiguration": {
       "overrideLambda": "string",
       "promptConfigurations": [
          {
             "additionalModelRequestFields": JSON value,
             "basePromptTemplate": "string",
             "foundationModel": "string",
             "inferenceConfiguration": {
                "maximumLength": number,
                "stopSequences": [ "string" ],
                "temperature": number,
                "topK": number,
                "topP": number
             },
             "parserMode": "string",
             "promptCreationMode": "string",
             "promptState": "string",
             "promptType": "string"
          }
       ]
    },
    "recommendedActions": [ "string" ],
    "updatedAt": "string",
```

```
        "version": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentVersion

Contains details about the version of the agent.

Type: AgentVersion object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetAgentVersion.

```
GET /agents/agentId/agentversions/agentVersion/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetDataSource

Service: Agents for Amazon Bedrock

Gets information about a data source.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId/datasources/dataSourceId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "dataSource": {
      "createdAt": "string",
      "dataDeletionPolicy": "string",
      "dataSourceConfiguration": {
```

```
        "confluenceConfiguration": {
           "crawlerConfiguration": {
              "filterConfiguration": {
                 "patternObjectFilter": {
                    "filters": [
                       {
                          "exclusionFilters": [ "string" ],
                          "inclusionFilters": [ "string" ],
                          "objectType": "string"
                       }
                    ]
                 },
                 "type": "string"
              }
           },
           "sourceConfiguration": {
              "authType": "string",
              "credentialsSecretArn": "string",
              "hostType": "string",
              "hostUrl": "string"
           }
        },
        "managedKnowledgeBaseConnectorConfiguration": {
           "connectorParameters": JSON value,
           "deletionProtectionConfiguration": {
              "deletionProtectionStatus": "string",
              "deletionProtectionThreshold": number
           },
           "mediaExtractionConfiguration": {
              "audioExtractionConfiguration": {
                 "audioExtractionStatus": "string"
              },
              "imageExtractionConfiguration": {
                 "imageExtractionStatus": "string"
              },
              "videoExtractionConfiguration": {
                 "videoExtractionStatus": "string"
              }
           },
           "syncSchedule": { ... }
        },
        "s3Configuration": {
           "bucketArn": "string",
           "bucketOwnerAccountId": "string",
```

```
            "inclusionPrefixes": [ "string" ]
         },
         "salesforceConfiguration": {
            "crawlerConfiguration": {
               "filterConfiguration": {
                  "patternObjectFilter": {
                     "filters": [
                        {
                           "exclusionFilters": [ "string" ],
                           "inclusionFilters": [ "string" ],
                           "objectType": "string"
                        }
                     ]
                  },
                  "type": "string"
               }
            },
            "sourceConfiguration": {
               "authType": "string",
               "credentialsSecretArn": "string",
               "hostUrl": "string"
            }
         },
         "sharePointConfiguration": {
            "crawlerConfiguration": {
               "filterConfiguration": {
                  "patternObjectFilter": {
                     "filters": [
                        {
                           "exclusionFilters": [ "string" ],
                           "inclusionFilters": [ "string" ],
                           "objectType": "string"
                        }
                     ]
                  },
                  "type": "string"
               }
            },
            "sourceConfiguration": {
               "authType": "string",
               "credentialsSecretArn": "string",
               "domain": "string",
               "hostType": "string",
               "siteUrls": [ "string" ],
```

```
            "tenantId": "string"
         }
      },
      "type": "string",
      "webConfiguration": {
         "crawlerConfiguration": {
            "crawlerLimits": {
               "maxPages": number,
               "rateLimit": number
            },
            "exclusionFilters": [ "string" ],
            "inclusionFilters": [ "string" ],
            "scope": "string",
            "userAgent": "string",
            "userAgentHeader": "string"
         },
         "sourceConfiguration": {
            "urlConfiguration": {
               "seedUrls": [
                  {
                     "url": "string"
                  }
               ]
            }
         }
      }
   },
   "dataSourceId": "string",
   "description": "string",
   "failureReasons": [ "string" ],
   "knowledgeBaseId": "string",
   "name": "string",
   "serverSideEncryptionConfiguration": {
      "kmsKeyArn": "string"
   },
   "status": "string",
   "updatedAt": "string",
   "vectorIngestionConfiguration": {
      "chunkingConfiguration": {
         "chunkingStrategy": "string",
         "fixedSizeChunkingConfiguration": {
            "maxTokens": number,
            "overlapPercentage": number
         },
```

```
            "hierarchicalChunkingConfiguration": {
               "levelConfigurations": [
                  {
                     "maxTokens": number
                  }
               ],
               "overlapTokens": number
            },
            "semanticChunkingConfiguration": {
               "breakpointPercentileThreshold": number,
               "bufferSize": number,
               "maxTokens": number
            }
         },
         "contextEnrichmentConfiguration": {
            "bedrockFoundationModelConfiguration": {
               "enrichmentStrategyConfiguration": {
                  "method": "string"
               },
               "modelArn": "string"
            },
            "type": "string"
         },
         "customTransformationConfiguration": {
            "intermediateStorage": {
               "s3Location": {
                  "uri": "string"
               }
            },
            "transformations": [
               {
                  "stepToApply": "string",
                  "transformationFunction": {
                     "transformationLambdaConfiguration": {
                        "lambdaArn": "string"
                     }
                  }
               }
            ]
         },
         "parsingConfiguration": {
            "bedrockDataAutomationConfiguration": {
               "parsingModality": "string"
            },
```

```
            "bedrockFoundationModelConfiguration": {
               "modelArn": "string",
               "parsingModality": "string",
               "parsingPrompt": {
                  "parsingPromptText": "string"
               }
            },
            "parsingStrategy": "string"
         }
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSource

Contains details about the data source.

Type: DataSource object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetFlow

Service: Agents for Amazon Bedrock

Retrieves information about a flow. For more information, see Manage a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
GET /flows/flowIdentifier/?includedData=includedData HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

includedData

Controls the scope of data returned. Set to `METADATA_ONLY` to return only resource metadata. Set to `ALL_DATA` or omit this field to return the full response.

Valid Values: `ALL_DATA | METADATA_ONLY`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
```

```
   "customerEncryptionKeyArn": "string",
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
                  "type": "string"
               }
            ],
            "name": "string",
            "outputs": [
               {
                  "name": "string",
                  "type": "string"
               }
            ],
            "type": "string"
         }
      ]
   },
   "description": "string",
   "executionRoleArn": "string",
   "id": "string",
   "name": "string",
   "status": "string",
   "updatedAt": "string",
   "validations": [
      {
         "details": { ... },
         "message": "string",
         "severity": "string",
```

```
            "type": "string"
        }
    ],
    "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

createdAt

The time at which the flow was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key that the flow is encrypted with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

definition

The definition of the nodes and connections between the nodes in the flow.

Type: FlowDefinition object

description

The description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create a flow. For more information, see Create a service row for flows in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the flow.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

status

The status of the flow. The following statuses are possible:

- NotPrepared – The flow has been created or updated, but hasn't been prepared. If you just created the flow, you can't test it. If you updated the flow, the DRAFT version won't contain the latest changes for testing. Send a PrepareFlow request to package the latest changes into the DRAFT version.
- Preparing – The flow is being prepared so that the DRAFT version contains the latest changes for testing.
- Prepared – The flow is prepared and the DRAFT version contains the latest changes for testing.

- Failed – The last API operation that you invoked on the flow failed. Send a GetFlow request and check the error message in the `validations` field.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

updatedAt

The time at which the flow was last updated.

Type: Timestamp

validations

A list of validation error messages related to the last failed operation on the flow.

Type: Array of FlowValidation objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

version

The version of the flow for which information was retrieved.

Type: String

Length Constraints: Fixed length of 5.

Pattern: DRAFT

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetFlowAlias

Service: Agents for Amazon Bedrock

Retrieves information about a flow. For more information, see Deploy a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
GET /flows/flowIdentifier/aliases/aliasIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

aliasIdentifier

The unique identifier of the alias for which to retrieve information.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(TSTALIASID|[0-9a-zA-Z]{10})`

Required: Yes

flowIdentifier

The unique identifier of the flow that the alias belongs to.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```

```
{
   "arn": "string",
   "concurrencyConfiguration": {
      "maxConcurrency": number,
      "type": "string"
   },
   "createdAt": "string",
   "description": "string",
   "flowId": "string",
   "id": "string",
   "name": "string",
   "routingConfiguration": [
      {
         "flowVersion": "string"
      }
   ],
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/(TSTALIASID|[0-9a-zA-Z]{10})`

concurrencyConfiguration

The configuration that specifies how nodes in the flow are executed in parallel.

Type: FlowAliasConcurrencyConfiguration object

createdAt

The time at which the flow was created.

Type: Timestamp

description

The description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

flowId

The unique identifier of the flow that the alias belongs to.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

id

The unique identifier of the alias of the flow.

Type: String

Pattern: `(TSTALIASID|[0-9a-zA-Z]{10})`

name

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

routingConfiguration

Contains information about the version that the alias is mapped to.

Type: Array of FlowAliasRoutingConfigurationListItem objects

Array Members: Fixed number of 1 item.

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetFlowVersion

Service: Agents for Amazon Bedrock

Retrieves information about a version of a flow. For more information, see Deploy a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
GET /flows/flowIdentifier/versions/flowVersion/?includedData=includedData HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow for which to get information.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

flowVersion

The version of the flow for which to get information.

Pattern: `[0-9]{1,5}`

Required: Yes

includedData

Controls the scope of data returned. Set to `METADATA_ONLY` to return only resource metadata. Set to `ALL_DATA` or omit this field to return the full response.

Valid Values: `ALL_DATA | METADATA_ONLY`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
                  "type": "string"
               }
            ],
            "name": "string",
            "outputs": [
               {
                  "name": "string",
                  "type": "string"
               }
            ],
            "type": "string"
         }
      ]
   },
   "description": "string",
   "executionRoleArn": "string",
```

```
    "id": "string",
    "name": "string",
    "status": "string",
    "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

createdAt

The time at which the flow was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key that the version of the flow is encrypted with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

definition

The definition of the nodes and connections between nodes in the flow.

Type: FlowDefinition object

description

The description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create a flow. For more information, see Create a service role for flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the version.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

status

The status of the flow.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

version

The version of the flow for which information was retrieved.

Type: String

Pattern: `[0-9]{1,5}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetIngestionJob

Service: Agents for Amazon Bedrock

Gets information about a data ingestion job. Data sources are ingested into your knowledge base so that Large Language Models (LLMs) can use your data.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId/datasources/dataSourceId/
ingestionjobs/ingestionJobId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source for the data ingestion job you want to get information on.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

ingestionJobId

The unique identifier of the data ingestion job you want to get information on.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the data ingestion job you want to get information on.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "ingestionJob": {
      "dataSourceId": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "ingestionJobId": "string",
      "knowledgeBaseId": "string",
      "startedAt": "string",
      "statistics": {
         "numberOfDocumentsDeleted": number,
         "numberOfDocumentsFailed": number,
         "numberOfDocumentsScanned": number,
         "numberOfDocumentsSkipped": number,
         "numberOfMetadataDocumentsModified": number,
         "numberOfMetadataDocumentsScanned": number,
         "numberOfModifiedDocumentsIndexed": number,
         "numberOfNewDocumentsIndexed": number
      },
      "status": "string",
      "textReadyAt": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ingestionJob

Contains details about the data ingestion job.

Type: IngestionJob object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetKnowledgeBase

Service: Agents for Amazon Bedrock

Gets information about a knowledge base.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base you want to get information on.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "knowledgeBase": {
      "createdAt": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "knowledgeBaseArn": "string",
      "knowledgeBaseConfiguration": {
         "kendraKnowledgeBaseConfiguration": {
            "kendraIndexArn": "string"
         },
         "managedKnowledgeBaseConfiguration": {
            "embeddingModelArn": "string",
```

```
            "embeddingModelConfiguration": {
               "bedrockEmbeddingModelConfiguration": {
                  "audio": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ],
                  "dimensions": number,
                  "embeddingDataType": "string",
                  "modelConfiguration": JSON value,
                  "video": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ]
               }
            },
            "embeddingModelType": "string",
            "serverSideEncryptionConfiguration": {
               "kmsKeyArn": "string"
            },
            "supplementalDataStorageConfiguration": {
               "storageLocations": [
                  {
                     "s3Location": {
                        "uri": "string"
                     },
                     "type": "string"
                  }
               ]
            }
         },
         "sqlKnowledgeBaseConfiguration": {
            "redshiftConfiguration": {
               "queryEngineConfiguration": {
                  "provisionedConfiguration": {
                     "authConfiguration": {
                        "databaseUser": "string",
                        "type": "string",
                        "usernamePasswordSecretArn": "string"
```

```
            },
            "clusterIdentifier": "string"
         },
         "serverlessConfiguration": {
            "authConfiguration": {
               "type": "string",
               "usernamePasswordSecretArn": "string"
            },
            "workgroupArn": "string"
         },
         "type": "string"
      },
      "queryGenerationConfiguration": {
         "executionTimeoutSeconds": number,
         "generationContext": {
            "curatedQueries": [
               {
                  "naturalLanguage": "string",
                  "sql": "string"
               }
            ],
            "tables": [
               {
                  "columns": [
                     {
                        "description": "string",
                        "inclusion": "string",
                        "name": "string"
                     }
                  ],
                  "description": "string",
                  "inclusion": "string",
                  "name": "string"
               }
            ]
         }
      },
      "storageConfigurations": [
         {
            "awsDataCatalogConfiguration": {
               "tableNames": [ "string" ]
            },
            "redshiftConfiguration": {
               "databaseName": "string"
```

```
                    },
                    "type": "string"
                 }
              ]
           },
           "type": "string"
        },
        "type": "string",
        "vectorKnowledgeBaseConfiguration": {
           "embeddingModelArn": "string",
           "embeddingModelConfiguration": {
              "bedrockEmbeddingModelConfiguration": {
                 "audio": [
                    {
                       "segmentationConfiguration": {
                          "fixedLengthDuration": number
                       }
                    }
                 ],
                 "dimensions": number,
                 "embeddingDataType": "string",
                 "modelConfiguration": JSON value,
                 "video": [
                    {
                       "segmentationConfiguration": {
                          "fixedLengthDuration": number
                       }
                    }
                 ]
              }
           },
           "supplementalDataStorageConfiguration": {
              "storageLocations": [
                 {
                    "s3Location": {
                       "uri": "string"
                    },
                    "type": "string"
                 }
              ]
           }
        }
     },
     "knowledgeBaseId": "string",
```

```
      "name": "string",
      "roleArn": "string",
      "status": "string",
      "storageConfiguration": {
         "mongoDbAtlasConfiguration": {
            "collectionName": "string",
            "credentialsSecretArn": "string",
            "databaseName": "string",
            "endpoint": "string",
            "endpointServiceName": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "textIndexName": "string",
            "vectorIndexName": "string"
         },
         "neptuneAnalyticsConfiguration": {
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "graphArn": "string"
         },
         "opensearchManagedClusterConfiguration": {
            "domainArn": "string",
            "domainEndpoint": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "opensearchServerlessConfiguration": {
            "collectionArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
```

```
            "pineconeConfiguration": {
               "connectionString": "string",
               "credentialsSecretArn": "string",
               "fieldMapping": {
                  "metadataField": "string",
                  "textField": "string"
               },
               "namespace": "string"
            },
            "rdsConfiguration": {
               "credentialsSecretArn": "string",
               "databaseName": "string",
               "fieldMapping": {
                  "customMetadataField": "string",
                  "metadataField": "string",
                  "primaryKeyField": "string",
                  "textField": "string",
                  "vectorField": "string"
               },
               "resourceArn": "string",
               "tableName": "string"
            },
            "redisEnterpriseCloudConfiguration": {
               "credentialsSecretArn": "string",
               "endpoint": "string",
               "fieldMapping": {
                  "metadataField": "string",
                  "textField": "string",
                  "vectorField": "string"
               },
               "vectorIndexName": "string"
            },
            "s3VectorsConfiguration": {
               "indexArn": "string",
               "indexName": "string",
               "vectorBucketArn": "string"
            },
            "type": "string"
         },
         "updatedAt": "string"
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

knowledgeBase

Contains details about the knowledge base.

Type: KnowledgeBase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetKnowledgeBaseDocuments

Service: Agents for Amazon Bedrock

Retrieves specific documents from a data source that is connected to a knowledge base. For more information, see Ingest changes directly into a knowledge base in the Amazon Bedrock User Guide.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/documents/getDocuments
 HTTP/1.1
Content-type: application/json

{
   "documentIdentifiers": [
      {
         "custom": {
            "id": "string"
         },
         "dataSourceType": "string",
         "s3": {
            "uri": "string"
         }
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source that contains the documents.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that is connected to the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

documentIdentifiers

A list of objects, each of which contains information to identify a document for which to retrieve information.

Type: Array of DocumentIdentifier objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "documentDetails": [
      {
         "dataSourceId": "string",
         "identifier": {
            "custom": {
               "id": "string"
            },
            "dataSourceType": "string",
            "s3": {
               "uri": "string"
            }
         },
         "knowledgeBaseId": "string",
         "status": "string",
         "statusReason": "string",
         "updatedAt": "string"
      }
   ]
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

documentDetails

A list of objects, each of which contains information about the documents that were retrieved.

Type: Array of KnowledgeBaseDocumentDetail objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetPrompt

Service: Agents for Amazon Bedrock

Retrieves information about the working draft (DRAFT version) of a prompt or a version of it, depending on whether you include the `promptVersion` field or not. For more information, see View information about prompts using Prompt management and View information about a version of your prompt in the Amazon Bedrock User Guide.

Request Syntax

```
GET /prompts/promptIdentifier/?includedData=includedData&promptVersion=promptVersion
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

includedData

Controls the scope of data returned. Set to `METADATA_ONLY` to return only resource metadata. Set to `ALL_DATA` or omit this field to return the full response.

Valid Values: `ALL_DATA | METADATA_ONLY`

promptIdentifier

The unique identifier of the prompt.

Pattern: `([0-9a-zA-Z]{10})|(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10})(?::[0-9]{1,5})?`

Required: Yes

promptVersion

The version of the prompt about which you want to retrieve information. Omit this field to return information about the working draft of the prompt.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "defaultVariant": "string",
   "description": "string",
   "id": "string",
   "name": "string",
   "updatedAt": "string",
   "variants": [
      {
         "additionalModelRequestFields": JSON value,
         "genAiResource": { ... },
         "inferenceConfiguration": { ... },
         "metadata": [
            {
               "key": "string",
               "value": "string"
            }
         ],
         "modelId": "string",
         "name": "string",
         "templateConfiguration": { ... },
         "templateType": "string"
      }
   ],
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the prompt or the prompt version (if you specified a version in the request).

Type: String

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?)`

createdAt

The time at which the prompt was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key that the prompt is encrypted with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

defaultVariant

The name of the default variant for the prompt. This value must match the `name` field in the relevant PromptVariant object.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

description

The descriptino of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

id

The unique identifier of the prompt.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

updatedAt

The time at which the prompt was last updated.

Type: Timestamp

variants

A list of objects, each containing details about a variant of the prompt.

Type: Array of PromptVariant objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

version

The version of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Get information about the working draft of a prompt

The following request returns information about the working draft of the prompt in the account with the identifier PROMPT12345:

```
GET /prompts/PROMPT12345 HTTP/1.1
```

List information about all the versions of a prompt

The following request returns information about version 2 of the prompt in the account with the identifier PROMPT12345:

```
GET /prompts/PROMPT12345?promptVersion=2 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetResourcePolicy

Service: Agents for Amazon Bedrock

Retrieves the resource policy associated with a knowledge base.

Request Syntax

```
GET /resourcepolicy/resourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the knowledge base to retrieve the resource policy for.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "policy": "string",
   "resourceArn": "string",
   "revisionId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

policy

The JSON-formatted resource policy associated with the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20480.

Pattern: `[\u0009\u000A\u000D\u0020-\u00FF]+`

resourceArn

The ARN of the knowledge base that the resource policy is associated with.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

revisionId

The revision identifier of the resource policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetVpcConfiguration

Service: Agents for Amazon Bedrock

Returns the details and current status of a single VPC configuration. Use this operation to poll for the outcome of an asynchronous create or delete.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId/vpcconfigurations/vpcConfigurationId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base that owns the VPC configuration.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

vpcConfigurationId

The unique identifier of the VPC configuration to retrieve.

Length Constraints: Fixed length of 32.

Pattern: `[a-z0-9](?:[a-z0-9-]{30}[a-z0-9])`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```

```
{
   "vpcConfiguration": {
      "createdAt": "string",
      "description": "string",
      "hostHeader": "string",
      "name": "string",
      "port": number,
      "protocol": "string",
      "resolutionMode": "string",
      "resourceTarget": "string",
      "status": "string",
      "statusMessage": "string",
      "subnetIds": [ "string" ],
      "tlsServerName": "string",
      "updatedAt": "string",
      "vpcConfigurationId": "string",
      "vpcId": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

vpcConfiguration

The VPC configuration, including its connection settings, resolution mode, and current lifecycle status.

Type: VpcConfiguration object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

IngestKnowledgeBaseDocuments

Service: Agents for Amazon Bedrock

Ingests documents directly into the knowledge base that is connected to the data source. The `dataSourceType` specified in the content for each document must match the type of the data source that you specify in the header. For more information, see Ingest changes directly into a knowledge base in the Amazon Bedrock User Guide.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/dataSourceId/documents HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "documents": [
      {
         "content": {
            "custom": {
               "customDocumentIdentifier": {
                  "id": "string"
               },
               "inlineContent": {
                  "byteContent": {
                     "data": blob,
                     "mimeType": "string"
                  },
                  "textContent": {
                     "data": "string"
                  },
                  "type": "string"
               },
               "s3Location": {
                  "bucketOwnerAccountId": "string",
                  "uri": "string"
               },
               "sourceType": "string"
            },
            "dataSourceType": "string",
            "s3": {
               "s3Location": {
                  "uri": "string"
```

```
                }
            }
        },
        "metadata": {
            "accessControlList": [
                {
                    "access": "string",
                    "name": "string",
                    "type": "string"
                }
            ],
            "inlineAttributes": [
                {
                    "key": "string",
                    "value": {
                        "booleanValue": boolean,
                        "numberValue": number,
                        "stringListValue": [ "string" ],
                        "stringValue": "string",
                        "type": "string"
                    }
                }
            ],
            "s3Location": {
                "bucketOwnerAccountId": "string",
                "uri": "string"
            },
            "type": "string"
        }
    }
]
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source connected to the knowledge base that you're adding documents to.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to ingest the documents into.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

documents

A list of objects, each of which contains information about the documents to add.

Type: Array of KnowledgeBaseDocument objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

Response Syntax

```
HTTP/1.1 202
Content-type: application/json
```

```
{
   "documentDetails": [
      {
         "dataSourceId": "string",
         "identifier": {
            "custom": {
               "id": "string"
            },
            "dataSourceType": "string",
            "s3": {
               "uri": "string"
            }
         },
         "knowledgeBaseId": "string",
         "status": "string",
         "statusReason": "string",
         "updatedAt": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

documentDetails

A list of objects, each of which contains information about the documents that were ingested.

Type: Array of KnowledgeBaseDocumentDetail objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAgentActionGroups

Service: Agents for Amazon Bedrock

Lists the action groups for an agent and information about each one.

Request Syntax

```
POST /agents/agentId/agentversions/agentVersion/actiongroups/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "actionGroupSummaries": [
      {
         "actionGroupId": "string",
         "actionGroupName": "string",
         "actionGroupState": "string",
         "description": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

actionGroupSummaries

A list of objects, each of which contains information about an action group.

Type: Array of ActionGroupSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgentActionGroups.

```
POST /agents/AGENT12345/agentversions/1/actiongroups/ HTTP/1.1
Content-type: application/json

{
   "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAgentAliases

Service: Agents for Amazon Bedrock

Lists the aliases of an agent and information about each one.

Request Syntax

```
POST /agents/agentId/agentaliases/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentAliasSummaries": [
      {
         "agentAliasId": "string",
         "agentAliasName": "string",
         "agentAliasStatus": "string",
         "aliasInvocationState": "string",
         "createdAt": "string",
         "description": "string",
         "routingConfiguration": [
            {
               "agentVersion": "string",
               "provisionedThroughput": "string"
            }
         ],
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentAliasSummaries

A list of objects, each of which contains information about an alias of the agent.

Type: Array of AgentAliasSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgentAliases.

```
POST /agents/ABCDEFGHIJ/agentaliases/ HTTP/1.1
Content-type: application/json

{
   "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAgentCollaborators

Service: Agents for Amazon Bedrock

Retrieve a list of an agent's collaborators.

Request Syntax

```
POST /agents/agentId/agentversions/agentVersion/agentcollaborators/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The agent's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The agent's version.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of agent collaborators to return in one page of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentCollaboratorSummaries": [
      {
         "agentDescriptor": {
            "aliasArn": "string"
         },
         "agentId": "string",
         "agentVersion": "string",
         "collaborationInstruction": "string",
         "collaboratorId": "string",
         "collaboratorName": "string",
         "createdAt": "string",
         "lastUpdatedAt": "string",
         "relayConversationHistory": "string"
      }
   ],
```

```
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentCollaboratorSummaries

A list of collaborator summaries.

Type: Array of AgentCollaboratorSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAgentKnowledgeBases

Service: Agents for Amazon Bedrock

Lists knowledge bases associated with an agent and information about each one.

Request Syntax

```
POST /agents/agentId/agentversions/agentVersion/knowledgebases/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent for which to return information about knowledge bases associated with it.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent for which to return information about knowledge bases associated with it.

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentKnowledgeBaseSummaries": [
      {
         "description": "string",
         "knowledgeBaseId": "string",
         "knowledgeBaseState": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentKnowledgeBaseSummaries

A list of objects, each of which contains information about a knowledge base associated with the agent.

Type: Array of AgentKnowledgeBaseSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAgents

Service: Agents for Amazon Bedrock

Lists the agents belonging to an account and information about each agent.

Request Syntax

```
POST /agents/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentSummaries": [
      {
         "agentId": "string",
         "agentName": "string",
         "agentStatus": "string",
         "description": "string",
         "guardrailConfiguration": {
            "guardrailIdentifier": "string",
            "guardrailVersion": "string"
         },
         "latestAgentVersion": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentSummaries

A list of objects, each of which contains information about an agent.

Type: Array of AgentSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgents.

```
POST /agents/ HTTP/1.1
Content-type: application/json

{
   "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAgentVersions

Service: Agents for Amazon Bedrock

Lists the versions of an agent and information about each version.

Request Syntax

```
POST /agents/agentId/agentversions/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentVersionSummaries": [
      {
         "agentName": "string",
         "agentStatus": "string",
         "agentVersion": "string",
         "createdAt": "string",
         "description": "string",
         "guardrailConfiguration": {
            "guardrailIdentifier": "string",
            "guardrailVersion": "string"
         },
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentVersionSummaries

A list of objects, each of which contains information about a version of the agent.

Type: Array of AgentVersionSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListAgentVersions.

```
POST /agents/agentId/agentversions/ HTTP/1.1
Content-type: application/json

{
   "maxResults": 10
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListDataSources

Service: Agents for Amazon Bedrock

Lists the data sources in a knowledge base and information about each one.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base for which to return a list of information.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "dataSourceSummaries": [
      {
         "dataSourceId": "string",
         "description": "string",
         "knowledgeBaseId": "string",
         "name": "string",
         "status": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSourceSummaries

A list of objects, each of which contains information about a data source.

Type: Array of DataSourceSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFlowAliases

Service: Agents for Amazon Bedrock

Returns a list of aliases for a flow.

Request Syntax

```
GET /flows/flowIdentifier/aliases?maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow for which aliases are being returned.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "flowAliasSummaries": [
      {
         "arn": "string",
         "concurrencyConfiguration": {
            "maxConcurrency": number,
            "type": "string"
         },
         "createdAt": "string",
         "description": "string",
         "flowId": "string",
         "id": "string",
         "name": "string",
         "routingConfiguration": [
            {
               "flowVersion": "string"
            }
         ],
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

flowAliasSummaries

A list, each member of which contains information about an alias.

Type: Array of FlowAliasSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFlows

Service: Agents for Amazon Bedrock

Returns a list of flows and information about each flow. For more information, see Manage a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
GET /flows/?maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "flowSummaries": [
      {
         "arn": "string",
         "createdAt": "string",
         "description": "string",
         "id": "string",
         "name": "string",
         "status": "string",
         "updatedAt": "string",
         "version": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

flowSummaries

A list, each member of which contains information about a flow.

Type: Array of FlowSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFlowVersions

Service: Agents for Amazon Bedrock

Returns a list of information about each flow. For more information, see Deploy a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
GET /flows/flowIdentifier/versions?maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "flowVersionSummaries": [
      {
         "arn": "string",
         "createdAt": "string",
         "id": "string",
         "status": "string",
         "version": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

flowVersionSummaries

A list, each member of which contains information about a flow.

Type: Array of FlowVersionSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListIngestionJobs

Service: Agents for Amazon Bedrock

Lists the data ingestion jobs for a data source. The list also includes information about each job.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/ingestionjobs/ HTTP/1.1
Content-type: application/json

{
   "filters": [
      {
         "attribute": "string",
         "operator": "string",
         "values": [ "string" ]
      }
   ],
   "maxResults": number,
   "nextToken": "string",
   "sortBy": {
      "attribute": "string",
      "order": "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source for the list of data ingestion jobs.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the list of data ingestion jobs.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

filters

Contains information about the filters for filtering the data.

Type: Array of IngestionJobFilter objects

Array Members: Fixed number of 1 item.

Required: No

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

sortBy

Contains details about how to sort the data.

Type: IngestionJobSortBy object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "ingestionJobSummaries": [
      {
         "dataSourceId": "string",
         "description": "string",
         "ingestionJobId": "string",
         "knowledgeBaseId": "string",
         "startedAt": "string",
         "statistics": {
            "numberOfDocumentsDeleted": number,
            "numberOfDocumentsFailed": number,
            "numberOfDocumentsScanned": number,
            "numberOfDocumentsSkipped": number,
            "numberOfMetadataDocumentsModified": number,
            "numberOfMetadataDocumentsScanned": number,
            "numberOfModifiedDocumentsIndexed": number,
            "numberOfNewDocumentsIndexed": number
         },
         "status": "string",
         "textReadyAt": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ingestionJobSummaries

A list of data ingestion jobs with information about each job.

Type: Array of IngestionJobSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListKnowledgeBaseDocuments

Service: Agents for Amazon Bedrock

Retrieves all the documents contained in a data source that is connected to a knowledge base. For more information, see Ingest changes directly into a knowledge base in the Amazon Bedrock User Guide.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/documents HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source that contains the documents.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that is connected to the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "documentDetails": [
      {
         "dataSourceId": "string",
         "identifier": {
            "custom": {
               "id": "string"
            },
            "dataSourceType": "string",
            "s3": {
               "uri": "string"
            }
```

```
            },
            "knowledgeBaseId": "string",
            "status": "string",
            "statusReason": "string",
            "updatedAt": "string"
        }
    ],
    "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

documentDetails

A list of objects, each of which contains information about the documents that were retrieved.

Type: Array of KnowledgeBaseDocumentDetail objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListKnowledgeBases

Service: Agents for Amazon Bedrock

Lists the knowledge bases in an account. The list also includesinformation about each knowledge base.

Request Syntax

```
POST /knowledgebases/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "knowledgeBaseSummaries": [
      {
         "description": "string",
         "knowledgeBaseId": "string",
         "name": "string",
         "status": "string",
         "updatedAt": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

knowledgeBaseSummaries

A list of knowledge bases with information about each knowledge base.

Type: Array of KnowledgeBaseSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListPrompts

Service: Agents for Amazon Bedrock

Returns either information about the working draft (DRAFT version) of each prompt in an account, or information about of all versions of a prompt, depending on whether you include the `promptIdentifier` field or not. For more information, see View information about prompts using Prompt management in the Amazon Bedrock User Guide.

Request Syntax

```
GET /prompts/?
maxResults=maxResults&nextToken=nextToken&promptIdentifier=promptIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

promptIdentifier

The unique identifier of the prompt for whose versions you want to return information. Omit this field to list information about all prompts in an account.

Pattern: `([0-9a-zA-Z]{10})|(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10})(?::[0-9]{1,5})?`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "nextToken": "string",
   "promptSummaries": [
      {
         "arn": "string",
         "createdAt": "string",
         "description": "string",
         "id": "string",
         "name": "string",
         "updatedAt": "string",
         "version": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

promptSummaries

A list, each member of which contains information about a prompt using Prompt management.

Type: Array of PromptSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

List information about all prompts in an account

The following request returns information about the working draft of each prompt in the account:

```
GET /prompts/ HTTP/1.1
```

List information about all the versions of a prompt

The following request returns information about all versions of the prompt in the account with the identifier PROMPT12345:

```
GET /prompts?promptIdentifier=PROMPT12345 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListTagsForResource

Service: Agents for Amazon Bedrock

List all the tags for the resource you specify.

Request Syntax

```
GET /tags/resourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource for which to list tags.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:(agent|agent-alias|knowledge-base|flow|prompt)/[A-Z0-9]{10}(?:/[A-Z0-9]{10})?$|^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:flow/([A-Z0-9]{10})/alias/([A-Z0-9]{10})$|^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:prompt/([A-Z0-9]{10})?(?::/d+)?$).*`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "tags": {
      "string" : "string"
   }
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

The key-value pairs for the tags associated with the resource.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListTagsForResource.

```
GET /tags/arn:aws:bedrock:us-west-2:123456789012:agent/ABCDEFGHIJ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)

- AWS SDK for Ruby V3

ListVpcConfigurations

Service: Agents for Amazon Bedrock

Returns a paginated list of the VPC configurations for a knowledge base. You can optionally filter by status. Use the `nextToken` parameter to retrieve additional results.

Request Syntax

```
GET /knowledgebases/knowledgeBaseId/vpcconfigurations/?
maxResults=maxResults&nextToken=nextToken&status=statusFilter HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base whose VPC configurations you want to list.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

maxResults

The maximum number of results to return in the response. If more results are available, the response returns a `nextToken`.

Valid Range: Minimum value of 1. Maximum value of 100.

nextToken

A pagination token to retrieve the next page of results, returned in a previous response when more results are available.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

statusFilter

The status to filter the results by. Only VPC configurations with the specified status are returned.

Valid Values: `CREATING | CREATED | DELETING | CREATE_FAILED | DELETE_FAILED`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "items": [
      {
         "createdAt": "string",
         "description": "string",
         "hostHeader": "string",
         "name": "string",
         "port": number,
         "protocol": "string",
         "resolutionMode": "string",
         "resourceTarget": "string",
         "status": "string",
         "statusMessage": "string",
         "tlsServerName": "string",
         "vpcConfigurationId": "string",
         "vpcId": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

items

A list of VPC configuration summaries.

Type: Array of VpcConfigurationSummary objects

nextToken

A pagination token to retrieve the next page of results, present when the total number of results exceeds the maximum number of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PrepareAgent

Service: Agents for Amazon Bedrock

Creates a DRAFT version of the agent that can be used for internal testing.

Request Syntax

```
POST /agents/agentId/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent for which to create a DRAFT version.

Pattern: [0-9a-zA-Z]{10}

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agentId": "string",
   "agentStatus": "string",
   "agentVersion": "string",
   "preparedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentId

The unique identifier of the agent for which the DRAFT version was created.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

agentStatus

The status of the DRAFT version and whether it is ready for use.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

agentVersion

The version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

preparedAt

The time at which the DRAFT version of the agent was last prepared.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of PrepareAgent.

```
POST /agents/ABCDEFGHIJ/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PrepareFlow

Service: Agents for Amazon Bedrock

Prepares the DRAFT version of a flow so that it can be invoked. For more information, see Test a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
POST /flows/flowIdentifier/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "id": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

status

The status of the flow. When you submit this request, the status will be `NotPrepared`. If preparation succeeds, the status becomes `Prepared`. If it fails, the status becomes `FAILED`.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PutResourcePolicy

Service: Agents for Amazon Bedrock

Associates a resource policy with a knowledge base. A resource policy allows other AWS accounts to access the knowledge base. For more information, see Cross-account access for knowledge bases.

Request Syntax

```
PUT /resourcepolicy/resourceArn HTTP/1.1
Content-type: application/json

{
   "expectedRevisionId": "string",
   "policy": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the knowledge base to attach the resource policy to.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

Required: Yes

Request Body

The request accepts the following data in JSON format.

expectedRevisionId

The expected revision identifier of the resource policy. Use this to prevent conflicts when multiple users update the same policy concurrently. Specify the `revisionId` from the most recent `GetResourcePolicy` or `PutResourcePolicy` response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: No

policy

The JSON-formatted resource policy to associate with the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20480.

Pattern: `[\u0009\u000A\u000D\u0020-\u00FF]+`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "resourceArn": "string",
   "revisionId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

resourceArn

The ARN of the knowledge base that the resource policy was attached to.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

revisionId

The revision identifier of the resource policy. Use this value in the `expectedRevisionId` field of a subsequent `PutResourcePolicy` or `DeleteResourcePolicy` request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StartIngestionJob

Service: Agents for Amazon Bedrock

Begins a data ingestion job. Data sources are ingested into your knowledge base so that Large Language Models (LLMs) can use your data.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/dataSourceId/ingestionjobs/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "description": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source you want to ingest into your knowledge base.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the data ingestion job.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

A description of the data ingestion job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "ingestionJob": {
      "dataSourceId": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "ingestionJobId": "string",
      "knowledgeBaseId": "string",
      "startedAt": "string",
      "statistics": {
         "numberOfDocumentsDeleted": number,
         "numberOfDocumentsFailed": number,
         "numberOfDocumentsScanned": number,
         "numberOfDocumentsSkipped": number,
         "numberOfMetadataDocumentsModified": number,
```

```
            "numberOfMetadataDocumentsScanned": number,
            "numberOfModifiedDocumentsIndexed": number,
            "numberOfNewDocumentsIndexed": number
        },
        "status": "string",
        "textReadyAt": "string",
        "updatedAt": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

ingestionJob

Contains information about the data ingestion job.

Type: IngestionJob object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StopIngestionJob

Service: Agents for Amazon Bedrock

Stops a currently running data ingestion job. You can send a `StartIngestionJob` request again to ingest the rest of your data when you are ready.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/
ingestionjobs/ingestionJobId/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source for the data ingestion job you want to stop.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

ingestionJobId

The unique identifier of the data ingestion job you want to stop.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the data ingestion job you want to stop.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "ingestionJob": {
      "dataSourceId": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "ingestionJobId": "string",
      "knowledgeBaseId": "string",
      "startedAt": "string",
      "statistics": {
         "numberOfDocumentsDeleted": number,
         "numberOfDocumentsFailed": number,
         "numberOfDocumentsScanned": number,
         "numberOfDocumentsSkipped": number,
         "numberOfMetadataDocumentsModified": number,
         "numberOfMetadataDocumentsScanned": number,
         "numberOfModifiedDocumentsIndexed": number,
         "numberOfNewDocumentsIndexed": number
      },
      "status": "string",
      "textReadyAt": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

ingestionJob

Contains information about the stopped data ingestion job.

Type: IngestionJob object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

TagResource

Service: Agents for Amazon Bedrock

Associate tags with a resource. For more information, see Tagging resources in the Amazon Bedrock User Guide.

Request Syntax

```
POST /tags/resourceArn HTTP/1.1
Content-type: application/json

{
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource to tag.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:(agent|agent-alias|knowledge-base|flow|prompt)/[A-Z0-9]{10}(?:/[A-Z0-9]{10})?$|^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:flow/([A-Z0-9]{10})/alias/([A-Z0-9]{10})$|^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:prompt/([A-Z0-9]{10})?(?::/d+)?$).*`

Required: Yes

Request Body

The request accepts the following data in JSON format.

tags

An object containing key-value pairs that define the tags to attach to the resource.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of TagResource.

```
POST /tags/arn:aws:bedrock:us-west-2:123456789012:agent/ABCDEFGHIJ HTTP/1.1
Content-type: application/json

{
   "tags": {
      "cost-center" : "Tech"
   }
}
```

Example response

This example illustrates one usage of TagResource.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UntagResource

Service: Agents for Amazon Bedrock

Remove tags from a resource.

Request Syntax

```
DELETE /tags/resourceArn?tagKeys=tagKeys HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource from which to remove tags.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:(agent|agent-alias|knowledge-base|flow|prompt)/[A-Z0-9]{10}(?:/[A-Z0-9]{10})?$|^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:flow/([A-Z0-9]{10})/alias/([A-Z0-9]{10})$|^arn:aws:bedrock:[a-zA-Z0-9-]+:/d{12}:prompt/([A-Z0-9]{10})?(?::/d+)?$).*`

Required: Yes

tagKeys

A list of keys of the tags to remove from the resource.

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UntagResource.

```
DELETE /tags/arn:aws:bedrock:us-west-2:123456789012:agent/ABCDEFGHIJ HTTP/1.1
```

Example response

This example illustrates one usage of UntagResource.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAgent

Service: Agents for Amazon Bedrock

Updates the configuration of an agent.

Request Syntax

```
PUT /agents/agentId/ HTTP/1.1
Content-type: application/json

{
   "agentCollaboration": "string",
   "agentName": "string",
   "agentResourceRoleArn": "string",
   "customerEncryptionKeyArn": "string",
   "customOrchestration": {
      "executor": { ... }
   },
   "description": "string",
   "foundationModel": "string",
   "guardrailConfiguration": {
      "guardrailIdentifier": "string",
      "guardrailVersion": "string"
   },
   "idleSessionTTLInSeconds": number,
   "instruction": "string",
   "memoryConfiguration": {
      "enabledMemoryTypes": [ "string" ],
      "sessionSummaryConfiguration": {
         "maxRecentSessions": number
      },
      "storageDays": number
   },
   "orchestrationType": "string",
   "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
         {
            "additionalModelRequestFields": JSON value,
            "basePromptTemplate": "string",
            "foundationModel": "string",
            "inferenceConfiguration": {
               "maximumLength": number,
```

```
                "stopSequences": [ "string" ],
                "temperature": number,
                "topK": number,
                "topP": number
            },
            "parserMode": "string",
            "promptCreationMode": "string",
            "promptState": "string",
            "promptType": "string"
        }
    ]
  }
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentCollaboration

The agent's collaboration role.

Type: String

Valid Values: `SUPERVISOR | SUPERVISOR_ROUTER | DISABLED`

Required: No

agentName

Specifies a new name for the agent.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key with which to encrypt the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

customOrchestration

Contains details of the custom orchestration configured for the agent.

Type: CustomOrchestration object

Required: No

description

Specifies a new description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

foundationModel

The identifier for the model that you want to be used for orchestration by the agent you create.

The `modelId` to provide depends on the type of model or throughput that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see Amazon Bedrock base model IDs (on-demand throughput) in the Amazon Bedrock User Guide.
- If you use an inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for cross-region inference in the Amazon Bedrock User Guide.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see Run inference using a Provisioned Throughput in the Amazon Bedrock User Guide.
- If you use a custom model, first purchase Provisioned Throughput for it. Then specify the ARN of the resulting provisioned model. For more information, see Use a custom model in Amazon Bedrock in the Amazon Bedrock User Guide.
- If you use an imported model, specify the ARN of the imported model. You can get the model ARN from a successful call to CreateModelImportJob or from the Imported models page in the Amazon Bedrock console.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

guardrailConfiguration

The unique Guardrail configuration assigned to the agent when it is updated.

Type: GuardrailConfiguration object

Required: No

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 5400.

Required: No

instruction

Specifies new instructions that tell the agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 4000.

Required: No

memoryConfiguration

Specifies the new memory configuration for the agent.

Type: MemoryConfiguration object

Required: No

orchestrationType

Specifies the type of orchestration strategy for the agent. This is set to `DEFAULT` orchestration type, by default.

Type: String

Valid Values: `DEFAULT | CUSTOM_ORCHESTRATION`

Required: No

promptOverrideConfiguration

Contains configurations to override prompts in different parts of an agent sequence. For more information, see Advanced prompts.

Type: PromptOverrideConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agent": {
      "agentArn": "string",
      "agentCollaboration": "string",
      "agentId": "string",
      "agentName": "string",
      "agentResourceRoleArn": "string",
      "agentStatus": "string",
      "agentVersion": "string",
      "clientToken": "string",
      "createdAt": "string",
      "customerEncryptionKeyArn": "string",
      "customOrchestration": {
         "executor": { ... }
      },
      "description": "string",
      "failureReasons": [ "string" ],
      "foundationModel": "string",
      "guardrailConfiguration": {
         "guardrailIdentifier": "string",
         "guardrailVersion": "string"
      },
      "idleSessionTTLInSeconds": number,
      "instruction": "string",
      "memoryConfiguration": {
         "enabledMemoryTypes": [ "string" ],
         "sessionSummaryConfiguration": {
```

```
            "maxRecentSessions": number
         },
         "storageDays": number
      },
      "orchestrationType": "string",
      "preparedAt": "string",
      "promptOverrideConfiguration": {
         "overrideLambda": "string",
         "promptConfigurations": [
            {
               "additionalModelRequestFields": JSON value,
               "basePromptTemplate": "string",
               "foundationModel": "string",
               "inferenceConfiguration": {
                  "maximumLength": number,
                  "stopSequences": [ "string" ],
                  "temperature": number,
                  "topK": number,
                  "topP": number
               },
               "parserMode": "string",
               "promptCreationMode": "string",
               "promptState": "string",
               "promptType": "string"
            }
         ]
      },
      "recommendedActions": [ "string" ],
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agent

Contains details about the agent that was updated.

Type: Agent object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateAgent.

```
PUT /agents/ABCDEFGHIJ/ HTTP/1.1
Content-type: application/json

{
  "agentName": "TestName",
  "agentResourceRoleArn": "arn:aws:iam::123456789012:role/
AmazonBedrockExecutionRoleForAgents_user",
  "instruction": "You are an IT agent who solves customer's problems",
  "description": "Description is here",
  "idleSessionTTLInSeconds": 900,
  "foundationModel": "anthropic.claude-v2"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAgentActionGroup

Service: Agents for Amazon Bedrock

Updates the configuration for an action group for an agent.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/actiongroups/actionGroupId/ HTTP/1.1
Content-type: application/json

{
   "actionGroupExecutor": { ... },
   "actionGroupName": "string",
   "actionGroupState": "string",
   "apiSchema": { ... },
   "description": "string",
   "functionSchema": { ... },
   "parentActionGroupSignature": "string",
   "parentActionGroupSignatureParams": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

actionGroupId

The unique identifier of the action group.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentId

The unique identifier of the agent for which to update the action group.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The unique identifier of the agent version for which to update the action group.

Length Constraints: Fixed length of 5.

Pattern: DRAFT

Required: Yes

Request Body

The request accepts the following data in JSON format.

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: ActionGroupExecutor object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

actionGroupName

Specifies a new name for the action group.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see Action group OpenAPI schemas.

Type: APISchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

description

Specifies a new name for the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Contains details about the function schema for the action group or the JSON or YAML-formatted payload defining the schema.

Type: FunctionSchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionGroupSignature

Update the built-in or computer use action for this action group. If you specify a value, you must leave the `description`, `apiSchema`, and `actionGroupExecutor` fields empty for this action group.

- To allow your agent to request the user for additional information when trying to complete a task, set this field to `AMAZON.UserInput`.
- To allow your agent to generate, run, and troubleshoot code when trying to complete a task, set this field to `AMAZON.CodeInterpreter`.
- To allow your agent to use an Anthropic computer use tool, specify one of the following values.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Anthropic Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. When operating computer use functionality, we recommend taking additional security precautions, such as executing computer actions in virtual environments with restricted data access and limited internet connectivity. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

- `ANTHROPIC.Computer` - Gives the agent permission to use the mouse and keyboard and take screenshots.
- `ANTHROPIC.TextEditor` - Gives the agent permission to view, create and edit files.
- `ANTHROPIC.Bash` - Gives the agent permission to run commands in a bash shell.

During orchestration, if your agent determines that it needs to invoke an API in an action group, but doesn't have enough information to complete the API request, it will invoke this action group instead and return an Observation reprompting the user for more information.

Type: String

Valid Values: `AMAZON.UserInput | AMAZON.CodeInterpreter | ANTHROPIC.Computer | ANTHROPIC.Bash | ANTHROPIC.TextEditor`

Required: No

parentActionGroupSignatureParams

The configuration settings for a computer use action.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Type: String to string map

Key Length Constraints: Minimum length of 0. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 100.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentActionGroup": {
      "actionGroupExecutor": { ... },
      "actionGroupId": "string",
      "actionGroupName": "string",
      "actionGroupState": "string",
      "agentId": "string",
      "agentVersion": "string",
      "apiSchema": { ... },
      "clientToken": "string",
      "createdAt": "string",
      "description": "string",
      "functionSchema": { ... },
      "parentActionGroupSignatureParams": {
         "string" : "string"
      },
      "parentActionSignature": "string",
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentActionGroup

Contains details about the action group that was updated.

Type: AgentActionGroup object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateAgentActionGroup.

```
PUT /agents/AGENT12345/agentversions/1/actiongroups/ACTION1234/ HTTP/1.1
Content-type: application/json

{
    "actionGroupName": "bedrock-temp-actions",
    "actionGroupState": "ENABLED",
    "description": "Testing = latest IT Management action",
    "apiSchema": {
        "s3": {
            "s3BucketName": "apischema-s3",
            "s3ObjectKey": "it_agent_openapi.json"
        }
    },
    "actionGroupExecutor": {
        "lambda": "arn:aws:lambda:us-west-2:123456789012:function:ItAgentLambda"
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAgentAlias

Service: Agents for Amazon Bedrock

Updates configurations for an alias of an agent.

Request Syntax

```
PUT /agents/agentId/agentaliases/agentAliasId/ HTTP/1.1
Content-type: application/json

{
   "agentAliasName": "string",
   "aliasInvocationState": "string",
   "description": "string",
   "routingConfiguration": [
      {
         "agentVersion": "string",
         "provisionedThroughput": "string"
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of the alias.

Length Constraints: Fixed length of 10.

Pattern: `(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

agentId

The unique identifier of the agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentAliasName

Specifies a new name for the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

aliasInvocationState

The invocation state for the agent alias. To pause the agent alias, set the value to `REJECT_INVOCATIONS`. To start the agent alias running again, set the value to `ACCEPT_INVOCATIONS`. Use the `GetAgentAlias`, or `ListAgentAliases`, operation to get the invocation state of an agent alias.

Type: String

Valid Values: `ACCEPT_INVOCATIONS | REJECT_INVOCATIONS`

Required: No

description

Specifies a new description for the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

routingConfiguration

Contains details about the routing configuration of the alias.

Type: Array of AgentAliasRoutingConfigurationListItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "agentAlias": {
      "agentAliasArn": "string",
      "agentAliasHistoryEvents": [
         {
            "endDate": "string",
            "routingConfiguration": [
               {
                  "agentVersion": "string",
                  "provisionedThroughput": "string"
               }
            ],
            "startDate": "string"
         }
      ],
      "agentAliasId": "string",
      "agentAliasName": "string",
      "agentAliasStatus": "string",
      "agentId": "string",
      "aliasInvocationState": "string",
      "clientToken": "string",
      "createdAt": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "routingConfiguration": [
         {
            "agentVersion": "string",
            "provisionedThroughput": "string"
         }
      ],
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

agentAlias

Contains details about the alias that was updated.

Type: AgentAlias object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateAgentAlias.

```
PUT /agents/ABCDEFGHIJ/agentaliases/ABCDEFGHIJ/ HTTP/1.1
Content-type: application/json

{
    "agentAliasName": "TestName",
    "description": "Updated description",
    "routingConfiguration": [
       {
          "agentVersion": "2"
       }
    ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAgentCollaborator

Service: Agents for Amazon Bedrock

Updates an agent's collaborator.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/agentcollaborators/collaboratorId/
 HTTP/1.1
Content-type: application/json

{
   "agentDescriptor": {
      "aliasArn": "string"
   },
   "collaborationInstruction": "string",
   "collaboratorName": "string",
   "relayConversationHistory": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The agent's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The agent's version.

Length Constraints: Fixed length of 5.

Pattern: DRAFT

Required: Yes

collaboratorId

The collaborator's ID.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

agentDescriptor

An agent descriptor for the agent collaborator.

Type: AgentDescriptor object

Required: Yes

collaborationInstruction

Instruction for the collaborator.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: Yes

collaboratorName

The collaborator's name.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

relayConversationHistory

A relay conversation history for the collaborator.

Type: String

Valid Values: `TO_COLLABORATOR | DISABLED`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentCollaborator": {
      "agentDescriptor": {
         "aliasArn": "string"
      },
      "agentId": "string",
      "agentVersion": "string",
      "clientToken": "string",
      "collaborationInstruction": "string",
      "collaboratorId": "string",
      "collaboratorName": "string",
      "createdAt": "string",
      "lastUpdatedAt": "string",
      "relayConversationHistory": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentCollaborator

Details about the collaborator.

Type: AgentCollaborator object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateAgentKnowledgeBase

Service: Agents for Amazon Bedrock

Updates the configuration for a knowledge base that has been associated with an agent.

Request Syntax

```
PUT /agents/agentId/agentversions/agentVersion/knowledgebases/knowledgeBaseId/ HTTP/1.1
Content-type: application/json

{
   "description": "string",
   "knowledgeBaseState": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

agentId

The unique identifier of the agent associated with the knowledge base that you want to update.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent associated with the knowledge base that you want to update.

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that has been associated with an agent.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

Specifies a new description for the knowledge base associated with an agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

knowledgeBaseState

Specifies whether the agent uses the knowledge base or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "agentKnowledgeBase": {
      "agentId": "string",
      "agentVersion": "string",
      "createdAt": "string",
      "description": "string",
      "knowledgeBaseId": "string",
      "knowledgeBaseState": "string",
      "updatedAt": "string"
   }
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

agentKnowledgeBase

Contains details about the knowledge base that has been associated with an agent.

Type: AgentKnowledgeBase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateDataSource

Service: Agents for Amazon Bedrock

Updates the configurations for a data source connector.

> **Important**
>
> You can't change the `chunkingConfiguration` after you create the data source connector. Specify the existing `chunkingConfiguration`.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId/datasources/dataSourceId HTTP/1.1
Content-type: application/json

{
   "dataDeletionPolicy": "string",
   "dataSourceConfiguration": {
      "confluenceConfiguration": {
         "crawlerConfiguration": {
            "filterConfiguration": {
               "patternObjectFilter": {
                  "filters": [
                     {
                        "exclusionFilters": [ "string" ],
                        "inclusionFilters": [ "string" ],
                        "objectType": "string"
                     }
                  ]
               },
               "type": "string"
            }
         },
         "sourceConfiguration": {
            "authType": "string",
            "credentialsSecretArn": "string",
            "hostType": "string",
            "hostUrl": "string"
         }
      },
      "managedKnowledgeBaseConnectorConfiguration": {
```

```
            "connectorParameters": JSON value,
            "deletionProtectionConfiguration": {
               "deletionProtectionStatus": "string",
               "deletionProtectionThreshold": number
            },
            "mediaExtractionConfiguration": {
               "audioExtractionConfiguration": {
                  "audioExtractionStatus": "string"
               },
               "imageExtractionConfiguration": {
                  "imageExtractionStatus": "string"
               },
               "videoExtractionConfiguration": {
                  "videoExtractionStatus": "string"
               }
            },
            "syncSchedule": { ... }
         },
         "s3Configuration": {
            "bucketArn": "string",
            "bucketOwnerAccountId": "string",
            "inclusionPrefixes": [ "string" ]
         },
         "salesforceConfiguration": {
            "crawlerConfiguration": {
               "filterConfiguration": {
                  "patternObjectFilter": {
                     "filters": [
                        {
                           "exclusionFilters": [ "string" ],
                           "inclusionFilters": [ "string" ],
                           "objectType": "string"
                        }
                     ]
                  },
                  "type": "string"
               }
            },
            "sourceConfiguration": {
               "authType": "string",
               "credentialsSecretArn": "string",
               "hostUrl": "string"
            }
         },
```

```
      "sharePointConfiguration": {
         "crawlerConfiguration": {
            "filterConfiguration": {
               "patternObjectFilter": {
                  "filters": [
                     {
                        "exclusionFilters": [ "string" ],
                        "inclusionFilters": [ "string" ],
                        "objectType": "string"
                     }
                  ]
               },
               "type": "string"
            }
         },
         "sourceConfiguration": {
            "authType": "string",
            "credentialsSecretArn": "string",
            "domain": "string",
            "hostType": "string",
            "siteUrls": [ "string" ],
            "tenantId": "string"
         }
      },
      "type": "string",
      "webConfiguration": {
         "crawlerConfiguration": {
            "crawlerLimits": {
               "maxPages": number,
               "rateLimit": number
            },
            "exclusionFilters": [ "string" ],
            "inclusionFilters": [ "string" ],
            "scope": "string",
            "userAgent": "string",
            "userAgentHeader": "string"
         },
         "sourceConfiguration": {
            "urlConfiguration": {
               "seedUrls": [
                  {
                     "url": "string"
                  }
               ]
```

```
            }
         }
      }
   },
   "description": "string",
   "name": "string",
   "serverSideEncryptionConfiguration": {
      "kmsKeyArn": "string"
   },
   "vectorIngestionConfiguration": {
      "chunkingConfiguration": {
         "chunkingStrategy": "string",
         "fixedSizeChunkingConfiguration": {
            "maxTokens": number,
            "overlapPercentage": number
         },
         "hierarchicalChunkingConfiguration": {
            "levelConfigurations": [
               {
                  "maxTokens": number
               }
            ],
            "overlapTokens": number
         },
         "semanticChunkingConfiguration": {
            "breakpointPercentileThreshold": number,
            "bufferSize": number,
            "maxTokens": number
         }
      },
      "contextEnrichmentConfiguration": {
         "bedrockFoundationModelConfiguration": {
            "enrichmentStrategyConfiguration": {
               "method": "string"
            },
            "modelArn": "string"
         },
         "type": "string"
      },
      "customTransformationConfiguration": {
         "intermediateStorage": {
            "s3Location": {
               "uri": "string"
            }
```

```
         },
         "transformations": [
            {
               "stepToApply": "string",
               "transformationFunction": {
                  "transformationLambdaConfiguration": {
                     "lambdaArn": "string"
                  }
               }
            }
         ]
      },
      "parsingConfiguration": {
         "bedrockDataAutomationConfiguration": {
            "parsingModality": "string"
         },
         "bedrockFoundationModelConfiguration": {
            "modelArn": "string",
            "parsingModality": "string",
            "parsingPrompt": {
               "parsingPromptText": "string"
            }
         },
         "parsingStrategy": "string"
      }
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the data source.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

dataDeletionPolicy

The data deletion policy for the data source that you want to update.

Type: String

Valid Values: `RETAIN | DELETE`

Required: No

dataSourceConfiguration

The connection configuration for the data source that you want to update.

Type: DataSourceConfiguration object

Required: Yes

description

Specifies a new description for the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

Specifies a new name for the data source.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

serverSideEncryptionConfiguration

Contains details about server-side encryption of the data source.

Type: ServerSideEncryptionConfiguration object

Required: No

vectorIngestionConfiguration

Contains details about how to ingest the documents in the data source.

Type: VectorIngestionConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "dataSource": {
      "createdAt": "string",
      "dataDeletionPolicy": "string",
      "dataSourceConfiguration": {
         "confluenceConfiguration": {
            "crawlerConfiguration": {
               "filterConfiguration": {
                  "patternObjectFilter": {
                     "filters": [
                        {
                           "exclusionFilters": [ "string" ],
                           "inclusionFilters": [ "string" ],
                           "objectType": "string"
                        }
                     ]
                  },
                  "type": "string"
               }
            },
            "sourceConfiguration": {
               "authType": "string",
               "credentialsSecretArn": "string",
               "hostType": "string",
               "hostUrl": "string"
            }
         },
```

```
        "managedKnowledgeBaseConnectorConfiguration": {
           "connectorParameters": JSON value,
           "deletionProtectionConfiguration": {
              "deletionProtectionStatus": "string",
              "deletionProtectionThreshold": number
           },
           "mediaExtractionConfiguration": {
              "audioExtractionConfiguration": {
                 "audioExtractionStatus": "string"
              },
              "imageExtractionConfiguration": {
                 "imageExtractionStatus": "string"
              },
              "videoExtractionConfiguration": {
                 "videoExtractionStatus": "string"
              }
           },
           "syncSchedule": { ... }
        },
        "s3Configuration": {
           "bucketArn": "string",
           "bucketOwnerAccountId": "string",
           "inclusionPrefixes": [ "string" ]
        },
        "salesforceConfiguration": {
           "crawlerConfiguration": {
              "filterConfiguration": {
                 "patternObjectFilter": {
                    "filters": [
                       {
                          "exclusionFilters": [ "string" ],
                          "inclusionFilters": [ "string" ],
                          "objectType": "string"
                       }
                    ]
                 },
                 "type": "string"
              }
           },
           "sourceConfiguration": {
              "authType": "string",
              "credentialsSecretArn": "string",
              "hostUrl": "string"
           }
```

```
        },
        "sharePointConfiguration": {
           "crawlerConfiguration": {
              "filterConfiguration": {
                 "patternObjectFilter": {
                    "filters": [
                       {
                          "exclusionFilters": [ "string" ],
                          "inclusionFilters": [ "string" ],
                          "objectType": "string"
                       }
                    ]
                 },
                 "type": "string"
              }
           },
           "sourceConfiguration": {
              "authType": "string",
              "credentialsSecretArn": "string",
              "domain": "string",
              "hostType": "string",
              "siteUrls": [ "string" ],
              "tenantId": "string"
           }
        },
        "type": "string",
        "webConfiguration": {
           "crawlerConfiguration": {
              "crawlerLimits": {
                 "maxPages": number,
                 "rateLimit": number
              },
              "exclusionFilters": [ "string" ],
              "inclusionFilters": [ "string" ],
              "scope": "string",
              "userAgent": "string",
              "userAgentHeader": "string"
           },
           "sourceConfiguration": {
              "urlConfiguration": {
                 "seedUrls": [
                    {
                       "url": "string"
                    }
```

```
                ]
             }
          }
       }
    },
    "dataSourceId": "string",
    "description": "string",
    "failureReasons": [ "string" ],
    "knowledgeBaseId": "string",
    "name": "string",
    "serverSideEncryptionConfiguration": {
       "kmsKeyArn": "string"
    },
    "status": "string",
    "updatedAt": "string",
    "vectorIngestionConfiguration": {
       "chunkingConfiguration": {
          "chunkingStrategy": "string",
          "fixedSizeChunkingConfiguration": {
             "maxTokens": number,
             "overlapPercentage": number
          },
          "hierarchicalChunkingConfiguration": {
             "levelConfigurations": [
                {
                   "maxTokens": number
                }
             ],
             "overlapTokens": number
          },
          "semanticChunkingConfiguration": {
             "breakpointPercentileThreshold": number,
             "bufferSize": number,
             "maxTokens": number
          }
       },
       "contextEnrichmentConfiguration": {
          "bedrockFoundationModelConfiguration": {
             "enrichmentStrategyConfiguration": {
                "method": "string"
             },
             "modelArn": "string"
          },
          "type": "string"
```

```
        },
        "customTransformationConfiguration": {
           "intermediateStorage": {
              "s3Location": {
                 "uri": "string"
              }
           },
           "transformations": [
              {
                 "stepToApply": "string",
                 "transformationFunction": {
                    "transformationLambdaConfiguration": {
                       "lambdaArn": "string"
                    }
                 }
              }
           ]
        },
        "parsingConfiguration": {
           "bedrockDataAutomationConfiguration": {
              "parsingModality": "string"
           },
           "bedrockFoundationModelConfiguration": {
              "modelArn": "string",
              "parsingModality": "string",
              "parsingPrompt": {
                 "parsingPromptText": "string"
              }
           },
           "parsingStrategy": "string"
        }
     }
  }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

dataSource

Contains details about the data source.

Type: DataSource object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateFlow

Service: Agents for Amazon Bedrock

Modifies a flow. Include both fields that you want to keep and fields that you want to change. For more information, see How it works and Create a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
PUT /flows/flowIdentifier/ HTTP/1.1
Content-type: application/json

{
   "customerEncryptionKeyArn": "string",
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
                  "type": "string"
               }
            ],
            "name": "string",
            "outputs": [
               {
                  "name": "string",
                  "type": "string"
               }
            ],
            "type": "string"
```

```
            }
         ]
      },
      "description": "string",
      "executionRoleArn": "string",
      "name": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

flowIdentifier

The unique identifier of the flow.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to encrypt the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

definition

A definition of the nodes and the connections between the nodes in the flow.

Type: FlowDefinition object

Required: No

description

A description for the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create and manage a flow. For more information, see Create a service role for flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

Required: Yes

name

A name for the flow.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "definition": {
```

```
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
                  "type": "string"
               }
            ],
            "name": "string",
            "outputs": [
               {
                  "name": "string",
                  "type": "string"
               }
            ],
            "type": "string"
         }
      ]
   },
   "description": "string",
   "executionRoleArn": "string",
   "id": "string",
   "name": "string",
   "status": "string",
   "updatedAt": "string",
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

createdAt

The time at which the flow was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key that the flow was encrypted with.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

definition

A definition of the nodes and the connections between nodes in the flow.

Type: FlowDefinition object

description

The description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

executionRoleArn

The Amazon Resource Name (ARN) of the service role with permissions to create a flow. For more information, see Create a service role for flows in Amazon Bedrock in the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the flow.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

status

The status of the flow. When you submit this request, the status will be `NotPrepared`. If updating fails, the status becomes `Failed`.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

updatedAt

The time at which the flow was last updated.

Type: Timestamp

version

The version of the flow. When you update a flow, the version updated is the `DRAFT` version.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateFlowAlias

Service: Agents for Amazon Bedrock

Modifies the alias of a flow. Include both fields that you want to keep and ones that you want to change. For more information, see Deploy a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

Request Syntax

```
PUT /flows/flowIdentifier/aliases/aliasIdentifier HTTP/1.1
Content-type: application/json

{
   "concurrencyConfiguration": {
      "maxConcurrency": number,
      "type": "string"
   },
   "description": "string",
   "name": "string",
   "routingConfiguration": [
      {
         "flowVersion": "string"
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

aliasIdentifier

The unique identifier of the alias.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(TSTALIASID|[0-9a-zA-Z]{10})`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

concurrencyConfiguration

The configuration that specifies how nodes in the flow are executed in parallel.

Type: FlowAliasConcurrencyConfiguration object

Required: No

description

A description for the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

routingConfiguration

Contains information about the version to which to map the alias.

Type: Array of FlowAliasRoutingConfigurationListItem objects

Array Members: Fixed number of 1 item.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "arn": "string",
   "concurrencyConfiguration": {
      "maxConcurrency": number,
      "type": "string"
   },
   "createdAt": "string",
   "description": "string",
   "flowId": "string",
   "id": "string",
   "name": "string",
   "routingConfiguration": [
      {
         "flowVersion": "string"
      }
   ],
   "updatedAt": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/(TSTALIASID|[0-9a-zA-Z]{10})`

concurrencyConfiguration

The configuration that specifies how nodes in the flow are executed in parallel.

Type: FlowAliasConcurrencyConfiguration object

createdAt

The time at which the flow was created.

Type: Timestamp

description

The description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

flowId

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

id

The unique identifier of the alias.

Type: String

Pattern: `(TSTALIASID|[0-9a-zA-Z]{10})`

name

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

routingConfiguration

Contains information about the version that the alias is mapped to.

Type: Array of FlowAliasRoutingConfigurationListItem objects

Array Members: Fixed number of 1 item.

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateKnowledgeBase

Service: Agents for Amazon Bedrock

Updates the configuration of a knowledge base with the fields that you specify. Because all fields will be overwritten, you must include the same values for fields that you want to keep the same.

You can change the following fields:

- `name`
- `description`
- `roleArn`

You can't change the `knowledgeBaseConfiguration` or `storageConfiguration` fields, so you must specify the same configurations as when you created the knowledge base. You can send a GetKnowledgeBase request and copy the same configurations.

Request Syntax

```
PUT /knowledgebases/knowledgeBaseId HTTP/1.1
Content-type: application/json

{
   "description": "string",
   "knowledgeBaseConfiguration": {
      "kendraKnowledgeBaseConfiguration": {
         "kendraIndexArn": "string"
      },
      "managedKnowledgeBaseConfiguration": {
         "embeddingModelArn": "string",
         "embeddingModelConfiguration": {
            "bedrockEmbeddingModelConfiguration": {
               "audio": [
                  {
                     "segmentationConfiguration": {
                        "fixedLengthDuration": number
                     }
                  }
               ],
               "dimensions": number,
               "embeddingDataType": "string",
               "modelConfiguration": JSON value,
```

```
            "video": [
               {
                  "segmentationConfiguration": {
                     "fixedLengthDuration": number
                  }
               }
            ]
         }
      },
      "embeddingModelType": "string",
      "serverSideEncryptionConfiguration": {
         "kmsKeyArn": "string"
      },
      "supplementalDataStorageConfiguration": {
         "storageLocations": [
            {
               "s3Location": {
                  "uri": "string"
               },
               "type": "string"
            }
         ]
      }
   },
   "sqlKnowledgeBaseConfiguration": {
      "redshiftConfiguration": {
         "queryEngineConfiguration": {
            "provisionedConfiguration": {
               "authConfiguration": {
                  "databaseUser": "string",
                  "type": "string",
                  "usernamePasswordSecretArn": "string"
               },
               "clusterIdentifier": "string"
            },
            "serverlessConfiguration": {
               "authConfiguration": {
                  "type": "string",
                  "usernamePasswordSecretArn": "string"
               },
               "workgroupArn": "string"
            },
            "type": "string"
         },
```

```
            "queryGenerationConfiguration": {
               "executionTimeoutSeconds": number,
               "generationContext": {
                  "curatedQueries": [
                     {
                        "naturalLanguage": "string",
                        "sql": "string"
                     }
                  ],
                  "tables": [
                     {
                        "columns": [
                           {
                              "description": "string",
                              "inclusion": "string",
                              "name": "string"
                           }
                        ],
                        "description": "string",
                        "inclusion": "string",
                        "name": "string"
                     }
                  ]
               }
            },
            "storageConfigurations": [
               {
                  "awsDataCatalogConfiguration": {
                     "tableNames": [ "string" ]
                  },
                  "redshiftConfiguration": {
                     "databaseName": "string"
                  },
                  "type": "string"
               }
            ]
         },
         "type": "string"
      },
      "type": "string",
      "vectorKnowledgeBaseConfiguration": {
         "embeddingModelArn": "string",
         "embeddingModelConfiguration": {
            "bedrockEmbeddingModelConfiguration": {
```

```
            "audio": [
               {
                  "segmentationConfiguration": {
                     "fixedLengthDuration": number
                  }
               }
            ],
            "dimensions": number,
            "embeddingDataType": "string",
            "modelConfiguration": JSON value,
            "video": [
               {
                  "segmentationConfiguration": {
                     "fixedLengthDuration": number
                  }
               }
            ]
         }
      },
      "supplementalDataStorageConfiguration": {
         "storageLocations": [
            {
               "s3Location": {
                  "uri": "string"
               },
               "type": "string"
            }
         ]
      }
   },
   "name": "string",
   "roleArn": "string",
   "storageConfiguration": {
      "mongoDbAtlasConfiguration": {
         "collectionName": "string",
         "credentialsSecretArn": "string",
         "databaseName": "string",
         "endpoint": "string",
         "endpointServiceName": "string",
         "fieldMapping": {
            "metadataField": "string",
            "textField": "string",
            "vectorField": "string"
```

```
            },
            "textIndexName": "string",
            "vectorIndexName": "string"
         },
         "neptuneAnalyticsConfiguration": {
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "graphArn": "string"
         },
         "opensearchManagedClusterConfiguration": {
            "domainArn": "string",
            "domainEndpoint": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "opensearchServerlessConfiguration": {
            "collectionArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "pineconeConfiguration": {
            "connectionString": "string",
            "credentialsSecretArn": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "namespace": "string"
         },
         "rdsConfiguration": {
            "credentialsSecretArn": "string",
            "databaseName": "string",
            "fieldMapping": {
               "customMetadataField": "string",
```

```
            "metadataField": "string",
            "primaryKeyField": "string",
            "textField": "string",
            "vectorField": "string"
         },
         "resourceArn": "string",
         "tableName": "string"
      },
      "redisEnterpriseCloudConfiguration": {
         "credentialsSecretArn": "string",
         "endpoint": "string",
         "fieldMapping": {
            "metadataField": "string",
            "textField": "string",
            "vectorField": "string"
         },
         "vectorIndexName": "string"
      },
      "s3VectorsConfiguration": {
         "indexArn": "string",
         "indexName": "string",
         "vectorBucketArn": "string"
      },
      "type": "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to update.

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

Specifies a new description for the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

knowledgeBaseConfiguration

Specifies the configuration for the embeddings model used for the knowledge base. You must use the same configuration as when the knowledge base was created.

Type: KnowledgeBaseConfiguration object

Required: Yes

name

Specifies a new name for the knowledge base.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

roleArn

Specifies a different Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

storageConfiguration

Specifies the configuration for the vector store used for the knowledge base. You must use the same configuration as when the knowledge base was created.

Type: StorageConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 202
Content-type: application/json

{
   "knowledgeBase": {
      "createdAt": "string",
      "description": "string",
      "failureReasons": [ "string" ],
      "knowledgeBaseArn": "string",
      "knowledgeBaseConfiguration": {
         "kendraKnowledgeBaseConfiguration": {
            "kendraIndexArn": "string"
         },
         "managedKnowledgeBaseConfiguration": {
            "embeddingModelArn": "string",
            "embeddingModelConfiguration": {
               "bedrockEmbeddingModelConfiguration": {
                  "audio": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ],
                  "dimensions": number,
                  "embeddingDataType": "string",
                  "modelConfiguration": JSON value,
                  "video": [
                     {
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ]
               }
            },
            "embeddingModelType": "string",
            "serverSideEncryptionConfiguration": {
               "kmsKeyArn": "string"
```

```
            },
            "supplementalDataStorageConfiguration": {
               "storageLocations": [
                  {
                     "s3Location": {
                        "uri": "string"
                     },
                     "type": "string"
                  }
               ]
            }
         },
         "sqlKnowledgeBaseConfiguration": {
            "redshiftConfiguration": {
               "queryEngineConfiguration": {
                  "provisionedConfiguration": {
                     "authConfiguration": {
                        "databaseUser": "string",
                        "type": "string",
                        "usernamePasswordSecretArn": "string"
                     },
                     "clusterIdentifier": "string"
                  },
                  "serverlessConfiguration": {
                     "authConfiguration": {
                        "type": "string",
                        "usernamePasswordSecretArn": "string"
                     },
                     "workgroupArn": "string"
                  },
                  "type": "string"
               },
               "queryGenerationConfiguration": {
                  "executionTimeoutSeconds": number,
                  "generationContext": {
                     "curatedQueries": [
                        {
                           "naturalLanguage": "string",
                           "sql": "string"
                        }
                     ],
                     "tables": [
                        {
                           "columns": [
```

```
                        {
                           "description": "string",
                           "inclusion": "string",
                           "name": "string"
                        }
                     ],
                     "description": "string",
                     "inclusion": "string",
                     "name": "string"
                  }
               ]
            }
         },
         "storageConfigurations": [
            {
               "awsDataCatalogConfiguration": {
                  "tableNames": [ "string" ]
               },
               "redshiftConfiguration": {
                  "databaseName": "string"
               },
               "type": "string"
            }
         ]
      },
      "type": "string"
   },
   "type": "string",
   "vectorKnowledgeBaseConfiguration": {
      "embeddingModelArn": "string",
      "embeddingModelConfiguration": {
         "bedrockEmbeddingModelConfiguration": {
            "audio": [
               {
                  "segmentationConfiguration": {
                     "fixedLengthDuration": number
                  }
               }
            ],
            "dimensions": number,
            "embeddingDataType": "string",
            "modelConfiguration": JSON value,
            "video": [
               {
```

```
                        "segmentationConfiguration": {
                           "fixedLengthDuration": number
                        }
                     }
                  ]
               }
            },
            "supplementalDataStorageConfiguration": {
               "storageLocations": [
                  {
                     "s3Location": {
                        "uri": "string"
                     },
                     "type": "string"
                  }
               ]
            }
         }
      },
      "knowledgeBaseId": "string",
      "name": "string",
      "roleArn": "string",
      "status": "string",
      "storageConfiguration": {
         "mongoDbAtlasConfiguration": {
            "collectionName": "string",
            "credentialsSecretArn": "string",
            "databaseName": "string",
            "endpoint": "string",
            "endpointServiceName": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "textIndexName": "string",
            "vectorIndexName": "string"
         },
         "neptuneAnalyticsConfiguration": {
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string"
            },
            "graphArn": "string"
```

```
        },
        "opensearchManagedClusterConfiguration": {
           "domainArn": "string",
           "domainEndpoint": "string",
           "fieldMapping": {
              "metadataField": "string",
              "textField": "string",
              "vectorField": "string"
           },
           "vectorIndexName": "string"
        },
        "opensearchServerlessConfiguration": {
           "collectionArn": "string",
           "fieldMapping": {
              "metadataField": "string",
              "textField": "string",
              "vectorField": "string"
           },
           "vectorIndexName": "string"
        },
        "pineconeConfiguration": {
           "connectionString": "string",
           "credentialsSecretArn": "string",
           "fieldMapping": {
              "metadataField": "string",
              "textField": "string"
           },
           "namespace": "string"
        },
        "rdsConfiguration": {
           "credentialsSecretArn": "string",
           "databaseName": "string",
           "fieldMapping": {
              "customMetadataField": "string",
              "metadataField": "string",
              "primaryKeyField": "string",
              "textField": "string",
              "vectorField": "string"
           },
           "resourceArn": "string",
           "tableName": "string"
        },
        "redisEnterpriseCloudConfiguration": {
           "credentialsSecretArn": "string",
```

```
            "endpoint": "string",
            "fieldMapping": {
               "metadataField": "string",
               "textField": "string",
               "vectorField": "string"
            },
            "vectorIndexName": "string"
         },
         "s3VectorsConfiguration": {
            "indexArn": "string",
            "indexName": "string",
            "vectorBucketArn": "string"
         },
         "type": "string"
      },
      "updatedAt": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response.

The following data is returned in JSON format by the service.

knowledgeBase

Contains details about the knowledge base.

Type: KnowledgeBase object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdatePrompt

Service: Agents for Amazon Bedrock

Modifies a prompt in your prompt library. Include both fields that you want to keep and fields that you want to replace. For more information, see Prompt management in Amazon Bedrock and Edit prompts in your prompt library in the Amazon Bedrock User Guide.

Request Syntax

```
PUT /prompts/promptIdentifier/ HTTP/1.1
Content-type: application/json

{
   "customerEncryptionKeyArn": "string",
   "defaultVariant": "string",
   "description": "string",
   "name": "string",
   "variants": [
      {
         "additionalModelRequestFields": JSON value,
         "genAiResource": { ... },
         "inferenceConfiguration": { ... },
         "metadata": [
            {
               "key": "string",
               "value": "string"
            }
         ],
         "modelId": "string",
         "name": "string",
         "templateConfiguration": { ... },
         "templateType": "string"
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

promptIdentifier

The unique identifier of the prompt.

Pattern: `([0-9a-zA-Z]{10})|(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10})(?::[0-9]{1,5})?`

Required: Yes

Request Body

The request accepts the following data in JSON format.

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to encrypt the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

defaultVariant

The name of the default variant for the prompt. This value must match the `name` field in the relevant PromptVariant object.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

description

A description for the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

name

A name for the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

variants

A list of objects, each containing details about a variant of the prompt.

Type: Array of PromptVariant objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "arn": "string",
   "createdAt": "string",
   "customerEncryptionKeyArn": "string",
   "defaultVariant": "string",
   "description": "string",
   "id": "string",
   "name": "string",
   "updatedAt": "string",
   "variants": [
      {
         "additionalModelRequestFields": JSON value,
         "genAiResource": { ... },
         "inferenceConfiguration": { ... },
         "metadata": [
            {
               "key": "string",
```

```
                "value": "string"
            }
        ],
        "modelId": "string",
        "name": "string",
        "templateConfiguration": { ... },
        "templateType": "string"
      }
   ],
   "version": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

arn

The Amazon Resource Name (ARN) of the prompt.

Type: String

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?)`

createdAt

The time at which the prompt was created.

Type: Timestamp

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to encrypt the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

defaultVariant

The name of the default variant for the prompt. This value must match the `name` field in the relevant PromptVariant object.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

description

The description of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

id

The unique identifier of the prompt.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

name

The name of the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

updatedAt

The time at which the prompt was last updated.

Type: Timestamp

variants

A list of objects, each containing details about a variant of the prompt.

Type: Array of PromptVariant objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

version

The version of the prompt. When you update a prompt, the version updated is the DRAFT version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 402

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ValidateFlowDefinition

Service: Agents for Amazon Bedrock

Validates the definition of a flow.

Request Syntax

```
POST /flows/validate-definition HTTP/1.1
Content-type: application/json

{
   "definition": {
      "connections": [
         {
            "configuration": { ... },
            "name": "string",
            "source": "string",
            "target": "string",
            "type": "string"
         }
      ],
      "nodes": [
         {
            "configuration": { ... },
            "inputs": [
               {
                  "category": "string",
                  "expression": "string",
                  "name": "string",
                  "type": "string"
               }
            ],
            "name": "string",
            "outputs": [
               {
                  "name": "string",
                  "type": "string"
               }
            ],
            "type": "string"
         }
      ]
   }
```

```
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

definition

The definition of a flow to validate.

Type: FlowDefinition object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "validations": [
      {
         "details": { ... },
         "message": "string",
         "severity": "string",
         "type": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

validations

Contains an array of objects, each of which contains an error identified by validation.

Type: Array of FlowValidation objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

fieldList

A list of objects containing fields that caused validation errors and their corresponding validation error messages.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Agents for Amazon Bedrock Runtime

The following actions are supported by Agents for Amazon Bedrock Runtime:

- AgenticRetrieveStream
- CheckIngestedDocumentAcl
- CreateInvocation
- CreateSession
- DeleteAgentMemory
- DeleteSession
- EndSession
- GenerateQuery
- GetAgentMemory
- GetDocumentContent
- GetExecutionFlowSnapshot
- GetFlowExecution
- GetIngestedDocumentAcl
- GetInvocationStep
- GetSession
- InvokeAgent
- InvokeFlow
- InvokeInlineAgent

- ListFlowExecutionEvents
- ListFlowExecutions
- ListInvocations
- ListInvocationSteps
- ListSessions
- ListTagsForResource
- OptimizePrompt
- PutInvocationStep
- Rerank
- Retrieve
- RetrieveAndGenerate
- RetrieveAndGenerateStream
- StartFlowExecution
- StopFlowExecution
- TagResource
- UntagResource
- UpdateSession

AgenticRetrieveStream

Service: Agents for Amazon Bedrock Runtime

Retrieves information from one or more knowledge bases using an agentic approach. Agentic retrieval uses a foundation model to intelligently decompose complex queries into sub-queries and iteratively retrieve relevant information from your knowledge bases. This approach improves retrieval accuracy for complex, multi-step questions that a single retrieval pass might not fully address.

The operation returns results through a stream that includes retrieval results, trace events for visibility into the process, and a generated response synthesized from the results by default, which can be turned off.

Request Syntax

```
POST /agenticRetrieveStream HTTP/1.1
Content-type: application/json

{
   "agenticRetrieveConfiguration": {
      "foundationModelConfiguration": {
         "bedrockFoundationModelConfiguration": {
            "modelConfiguration": {
               "modelArn": "string"
            }
         },
         "type": "string"
      },
      "foundationModelType": "string",
      "maxAgentIteration": number,
      "rerankingConfiguration": {
         "bedrockRerankingConfiguration": {
            "modelConfiguration": {
               "modelArn": "string"
            }
         },
         "type": "string"
      },
      "rerankingModelType": "string"
   },
   "generateResponse": boolean,
   "memoryConfiguration": {
```

```
      "memoryId": "string",
      "persistenceMode": "string",
      "retrievalConfigs": [
         {
            "metadataFilters": [
               {
                  "left": { ... },
                  "operator": "string",
                  "right": { ... }
               }
            ],
            "namespace": "string",
            "namespacePath": "string",
            "strategyId": "string"
         }
      ],
      "sessionBinding": {
         "actorId": "string",
         "sessionId": "string"
      }
   },
   "messages": [
      {
         "content": {
            "text": "string"
         },
         "role": "string"
      }
   ],
   "nextToken": "string",
   "policyConfiguration": {
      "bedrockGuardrailConfiguration": {
         "guardrailId": "string",
         "guardrailVersion": "string"
      }
   },
   "retrievers": [
      {
         "configuration": { ... },
         "description": "string"
      }
   ],
   "userContext": {
      "userId": "string"
```

```
    }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

agenticRetrieveConfiguration

Configuration settings for the agentic retrieval operation.

Type: AgenticRetrieveConfiguration object

Required: Yes

generateResponse

Whether to generate a response based on the retrieved results.

Type: Boolean

Required: No

memoryConfiguration

The configuration for using an Amazon Bedrock AgentCore Memory resource with this retrieval.

Type: AgenticRetrieveMemoryConfiguration object

Required: No

messages

The list of messages for the agentic retrieval conversation.

Type: Array of AgenticRetrieveMessage objects

Array Members: Minimum number of 1 item.

Required: Yes

nextToken

Opaque continuation token for paginated results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

policyConfiguration

Policy configuration for guardrails and content filtering.

Type: AgenticRetrievePolicyConfiguration object

Required: No

retrievers

The list of retrievers to use for agentic retrieval.

Type: Array of AgenticRetriever objects

Array Members: Minimum number of 1 item.

Required: Yes

userContext

Contains information about the user making the request. This is used for access control filtering to ensure that retrieval results only include documents the user is authorized to access.

Type: UserContext object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "accessDeniedException": {
   },
   "badGatewayException": {
   },
   "conflictException": {
```

```
    },
    "dependencyFailedException": {
    },
    "internalServerException": {
    },
    "resourceNotFoundException": {
    },
    "responseEvent": {
       "text": "string"
    },
    "result": {
       "generatedResponse": {
          "answer": "string",
          "citations": [
             {
                "endIndex": number,
                "references": [
                   {
                      "resultIndex": number
                   }
                ],
                "startIndex": number
             }
          ]
       },
       "nextToken": "string",
       "results": [
          {
             "content": {
                "byteContent": blob,
                "mimeType": "string",
                "text": "string"
             },
             "metadata": {
                "string" : JSON value
             },
             "sourceRetriever": {
                "identifier": "string"
             }
          }
       ]
    },
    "serviceQuotaExceededException": {
    },
```

```
    "throttlingException": {
    },
    "traceEvent": {
       "attributes": {
          "actions": [
             {
                "fullDocumentExpansion": {
                   "documentId": "string",
                   "sourceRetriever": {
                      "identifier": "string"
                   }
                },
                "memoryRetrieve": {
                   "inputQuery": {
                      "text": "string"
                   },
                   "memoryId": "string",
                   "namespace": "string",
                   "namespacePath": "string",
                   "strategyId": "string"
                },
                "retrieve": {
                   "inputQuery": {
                      "text": "string"
                   },
                   "sourceRetrievers": [
                      {
                         "identifier": "string"
                      }
                   ]
                }
             }
          ],
          "failures": [
             {
                "message": "string"
             }
          ],
          "message": "string",
          "retrievalMetadata": [
             {
                "identifier": "string",
                "retrievalType": "string"
             }
```

```
         ],
         "retrievalResponse": [
            {
               "content": {
                  "byteContent": blob,
                  "mimeType": "string",
                  "text": "string"
               },
               "metadata": {
                  "string" : JSON value
               },
               "sourceRetriever": {
                  "identifier": "string"
               }
            }
         ],
         "status": "string",
         "step": "string",
         "warnings": [
            { ... }
         ]
      },
      "id": "string",
      "timestamp": number
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

accessDeniedException

Access to the resource was denied.

Type: Exception

HTTP Status Code: 403

badGatewayException

A bad gateway error occurred.

Type: Exception

HTTP Status Code: 502

conflictException

A conflict occurred with the current state of the resource.

Type: Exception

HTTP Status Code: 409

dependencyFailedException

A dependency failed during the operation.

Type: Exception

HTTP Status Code: 424

internalServerException

An internal server error occurred.

Type: Exception

HTTP Status Code: 500

resourceNotFoundException

The specified resource was not found.

Type: Exception

HTTP Status Code: 404

responseEvent

A chunk of the generated answer. Emitted only when generateResponse is true.

Type: AgenticRetrieveResponseEvent object

result

A retrieval result event containing the retrieved items.

Type: AgenticRetrieveResultEvent object

serviceQuotaExceededException

The service quota has been exceeded.

Type: Exception

HTTP Status Code: 400

throttlingException

The request was throttled.

Type: Exception

HTTP Status Code: 429

traceEvent

A trace event providing visibility into the retrieval process.

Type: AgenticRetrieveTraceEvent object

validationException

The request validation failed.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Send a basic query using a managed model

The following example uses the managed foundation model type to query a knowledge base with agentic retrieval. When you set `foundationModelType` to `MANAGED`, the service selects the optimal model for you.

Sample Request

```
POST /agenticRetrieveStream HTTP/1.1
Content-type: application/json

{
    "messages": [
        {
            "content": {
                "text": "What are the main benefits of using Amazon Bedrock?"
            },
            "role": "user"
        }
    ],
    "retrievers": [
        {
            "description": "Knowledge base about Amazon Bedrock documentation",
            "configuration": {
                "knowledgeBase": {
                    "knowledgeBaseId": "KB12345678"
                }
            }
```

```
        }
    ],
    "agenticRetrieveConfiguration": {
        "foundationModelType": "MANAGED",
        "rerankingModelType": "MANAGED"
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CheckIngestedDocumentAcl

Service: Agents for Amazon Bedrock Runtime

Checks whether a user has access to a specific document by verifying against the ingested access control list (ACL) in a knowledge base. Use this operation to validate that document-level access control is working as expected after ingestion. To use this operation, you must have the `bedrock:CheckIngestedDocumentAcl` permission.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/check-ingested-document-acl HTTP/1.1
Content-type: application/json

{
   "documentId": "string",
   "userContext": {
      "userId": "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source that contains the document.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that contains the document.

Length Constraints: Minimum length of 10. Maximum length of 2048.

Pattern: `[0-9a-zA-Z]{10}$|^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

documentId

The unique identifier of the document to check access for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1825.

Pattern: \P{C}*

Required: Yes

userContext

The context object containing identity information for access control filtering, including user ID and optional group memberships used to evaluate the document access control list (ACL).

Type: UserContext object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "hasAccess": boolean
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

hasAccess

Specifies whether the user has access to the document based on the ingested access control list (ACL). Returns `true` if the user is allowed access, and `false` otherwise.

Type: Boolean

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Check if a user has access to a document

The following example checks whether a specific user has access to a document in a knowledge base based on the ingested ACL.

Sample Request

```
POST /knowledgebases/KB12345678/datasources/DS12345678/check-ingested-document-acl
 HTTP/1.1
Content-type: application/json

{
    "documentId": "doc-001",
    "userContext": {
        "userId": "user@example.com",
        "userGroups": [
            {
                "id": "engineering",
                "type": "KNOWLEDGE_BASE"
            }
        ]
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateInvocation

Service: Agents for Amazon Bedrock Runtime

Creates a new invocation within a session. An invocation groups the related invocation steps that store the content from a conversation. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Related APIs

- ListInvocations
- ListSessions
- GetSession

Request Syntax

```
PUT /sessions/sessionIdentifier/invocations/ HTTP/1.1
Content-type: application/json

{
   "description": "string",
   "invocationId": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

sessionIdentifier

The unique identifier for the associated session for the invocation. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

description

A description for the interactions in the invocation. For example, "User asking about weather in Seattle".

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

invocationId

A unique identifier for the invocation in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "createdAt": "string",
   "invocationId": "string",
   "sessionId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp for when the invocation was created.

Type: Timestamp

invocationId

The unique identifier for the invocation.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionId

The unique identifier for the session associated with the invocation.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Request example

This example illustrates one usage of CreateInvocation.

```
PUT bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/invocations/ HTTP/1.1
Content-type: application/json

{
    "description": "User asking about weather in Seattle",
```

```
    "invocationId": "12345abc-1234-abcd-1234-abcdef123456"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateSession

Service: Agents for Amazon Bedrock Runtime

Creates a session to temporarily store conversations for generative AI (GenAI) applications built with open-source frameworks such as LangGraph and LlamaIndex. Sessions enable you to save the state of conversations at checkpoints, with the added security and infrastructure of AWS. For more information, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

By default, Amazon Bedrock uses AWS-managed keys for session encryption, including session metadata, or you can use your own KMS key. For more information, see Amazon Bedrock session encryption.

> **Note**
>
> You use a session to store state and conversation history for generative AI applications built with open-source frameworks. For Amazon Bedrock Agents, the service automatically manages conversation context and associates them with the agent-specific sessionId you specify in the InvokeAgent API operation.

Related APIs:

- ListSessions
- GetSession
- EndSession
- DeleteSession

Request Syntax

```
PUT /sessions/ HTTP/1.1
Content-type: application/json

{
   "encryptionKeyArn": "string",
   "sessionMetadata": {
      "string" : "string"
   },
```

```
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

encryptionKeyArn

The Amazon Resource Name (ARN) of the KMS key to use to encrypt the session data. The user or role creating the session must have permission to use the key. For more information, see Amazon Bedrock session encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

sessionMetadata

A map of key-value pairs containing attributes to be persisted across the session. For example, the user's ID, their language preference, and the type of device they are using.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 5000.

Required: No

tags

Specify the key-value pairs for the tags that you want to attach to the session.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "createdAt": "string",
   "sessionArn": "string",
   "sessionId": "string",
   "sessionStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp for when the session was created.

Type: Timestamp

sessionArn

The Amazon Resource Name (ARN) of the created session.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionId

The unique identifier for the session.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionStatus

The current status of the session.

Type: String

Valid Values: `ACTIVE | EXPIRED | ENDED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of CreateSession.

```
PUT bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/ HTTP/1.1
Content-type: application/json

{
    "encryptionKeyArn": "key Amazon Resource Name (ARN)",
    "sessionMetadata": {
        "deviceType": "mobile",
        "language": "english"
    },
    "tags": {
        "Environment": "Production",
        "Project": "Demo"
```

```
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteAgentMemory

Service: Agents for Amazon Bedrock Runtime

Deletes memory from the specified memory identifier.

Request Syntax

```
DELETE /agents/agentId/agentAliases/agentAliasId/memories?
memoryId=memoryId&sessionId=sessionId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of an alias of an agent.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

agentId

The unique identifier of the agent to which the alias belongs.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

memoryId

The unique identifier of the memory.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

sessionId

The unique session identifier of the memory.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 202
```

Response Elements

If the action is successful, the service sends back an HTTP 202 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteSession

Service: Agents for Amazon Bedrock Runtime

Deletes a session that you ended. You can't delete a session with an ACTIVE status. To delete an active session, you must first end it with the EndSession API operation. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
DELETE /sessions/sessionIdentifier/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

sessionIdentifier

The unique identifier for the session to be deleted. You can specify either the session's sessionId or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of DeleteSession.

```
DELETE bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

EndSession

Service: Agents for Amazon Bedrock Runtime

Ends the session. After you end a session, you can still access its content but you can't add to it. To delete the session and it's content, you use the DeleteSession API operation. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
PATCH /sessions/sessionIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

sessionIdentifier

The unique identifier for the session to end. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "sessionArn": "string",
   "sessionId": "string",
   "sessionStatus": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

sessionArn

The Amazon Resource Name (ARN) of the session you ended.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionId

The unique identifier of the session you ended.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionStatus

The current status of the session you ended.

Type: String

Valid Values: `ACTIVE | EXPIRED | ENDED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of EndSession.

```
PATCH bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GenerateQuery

Service: Agents for Amazon Bedrock Runtime

Generates an SQL query from a natural language query. For more information, see Generate a query for structured data in the Amazon Bedrock User Guide.

Request Syntax

```
POST /generateQuery HTTP/1.1
Content-type: application/json

{
   "queryGenerationInput": {
      "text": "string",
      "type": "string"
   },
   "transformationConfiguration": {
      "mode": "string",
      "textToSqlConfiguration": {
         "knowledgeBaseConfiguration": {
            "knowledgeBaseArn": "string"
         },
         "type": "string"
      }
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

queryGenerationInput

Specifies information about a natural language query to transform into SQL.

Type: QueryGenerationInput object

Required: Yes

transformationConfiguration

Specifies configurations for transforming the natural language query into SQL.

Type: TransformationConfiguration object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "queries": [
      {
         "sql": "string",
         "type": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

queries

A list of objects, each of which defines a generated query that can correspond to the natural language queries.

Type: Array of GeneratedQuery objects

Array Members: Minimum number of 0 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAgentMemory

Service: Agents for Amazon Bedrock Runtime

Gets the sessions stored in the memory of the agent.

Request Syntax

```
GET /agents/agentId/agentAliases/agentAliasId/memories?
maxItems=maxItems&memoryId=memoryId&memoryType=memoryType&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The unique identifier of an alias of an agent.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

agentId

The unique identifier of the agent to which the alias belongs.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

maxItems

The maximum number of items to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

memoryId

The unique identifier of the memory.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: Yes

memoryType

The type of memory.

Valid Values: `SESSION_SUMMARY`

Required: Yes

nextToken

If the total number of results is greater than the maxItems value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "memoryContents": [
      { ... }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

memoryContents

Contains details of the sessions stored in the memory

Type: Array of Memory objects

nextToken

If the total number of results is greater than the maxItems value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetDocumentContent

Service: Agents for Amazon Bedrock Runtime

Retrieves the content of an ingested document from a knowledge base. Returns a pre-signed URL for secure document access.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/documents/documentId/
content HTTP/1.1
Content-type: application/json

{
   "outputFormat": "string",
   "userContext": {
      "userId": "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source that contains the document.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

documentId

The unique identifier of the document to retrieve content for.

Length Constraints: Minimum length of 1. Maximum length of 1825.

Pattern: `\P{C}*`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that contains the document.

Length Constraints: Minimum length of 10. Maximum length of 2048.

Pattern: `[0-9a-zA-Z]{10}$|^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

outputFormat

The output format for the document content. `RAW` returns the original file. `EXTRACTED` returns parsed text as JSON. Defaults to `RAW`.

Type: String

Valid Values: `RAW | EXTRACTED`

Required: No

userContext

Contains information about the user making the request. This is used for access control filtering to ensure that results only include documents the user is authorized to access.

Type: UserContext object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "documentContentLength": number,
```

```
    "mimeType": "string",
    "presignedUrl": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

documentContentLength

The size of the document content in bytes available at the pre-signed URL.

Type: Long

mimeType

The MIME type of the document content. For `RAW` format, this is the original file type (for example, `application/pdf`). For `EXTRACTED` format, this is always `application/json`.

Type: String

presignedUrl

A pre-signed URL for downloading the document content. The URL expires after 5 minutes.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Retrieve the original document content

The following example retrieves the original (raw) content of a document from a knowledge base. The response includes a pre-signed URL that you can use to download the document.

Sample Request

```
POST /knowledgebases/KB12345678/datasources/DS12345678/documents/DOC12345678/content
 HTTP/1.1
Content-type: application/json

{
    "outputFormat": "RAW"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetExecutionFlowSnapshot

Service: Agents for Amazon Bedrock Runtime

Retrieves the flow definition snapshot used for a flow execution. The snapshot represents the flow metadata and definition as it existed at the time the execution was started. Note that even if the flow is edited after an execution starts, the snapshot connected to the execution remains unchanged.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Request Syntax

```
GET /flows/flowIdentifier/aliases/flowAliasIdentifier/executions/executionIdentifier/
flowsnapshot HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

executionIdentifier

The unique identifier of the flow execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

Required: Yes

flowAliasIdentifier

The unique identifier of the flow alias used for the flow execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "customerEncryptionKeyArn": "string",
   "definition": "string",
   "executionRoleArn": "string",
   "flowAliasIdentifier": "string",
   "flowIdentifier": "string",
   "flowVersion": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the customer managed AWS KMS key that's used to encrypt the flow snapshot.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

definition

The flow definition used for the flow execution, including the nodes, connections, and configuration at the time when the execution started.

The definition returns as a string that follows the structure of a FlowDefinition object.

Type: String

executionRoleArn

The Amazon Resource Name (ARN) of the IAM service role that's used by the flow execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/(service-role/)?.+`

flowAliasIdentifier

The unique identifier of the flow alias used for the flow execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

flowIdentifier

The unique identifier of the flow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

flowVersion

The version of the flow used for the flow execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetFlowExecution

Service: Agents for Amazon Bedrock Runtime

Retrieves details about a specific flow execution, including its status, start and end times, and any errors that occurred during execution.

Request Syntax

```
GET /flows/flowIdentifier/aliases/flowAliasIdentifier/executions/executionIdentifier
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

executionIdentifier

The unique identifier of the flow execution to retrieve.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

Required: Yes

flowAliasIdentifier

The unique identifier of the flow alias used for the execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "endedAt": "string",
   "errors": [
      {
         "error": "string",
         "message": "string",
         "nodeName": "string"
      }
   ],
   "executionArn": "string",
   "flowAliasIdentifier": "string",
   "flowIdentifier": "string",
   "flowVersion": "string",
   "startedAt": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

endedAt

The timestamp when the flow execution ended. This field is only populated when the execution has completed, failed, timed out, or been aborted.

Type: Timestamp

errors

A list of errors that occurred during the flow execution. Each error includes an error code, message, and the node where the error occurred, if applicable.

Type: Array of FlowExecutionError objects

executionArn

The Amazon Resource Name (ARN) that uniquely identifies the flow execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

flowAliasIdentifier

The unique identifier of the flow alias used for the execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

flowIdentifier

The unique identifier of the flow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

flowVersion

The version of the flow used for the execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

startedAt

The timestamp when the flow execution started.

Type: Timestamp

status

The current status of the flow execution.

Flow executions time out after 24 hours.

Type: String

Valid Values: `Running | Succeeded | Failed | TimedOut | Aborted`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetIngestedDocumentAcl

Service: Agents for Amazon Bedrock Runtime

Retrieves the ingested access control list (ACL) for a specific document in a knowledge base. Use this operation to inspect the allow and deny lists that were ingested for a document to troubleshoot access control issues. To use this operation, you must have the `bedrock:GetIngestedDocumentAcl` permission.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/datasources/dataSourceId/get-ingested-document-acl
 HTTP/1.1
Content-type: application/json

{
   "documentId": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

dataSourceId

The unique identifier of the data source that contains the document.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base that contains the document.

Length Constraints: Minimum length of 10. Maximum length of 2048.

Pattern: `[0-9a-zA-Z]{10}$|^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

documentId

The unique identifier of the document to retrieve the ingested access control list (ACL) for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1825.

Pattern: \P{C}*

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "documentAcl": {
      "allowList": {
         "conditions": [
            {
               "conditionOperator": "string",
               "groups": [
                  {
                     "id": "string",
                     "type": "string"
                  }
               ],
               "users": [
                  {
                     "id": "string",
                     "type": "string"
                  }
               ]
            }
         ],
         "memberRelation": "string"
```

```
      },
      "denyList": {
         "conditions": [
            {
               "conditionOperator": "string",
               "groups": [
                  {
                     "id": "string",
                     "type": "string"
                  }
               ],
               "users": [
                  {
                     "id": "string",
                     "type": "string"
                  }
               ]
            }
         ],
         "memberRelation": "string"
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

documentAcl

The ingested document access control list (ACL) containing allow and deny membership information.

Type: DocumentAcl object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Retrieve the ingested ACL for a document

The following example retrieves the full ingested access control list (ACL) for a document, including both allow and deny membership lists.

Sample Request

```
POST /knowledgebases/KB12345678/datasources/DS12345678/get-ingested-document-acl
 HTTP/1.1
Content-type: application/json

{
    "documentId": "doc-001"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetInvocationStep

Service: Agents for Amazon Bedrock Runtime

Retrieves the details of a specific invocation step within an invocation in a session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
POST /sessions/sessionIdentifier/invocationSteps/invocationStepId HTTP/1.1
Content-type: application/json

{
   "invocationIdentifier": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

invocationStepId

The unique identifier (in UUID format) for the specific invocation step to retrieve.

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

sessionIdentifier

The unique identifier for the invocation step's associated session. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

invocationIdentifier

The unique identifier for the invocation in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "invocationStep": {
      "invocationId": "string",
      "invocationStepId": "string",
      "invocationStepTime": "string",
      "payload": { ... },
      "sessionId": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationStep

The complete details of the requested invocation step.

Type: InvocationStep object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetInvocationStep.

```
POST bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/invocationSteps/11111111-2222-3333-4444-555555555555 HTTP/1.1
```

```
Content-type: application/json

{
   "invocationIdentifier": "abc-1234-abcd-1234-abcdef123456"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetSession

Service: Agents for Amazon Bedrock Runtime

Retrieves details about a specific session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
GET /sessions/sessionIdentifier/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

sessionIdentifier

A unique identifier for the session to retrieve. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "encryptionKeyArn": "string",
   "lastUpdatedAt": "string",
   "sessionArn": "string",
   "sessionId": "string",
   "sessionMetadata": {
```

```
      "string" : "string"
   },
   "sessionStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp for when the session was created.

Type: Timestamp

encryptionKeyArn

The Amazon Resource Name (ARN) of the AWS Key Management Service key used to encrypt the session data. For more information, see Amazon Bedrock session encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

lastUpdatedAt

The timestamp for when the session was last modified.

Type: Timestamp

sessionArn

The Amazon Resource Name (ARN) of the session.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionId

The unique identifier for the session in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionMetadata

A map of key-value pairs containing attributes persisted across the session.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 5000.

sessionStatus

The current status of the session.

Type: String

Valid Values: `ACTIVE | EXPIRED | ENDED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of GetSession.

```
GET bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/ HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeAgent

Service: Agents for Amazon Bedrock Runtime

> **Note**
>
> Amazon Bedrock Agents (now Amazon Bedrock Agents Classic) is no longer open to new customers. For capabilities similar to Bedrock Agents Classic, explore Amazon Bedrock AgentCore. Existing customers can continue to use the service as normal. For more information, see Amazon Bedrock Agents Classic availability change.

> **Note**
>
> The AWS CLI doesn't support streaming operations in Amazon Bedrock, including `InvokeAgent`.

Sends a prompt for the agent to process and respond to. Note the following fields for the request:

- To continue the same conversation with an agent, use the same `sessionId` value in the request.
- To activate trace enablement, turn `enableTrace` to `true`. Trace enablement helps you follow the agent's reasoning process that led it to the information it processed, the actions it took, and the final result it yielded. For more information, see Trace enablement.
- End a conversation by setting `endSession` to `true`.
- In the `sessionState` object, you can include attributes for the session or prompt or, if you configured an action group to return control, results from invocation of the action group.

The response contains both **chunk** and **trace** attributes.

The final response is returned in the `bytes` field of the `chunk` object. The `InvokeAgent` returns one chunk for the entire interaction.

- The `attribution` object contains citations for parts of the response.
- If you set `enableTrace` to `true` in the request, you can trace the agent's steps and reasoning process that led it to the response.
- If the action predicted was configured to return control, the response returns parameters for the action, elicited from the user, in the `returnControl` field.

- Errors are also surfaced in the response.

Request Syntax

```
POST /agents/agentId/agentAliases/agentAliasId/sessions/sessionId/text HTTP/1.1
x-amz-source-arn: sourceArn
Content-type: application/json

{
   "bedrockModelConfigurations": {
      "performanceConfig": {
         "latency": "string"
      }
   },
   "enableTrace": boolean,
   "endSession": boolean,
   "inputText": "string",
   "memoryId": "string",
   "promptCreationConfigurations": {
      "excludePreviousThinkingSteps": boolean,
      "previousConversationTurnsToInclude": number
   },
   "sessionState": {
      "conversationHistory": {
         "messages": [
            {
               "content": [
                  { ... }
               ],
               "role": "string"
            }
         ]
      },
      "files": [
         {
            "name": "string",
            "source": {
               "byteContent": {
                  "data": blob,
                  "mediaType": "string"
               },
               "s3Location": {
                  "uri": "string"
```

```
            },
            "sourceType": "string"
         },
         "useCase": "string"
      }
   ],
   "invocationId": "string",
   "knowledgeBaseConfigurations": [
      {
         "knowledgeBaseId": "string",
         "retrievalConfiguration": {
            "managedSearchConfiguration": {
               "filter": { ... },
               "numberOfResults": number,
               "rerankingConfiguration": {
                  "bedrockRerankingConfiguration": {
                     "metadataConfiguration": {
                        "selectionMode": "string",
                        "selectiveModeConfiguration": { ... }
                     },
                     "modelConfiguration": {
                        "additionalModelRequestFields": {
                           "string" : JSON value
                        },
                        "modelArn": "string"
                     },
                     "numberOfRerankedResults": number
                  },
                  "type": "string"
               },
               "rerankingModelType": "string"
            },
            "vectorSearchConfiguration": {
               "filter": { ... },
               "implicitFilterConfiguration": {
                  "metadataAttributes": [
                     {
                        "description": "string",
                        "key": "string",
                        "type": "string"
                     }
                  ],
                  "modelArn": "string"
               },
```

```
                    "numberOfResults": number,
                    "overrideSearchType": "string",
                    "rerankingConfiguration": {
                       "bedrockRerankingConfiguration": {
                          "metadataConfiguration": {
                             "selectionMode": "string",
                             "selectiveModeConfiguration": { ... }
                          },
                          "modelConfiguration": {
                             "additionalModelRequestFields": {
                                "string" : JSON value
                             },
                             "modelArn": "string"
                          },
                          "numberOfRerankedResults": number
                       },
                       "type": "string"
                    }
                 }
              }
           }
        ],
        "promptSessionAttributes": {
           "string" : "string"
        },
        "returnControlInvocationResults": [
           { ... }
        ],
        "sessionAttributes": {
           "string" : "string"
        }
    },
    "streamingConfigurations": {
        "applyGuardrailInterval": number,
        "streamFinalResponse": boolean
    }
}
```

URI Request Parameters

The request uses the following URI parameters.

agentAliasId

The alias of the agent to use.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

agentId

The unique identifier of the agent to use.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

sessionId

The unique identifier of the session. Use the same value across requests to continue the same conversation.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: Yes

sourceArn

The ARN of the resource making the request.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent/[0-9a-zA-Z]{10}`

Request Body

The request accepts the following data in JSON format.

bedrockModelConfigurations

Model performance settings for the request.

Type: BedrockModelConfigurations object

Required: No

enableTrace

Specifies whether to turn on the trace or not to track the agent's reasoning process. For more information, see Trace enablement.

Type: Boolean

Required: No

endSession

Specifies whether to end the session with the agent or not.

Type: Boolean

Required: No

inputText

The prompt text to send the agent.

> **Note**
>
> If you include `returnControlInvocationResults` in the `sessionState` field, the `inputText` field will be ignored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: No

memoryId

The unique identifier of the agent memory.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

promptCreationConfigurations

Specifies parameters that control how the service populates the agent prompt for an `InvokeAgent` request. You can control which aspects of previous invocations in the same agent session the service uses to populate the agent prompt. This gives you more granular control over the contextual history that is used to process the current request.

Type: PromptCreationConfigurations object

Required: No

sessionState

Contains parameters that specify various attributes of the session. For more information, see Control session context.

> **Note**
>
> If you include `returnControlInvocationResults` in the `sessionState` field, the `inputText` field will be ignored.

Type: SessionState object

Required: No

streamingConfigurations

Specifies the configurations for streaming.

> **Note**
>
> To use agent streaming, you need permissions to perform the `bedrock:InvokeModelWithResponseStream` action.

Type: StreamingConfigurations object

Required: No

Response Syntax

```
HTTP/1.1 200
x-amzn-bedrock-agent-content-type: contentType
x-amz-bedrock-agent-session-id: sessionId
x-amz-bedrock-agent-memory-id: memoryId
Content-type: application/json

{
   "accessDeniedException": {
   },
   "badGatewayException": {
   },
   "chunk": {
      "attribution": {
         "citations": [
            {
               "generatedResponsePart": {
                  "textResponsePart": {
                     "span": {
                        "end": number,
                        "start": number
                     },
                     "text": "string"
                  }
               },
               "retrievedReferences": [
                  {
                     "content": {
                        "audio": {
                           "s3Uri": "string",
                           "transcription": "string"
                        },
                        "byteContent": "string",
                        "row": [
                           {
                              "columnName": "string",
                              "columnValue": "string",
                              "type": "string"
                           }
                        ],
                        "text": "string",
                        "type": "string",
```

```
                        "video": {
                           "s3Uri": "string",
                           "summary": "string"
                        }
                     },
                     "location": {
                        "confluenceLocation": {
                           "url": "string"
                        },
                        "customDocumentLocation": {
                           "id": "string"
                        },
                        "googleDriveLocation": {
                           "url": "string"
                        },
                        "kendraDocumentLocation": {
                           "uri": "string"
                        },
                        "oneDriveLocation": {
                           "url": "string"
                        },
                        "s3Location": {
                           "uri": "string"
                        },
                        "salesforceLocation": {
                           "url": "string"
                        },
                        "sharePointLocation": {
                           "url": "string"
                        },
                        "sqlLocation": {
                           "query": "string"
                        },
                        "type": "string",
                        "webLocation": {
                           "url": "string"
                        }
                     },
                     "metadata": {
                        "string" : JSON value
                     }
                  }
               ]
            }
```

```
            ]
        },
        "bytes": blob
    },
    "conflictException": {
    },
    "dependencyFailedException": {
    },
    "files": {
        "files": [
            {
                "bytes": blob,
                "name": "string",
                "type": "string"
            }
        ]
    },
    "internalServerException": {
    },
    "modelNotReadyException": {
    },
    "resourceNotFoundException": {
    },
    "returnControl": {
        "invocationId": "string",
        "invocationInputs": [
            { ... }
        ]
    },
    "serviceQuotaExceededException": {
    },
    "throttlingException": {
    },
    "trace": {
        "agentAliasId": "string",
        "agentId": "string",
        "agentVersion": "string",
        "callerChain": [
            { ... }
        ],
        "collaboratorName": "string",
        "eventTime": "string",
        "sessionId": "string",
        "trace": { ... }
```

```
    },
    "validationException": {
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType

The MIME type of the input data in the request. The default value is `application/json`.

memoryId

The unique identifier of the agent memory.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

sessionId

The unique identifier of the session with the agent.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

The following data is returned in JSON format by the service.

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

chunk

Contains a part of an agent response and citations for it.

Type: PayloadPart object

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

files

Contains intermediate response for code interpreter if any files have been generated.

Type: FilePart object

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

modelNotReadyException

The model specified in the request is not ready to serve Inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

Type: Exception

HTTP Status Code: 424

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception

HTTP Status Code: 404

returnControl

Contains the parameters and information that the agent elicited from the customer to carry out an action. This information is returned to the system and can be used in your own setup for fulfilling the action.

Type: ReturnControlPayload object

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception

HTTP Status Code: 400

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

trace

Contains information about the agent and session, alongside the agent's reasoning process and results from calling actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace events.

Type: TracePart object

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ModelNotReadyException

The model specified in the request is not ready to serve inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

HTTP Status Code: 424

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example simple request

The following example inquires the agent to get the weather for Seattle.

```
POST https://bedrock-agent-runtime.us-east-1.amazonaws.com/agents/AGENT12345/
agentAliases/TSTALIASID/sessions/abb/text

{
    "inputText": "give me the weather for seattle",
    "enableTrace": true
}
```

Example response (action group defined with OpenAPI schema, control returned)

The following example shows a response from an agent that has invoked an action group that was configured as follows:

- Defined with an OpenAPI schema
- Configured to return control to the agent developer

```
HTTP/1.1 200
x-amzn-bedrock-agent-content-type: application/json
x-amz-bedrock-agent-session-id: session0
Content-type: application/json

{
    "invocationInputs": [{
        "apiInvocationInput": {
            "actionGroup": "WeatherAPIs",
            "apiPath": "/get-weather",
            "httpMethod": "get",
            "parameters": [
                {
                    "name": "location",
                    "type": "string",
                    "value": "seattle"
                },
                {
                    "name": "date",
                    "type": "string",
```

```
                    "value": "2024-09-15"
                }
            ]
        }
    }],
    "invocationId": "337cb2f6-ec74-4b49-8141-00b8091498ad"
}
```

Example request using results from returned control (action group defined with OpenAPI schema)

The following example shows a request in which the results returned in the `InvokeAgent` response from an agent are passed to the `sessionState` of a new request. The results were returned from an agent that has invoked an action group that was configured as follows:

- Defined with an OpenAPI schema
- Configured to return control to the agent developer

The `invocationId` must match the `invocationId` that was returned in the response.

```
POST https: //bedrock-agent-runtime.us-east-1.amazonaws.com/agents/AGENT12345/
agentAliases/TSTALIASID/sessions/abb/text

{
    "enableTrace": true,
    "sessionState": {
        "invocationId": "337cb2f6-ec74-4b49-8141-00b8091498ad",
        "returnControlInvocationResults": [{
            "apiResult": {
                "actionGroup": "WeatherAPIs",
                "httpMethod": "get",
                "apiPath": "/get-weather",
                "responseBody": {
                    "application/json": {
                        "body": "It's rainy in Seattle today."
                    }
                }
            }
        }]
    }
}
```

Example response (action group defined with function details, control returned)

The following example shows a response from an agent that has invoked an action group that was configured as follows:

- Defined with function details
- Configured to return control to the agent developer

```
HTTP/1.1 200
x-amzn-bedrock-agent-content-type: application/json
x-amz-bedrock-agent-session-id: session0
Content-type: application/json

{
    "invocationInputs": [{
        "functionInvocationInput": {
            "actionGroup": "WeatherAPIs",
            "function": "getWeather",
            "parameters": [
                {
                    "name": "location",
                    "type": "string",
                    "value": "seattle"
                },
                {
                    "name": "date",
                    "type": "string",
                    "value": "2024-09-15"
                }
            ]
        }
    }],
    "invocationId": "79e0feaa-c6f7-49bf-814d-b7c498505172"
}
```

Example request using results from returned control (action group defined with function details)

The following example shows a request in which the results returned in the `InvokeAgent` response from an agent are passed to the `sessionState` of a new request. The results were returned from an agent that has invoked an action group that was configured as follows:

- Defined with function details
- Configured to return control to the agent developer

The `invocationId` must match the `invocationId` that was returned in the response.

```
POST https://bedrock-agent-runtime.us-east-1.amazonaws.com/agents/AGENT12345/
agentAliases/TSTALIASID/sessions/abb/text

{
    "enableTrace": true,
    "sessionState": {
        "invocationId": "79e0feaa-c6f7-49bf-814d-b7c498505172",
        "returnControlInvocationResults": [{
            "functionResult": {
                "actionGroup": "WeatherAPIs",
                "function": "getWeather",
                "responseBody": {
                    "TEXT": {
                        "body": "It's rainy in Seattle today."
                    }
                }
            }
        }]
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeFlow

Service: Agents for Amazon Bedrock Runtime

Invokes an alias of a flow to run the inputs that you specify and return the output of each node as a stream. If there's an error, the error is returned. For more information, see Test a flow in Amazon Bedrock in the Amazon Bedrock User Guide.

> **Note**
>
> The AWS CLI doesn't support streaming operations in Amazon Bedrock, including `InvokeFlow`.

Request Syntax

```
POST /flows/flowIdentifier/aliases/flowAliasIdentifier HTTP/1.1
Content-type: application/json

{
   "enableTrace": boolean,
   "executionId": "string",
   "inputs": [
      {
         "content": { ... },
         "nodeInputName": "string",
         "nodeName": "string",
         "nodeOutputName": "string"
      }
   ],
   "modelPerformanceConfiguration": {
      "performanceConfig": {
         "latency": "string"
      }
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

flowAliasIdentifier

The unique identifier of the flow alias.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

enableTrace

Specifies whether to return the trace for the flow or not. Traces track inputs and outputs for nodes in the flow. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Type: Boolean

Required: No

executionId

The unique identifier for the current flow execution. If you don't provide a value, Amazon Bedrock creates the identifier for you.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

inputs

A list of objects, each containing information about an input into the flow.

Type: Array of FlowInput objects

Array Members: Fixed number of 1 item.

Required: Yes

modelPerformanceConfiguration

Model performance settings for the request.

Type: ModelPerformanceConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 200
x-amz-bedrock-flow-execution-id: executionId
Content-type: application/json

{
   "accessDeniedException": {
   },
   "badGatewayException": {
   },
   "conflictException": {
   },
   "dependencyFailedException": {
   },
   "flowCompletionEvent": {
      "completionReason": "string"
   },
   "flowMultiTurnInputRequestEvent": {
      "content": { ... },
      "nodeName": "string",
      "nodeType": "string"
```

```
    },
    "flowOutputEvent": {
       "content": { ... },
       "nodeName": "string",
       "nodeType": "string"
    },
    "flowTraceEvent": {
       "trace": { ... }
    },
    "internalServerException": {
    },
    "resourceNotFoundException": {
    },
    "serviceQuotaExceededException": {
    },
    "throttlingException": {
    },
    "validationException": {
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

executionId

The unique identifier for the current flow execution.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

The following data is returned in JSON format by the service.

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

flowCompletionEvent

Contains information about why the flow completed.

Type: FlowCompletionEvent object

flowMultiTurnInputRequestEvent

The event stream containing the multi-turn input request information from the flow.

Type: FlowMultiTurnInputRequestEvent object

flowOutputEvent

Contains information about an output from flow invocation.

Type: FlowOutputEvent object

flowTraceEvent

Contains information about a trace, which tracks an input or output for a node in the flow.

Type: FlowTraceEvent object

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception

HTTP Status Code: 404

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception

HTTP Status Code: 400

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeInlineAgent

Service: Agents for Amazon Bedrock Runtime

Invokes an inline Amazon Bedrock agent using the configurations you provide with the request.

- Specify the following fields for security purposes.
  - (Optional) `customerEncryptionKeyArn` – The Amazon Resource Name (ARN) of a AWS KMS key to encrypt the creation of the agent.
  - (Optional) `idleSessionTTLinSeconds` – Specify the number of seconds for which the agent should maintain session information. After this time expires, the subsequent `InvokeInlineAgent` request begins a new session.
- To override the default prompt behavior for agent orchestration and to use advanced prompts, include a `promptOverrideConfiguration` object. For more information, see Advanced prompts.
- The agent instructions will not be honored if your agent has only one knowledge base, uses default prompts, has no action group, and user input is disabled.

> **Note**
>
> The AWS CLI doesn't support streaming operations in Amazon Bedrock, including `InvokeInlineAgent`.

Request Syntax

```
POST /agents/sessionId HTTP/1.1
Content-type: application/json

{
   "actionGroups": [
      {
         "actionGroupExecutor": { ... },
         "actionGroupName": "string",
         "apiSchema": { ... },
         "description": "string",
         "functionSchema": { ... },
         "parentActionGroupSignature": "string",
         "parentActionGroupSignatureParams": {
```

```
            "string" : "string"
         }
      }
   ],
   "agentCollaboration": "string",
   "agentName": "string",
   "bedrockModelConfigurations": {
      "performanceConfig": {
         "latency": "string"
      }
   },
   "collaboratorConfigurations": [
      {
         "agentAliasArn": "string",
         "collaboratorInstruction": "string",
         "collaboratorName": "string",
         "relayConversationHistory": "string"
      }
   ],
   "collaborators": [
      {
         "actionGroups": [
            {
               "actionGroupExecutor": { ... },
               "actionGroupName": "string",
               "apiSchema": { ... },
               "description": "string",
               "functionSchema": { ... },
               "parentActionGroupSignature": "string",
               "parentActionGroupSignatureParams": {
                  "string" : "string"
               }
            }
         ],
         "agentCollaboration": "string",
         "agentName": "string",
         "collaboratorConfigurations": [
            {
               "agentAliasArn": "string",
               "collaboratorInstruction": "string",
               "collaboratorName": "string",
               "relayConversationHistory": "string"
            }
         ],
```

```
        "customerEncryptionKeyArn": "string",
        "foundationModel": "string",
        "guardrailConfiguration": {
           "guardrailIdentifier": "string",
           "guardrailVersion": "string"
        },
        "idleSessionTTLInSeconds": number,
        "instruction": "string",
        "knowledgeBases": [
           {
              "description": "string",
              "knowledgeBaseId": "string",
              "retrievalConfiguration": {
                 "managedSearchConfiguration": {
                    "filter": { ... },
                    "numberOfResults": number,
                    "rerankingConfiguration": {
                       "bedrockRerankingConfiguration": {
                          "metadataConfiguration": {
                             "selectionMode": "string",
                             "selectiveModeConfiguration": { ... }
                          },
                          "modelConfiguration": {
                             "additionalModelRequestFields": {
                                "string" : JSON value
                             },
                             "modelArn": "string"
                          },
                          "numberOfRerankedResults": number
                       },
                       "type": "string"
                    },
                    "rerankingModelType": "string"
                 },
                 "vectorSearchConfiguration": {
                    "filter": { ... },
                    "implicitFilterConfiguration": {
                       "metadataAttributes": [
                          {
                             "description": "string",
                             "key": "string",
                             "type": "string"
                          }
                       ],
```

```
                        "modelArn": "string"
                    },
                    "numberOfResults": number,
                    "overrideSearchType": "string",
                    "rerankingConfiguration": {
                        "bedrockRerankingConfiguration": {
                            "metadataConfiguration": {
                                "selectionMode": "string",
                                "selectiveModeConfiguration": { ... }
                            },
                            "modelConfiguration": {
                                "additionalModelRequestFields": {
                                    "string" : JSON value
                                },
                                "modelArn": "string"
                            },
                            "numberOfRerankedResults": number
                        },
                        "type": "string"
                    }
                }
            }
        ],
        "promptOverrideConfiguration": {
            "overrideLambda": "string",
            "promptConfigurations": [
                {
                    "additionalModelRequestFields": JSON value,
                    "basePromptTemplate": "string",
                    "foundationModel": "string",
                    "inferenceConfiguration": {
                        "maximumLength": number,
                        "stopSequences": [ "string" ],
                        "temperature": number,
                        "topK": number,
                        "topP": number
                    },
                    "parserMode": "string",
                    "promptCreationMode": "string",
                    "promptState": "string",
                    "promptType": "string"
                }
            ]
```

```
            }
        }
    ],
    "customerEncryptionKeyArn": "string",
    "customOrchestration": {
        "executor": { ... }
    },
    "enableTrace": boolean,
    "endSession": boolean,
    "foundationModel": "string",
    "guardrailConfiguration": {
        "guardrailIdentifier": "string",
        "guardrailVersion": "string"
    },
    "idleSessionTTLInSeconds": number,
    "inlineSessionState": {
        "conversationHistory": {
            "messages": [
                {
                    "content": [
                        { ... }
                    ],
                    "role": "string"
                }
            ]
        },
        "files": [
            {
                "name": "string",
                "source": {
                    "byteContent": {
                        "data": blob,
                        "mediaType": "string"
                    },
                    "s3Location": {
                        "uri": "string"
                    },
                    "sourceType": "string"
                },
                "useCase": "string"
            }
        ],
        "invocationId": "string",
        "promptSessionAttributes": {
```

```
            "string" : "string"
         },
         "returnControlInvocationResults": [
            { ... }
         ],
         "sessionAttributes": {
            "string" : "string"
         }
      },
      "inputText": "string",
      "instruction": "string",
      "knowledgeBases": [
         {
            "description": "string",
            "knowledgeBaseId": "string",
            "retrievalConfiguration": {
               "managedSearchConfiguration": {
                  "filter": { ... },
                  "numberOfResults": number,
                  "rerankingConfiguration": {
                     "bedrockRerankingConfiguration": {
                        "metadataConfiguration": {
                           "selectionMode": "string",
                           "selectiveModeConfiguration": { ... }
                        },
                        "modelConfiguration": {
                           "additionalModelRequestFields": {
                              "string" : JSON value
                           },
                           "modelArn": "string"
                        },
                        "numberOfRerankedResults": number
                     },
                     "type": "string"
                  },
                  "rerankingModelType": "string"
               },
               "vectorSearchConfiguration": {
                  "filter": { ... },
                  "implicitFilterConfiguration": {
                     "metadataAttributes": [
                        {
                           "description": "string",
                           "key": "string",
```

```
                        "type": "string"
                     }
                  ],
                  "modelArn": "string"
               },
               "numberOfResults": number,
               "overrideSearchType": "string",
               "rerankingConfiguration": {
                  "bedrockRerankingConfiguration": {
                     "metadataConfiguration": {
                        "selectionMode": "string",
                        "selectiveModeConfiguration": { ... }
                     },
                     "modelConfiguration": {
                        "additionalModelRequestFields": {
                           "string" : JSON value
                        },
                        "modelArn": "string"
                     },
                     "numberOfRerankedResults": number
                  },
                  "type": "string"
               }
            }
         }
      }
   ],
   "orchestrationType": "string",
   "promptCreationConfigurations": {
      "excludePreviousThinkingSteps": boolean,
      "previousConversationTurnsToInclude": number
   },
   "promptOverrideConfiguration": {
      "overrideLambda": "string",
      "promptConfigurations": [
         {
            "additionalModelRequestFields": JSON value,
            "basePromptTemplate": "string",
            "foundationModel": "string",
            "inferenceConfiguration": {
               "maximumLength": number,
               "stopSequences": [ "string" ],
               "temperature": number,
               "topK": number,
```

```
                "topP": number
            },
            "parserMode": "string",
            "promptCreationMode": "string",
            "promptState": "string",
            "promptType": "string"
         }
      ]
   },
   "streamingConfigurations": {
      "applyGuardrailInterval": number,
      "streamFinalResponse": boolean
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

sessionId

The unique identifier of the session. Use the same value across requests to continue the same conversation.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: Yes

Request Body

The request accepts the following data in JSON format.

actionGroups

A list of action groups with each action group defining the action the inline agent needs to carry out.

Type: Array of AgentActionGroup objects

Required: No

agentCollaboration

Defines how the inline collaborator agent handles information across multiple collaborator agents to coordinate a final response. The inline collaborator agent can also be the supervisor.

Type: String

Valid Values: `SUPERVISOR | SUPERVISOR_ROUTER | DISABLED`

Required: No

agentName

The name for the agent.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

bedrockModelConfigurations

Model settings for the request.

Type: InlineBedrockModelConfigurations object

Required: No

collaboratorConfigurations

Settings for an inline agent collaborator called with InvokeInlineAgent.

Type: Array of CollaboratorConfiguration objects

Required: No

collaborators

List of collaborator inline agents.

Type: Array of Collaborator objects

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key to use to encrypt your inline agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

customOrchestration

Contains details of the custom orchestration configured for the agent.

Type: CustomOrchestration object

Required: No

enableTrace

Specifies whether to turn on the trace or not to track the agent's reasoning process. For more information, see Using trace.

Type: Boolean

Required: No

endSession

Specifies whether to end the session with the inline agent or not.

Type: Boolean

Required: No

foundationModel

The model identifier (ID) of the model to use for orchestration by the inline agent. For example, `meta.llama3-1-70b-instruct-v1:0`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}`

```
([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}
[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|
(([0-9a-zA-Z][_-]?)+))$|(^arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|
[0-9a-z-]{1,20}):(|[0-9]{12}):inference-profile/[a-zA-Z0-9-:.]+)
```

Required: Yes

guardrailConfiguration

The guardrails to assign to the inline agent.

Type: GuardrailConfigurationWithArn object

Required: No

idleSessionTTLInSeconds

The number of seconds for which the inline agent should maintain session information. After this time expires, the subsequent `InvokeInlineAgent` request begins a new session.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and the data provided before the timeout is deleted.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

inlineSessionState

Parameters that specify the various attributes of a sessions. You can include attributes for the session or prompt or, if you configured an action group to return control, results from invocation of the action group. For more information, see Control session context.

> **Note**
>
> If you include `returnControlInvocationResults` in the `sessionState` field, the `inputText` field will be ignored.

Type: InlineSessionState object

Required: No

inputText

The prompt text to send to the agent.

> **Note**
>
> If you include `returnControlInvocationResults` in the `sessionState` field, the `inputText` field will be ignored.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: No

instruction

The instructions that tell the inline agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40.

Required: Yes

knowledgeBases

Contains information of the knowledge bases to associate with.

Type: Array of KnowledgeBase objects

Required: No

orchestrationType

Specifies the type of orchestration strategy for the agent. This is set to DEFAULT orchestration type, by default.

Type: String

Valid Values: `DEFAULT | CUSTOM_ORCHESTRATION`

Required: No

promptCreationConfigurations

Specifies parameters that control how the service populates the agent prompt for an `InvokeInlineAgent` request. You can control which aspects of previous invocations in the same agent session the service uses to populate the agent prompt. This gives you more granular control over the contextual history that is used to process the current request.

Type: PromptCreationConfigurations object

Required: No

promptOverrideConfiguration

Configurations for advanced prompts used to override the default prompts to enhance the accuracy of the inline agent.

Type: PromptOverrideConfiguration object

Required: No

streamingConfigurations

Specifies the configurations for streaming.

> **Note**
>
> To use agent streaming, you need permissions to perform the `bedrock:InvokeModelWithResponseStream` action.

Type: StreamingConfigurations object

Required: No

Response Syntax

```
HTTP/1.1 200
x-amzn-bedrock-agent-content-type: contentType
x-amz-bedrock-agent-session-id: sessionId
Content-type: application/json

{
   "accessDeniedException": {
```

```
    },
    "badGatewayException": {
    },
    "chunk": {
       "attribution": {
          "citations": [
             {
                "generatedResponsePart": {
                   "textResponsePart": {
                      "span": {
                         "end": number,
                         "start": number
                      },
                      "text": "string"
                   }
                },
                "retrievedReferences": [
                   {
                      "content": {
                         "audio": {
                            "s3Uri": "string",
                            "transcription": "string"
                         },
                         "byteContent": "string",
                         "row": [
                            {
                               "columnName": "string",
                               "columnValue": "string",
                               "type": "string"
                            }
                         ],
                         "text": "string",
                         "type": "string",
                         "video": {
                            "s3Uri": "string",
                            "summary": "string"
                         }
                      },
                      "location": {
                         "confluenceLocation": {
                            "url": "string"
                         },
                         "customDocumentLocation": {
                            "id": "string"
```

```
                        },
                        "googleDriveLocation": {
                           "url": "string"
                        },
                        "kendraDocumentLocation": {
                           "uri": "string"
                        },
                        "oneDriveLocation": {
                           "url": "string"
                        },
                        "s3Location": {
                           "uri": "string"
                        },
                        "salesforceLocation": {
                           "url": "string"
                        },
                        "sharePointLocation": {
                           "url": "string"
                        },
                        "sqlLocation": {
                           "query": "string"
                        },
                        "type": "string",
                        "webLocation": {
                           "url": "string"
                        }
                     },
                     "metadata": {
                        "string" : JSON value
                     }
                  }
               ]
            }
         ]
      },
      "bytes": blob
   },
   "conflictException": {
   },
   "dependencyFailedException": {
   },
   "files": {
      "files": [
         {
```

```
            "bytes": blob,
            "name": "string",
            "type": "string"
         }
      ]
   },
   "internalServerException": {
   },
   "resourceNotFoundException": {
   },
   "returnControl": {
      "invocationId": "string",
      "invocationInputs": [
         { ... }
      ]
   },
   "serviceQuotaExceededException": {
   },
   "throttlingException": {
   },
   "trace": {
      "callerChain": [
         { ... }
      ],
      "collaboratorName": "string",
      "eventTime": "string",
      "sessionId": "string",
      "trace": { ... }
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType

The MIME type of the input data in the request. The default value is application/json.

sessionId

The unique identifier of the session with the agent.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

The following data is returned in JSON format by the service.

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

chunk

Contains a part of an agent response and citations for it.

Type: InlineAgentPayloadPart object

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

files

Contains intermediate response for code interpreter if any files have been generated.

Type: InlineAgentFilePart object

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception

HTTP Status Code: 404

returnControl

Contains the parameters and information that the agent elicited from the customer to carry out an action. This information is returned to the system and can be used in your own setup for fulfilling the action.

Type: InlineAgentReturnControlPayload object

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception

HTTP Status Code: 400

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

trace

Contains information about the agent and session, alongside the agent's reasoning process and results from calling actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace events.

Type: InlineAgentTracePart object

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFlowExecutionEvents

Service: Agents for Amazon Bedrock Runtime

Lists events that occurred during a flow execution. Events provide detailed information about the execution progress, including node inputs and outputs, flow inputs and outputs, condition results, and failure events.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Request Syntax

```
GET /flows/flowIdentifier/aliases/flowAliasIdentifier/executions/executionIdentifier/
events?eventType=eventType&maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

eventType

The type of events to retrieve. Specify `Node` for node-level events or `Flow` for flow-level events.

Valid Values: `Node | Flow`

Required: Yes

executionIdentifier

The unique identifier of the flow execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

Required: Yes

flowAliasIdentifier

The unique identifier of the flow alias used for the execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

maxResults

The maximum number of events to return in a single response. If more events exist than the specified maxResults value, a token is included in the response so that the remaining results can be retrieved.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

A token to retrieve the next set of results. This value is returned in the response if more results are available.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "flowExecutionEvents": [
      { ... }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

flowExecutionEvents

A list of events that occurred during the flow execution. Events can include node inputs and outputs, flow inputs and outputs, condition results, and failure events.

Type: Array of FlowExecutionEvent objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

A token to retrieve the next set of results. This value is returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListFlowExecutions

Service: Agents for Amazon Bedrock Runtime

Lists all executions of a flow. Results can be paginated and include summary information about each execution, such as status, start and end times, and the execution's Amazon Resource Name (ARN).

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Request Syntax

```
GET /flows/flowIdentifier/executions?
flowAliasIdentifier=flowAliasIdentifier&maxResults=maxResults&nextToken=nextToken
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

flowAliasIdentifier

The unique identifier of the flow alias to list executions for.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

flowIdentifier

The unique identifier of the flow to list executions for.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

maxResults

The maximum number of flow executions to return in a single response. If more executions exist than the specified `maxResults` value, a token is included in the response so that the remaining results can be retrieved.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

A token to retrieve the next set of results. This value is returned in the response if more results are available.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "flowExecutionSummaries": [
      {
         "createdAt": "string",
         "endedAt": "string",
         "executionArn": "string",
         "flowAliasIdentifier": "string",
         "flowIdentifier": "string",
         "flowVersion": "string",
         "status": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

flowExecutionSummaries

A list of flow execution summaries. Each summary includes the execution ARN, flow identifier, flow alias identifier, flow version, status, and timestamps.

Type: Array of FlowExecutionSummary objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

nextToken

A token to retrieve the next set of results. This value is returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListInvocations

Service: Agents for Amazon Bedrock Runtime

Lists all invocations associated with a specific session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
POST /sessions/sessionIdentifier/invocations/?maxResults=maxResults&nextToken=nextToken
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sessionIdentifier

The unique identifier for the session to list invocations for. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "invocationSummaries": [
      {
         "createdAt": "string",
         "invocationId": "string",
         "sessionId": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationSummaries

A list of invocation summaries associated with the session.

Type: Array of InvocationSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListInvocations.

```
POST bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/invocations/?maxResults=10&nextToken=abc123 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListInvocationSteps

Service: Agents for Amazon Bedrock Runtime

Lists all invocation steps associated with a session and optionally, an invocation within the session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
POST /sessions/sessionIdentifier/invocationSteps/?
maxResults=maxResults&nextToken=nextToken HTTP/1.1
Content-type: application/json

{
   "invocationIdentifier": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

sessionIdentifier

The unique identifier for the session associated with the invocation steps. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

invocationIdentifier

The unique identifier (in UUID format) for the invocation to list invocation steps for.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "invocationStepSummaries": [
      {
         "invocationId": "string",
         "invocationStepId": "string",
         "invocationStepTime": "string",
         "sessionId": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationStepSummaries

A list of summaries for each invocation step associated with a session and if you specified it, an invocation within the session.

Type: Array of InvocationStepSummary objects

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListInvocationSteps.

```
POST bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/invocationSteps/?maxResults=10&nextToken=abc123 HTTP/1.1
Content-type: application/json

{
    "invocationIdentifier": "abc-1234-abcd-1234-abcdef123456"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListSessions

Service: Agents for Amazon Bedrock Runtime

Lists all sessions in your AWS account. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
POST /sessions/?maxResults=maxResults&nextToken=nextToken HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of results to return in the response. If the total number of results is greater than this value, use the token returned in the response in the `nextToken` field when making another request to return the next batch of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, enter the token returned in the `nextToken` field in the response in this field to return the next batch of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "nextToken": "string",
   "sessionSummaries": [
      {
         "createdAt": "string",
         "lastUpdatedAt": "string",
         "sessionArn": "string",
         "sessionId": "string",
         "sessionStatus": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

If the total number of results is greater than the `maxResults` value provided in the request, use this token when making another request in the `nextToken` field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sessionSummaries

A list of summaries for each session in your AWS account.

Type: Array of SessionSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListSessions.

```
POST bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/?
maxResults=10&nextToken=nextToken HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListTagsForResource

Service: Agents for Amazon Bedrock Runtime

List all the tags for the resource you specify.

Request Syntax

```
GET /tags/resourceArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource for which to list tags.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-zA-Z0-9-]+:[0-9]{12}:(session)/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}$).*`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "tags": {
      "string" : "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

The key-value pairs for the tags associated with the resource.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of ListTagsForResource.

```
GET /tags/arn:aws:bedrock:us-west-2:123456789012:session/12345abc-1234-abcd-1234-
abcdef123456 HTTP/1.1
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

OptimizePrompt

Service: Agents for Amazon Bedrock Runtime

Optimizes a prompt for the task that you specify. For more information, see Optimize a prompt in the Amazon Bedrock User Guide.

Request Syntax

```
POST /optimize-prompt HTTP/1.1
Content-type: application/json

{
   "input": { ... },
   "targetModelId": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

input

Contains the prompt to optimize.

Type: InputPrompt object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

targetModelId

The unique identifier of the model that you want to optimize the prompt for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-`

model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:provisioned-model/[a-z0-9]{12})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_-]?)+)

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "accessDeniedException": {
   },
   "analyzePromptEvent": {
      "message": "string"
   },
   "badGatewayException": {
   },
   "dependencyFailedException": {
   },
   "internalServerException": {
   },
   "optimizedPromptEvent": {
      "optimizedPrompt": { ... }
   },
   "throttlingException": {
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

analyzePromptEvent

An event in which the prompt was analyzed in preparation for optimization.

Type: AnalyzePromptEvent object

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

optimizedPromptEvent

An event in which the prompt was optimized.

Type: OptimizedPromptEvent object

throttlingException

Your request was throttled because of service-wide limitations. Resubmit your request later or in a different region. You can also purchase Provisioned Throughput to increase the rate or number of tokens you can process.

Type: Exception

HTTP Status Code: 429

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example

This example illustrates one usage of OptimizePrompt.

```
POST /optimize-prompt HTTP/1.1
Content-type: application/json

{
    "input": {
        "textPrompt": {
            "text": "Please summarize this text: "
        }
    },
    "targetModelId": "anthropic.claude-3-sonnet-20240229-v1:0"
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

PutInvocationStep

Service: Agents for Amazon Bedrock Runtime

Add an invocation step to an invocation in a session. An invocation step stores fine-grained state checkpoints, including text and images, for each interaction. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Related APIs:

- GetInvocationStep
- ListInvocationSteps
- ListInvocations
- ListSessions

Request Syntax

```
PUT /sessions/sessionIdentifier/invocationSteps/ HTTP/1.1
Content-type: application/json

{
   "invocationIdentifier": "string",
   "invocationStepId": "string",
   "invocationStepTime": "string",
   "payload": { ... }
}
```

URI Request Parameters

The request uses the following URI parameters.

sessionIdentifier

The unique identifier for the session to add the invocation step to. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

invocationIdentifier

The unique identifier (in UUID format) of the invocation to add the invocation step to.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

invocationStepId

The unique identifier of the invocation step in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: No

invocationStepTime

The timestamp for when the invocation step occurred.

Type: Timestamp

Required: Yes

payload

The payload for the invocation step, including text and images for the interaction.

Type: InvocationStepPayload object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

Response Syntax

```
HTTP/1.1 201
Content-type: application/json
```

```
{
   "invocationStepId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

invocationStepId

The unique identifier of the invocation step in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of PutInvocationStep.

```
PUT bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/invocationSteps/ HTTP/1.1
Content-type: application/json

{
    "invocationIdentifier": "11111111-2222-3333-4444-555555555555",
    "invocationStepId": "abc-1234-abcd-1234-abcdef123456",
    "invocationStepTime": "2023-08-08T12:00:00Z",
    "payload": {
        "contentBlocks": [
            {
                "text": "What's the weather in Seattle?"
```

```
            },
            {
                "image": {
                    "format": "png",
                    "source": {
                        "bytes": "base64_encoded_image_data"
                    }
                }
            }
        ]
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Rerank

Service: Agents for Amazon Bedrock Runtime

Reranks the relevance of sources based on queries. For more information, see Improve the relevance of query responses with a reranker model.

Request Syntax

```
POST /rerank HTTP/1.1
Content-type: application/json

{
   "nextToken": "string",
   "queries": [
      {
         "textQuery": {
            "text": "string"
         },
         "type": "string"
      }
   ],
   "rerankingConfiguration": {
      "bedrockRerankingConfiguration": {
         "modelConfiguration": {
            "additionalModelRequestFields": {
               "string" : JSON value
            },
            "modelArn": "string"
         },
         "numberOfResults": number
      },
      "type": "string"
   },
   "sources": [
      {
         "inlineDocumentSource": {
            "jsonDocument": JSON value,
            "textDocument": {
               "text": "string"
            },
            "type": "string"
         },
         "type": "string"
```

```
        }
    ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

nextToken

If the total number of results was greater than could fit in a response, a token is returned in the `nextToken` field. You can enter that token in this field to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

queries

An array of objects, each of which contains information about a query to submit to the reranker model.

Type: Array of RerankQuery objects

Array Members: Fixed number of 1 item.

Required: Yes

rerankingConfiguration

Contains configurations for reranking.

Type: RerankingConfiguration object

Required: Yes

sources

An array of objects, each of which contains information about the sources to rerank.

Type: Array of RerankSource objects

Array Members: Minimum number of 1 item. Maximum number of 1000 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "nextToken": "string",
   "results": [
      {
         "document": {
            "jsonDocument": JSON value,
            "textDocument": {
               "text": "string"
            },
            "type": "string"
         },
         "index": number,
         "relevanceScore": number
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

If the total number of results is greater than can fit in the response, use this token in the `nextToken` field when making another request to return the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

results

An array of objects, each of which contains information about the results of reranking.

Type: Array of RerankResult objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4

- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Retrieve

Service: Agents for Amazon Bedrock Runtime

Queries a knowledge base and retrieves information from it.

Request Syntax

```
POST /knowledgebases/knowledgeBaseId/retrieve HTTP/1.1
Content-type: application/json

{
   "guardrailConfiguration": {
      "guardrailId": "string",
      "guardrailVersion": "string"
   },
   "nextToken": "string",
   "retrievalConfiguration": {
      "managedSearchConfiguration": {
         "filter": { ... },
         "numberOfResults": number,
         "rerankingConfiguration": {
            "bedrockRerankingConfiguration": {
               "metadataConfiguration": {
                  "selectionMode": "string",
                  "selectiveModeConfiguration": { ... }
               },
               "modelConfiguration": {
                  "additionalModelRequestFields": {
                     "string" : JSON value
                  },
                  "modelArn": "string"
               },
               "numberOfRerankedResults": number
            },
            "type": "string"
         },
         "rerankingModelType": "string"
      },
      "vectorSearchConfiguration": {
         "filter": { ... },
         "implicitFilterConfiguration": {
            "metadataAttributes": [
               {
```

```
                    "description": "string",
                    "key": "string",
                    "type": "string"
                 }
              ],
              "modelArn": "string"
           },
           "numberOfResults": number,
           "overrideSearchType": "string",
           "rerankingConfiguration": {
              "bedrockRerankingConfiguration": {
                 "metadataConfiguration": {
                    "selectionMode": "string",
                    "selectiveModeConfiguration": { ... }
                 },
                 "modelConfiguration": {
                    "additionalModelRequestFields": {
                       "string" : JSON value
                    },
                    "modelArn": "string"
                 },
                 "numberOfRerankedResults": number
              },
              "type": "string"
           }
        }
     },
     "retrievalQuery": {
        "image": {
           "format": "string",
           "inlineContent": blob
        },
        "text": "string",
        "type": "string"
     },
     "userContext": {
        "userId": "string"
     }
}
```

URI Request Parameters

The request uses the following URI parameters.

knowledgeBaseId

The unique identifier of the knowledge base to query.

Length Constraints: Minimum length of 10. Maximum length of 2048.

Pattern: `[0-9a-zA-Z]{10}$|^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:knowledge-base/[0-9a-zA-Z]{10}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

guardrailConfiguration

Guardrail settings.

Type: GuardrailConfiguration object

Required: No

nextToken

If there are more results than can fit in the response, the response returns a `nextToken`. Use this token in the `nextToken` field of another request to retrieve the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

retrievalConfiguration

Contains configurations for the knowledge base query and retrieval process. For more information, see Query configurations.

Type: KnowledgeBaseRetrievalConfiguration object

Required: No

retrievalQuery

Contains the query to send the knowledge base.

Type: KnowledgeBaseQuery object

Required: Yes

userContext

Contains information about the user making the request. This is used for access control filtering to ensure that retrieval results only include documents the user is authorized to access.

Type: UserContext object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "guardrailAction": "string",
   "nextToken": "string",
   "retrievalResults": [
      {
         "content": {
            "audio": {
               "s3Uri": "string",
               "transcription": "string"
            },
            "byteContent": "string",
            "row": [
               {
                  "columnName": "string",
                  "columnValue": "string",
                  "type": "string"
               }
            ],
            "text": "string",
            "type": "string",
            "video": {
               "s3Uri": "string",
```

```
            "summary": "string"
         }
      },
      "documentId": "string",
      "location": {
         "confluenceLocation": {
            "url": "string"
         },
         "customDocumentLocation": {
            "id": "string"
         },
         "googleDriveLocation": {
            "url": "string"
         },
         "kendraDocumentLocation": {
            "uri": "string"
         },
         "oneDriveLocation": {
            "url": "string"
         },
         "s3Location": {
            "uri": "string"
         },
         "salesforceLocation": {
            "url": "string"
         },
         "sharePointLocation": {
            "url": "string"
         },
         "sqlLocation": {
            "query": "string"
         },
         "type": "string",
         "webLocation": {
            "url": "string"
         }
      },
      "metadata": {
         "string" : JSON value
      },
      "score": number
   }
]
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

guardrailAction

Specifies if there is a guardrail intervention in the response.

Type: String

Valid Values: `INTERVENED | NONE`

nextToken

If there are more results than can fit in the response, the response returns a `nextToken`. Use this token in the `nextToken` field of another request to retrieve the next batch of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

retrievalResults

A list of results from querying the knowledge base.

Type: Array of KnowledgeBaseRetrievalResult objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Send a basic query

The following example queries a knowledge base.

Sample Request

```
POST /knowledgebases/KB12345678/retrieve HTTP/1.1
Content-type: application/json

{
    "retrievalQuery": {
        "text": "What is AWS?"
    }
}
```

Send a query and include filters

To include filters in a knowledge base query, at least one of the data source files must include a `.metadata.json` file. For example, if you had a data source of articles called `articles.pdf`, accompanied by a metadata file called `articles.pdf.metadata.json`, you could tag it for `genre`, `year`, and `author`. In the `Retrieve` request, you could apply the following filter to return all entertainment articles written after 2018, in addition to `cooking` or `sports` articles written by authors starting with C.

Sample Request

```
POST /knowledgebases/KB12345678/retrieve HTTP/1.1
Content-type: application/json

{
    "retrievalQuery": {
        "text": "What is AWS?",
    },
    "vectorSearchConfiguration": {
        "numberOfResults": 5,
        "filter": {
            "orAll": [
                {
                    "andAll": [
                        {
                            "equals": {
                                "key": "genre",
                                "value": "entertainment"
                            }
                        },
                        {
                            "greaterThan": {
                                "key": "year",
                                "value": 2018
                            }
                        }
                    ]
                },
                {
                    "andAll": [
                        {
                            "in": {
                                "key": "genre",
                                "value": ["cooking", "sports"]
                            }
                        },
                        {
                            "startsWith": {
                                "key": "author",
                                "value": "C"
                            }
                        }
                    ]
```

```
                }
            ]
        }
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

RetrieveAndGenerate

Service: Agents for Amazon Bedrock Runtime

Queries a knowledge base and generates responses based on the retrieved results and using the specified foundation model or inference profile. The response only cites sources that are relevant to the query.

> **Note**
>
> This API cannot be used with managed knowledge bases. Use AgenticRetrieveStream or Retrieve with managed knowledge bases.

Request Syntax

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json

{
   "input": {
      "text": "string"
   },
   "retrieveAndGenerateConfiguration": {
      "externalSourcesConfiguration": {
         "generationConfiguration": {
            "additionalModelRequestFields": {
               "string" : JSON value
            },
            "guardrailConfiguration": {
               "guardrailId": "string",
               "guardrailVersion": "string"
            },
            "inferenceConfig": {
               "textInferenceConfig": {
                  "maxTokens": number,
                  "stopSequences": [ "string" ],
                  "temperature": number,
                  "topP": number
               }
            },
            "performanceConfig": {
               "latency": "string"
```

```
            },
            "promptTemplate": {
               "textPromptTemplate": "string"
            }
         },
         "modelArn": "string",
         "sources": [
            {
               "byteContent": {
                  "contentType": "string",
                  "data": blob,
                  "identifier": "string"
               },
               "s3Location": {
                  "uri": "string"
               },
               "sourceType": "string"
            }
         ]
      },
      "knowledgeBaseConfiguration": {
         "generationConfiguration": {
            "additionalModelRequestFields": {
               "string" : JSON value
            },
            "guardrailConfiguration": {
               "guardrailId": "string",
               "guardrailVersion": "string"
            },
            "inferenceConfig": {
               "textInferenceConfig": {
                  "maxTokens": number,
                  "stopSequences": [ "string" ],
                  "temperature": number,
                  "topP": number
               }
            },
            "performanceConfig": {
               "latency": "string"
            },
            "promptTemplate": {
               "textPromptTemplate": "string"
            }
         },
```

```
        "knowledgeBaseId": "string",
        "modelArn": "string",
        "orchestrationConfiguration": {
           "additionalModelRequestFields": {
              "string" : JSON value
           },
           "inferenceConfig": {
              "textInferenceConfig": {
                 "maxTokens": number,
                 "stopSequences": [ "string" ],
                 "temperature": number,
                 "topP": number
              }
           },
           "performanceConfig": {
              "latency": "string"
           },
           "promptTemplate": {
              "textPromptTemplate": "string"
           },
           "queryTransformationConfiguration": {
              "type": "string"
           }
        },
        "retrievalConfiguration": {
           "managedSearchConfiguration": {
              "filter": { ... },
              "numberOfResults": number,
              "rerankingConfiguration": {
                 "bedrockRerankingConfiguration": {
                    "metadataConfiguration": {
                       "selectionMode": "string",
                       "selectiveModeConfiguration": { ... }
                    },
                    "modelConfiguration": {
                       "additionalModelRequestFields": {
                          "string" : JSON value
                       },
                       "modelArn": "string"
                    },
                    "numberOfRerankedResults": number
                 },
                 "type": "string"
              },
```

```
                "rerankingModelType": "string"
            },
            "vectorSearchConfiguration": {
               "filter": { ... },
               "implicitFilterConfiguration": {
                  "metadataAttributes": [
                     {
                        "description": "string",
                        "key": "string",
                        "type": "string"
                     }
                  ],
                  "modelArn": "string"
               },
               "numberOfResults": number,
               "overrideSearchType": "string",
               "rerankingConfiguration": {
                  "bedrockRerankingConfiguration": {
                     "metadataConfiguration": {
                        "selectionMode": "string",
                        "selectiveModeConfiguration": { ... }
                     },
                     "modelConfiguration": {
                        "additionalModelRequestFields": {
                           "string" : JSON value
                        },
                        "modelArn": "string"
                     },
                     "numberOfRerankedResults": number
                  },
                  "type": "string"
               }
            }
         }
      },
      "type": "string"
   },
   "sessionConfiguration": {
      "kmsKeyArn": "string"
   },
   "sessionId": "string",
   "userContext": {
      "userId": "string"
   }
```

```
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

input

Contains the query to be made to the knowledge base.

Type: RetrieveAndGenerateInput object

Required: Yes

retrieveAndGenerateConfiguration

Contains configurations for the knowledge base query and retrieval process. For more information, see Query configurations.

Type: RetrieveAndGenerateConfiguration object

Required: No

sessionConfiguration

Contains details about the session with the knowledge base.

Type: RetrieveAndGenerateSessionConfiguration object

Required: No

sessionId

The unique identifier of the session. When you first make a `RetrieveAndGenerate` request, Amazon Bedrock automatically generates this value. You must reuse this value for all subsequent requests in the same conversational session. This value allows Amazon Bedrock to maintain context and knowledge from previous interactions. You can't explicitly set the `sessionId` yourself.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

userContext

Contains information about the user making the request. This is used for access control filtering to ensure that retrieval results only include documents the user is authorized to access.

Type: UserContext object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "citations": [
      {
         "generatedResponsePart": {
            "textResponsePart": {
               "span": {
                  "end": number,
                  "start": number
               },
               "text": "string"
            }
         },
         "retrievedReferences": [
            {
               "content": {
                  "audio": {
                     "s3Uri": "string",
                     "transcription": "string"
                  },
                  "byteContent": "string",
                  "row": [
                     {
                        "columnName": "string",
                        "columnValue": "string",
                        "type": "string"
                     }
```

```
            ],
            "text": "string",
            "type": "string",
            "video": {
               "s3Uri": "string",
               "summary": "string"
            }
         },
         "location": {
            "confluenceLocation": {
               "url": "string"
            },
            "customDocumentLocation": {
               "id": "string"
            },
            "googleDriveLocation": {
               "url": "string"
            },
            "kendraDocumentLocation": {
               "uri": "string"
            },
            "oneDriveLocation": {
               "url": "string"
            },
            "s3Location": {
               "uri": "string"
            },
            "salesforceLocation": {
               "url": "string"
            },
            "sharePointLocation": {
               "url": "string"
            },
            "sqlLocation": {
               "query": "string"
            },
            "type": "string",
            "webLocation": {
               "url": "string"
            }
         },
         "metadata": {
            "string" : JSON value
         }
```

```
            }
         ]
      }
   ],
   "guardrailAction": "string",
   "output": {
      "text": "string"
   },
   "sessionId": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

citations

A list of segments of the generated response that are based on sources in the knowledge base, alongside information about the sources.

Type: Array of Citation objects

guardrailAction

Specifies if there is a guardrail intervention in the response.

Type: String

Valid Values: `INTERVENED | NONE`

output

Contains the response generated from querying the knowledge base.

Type: RetrieveAndGenerateOutput object

sessionId

The unique identifier of the session. When you first make a `RetrieveAndGenerate` request, Amazon Bedrock automatically generates this value. You must reuse this value for all subsequent requests in the same conversational session. This value allows Amazon Bedrock to maintain context and knowledge from previous interactions. You can't explicitly set the `sessionId` yourself.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Send a basic query

The following example uses the minimally required fields to generate a response after querying a knowledge base.

Sample Request

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json
```

```
{
    "input": {
        "text": "What is AWS?"
    },
    "retrieveAndGenerateConfiguration": {
        "knowledgeBaseConfiguration": {
            "knowledgeBaseId": "KB12345678",
            "modelArn": "anthropic.claude-v2:1"
        },
        "type": "KNOWLEDGE_BASE"
    }
}
```

Send a query and include filters

To include filters in a knowledge base query, at least one of the data source files must include a `.metadata.json` file. For example, if you had a data source of articles called `articles.pdf`, accompanied by a metadata file called `articles.pdf.metadata.json`, you could tag it for `genre`, `year`, and `author`. In the `Retrieve` request, you could apply the following filter to return all entertainment articles written after 2018, in addition to `cooking` or `sports` articles written by authors starting with C.

Sample Request

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json

{
    "input": {
        "text": "What is AWS?",
    },
    "retrieveAndGenerateConfiguration": {
        "knowledgeBaseConfiguration": {
            "knowledgeBaseId": "KB12345678",
            "modelArn": "anthropic.claude-v2:1",
            "retrievalConfiguration": {
                "vectorSearchConfiguration": {
                    "numberOfResults": 5,
                    "filter": {
                        "orAll": [
                            {
                                "andAll": [
```

```
                                    {
                                        "equals": {
                                            "key": "genre",
                                            "value": "entertainment"
                                        }
                                    },
                                    {
                                        "greaterThan": {
                                            "key": "year",
                                            "value": 2018
                                        }
                                    }
                                ]
                            },
                            {
                                "andAll": [
                                    {
                                        "in": {
                                            "key": "genre",
                                            "value": ["cooking", "sports"]
                                        }
                                    },
                                    {
                                        "startsWith": {
                                            "key": "author",
                                            "value": "C"
                                        }
                                    }
                                ]
                            }
                        ]
                    }
                }
            }
        },
        "type": "KNOWLEDGE_BASE"
    }
}
```

Use an inference profile when generating a response

The following example uses an inference profile in generating a response after querying a knowledge base.

Sample Request

```
POST /retrieveAndGenerate HTTP/1.1
Content-type: application/json

{
    "input": {
        "text": "What is AWS?"
    },
    "retrieveAndGenerateConfiguration": {
        "knowledgeBaseConfiguration": {
            "knowledgeBaseId": "KB12345678",
            "modelArn": "arn:aws:bedrock:us-west-2:123456789012:inference-profile/
us.anthropic.claude-3-5-sonnet-20240620-v1:0"
        },
        "type": "KNOWLEDGE_BASE"
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

RetrieveAndGenerateStream

Service: Agents for Amazon Bedrock Runtime

Queries a knowledge base and generates responses based on the retrieved results, with output in streaming format.

> **Note**
>
> This API cannot be used with managed knowledge bases. Use AgenticRetrieveStream or Retrieve with managed knowledge bases.

> **Note**
>
> The AWS CLI doesn't support streaming operations in Amazon Bedrock, including `InvokeModelWithResponseStream`.

This operation requires permission for the `bedrock:RetrieveAndGenerate` action.

Request Syntax

```
POST /retrieveAndGenerateStream HTTP/1.1
Content-type: application/json

{
   "input": {
      "text": "string"
   },
   "retrieveAndGenerateConfiguration": {
      "externalSourcesConfiguration": {
         "generationConfiguration": {
            "additionalModelRequestFields": {
               "string" : JSON value
            },
            "guardrailConfiguration": {
               "guardrailId": "string",
               "guardrailVersion": "string"
            },
            "inferenceConfig": {
               "textInferenceConfig": {
```

```
                "maxTokens": number,
                "stopSequences": [ "string" ],
                "temperature": number,
                "topP": number
             }
          },
          "performanceConfig": {
             "latency": "string"
          },
          "promptTemplate": {
             "textPromptTemplate": "string"
          }
       },
       "modelArn": "string",
       "sources": [
          {
             "byteContent": {
                "contentType": "string",
                "data": blob,
                "identifier": "string"
             },
             "s3Location": {
                "uri": "string"
             },
             "sourceType": "string"
          }
       ]
    },
    "knowledgeBaseConfiguration": {
       "generationConfiguration": {
          "additionalModelRequestFields": {
             "string" : JSON value
          },
          "guardrailConfiguration": {
             "guardrailId": "string",
             "guardrailVersion": "string"
          },
          "inferenceConfig": {
             "textInferenceConfig": {
                "maxTokens": number,
                "stopSequences": [ "string" ],
                "temperature": number,
                "topP": number
             }
```

```
            },
            "performanceConfig": {
               "latency": "string"
            },
            "promptTemplate": {
               "textPromptTemplate": "string"
            }
         },
         "knowledgeBaseId": "string",
         "modelArn": "string",
         "orchestrationConfiguration": {
            "additionalModelRequestFields": {
               "string" : JSON value
            },
            "inferenceConfig": {
               "textInferenceConfig": {
                  "maxTokens": number,
                  "stopSequences": [ "string" ],
                  "temperature": number,
                  "topP": number
               }
            },
            "performanceConfig": {
               "latency": "string"
            },
            "promptTemplate": {
               "textPromptTemplate": "string"
            },
            "queryTransformationConfiguration": {
               "type": "string"
            }
         },
         "retrievalConfiguration": {
            "managedSearchConfiguration": {
               "filter": { ... },
               "numberOfResults": number,
               "rerankingConfiguration": {
                  "bedrockRerankingConfiguration": {
                     "metadataConfiguration": {
                        "selectionMode": "string",
                        "selectiveModeConfiguration": { ... }
                     },
                     "modelConfiguration": {
                        "additionalModelRequestFields": {
```

```
                        "string" : JSON value
                     },
                     "modelArn": "string"
                  },
                  "numberOfRerankedResults": number
               },
               "type": "string"
            },
            "rerankingModelType": "string"
         },
         "vectorSearchConfiguration": {
            "filter": { ... },
            "implicitFilterConfiguration": {
               "metadataAttributes": [
                  {
                     "description": "string",
                     "key": "string",
                     "type": "string"
                  }
               ],
               "modelArn": "string"
            },
            "numberOfResults": number,
            "overrideSearchType": "string",
            "rerankingConfiguration": {
               "bedrockRerankingConfiguration": {
                  "metadataConfiguration": {
                     "selectionMode": "string",
                     "selectiveModeConfiguration": { ... }
                  },
                  "modelConfiguration": {
                     "additionalModelRequestFields": {
                        "string" : JSON value
                     },
                     "modelArn": "string"
                  },
                  "numberOfRerankedResults": number
               },
               "type": "string"
            }
         }
      },
      "type": "string"
```

```
    },
    "sessionConfiguration": {
       "kmsKeyArn": "string"
    },
    "sessionId": "string",
    "userContext": {
       "userId": "string"
    }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

input

Contains the query to be made to the knowledge base.

Type: RetrieveAndGenerateInput object

Required: Yes

retrieveAndGenerateConfiguration

Contains configurations for the knowledge base query and retrieval process. For more information, see Query configurations.

Type: RetrieveAndGenerateConfiguration object

Required: No

sessionConfiguration

Contains details about the session with the knowledge base.

Type: RetrieveAndGenerateSessionConfiguration object

Required: No

sessionId

The unique identifier of the session. When you first make a `RetrieveAndGenerate` request, Amazon Bedrock automatically generates this value. You must reuse this value for all subsequent requests in the same conversational session. This value allows Amazon Bedrock to maintain context and knowledge from previous interactions. You can't explicitly set the `sessionId` yourself.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

userContext

Contains information about the user making the request. This is used for access control filtering to ensure that retrieval results only include documents the user is authorized to access.

Type: UserContext object

Required: No

Response Syntax

```
HTTP/1.1 200
x-amzn-bedrock-knowledge-base-session-id: sessionId
Content-type: application/json

{
   "accessDeniedException": {
   },
   "badGatewayException": {
   },
   "citation": {
      "citation": {
         "generatedResponsePart": {
            "textResponsePart": {
               "span": {
                  "end": number,
                  "start": number
```

```
            },
            "text": "string"
         }
      },
      "retrievedReferences": [
         {
            "content": {
               "audio": {
                  "s3Uri": "string",
                  "transcription": "string"
               },
               "byteContent": "string",
               "row": [
                  {
                     "columnName": "string",
                     "columnValue": "string",
                     "type": "string"
                  }
               ],
               "text": "string",
               "type": "string",
               "video": {
                  "s3Uri": "string",
                  "summary": "string"
               }
            },
            "location": {
               "confluenceLocation": {
                  "url": "string"
               },
               "customDocumentLocation": {
                  "id": "string"
               },
               "googleDriveLocation": {
                  "url": "string"
               },
               "kendraDocumentLocation": {
                  "uri": "string"
               },
               "oneDriveLocation": {
                  "url": "string"
               },
               "s3Location": {
                  "uri": "string"
```

```
                },
                "salesforceLocation": {
                   "url": "string"
                },
                "sharePointLocation": {
                   "url": "string"
                },
                "sqlLocation": {
                   "query": "string"
                },
                "type": "string",
                "webLocation": {
                   "url": "string"
                }
             },
             "metadata": {
                "string" : JSON value
             }
          }
       ]
    },
    "generatedResponsePart": {
       "textResponsePart": {
          "span": {
             "end": number,
             "start": number
          },
          "text": "string"
       }
    },
    "retrievedReferences": [
       {
          "content": {
             "audio": {
                "s3Uri": "string",
                "transcription": "string"
             },
             "byteContent": "string",
             "row": [
                {
                   "columnName": "string",
                   "columnValue": "string",
                   "type": "string"
                }
```

```
            ],
            "text": "string",
            "type": "string",
            "video": {
               "s3Uri": "string",
               "summary": "string"
            }
         },
         "location": {
            "confluenceLocation": {
               "url": "string"
            },
            "customDocumentLocation": {
               "id": "string"
            },
            "googleDriveLocation": {
               "url": "string"
            },
            "kendraDocumentLocation": {
               "uri": "string"
            },
            "oneDriveLocation": {
               "url": "string"
            },
            "s3Location": {
               "uri": "string"
            },
            "salesforceLocation": {
               "url": "string"
            },
            "sharePointLocation": {
               "url": "string"
            },
            "sqlLocation": {
               "query": "string"
            },
            "type": "string",
            "webLocation": {
               "url": "string"
            }
         },
         "metadata": {
            "string" : JSON value
         }
```

```
            }
         ]
      },
      "conflictException": {
      },
      "dependencyFailedException": {
      },
      "guardrail": {
         "action": "string"
      },
      "internalServerException": {
      },
      "output": {
         "text": "string"
      },
      "resourceNotFoundException": {
      },
      "serviceQuotaExceededException": {
      },
      "throttlingException": {
      },
      "validationException": {
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

sessionId

The session ID.

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

The following data is returned in JSON format by the service.

accessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 403

badGatewayException

The request failed due to a bad gateway error.

Type: Exception

HTTP Status Code: 502

citation

A citation event.

Type: CitationEvent object

conflictException

Error occurred because of a conflict while performing an operation.

Type: Exception

HTTP Status Code: 409

dependencyFailedException

The request failed due to a dependency error.

Type: Exception

HTTP Status Code: 424

guardrail

A guardrail event.

Type: GuardrailEvent object

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

output

An output event.

Type: RetrieveAndGenerateOutputEvent object

resourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 404

serviceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

Type: Exception

HTTP Status Code: 400

throttlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 429

validationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2

- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StartFlowExecution

Service: Agents for Amazon Bedrock Runtime

Starts an execution of an Amazon Bedrock flow. Unlike flows that run until completion or time out after five minutes, flow executions let you run flows asynchronously for longer durations. Flow executions also yield control so that your application can perform other tasks.

This operation returns an Amazon Resource Name (ARN) that you can use to track and manage your flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Request Syntax

```
POST /flows/flowIdentifier/aliases/flowAliasIdentifier/executions HTTP/1.1
Content-type: application/json

{
   "flowExecutionName": "string",
   "inputs": [
      {
         "content": { ... },
         "nodeInputName": "string",
         "nodeName": "string",
         "nodeOutputName": "string"
      }
   ],
   "modelPerformanceConfiguration": {
      "performanceConfig": {
         "latency": "string"
      }
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

flowAliasIdentifier

The unique identifier of the flow alias to use for the flow execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow to execute.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

flowExecutionName

The unique name for the flow execution. If you don't provide one, a system-generated name is used.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-zA-Z0-9-]{1,36}`

Required: No

inputs

The input data required for the flow execution. This must match the input schema defined in the flow.

Type: Array of FlowInput objects

Array Members: Fixed number of 1 item.

Required: Yes

modelPerformanceConfiguration

The performance settings for the foundation model used in the flow execution.

Type: ModelPerformanceConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "executionArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

executionArn

The Amazon Resource Name (ARN) that uniquely identifies the flow execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StopFlowExecution

Service: Agents for Amazon Bedrock Runtime

Stops an Amazon Bedrock flow's execution. This operation prevents further processing of the flow and changes the execution status to `Aborted`.

Request Syntax

```
POST /flows/flowIdentifier/aliases/flowAliasIdentifier/executions/executionIdentifier/
stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

executionIdentifier

The unique identifier of the flow execution to stop.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

Required: Yes

flowAliasIdentifier

The unique identifier of the flow alias used for the execution.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "executionArn": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

executionArn

The Amazon Resource Name (ARN) that uniquely identifies the flow execution that was stopped.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

status

The updated status of the flow execution after the stop request. This will typically be ABORTED if the execution was successfully stopped.

Type: String

Valid Values: `Running | Succeeded | Failed | TimedOut | Aborted`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

BadGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 502

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

DependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

resourceName

The name of the dependency that caused the issue, such as Amazon Bedrock, Lambda, or AWS STS.

HTTP Status Code: 424

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3

- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

TagResource

Service: Agents for Amazon Bedrock Runtime

Associate tags with a resource. For more information, see Tagging resources in the Amazon Bedrock User Guide.

Request Syntax

```
POST /tags/resourceArn HTTP/1.1
Content-type: application/json

{
   "tags": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource to tag.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-zA-Z0-9-]+:[0-9]{12}:(session)/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}$).*`

Required: Yes

Request Body

The request accepts the following data in JSON format.

tags

An object containing key-value pairs that define the tags to attach to the resource.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Key Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of TagResource.

```
POST /tags/arn:aws:bedrock:us-west-2:123456789012:session/12345abc-1234-abcd-1234-
abcdef123456 HTTP/1.1
Content-type: application/json

{
   "tags": {
      "cost-center" : "Tech"
   }
}
```

Example response

This example illustrates one usage of TagResource.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UntagResource

Service: Agents for Amazon Bedrock Runtime

Remove tags from a resource.

Request Syntax

```
DELETE /tags/resourceArn?tagKeys=tagKeys HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

resourceArn

The Amazon Resource Name (ARN) of the resource from which to remove tags.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-zA-Z0-9-]+:[0-9]{12}:(session)/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}$).*`

Required: Yes

tagKeys

A list of keys of the tags to remove from the resource.

Array Members: Minimum number of 1 item. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UntagResource.

```
DELETE /tags/arn:aws:bedrock:us-west-2:123456789012:session/12345abc-1234-abcd-1234-
abcdef123456 HTTP/1.1
```

Example response

This example illustrates one usage of UntagResource.

```
HTTP/1.1 200
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateSession

Service: Agents for Amazon Bedrock Runtime

Updates the metadata or encryption settings of a session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Request Syntax

```
PUT /sessions/sessionIdentifier/ HTTP/1.1
Content-type: application/json

{
   "sessionMetadata": {
      "string" : "string"
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

sessionIdentifier

The unique identifier of the session to modify. You can specify either the session's `sessionId` or its Amazon Resource Name (ARN).

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})|([a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

sessionMetadata

A map of key-value pairs containing attributes to be persisted across the session. For example the user's ID, their language preference, and the type of device they are using.

Type: String to string map

Map Entries: Minimum number of 0 items. Maximum number of 50 items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 5000.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "createdAt": "string",
   "lastUpdatedAt": "string",
   "sessionArn": "string",
   "sessionId": "string",
   "sessionStatus": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

createdAt

The timestamp for when the session was created.

Type: Timestamp

lastUpdatedAt

The timestamp for when the session was last modified.

Type: Timestamp

sessionArn

The Amazon Resource Name (ARN) of the session that was updated.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionId

The unique identifier of the session you updated.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

sessionStatus

The status of the session you updated.

Type: String

Valid Values: `ACTIVE | EXPIRED | ENDED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

HTTP Status Code: 403

ConflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

HTTP Status Code: 409

InternalServerException

An internal server error occurred. Retry your request.

reason

The reason for the exception. If the reason is `BEDROCK_MODEL_INVOCATION_SERVICE_UNAVAILABLE`, the model invocation service is unavailable. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Example request

This example illustrates one usage of UpdateSession.

```
PUT bedrock-agent-runtime.us-east-1.amazonaws.com/sessions/12345abc-1234-abcd-1234-
abcdef123456/ HTTP/1.1
Content-type: application/json

{
   "sessionMetadata": {
      "userLanguage" : "French"
   }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Data Automation for Amazon Bedrock

The following actions are supported by Data Automation for Amazon Bedrock:

- CopyBlueprintStage
- CreateBlueprint
- CreateBlueprintVersion
- CreateDataAutomationLibrary
- CreateDataAutomationProject
- DeleteBlueprint
- DeleteDataAutomationLibrary
- DeleteDataAutomationProject
- GetBlueprint
- GetBlueprintOptimizationStatus
- GetDataAutomationLibrary
- GetDataAutomationLibraryEntity
- GetDataAutomationLibraryIngestionJob
- GetDataAutomationProject
- InvokeBlueprintOptimizationAsync
- InvokeDataAutomationLibraryIngestionJob
- ListBlueprints

- ListDataAutomationLibraries
- ListDataAutomationLibraryEntities
- ListDataAutomationLibraryIngestionJobs
- ListDataAutomationProjects
- ListTagsForResource
- TagResource
- UntagResource
- UpdateBlueprint
- UpdateDataAutomationLibrary
- UpdateDataAutomationProject

CopyBlueprintStage

Service: Data Automation for Amazon Bedrock

Copies a blueprint from one stage to another.

Request Syntax

```
PUT /blueprints/blueprintArn/copy-stage HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "sourceStage": "string",
   "targetStage": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

blueprintArn

The blueprint's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

To avoid creating duplicate resources, specify a unique token for this operation.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

sourceStage

The stage to copy from.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: Yes

targetStage

The stage to copy to.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateBlueprint

Service: Data Automation for Amazon Bedrock

Creates a blueprint. A blueprint has a schema that maps data fields to prompts that tell a model how to get the data out of an image or document. For schema examples, use ListBlueprints to get a list of public blueprints and GetBlueprint to view a blueprint.

Request Syntax

```
PUT /blueprints/ HTTP/1.1
Content-type: application/json

{
   "blueprintName": "string",
   "blueprintStage": "string",
   "clientToken": "string",
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "schema": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "type": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

blueprintName

A name for the blueprint.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

blueprintStage

To indicate that the blueprint is not final, set to `DEVELOPMENT`. Otherwise, the blueprint is created in the `LIVE` stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

clientToken

To avoid creating duplicate resources, specify a unique token for this blueprint.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

encryptionConfiguration

Encryption settings for the blueprint.

Type: EncryptionConfiguration object

Required: No

schema

The blueprint's schema.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: Yes

tags

Tags assigned to a blueprint at time of creation.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

type

The type of blueprint to create.

Type: String

Valid Values: `DOCUMENT | IMAGE | AUDIO`

Required: Yes

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "blueprint": {
      "blueprintArn": "string",
      "blueprintName": "string",
      "blueprintStage": "string",
      "blueprintVersion": "string",
      "creationTime": "string",
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string",
      "lastModifiedTime": "string",
      "optimizationSamples": [
         {
            "assetS3Object": {
               "s3Uri": "string",
               "version": "string"
```

```
            },
            "groundTruthS3Object": {
               "s3Uri": "string",
               "version": "string"
            }
         }
      ],
      "optimizationTime": "string",
      "schema": "string",
      "type": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

blueprint

Details about the blueprint.

Type: Blueprint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateBlueprintVersion

Service: Data Automation for Amazon Bedrock

Creates an immutable snapshot of a blueprint.

Request Syntax

```
POST /blueprints/blueprintArn/versions/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

blueprintArn

The blueprint's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

To avoid creating duplicate resources, specify a unique token for this blueprint version.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "blueprint": {
      "blueprintArn": "string",
      "blueprintName": "string",
      "blueprintStage": "string",
      "blueprintVersion": "string",
      "creationTime": "string",
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string",
      "lastModifiedTime": "string",
      "optimizationSamples": [
         {
            "assetS3Object": {
               "s3Uri": "string",
               "version": "string"
            },
            "groundTruthS3Object": {
               "s3Uri": "string",
               "version": "string"
            }
         }
      ],
      "optimizationTime": "string",
      "schema": "string",
      "type": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

blueprint

Details about the blueprint.

Type: Blueprint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateDataAutomationLibrary

Service: Data Automation for Amazon Bedrock

Creates a data automation library.

Request Syntax

```
PUT /data-automation-libraries/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "libraryDescription": "string",
   "libraryName": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

To avoid creating duplicate resources, specify a unique token for this library.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

encryptionConfiguration

Encryption settings for the library.

Type: EncryptionConfiguration object

Required: No

libraryDescription

A description for the library.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Pattern: `[a-zA-Z0-9\s!"\#\$%'&\(\)\*\+\,\-\./:;=\?@\[\\\]\^_`\{\|\}~><À-ÖØ-Üßà-öø-üß¿¡Œ-œ°£¥₹€§©ª®™¹±-µ##-##₽###₣###₺]*`

Required: No

libraryName

A name for the library.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

tags

Tags assigned to the library at time of creation.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "libraryArn": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

libraryArn

The library's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

status

The library's status.

Type: String

Valid Values: `ACTIVE | DELETING`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CreateDataAutomationProject

Service: Data Automation for Amazon Bedrock

Creates a data automation project.

Request Syntax

```
PUT /data-automation-projects/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "customOutputConfiguration": {
      "blueprints": [
         {
            "blueprintArn": "string",
            "blueprintStage": "string",
            "blueprintVersion": "string"
         }
      ],
      "document": {
         "fallbackBlueprints": [
            {
               "blueprintArn": "string",
               "blueprintStage": "string",
               "blueprintVersion": "string"
            }
         ]
      }
   },
   "dataAutomationLibraryConfiguration": {
      "libraries": [
         {
            "libraryArn": "string"
         }
      ]
   },
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
```

```
   "overrideConfiguration": {
      "audio": {
         "languageConfiguration": {
            "generativeOutputLanguage": "string",
            "identifyMultipleLanguages": boolean,
            "inputLanguages": [ "string" ]
         },
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         }
      },
      "document": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         },
         "splitter": {
            "state": "string"
         }
      },
      "image": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
```

```
                "redactionMaskMode": "string"
             }
          }
       },
       "modalityRouting": {
          "jpeg": "string",
          "mov": "string",
          "mp4": "string",
          "png": "string"
       },
       "video": {
          "modalityProcessing": {
             "state": "string"
          },
          "sensitiveDataConfiguration": {
             "detectionMode": "string",
             "detectionScope": [ "string" ],
             "piiEntitiesConfiguration": {
                "piiEntityTypes": [ "string" ],
                "redactionMaskMode": "string"
             }
          }
       }
    },
    "projectDescription": "string",
    "projectName": "string",
    "projectStage": "string",
    "projectType": "string",
    "standardOutputConfiguration": {
       "audio": {
          "extraction": {
             "category": {
                "state": "string",
                "typeConfiguration": {
                   "transcript": {
                      "channelLabeling": {
                         "state": "string"
                      },
                      "speakerLabeling": {
                         "state": "string"
                      }
                   }
                },
                "types": [ "string" ]
```

```
            }
         },
         "generativeField": {
            "state": "string",
            "types": [ "string" ]
         }
      },
      "document": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
            "granularity": {
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string"
         },
         "outputFormat": {
            "additionalFileFormat": {
               "state": "string"
            },
            "textFormat": {
               "types": [ "string" ]
            }
         }
      },
      "image": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
            "category": {
               "state": "string",
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string",
            "types": [ "string" ]
         }
      },
      "video": {
```

```
            "extraction": {
               "boundingBox": {
                  "state": "string"
               },
               "category": {
                  "state": "string",
                  "types": [ "string" ]
               }
            },
            "generativeField": {
               "state": "string",
               "types": [ "string" ]
            }
         }
      },
      "tags": [
         {
            "key": "string",
            "value": "string"
         }
      ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientToken

To avoid creating duplicate resources, specify a unique token for this project.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

customOutputConfiguration

Blueprints to enabled for the project.

Type: CustomOutputConfiguration object

Required: No

dataAutomationLibraryConfiguration

Library configuration for the project.

Type: DataAutomationLibraryConfiguration object

Required: No

encryptionConfiguration

Encryption settings for the project.

Type: EncryptionConfiguration object

Required: No

overrideConfiguration

Additional settings to apply to the project.

Type: OverrideConfiguration object

Required: No

projectDescription

A description for the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Required: No

projectName

A name for the project.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

projectStage

To indicate that the project is not final, set to `DEVELOPMENT`. Otherwise, the project is created in the `LIVE` stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

projectType

The type of bedrock data automation API that is compatible with this project.

Type: String

Valid Values: `ASYNC | SYNC`

Required: No

standardOutputConfiguration

The standard output configuration for the project. If you omit settings for any types, default settings are applied. You can specify an empty configuration to create a project with default settings and view the settings with GetDataAutomationProject.

Type: StandardOutputConfiguration object

Required: Yes

tags

Tags assigned to the project at time of creation.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "projectArn": "string",
   "projectStage": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

projectArn

The project's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

projectStage

The project's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

status

The project's status.

Type: String

Valid Values: `COMPLETED | IN_PROGRESS | FAILED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteBlueprint

Service: Data Automation for Amazon Bedrock

Deletes a blueprint or blueprint version.

Request Syntax

```
DELETE /blueprints/blueprintArn/?blueprintVersion=blueprintVersion HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

blueprintArn

The blueprint's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

blueprintVersion

To delete a blueprint version, specify the version.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3

- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteDataAutomationLibrary

Service: Data Automation for Amazon Bedrock

Deletes a data automation library.

Request Syntax

```
DELETE /data-automation-libraries/libraryArn/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "libraryArn": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

libraryArn

The library's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

status

The library's status.

Type: String

Valid Values: `ACTIVE | DELETING`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

DeleteDataAutomationProject

Service: Data Automation for Amazon Bedrock

Deletes a data automation project.

Request Syntax

```
DELETE /data-automation-projects/projectArn/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

projectArn

The project's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "projectArn": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

projectArn

The project's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

status

The project's status.

Type: String

Valid Values: `COMPLETED | IN_PROGRESS | FAILED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetBlueprint

Service: Data Automation for Amazon Bedrock

Gets details about a blueprint or blueprint version.

Request Syntax

```
POST /blueprints/blueprintArn/ HTTP/1.1
Content-type: application/json

{
   "blueprintStage": "string",
   "blueprintVersion": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

blueprintArn

The blueprint's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

blueprintStage

To view the development stage, set to `DEVELOPMENT`.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

blueprintVersion

To view a blueprint version, specify the version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "blueprint": {
      "blueprintArn": "string",
      "blueprintName": "string",
      "blueprintStage": "string",
      "blueprintVersion": "string",
      "creationTime": "string",
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string",
      "lastModifiedTime": "string",
      "optimizationSamples": [
         {
            "assetS3Object": {
               "s3Uri": "string",
               "version": "string"
            },
            "groundTruthS3Object": {
               "s3Uri": "string",
               "version": "string"
            }
```

```
            }
        ],
        "optimizationTime": "string",
        "schema": "string",
        "type": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

blueprint

Details about the blueprint.

Type: Blueprint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetBlueprintOptimizationStatus

Service: Data Automation for Amazon Bedrock

Gets the status of a blueprint optimization job.

Request Syntax

```
POST /getBlueprintOptimizationStatus/invocationArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

invocationArn

The ARN of the optimization job invocation.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:blueprint-optimization-invocation/[a-zA-Z0-9-_]+`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "errorMessage": "string",
   "errorType": "string",
   "outputConfiguration": {
      "s3Object": {
         "s3Uri": "string",
         "version": "string"
      }
```

```
    },
    "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

errorMessage

The error message if the job failed.

Type: String

errorType

The type of error if the job failed.

Type: String

outputConfiguration

The location of the optimization results.

Type: BlueprintOptimizationOutputConfiguration object

status

The job's status.

Type: String

Valid Values: `Created | InProgress | Success | ServiceError | ClientError`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetDataAutomationLibrary

Service: Data Automation for Amazon Bedrock

Gets details about a data automation library.

Request Syntax

```
POST /data-automation-libraries/libraryArn/ HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "library": {
      "creationTime": "string",
      "entityTypes": [
         {
            "entityMetadata": "string",
            "entityType": "string"
         }
```

```
      ],
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string",
      "libraryArn": "string",
      "libraryDescription": "string",
      "libraryName": "string",
      "status": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

library

Details about the library.

Type: DataAutomationLibrary object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetDataAutomationLibraryEntity

Service: Data Automation for Amazon Bedrock

Gets details about a data automation library entity.

Request Syntax

```
POST /data-automation-libraries/libraryArn/entityType/entityType/entities/entityId
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

entityId

The entity ID.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

entityType

The entity type.

Valid Values: VOCABULARY

Required: Yes

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "entity": { ... }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

entity

Details about the entity.

Type: EntityDetails object

Note: This object is a Union. Only one member of this object can be specified or returned.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetDataAutomationLibraryIngestionJob

Service: Data Automation for Amazon Bedrock

Gets details about a data automation library ingestion job.

Request Syntax

```
POST /data-automation-libraries/libraryArn/library-ingestion-jobs/jobArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobArn

The job's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library-ingestion-job/[a-zA-Z0-9-]{12,36}`

Required: Yes

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

```
Content-type: application/json

{
   "job": {
      "completionTime": "string",
      "creationTime": "string",
      "entityType": "string",
      "errorMessage": "string",
      "errorType": "string",
      "jobArn": "string",
      "jobStatus": "string",
      "operationType": "string",
      "outputConfiguration": {
         "s3Uri": "string"
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

job

Details about the job.

Type: DataAutomationLibraryIngestionJob object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetDataAutomationProject

Service: Data Automation for Amazon Bedrock

Gets details about a data automation project.

Request Syntax

```
POST /data-automation-projects/projectArn/ HTTP/1.1
Content-type: application/json

{
   "projectStage": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

projectArn

The project's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

projectStage

To view the development stage, set to `DEVELOPMENT`.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "project": {
      "creationTime": "string",
      "customOutputConfiguration": {
         "blueprints": [
            {
               "blueprintArn": "string",
               "blueprintStage": "string",
               "blueprintVersion": "string"
            }
         ],
         "document": {
            "fallbackBlueprints": [
               {
                  "blueprintArn": "string",
                  "blueprintStage": "string",
                  "blueprintVersion": "string"
               }
            ]
         }
      },
      "dataAutomationLibraryConfiguration": {
         "libraries": [
            {
               "libraryArn": "string"
            }
         ]
      },
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string",
      "lastModifiedTime": "string",
      "overrideConfiguration": {
         "audio": {
            "languageConfiguration": {
               "generativeOutputLanguage": "string",
               "identifyMultipleLanguages": boolean,
```

```
            "inputLanguages": [ "string" ]
         },
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         }
      },
      "document": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         },
         "splitter": {
            "state": "string"
         }
      },
      "image": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         }
      },
      "modalityRouting": {
```

```
            "jpeg": "string",
            "mov": "string",
            "mp4": "string",
            "png": "string"
         },
         "video": {
            "modalityProcessing": {
               "state": "string"
            },
            "sensitiveDataConfiguration": {
               "detectionMode": "string",
               "detectionScope": [ "string" ],
               "piiEntitiesConfiguration": {
                  "piiEntityTypes": [ "string" ],
                  "redactionMaskMode": "string"
               }
            }
         }
      },
      "projectArn": "string",
      "projectDescription": "string",
      "projectName": "string",
      "projectStage": "string",
      "projectType": "string",
      "standardOutputConfiguration": {
         "audio": {
            "extraction": {
               "category": {
                  "state": "string",
                  "typeConfiguration": {
                     "transcript": {
                        "channelLabeling": {
                           "state": "string"
                        },
                        "speakerLabeling": {
                           "state": "string"
                        }
                     }
                  },
                  "types": [ "string" ]
               }
            },
            "generativeField": {
               "state": "string",
```

```
            "types": [ "string" ]
         }
      },
      "document": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
            "granularity": {
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string"
         },
         "outputFormat": {
            "additionalFileFormat": {
               "state": "string"
            },
            "textFormat": {
               "types": [ "string" ]
            }
         }
      },
      "image": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
            "category": {
               "state": "string",
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string",
            "types": [ "string" ]
         }
      },
      "video": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
```

```
                "category": {
                   "state": "string",
                   "types": [ "string" ]
                }
             },
             "generativeField": {
                "state": "string",
                "types": [ "string" ]
             }
          }
       },
       "status": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

project

Details about the project.

Type: DataAutomationProject object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeBlueprintOptimizationAsync

Service: Data Automation for Amazon Bedrock

Starts an asynchronous job to optimize a blueprint using sample data.

Request Syntax

```
POST /invokeBlueprintOptimizationAsync HTTP/1.1
Content-type: application/json

{
   "blueprint": {
      "blueprintArn": "string",
      "stage": "string"
   },
   "dataAutomationProfileArn": "string",
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "outputConfiguration": {
      "s3Object": {
         "s3Uri": "string",
         "version": "string"
      }
   },
   "samples": [
      {
         "assetS3Object": {
            "s3Uri": "string",
            "version": "string"
         },
         "groundTruthS3Object": {
            "s3Uri": "string",
            "version": "string"
         }
      }
   ],
   "tags": [
      {
         "key": "string",
```

```
            "value": "string"
        }
    ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

blueprint

The blueprint to optimize.

Type: BlueprintOptimizationObject object

Required: Yes

dataAutomationProfileArn

The ARN of the data automation profile to use.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-profile/[a-zA-Z0-9-_.]+`

Required: Yes

encryptionConfiguration

Encryption settings for the optimization job.

Type: EncryptionConfiguration object

Required: No

outputConfiguration

Where to store the optimization results.

Type: BlueprintOptimizationOutputConfiguration object

Required: Yes

samples

Sample assets and ground truth data to use for optimization.

Type: Array of BlueprintOptimizationSample objects

Required: Yes

tags

Tags to assign to the optimization job.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "invocationArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationArn

The ARN of the optimization job invocation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:blueprint-optimization-invocation/[a-zA-Z0-9-_]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeDataAutomationLibraryIngestionJob

Service: Data Automation for Amazon Bedrock

Invokes a data automation library ingestion job.

Request Syntax

```
PUT /data-automation-libraries/libraryArn/library-ingestion-jobs/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "entityType": "string",
   "inputConfiguration": {
      "inlinePayload": { ... },
      "s3Object": {
         "s3Uri": "string",
         "version": "string"
      }
   },
   "notificationConfiguration": {
      "eventBridgeConfiguration": {
         "eventBridgeEnabled": boolean
      }
   },
   "operationType": "string",
   "outputConfiguration": {
      "s3Uri": "string"
   },
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request uses the following URI parameters.

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

To avoid creating duplicate resources, specify a unique token for this job.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

entityType

The entity type.

Type: String

Valid Values: VOCABULARY

Required: Yes

inputConfiguration

Input configuration for the job.

Type: InputConfiguration object

Required: Yes

notificationConfiguration

Notification configuration for the job.

Type: NotificationConfiguration object

Required: No

operationType

The operation type.

Type: String

Valid Values: `UPSERT | DELETE`

Required: Yes

outputConfiguration

Output configuration for the job.

Type: OutputConfiguration object

Required: Yes

tags

Tags assigned to the job at time of creation.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

jobArn

The job's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library-ingestion-job/[a-zA-Z0-9-]{12,36}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListBlueprints

Service: Data Automation for Amazon Bedrock

Gets a list a blueprints in your account, or public blueprints provided by the service, or versions of a blueprint. By default, the service returns a list of blueprints without blueprint versions that are in your account and in the `LIVE` stage.

Request Syntax

```
POST /blueprints/ HTTP/1.1
Content-type: application/json

{
   "blueprintArn": "string",
   "blueprintStageFilter": "string",
   "maxResults": number,
   "nextToken": "string",
   "projectFilter": {
      "projectArn": "string",
      "projectStage": "string"
   },
   "resourceOwner": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

blueprintArn

The blueprint's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: No

blueprintStageFilter

The blueprints' stage. To view blueprints in development, specify `DEVELOPMENT`.

Type: String

Valid Values: `DEVELOPMENT | LIVE | ALL`

Required: No

maxResults

The maximum number of blueprints to return in one page of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

projectFilter

To view blueprints associated with a project, specify the project.

Type: DataAutomationProjectFilter object

Required: No

resourceOwner

The blueprints' owner. To view public blueprints, specify `SERVICE`.

Type: String

Valid Values: `SERVICE | ACCOUNT`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "blueprints": [
      {
         "blueprintArn": "string",
         "blueprintName": "string",
         "blueprintStage": "string",
         "blueprintVersion": "string",
         "creationTime": "string",
         "lastModifiedTime": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

blueprints

A list of blueprint summaries.

Type: Array of BlueprintSummary objects

nextToken

A pagination token returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListDataAutomationLibraries

Service: Data Automation for Amazon Bedrock

Gets a list of data automation libraries in your account.

Request Syntax

```
POST /data-automation-libraries/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string",
   "projectFilter": {
      "projectArn": "string",
      "projectStage": "string"
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of libraries to return in one page of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

projectFilter

To view libraries associated with a project, specify the project.

Type: DataAutomationProjectFilter object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "libraries": [
      {
         "creationTime": "string",
         "libraryArn": "string",
         "libraryName": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

libraries

A list of library summaries.

Type: Array of DataAutomationLibrarySummary objects

nextToken

A pagination token returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListDataAutomationLibraryEntities

Service: Data Automation for Amazon Bedrock

Gets a list of data automation library entities.

Request Syntax

```
POST /data-automation-libraries/libraryArn/entityType/entityType/entities/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

entityType

The entity type.

Valid Values: VOCABULARY

Required: Yes

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of entities to return in one page of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "entities": [
      { ... }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

entities

A list of entity summaries.

Type: Array of DataAutomationLibraryEntitySummary objects

nextToken

A pagination token returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListDataAutomationLibraryIngestionJobs

Service: Data Automation for Amazon Bedrock

Gets a list of data automation library ingestion jobs.

Request Syntax

```
POST /data-automation-libraries/libraryArn/library-ingestion-jobs/ HTTP/1.1
Content-type: application/json

{
   "maxResults": number,
   "nextToken": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

maxResults

The maximum number of jobs to return in one page of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "jobs": [
      {
         "completionTime": "string",
         "creationTime": "string",
         "entityType": "string",
         "jobArn": "string",
         "jobStatus": "string",
         "operationType": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

jobs

A list of job summaries.

Type: Array of DataAutomationLibraryIngestionJobSummary objects

nextToken

A pagination token returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListDataAutomationProjects

Service: Data Automation for Amazon Bedrock

Gets a list of data automation projects in your account, or public projects provided by the service.

Request Syntax

```
POST /data-automation-projects/ HTTP/1.1
Content-type: application/json

{
   "blueprintFilter": {
      "blueprintArn": "string",
      "blueprintStage": "string",
      "blueprintVersion": "string"
   },
   "libraryFilter": {
      "libraryArn": "string"
   },
   "maxResults": number,
   "nextToken": "string",
   "projectStageFilter": "string",
   "projectTypeFilter": "string",
   "resourceOwner": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

blueprintFilter

To view projects associated with a blueprint, specify the blueprint.

Type: BlueprintFilter object

Required: No

libraryFilter

To view projects associated with a library, specify the library.

Type: DataAutomationLibraryFilter object

Required: No

maxResults

The maximum number of projects to return in one page of results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

Required: No

projectStageFilter

The projects' stage. To view projects in development, specify `DEVELOPMENT`.

Type: String

Valid Values: `DEVELOPMENT | LIVE | ALL`

Required: No

projectTypeFilter

The type of projects to return. Specify `ASYNC` for asynchronous projects, `SYNC` for synchronous projects, or `ALL` to return both types.

Type: String

Valid Values: ASYNC | SYNC | ALL

Required: No

resourceOwner

The projects' owner. To view public projects, specify SERVICE.

Type: String

Valid Values: SERVICE | ACCOUNT

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "nextToken": "string",
   "projects": [
      {
         "creationTime": "string",
         "projectArn": "string",
         "projectName": "string",
         "projectStage": "string",
         "projectType": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

A pagination token returned if more results are available.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

projects

A list of project summaries.

Type: Array of DataAutomationProjectSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListTagsForResource

Service: Data Automation for Amazon Bedrock

Lists a resources tags.

Request Syntax

```
POST /listTagsForResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The ARN of the resource you want to check the tags of

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-z0-9-]*:[0-9]{12}:(blueprint|data-automation-project|blueprint-optimization-invocation|data-automation-library|data-automation-library-ingestion-job)/[a-zA-Z0-9-]{12,36}`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json
```

```
{
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

The requested resource's tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

TagResource

Service: Data Automation for Amazon Bedrock

Lets you add a tag to an already created resource.

Request Syntax

```
POST /tagResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The ARN of the resource you want to tag

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-z0-9-]*:[0-9]{12}:(blueprint|data-automation-project|blueprint-optimization-invocation|data-automation-library|data-automation-library-ingestion-job)/[a-zA-Z0-9-]{12,36}`

Required: Yes

tags

The tags you want to add to a resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UntagResource

Service: Data Automation for Amazon Bedrock

Remove the tags from a resource.

Request Syntax

```
POST /untagResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string",
   "tagKeys": [ "string" ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN

The ARN of the resource you want to untag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-z0-9-]*:[0-9]{12}:(blueprint|data-automation-project|blueprint-optimization-invocation|data-automation-library|data-automation-library-ingestion-job)/[a-zA-Z0-9-]{12,36}`

Required: Yes

tagKeys

The keys of the tag you want to untag.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateBlueprint

Service: Data Automation for Amazon Bedrock

Updates a blueprint.

Request Syntax

```
PUT /blueprints/blueprintArn/ HTTP/1.1
Content-type: application/json

{
   "blueprintStage": "string",
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "schema": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

blueprintArn

The blueprint's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

Request Body

The request accepts the following data in JSON format.

blueprintStage

To update the development stage, set to `DEVELOPMENT`.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

encryptionConfiguration

Encryption for the updated blueprint.

Type: EncryptionConfiguration object

Required: No

schema

The blueprint's schema.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "blueprint": {
      "blueprintArn": "string",
      "blueprintName": "string",
      "blueprintStage": "string",
      "blueprintVersion": "string",
      "creationTime": "string",
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string",
```

```
      "lastModifiedTime": "string",
      "optimizationSamples": [
         {
            "assetS3Object": {
               "s3Uri": "string",
               "version": "string"
            },
            "groundTruthS3Object": {
               "s3Uri": "string",
               "version": "string"
            }
         }
      ],
      "optimizationTime": "string",
      "schema": "string",
      "type": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

blueprint

Details about the blueprint.

Type: Blueprint object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateDataAutomationLibrary

Service: Data Automation for Amazon Bedrock

Updates a data automation library.

Request Syntax

```
PUT /data-automation-libraries/libraryArn/ HTTP/1.1
Content-type: application/json

{
   "clientToken": "string",
   "libraryDescription": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

libraryArn

The library's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

To avoid creating duplicate resources, specify a unique token for this operation.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

libraryDescription

A description for the library.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Pattern: `[a-zA-Z0-9\s!"\#\$%'&\(\)\*\+\,\-\./:;=\?@\[\\\]\^_`\{\|\}~><À-ÖØ-Üßà-öø-üß¿¡Œ-œ°£¥₹€§©ª®™¹±-µ##-##₽###₣###₺]*`

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "libraryArn": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

libraryArn

The library's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

status

The library's status.

Type: String

Valid Values: `ACTIVE | DELETING`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UpdateDataAutomationProject

Service: Data Automation for Amazon Bedrock

Updates a data automation project.

Request Syntax

```
PUT /data-automation-projects/projectArn/ HTTP/1.1
Content-type: application/json

{
   "customOutputConfiguration": {
      "blueprints": [
         {
            "blueprintArn": "string",
            "blueprintStage": "string",
            "blueprintVersion": "string"
         }
      ],
      "document": {
         "fallbackBlueprints": [
            {
               "blueprintArn": "string",
               "blueprintStage": "string",
               "blueprintVersion": "string"
            }
         ]
      }
   },
   "dataAutomationLibraryConfiguration": {
      "libraries": [
         {
            "libraryArn": "string"
         }
      ]
   },
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "overrideConfiguration": {
```

```
      "audio": {
         "languageConfiguration": {
            "generativeOutputLanguage": "string",
            "identifyMultipleLanguages": boolean,
            "inputLanguages": [ "string" ]
         },
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         }
      },
      "document": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         },
         "splitter": {
            "state": "string"
         }
      },
      "image": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
```

```
            }
         }
      },
      "modalityRouting": {
         "jpeg": "string",
         "mov": "string",
         "mp4": "string",
         "png": "string"
      },
      "video": {
         "modalityProcessing": {
            "state": "string"
         },
         "sensitiveDataConfiguration": {
            "detectionMode": "string",
            "detectionScope": [ "string" ],
            "piiEntitiesConfiguration": {
               "piiEntityTypes": [ "string" ],
               "redactionMaskMode": "string"
            }
         }
      }
   },
   "projectDescription": "string",
   "projectStage": "string",
   "standardOutputConfiguration": {
      "audio": {
         "extraction": {
            "category": {
               "state": "string",
               "typeConfiguration": {
                  "transcript": {
                     "channelLabeling": {
                        "state": "string"
                     },
                     "speakerLabeling": {
                        "state": "string"
                     }
                  }
               },
               "types": [ "string" ]
            }
         },
         "generativeField": {
```

```
            "state": "string",
            "types": [ "string" ]
         }
      },
      "document": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
            "granularity": {
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string"
         },
         "outputFormat": {
            "additionalFileFormat": {
               "state": "string"
            },
            "textFormat": {
               "types": [ "string" ]
            }
         }
      },
      "image": {
         "extraction": {
            "boundingBox": {
               "state": "string"
            },
            "category": {
               "state": "string",
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string",
            "types": [ "string" ]
         }
      },
      "video": {
         "extraction": {
            "boundingBox": {
               "state": "string"
```

```
            },
            "category": {
               "state": "string",
               "types": [ "string" ]
            }
         },
         "generativeField": {
            "state": "string",
            "types": [ "string" ]
         }
      }
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

projectArn

The project's ARN.

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

Required: Yes

Request Body

The request accepts the following data in JSON format.

customOutputConfiguration

Blueprints to enable for the project.

Type: CustomOutputConfiguration object

Required: No

dataAutomationLibraryConfiguration

Library configuration for the project.

Type: DataAutomationLibraryConfiguration object

Required: No

encryptionConfiguration

Encryption for the created project.

Type: EncryptionConfiguration object

Required: No

overrideConfiguration

Additional settings to apply to the project.

Type: OverrideConfiguration object

Required: No

projectDescription

A description for the project.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Required: No

projectStage

To update the development stage, set to `DEVELOPMENT`.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

standardOutputConfiguration

The standard output configuration for the project. If you omit settings for any types, default settings are applied.

Type: StandardOutputConfiguration object

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "projectArn": "string",
   "projectStage": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

projectArn

The project's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

projectStage

The project's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

status

The project's status.

Type: String

Valid Values: `COMPLETED | IN_PROGRESS | FAILED`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 403

ConflictException

HTTP Status Code: 409

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 404

ServiceQuotaExceededException

HTTP Status Code: 402

ThrottlingException

HTTP Status Code: 429

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Runtime for Amazon Bedrock Data Automation

The following actions are supported by Runtime for Amazon Bedrock Data Automation:

- GetDataAutomationStatus
- InvokeDataAutomation
- InvokeDataAutomationAsync
- ListTagsForResource
- TagResource
- UntagResource

GetDataAutomationStatus

Service: Runtime for Amazon Bedrock Data Automation

Gets details about a data automation invocation. To process data, start an invocation with InvokeDataAutomationAsync. When the status is `Success`, retrieve the output from the S3 bucket.

Request Syntax

```
{
   "invocationArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

invocationArn

The invocation ARN returned by InvokeDataAutomationAsync.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:(insights-invocation|data-automation-invocation)/[a-zA-Z0-9-_]+`

Required: Yes

Response Syntax

```
{
   "errorMessage": "string",
   "errorType": "string",
   "outputConfiguration": {
      "s3Uri": "string"
   },
   "status": "string"
```

```
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

errorMessage

If an error ocurred, the error message.

Type: String

errorType

If an error ocurred, the error's type.

Type: String

outputConfiguration

The path of the output metadata object.

Type: OutputConfiguration object

status

The invocation's status.

Type: String

Valid Values: `Created | InProgress | Success | ServiceError | ClientError`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 400

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 400

ThrottlingException

HTTP Status Code: 400

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeDataAutomation

Service: Runtime for Amazon Bedrock Data Automation

Process data with a data automation project or blueprints. The service processes data synchronously. Output can either be returned inline or store in the output in the S3 bucket that you specify.

Request Syntax

```
{
   "blueprints": [
      {
         "blueprintArn": "string",
         "stage": "string",
         "version": "string"
      }
   ],
   "dataAutomationConfiguration": {
      "dataAutomationProjectArn": "string",
      "stage": "string"
   },
   "dataAutomationProfileArn": "string",
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "inputConfiguration": {
      "bytes": blob,
      "s3Uri": "string"
   },
   "outputConfiguration": {
      "s3Uri": "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

blueprints

A list of blueprints to apply to the data.

Type: Array of Blueprint objects

Array Members: Minimum number of 1 item. Maximum number of 40 items.

Required: No

dataAutomationConfiguration

The data automation project to process the data with.

Type: DataAutomationConfiguration object

Required: No

dataAutomationProfileArn

The ARN of the profile calling your request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-profile/[a-zA-Z0-9-_.]+`

Required: Yes

encryptionConfiguration

Encryption settings for the invocation.

Type: EncryptionConfiguration object

Required: No

inputConfiguration

Where to read the input data from.

Type: SyncInputConfiguration object

Required: Yes

outputConfiguration

Where to store the output.

Type: OutputConfiguration object

Required: No

Response Syntax

```
{
   "outputConfiguration": {
      "s3Uri": "string"
   },
   "outputSegments": [
      {
         "customOutput": "string",
         "customOutputStatus": "string",
         "standardOutput": "string"
      }
   ],
   "semanticModality": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

outputConfiguration

Where to store the output.

Type: OutputConfiguration object

outputSegments

List of outputs for each logical sub-doc.

Type: Array of OutputSegment objects

semanticModality

Detected semantic modality.

Type: String

Valid Values: `DOCUMENT | IMAGE | AUDIO | VIDEO`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 400

InternalServerException

HTTP Status Code: 500

ServiceUnavailableException

HTTP Status Code: 500

ThrottlingException

HTTP Status Code: 400

TooManyRequestsException

HTTP Status Code: 400

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeDataAutomationAsync

Service: Runtime for Amazon Bedrock Data Automation

Process data with a data automation project or blueprints. The service processes data asynchronously and stores the output in the S3 bucket that you specify. To check if an invocation is complete, use GetDataAutomationStatus.

Request Syntax

```
{
   "blueprints": [
      {
         "blueprintArn": "string",
         "stage": "string",
         "version": "string"
      }
   ],
   "clientToken": "string",
   "dataAutomationConfiguration": {
      "dataAutomationProjectArn": "string",
      "stage": "string"
   },
   "dataAutomationProfileArn": "string",
   "encryptionConfiguration": {
      "kmsEncryptionContext": {
         "string" : "string"
      },
      "kmsKeyId": "string"
   },
   "inputConfiguration": {
      "assetProcessingConfiguration": {
         "video": {
            "segmentConfiguration": { ... }
         }
      },
      "s3Uri": "string"
   },
   "notificationConfiguration": {
      "eventBridgeConfiguration": {
         "eventBridgeEnabled": boolean
      }
   },
   "outputConfiguration": {
```

```
      "s3Uri": "string"
   },
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

blueprints

A list of blueprints to apply to the data.

Type: Array of Blueprint objects

Array Members: Minimum number of 1 item. Maximum number of 40 items.

Required: No

clientToken

To avoid making duplicate calls, specify a unique token for this invocation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9])*`

Required: No

dataAutomationConfiguration

The data automation project to process the data with.

Type: DataAutomationConfiguration object

Required: No

dataAutomationProfileArn

The ARN of the profile calling your request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-profile/[a-zA-Z0-9-_.]+`

Required: Yes

encryptionConfiguration

Encryption settings for the invocation.

Type: EncryptionConfiguration object

Required: No

inputConfiguration

Where to read the input data from.

Type: InputConfiguration object

Required: Yes

notificationConfiguration

Notification settings for the invocation.

Type: NotificationConfiguration object

Required: No

outputConfiguration

Where to store the output.

Type: OutputConfiguration object

Required: Yes

tags

Tags associated with your inference at time of creation.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
{
   "invocationArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationArn

The ARN of the invocation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:(insights-invocation|data-automation-invocation)/[a-zA-Z0-9-_]+`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 400

InternalServerException

HTTP Status Code: 500

ServiceQuotaExceededException

HTTP Status Code: 400

ThrottlingException

HTTP Status Code: 400

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListTagsForResource

Service: Runtime for Amazon Bedrock Data Automation

Lists a resources tags.

Request Syntax

```
{
   "resourceARN": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

resourceARN

The ARN of the resource you want to check the tags of

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-invocation/[a-zA-Z0-9-_]+`

Required: Yes

Response Syntax

```
{
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

The requested resource's tags.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 400

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 400

ThrottlingException

HTTP Status Code: 400

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

TagResource

Service: Runtime for Amazon Bedrock Data Automation

Lets you add a tag to an already created resource.

Request Syntax

```
{
   "resourceARN": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

resourceARN

The ARN of the resource you want to tag

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-invocation/[a-zA-Z0-9-_]+`

Required: Yes

tags

The tags you want to add to a resource.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 400

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 400

ServiceQuotaExceededException

HTTP Status Code: 400

ThrottlingException

HTTP Status Code: 400

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

UntagResource

Service: Runtime for Amazon Bedrock Data Automation

Remove the tags from a resource.

Request Syntax

```
{
   "resourceARN": "string",
   "tagKeys": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

resourceARN

The ARN of the resource you want to untag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-invocation/[a-zA-Z0-9-_]+`

Required: Yes

tagKeys

The keys of the tag you want to untag.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!aws:)[\p{L}\p{Z}\p{N}_.:/=+\-@]*`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

HTTP Status Code: 400

InternalServerException

HTTP Status Code: 500

ResourceNotFoundException

HTTP Status Code: 400

ThrottlingException

HTTP Status Code: 400

ValidationException

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2

- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Amazon Bedrock Runtime

The following actions are supported by Amazon Bedrock Runtime:

- ApplyGuardrail
- Converse
- ConverseStream
- CountTokens
- GetAsyncInvoke
- InvokeGuardrailChecks
- InvokeModel
- InvokeModelWithBidirectionalStream
- InvokeModelWithResponseStream
- ListAsyncInvokes
- StartAsyncInvoke

ApplyGuardrail

Service: Amazon Bedrock Runtime

The action to apply a guardrail.

For troubleshooting some of the common errors you might encounter when using the `ApplyGuardrail` API, see Troubleshooting Amazon Bedrock API Error Codes in the Amazon Bedrock User Guide

Request Syntax

```
POST /guardrail/guardrailIdentifier/version/guardrailVersion/apply HTTP/1.1
Content-type: application/json

{
   "content": [
      { ... }
   ],
   "outputScope": "string",
   "source": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

guardrailIdentifier

The guardrail identifier used in the request to apply the guardrail.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(|([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

guardrailVersion

The guardrail version used in the request to apply the guardrail.

Pattern: `(|([1-9][0-9]{0,7})|(DRAFT))`

Required: Yes

Request Body

The request accepts the following data in JSON format.

content

The content details used in the request to apply the guardrail.

Type: Array of GuardrailContentBlock objects

Required: Yes

outputScope

Specifies the scope of the output that you get in the response. Set to FULL to return the entire output, including any detected and non-detected entries in the response for enhanced debugging.

Note that the full output scope doesn't apply to word filters or regex in sensitive information filters. It does apply to all other filtering policies, including sensitive information with filters that can detect personally identifiable information (PII).

Type: String

Valid Values: INTERVENTIONS | FULL

Required: No

source

The source of data used in the request to apply the guardrail.

Type: String

Valid Values: INPUT | OUTPUT

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "action": "string",
```

```
   "actionReason": "string",
   "assessments": [
      {
         "appliedGuardrailDetails": {
            "guardrailArn": "string",
            "guardrailId": "string",
            "guardrailOrigin": [ "string" ],
            "guardrailOwnership": "string",
            "guardrailVersion": "string"
         },
         "automatedReasoningPolicy": {
            "findings": [
               { ... }
            ]
         },
         "contentPolicy": {
            "filters": [
               {
                  "action": "string",
                  "confidence": "string",
                  "detected": boolean,
                  "filterStrength": "string",
                  "type": "string"
               }
            ]
         },
         "contextualGroundingPolicy": {
            "filters": [
               {
                  "action": "string",
                  "detected": boolean,
                  "score": number,
                  "threshold": number,
                  "type": "string"
               }
            ]
         },
         "invocationMetrics": {
            "guardrailCoverage": {
               "images": {
                  "guarded": number,
                  "total": number
               },
               "textCharacters": {
```

```
                "guarded": number,
                "total": number
            }
        },
        "guardrailProcessingLatency": number,
        "usage": {
            "automatedReasoningPolicies": number,
            "automatedReasoningPolicyUnits": number,
            "contentPolicyImageUnits": number,
            "contentPolicyUnits": number,
            "contextualGroundingPolicyUnits": number,
            "sensitiveInformationPolicyFreeUnits": number,
            "sensitiveInformationPolicyUnits": number,
            "topicPolicyUnits": number,
            "wordPolicyUnits": number
        }
    },
    "sensitiveInformationPolicy": {
        "piiEntities": [
            {
                "action": "string",
                "detected": boolean,
                "match": "string",
                "type": "string"
            }
        ],
        "regexes": [
            {
                "action": "string",
                "detected": boolean,
                "match": "string",
                "name": "string",
                "regex": "string"
            }
        ]
    },
    "topicPolicy": {
        "topics": [
            {
                "action": "string",
                "detected": boolean,
                "name": "string",
                "type": "string"
            }
```

```
            ]
         },
         "wordPolicy": {
            "customWords": [
               {
                  "action": "string",
                  "detected": boolean,
                  "match": "string"
               }
            ],
            "managedWordLists": [
               {
                  "action": "string",
                  "detected": boolean,
                  "match": "string",
                  "type": "string"
               }
            ]
         }
      }
   ],
   "guardrailCoverage": {
      "images": {
         "guarded": number,
         "total": number
      },
      "textCharacters": {
         "guarded": number,
         "total": number
      }
   },
   "outputs": [
      {
         "text": "string"
      }
   ],
   "usage": {
      "automatedReasoningPolicies": number,
      "automatedReasoningPolicyUnits": number,
      "contentPolicyImageUnits": number,
      "contentPolicyUnits": number,
      "contextualGroundingPolicyUnits": number,
      "sensitiveInformationPolicyFreeUnits": number,
      "sensitiveInformationPolicyUnits": number,
```

```
        "topicPolicyUnits": number,
        "wordPolicyUnits": number
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

action

The action taken in the response from the guardrail.

Type: String

Valid Values: `NONE | GUARDRAIL_INTERVENED`

actionReason

The reason for the action taken when harmful content is detected.

Type: String

assessments

The assessment details in the response from the guardrail.

Type: Array of GuardrailAssessment objects

guardrailCoverage

The guardrail coverage details in the apply guardrail response.

Type: GuardrailCoverage object

outputs

The output details in the response from the guardrail.

Type: Array of GuardrailOutputContent objects

usage

The usage details in the response from the guardrail.

Type: GuardrailUsage object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

HTTP Status Code: 400

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Converse

Service: Amazon Bedrock Runtime

Sends messages to the specified Amazon Bedrock model. `Converse` provides a consistent interface that works with all models that support messages. This allows you to write code once and use it with different models. If a model has unique inference parameters, you can also pass those unique parameters to the model.

Amazon Bedrock doesn't store any text, images, or documents that you provide as content. The data is only used to generate the response.

You can submit a prompt by including it in the `messages` field, specifying the `modelId` of a foundation model or inference profile to run inference on it, and including any other fields that are relevant to your use case.

You can also submit a prompt from Prompt management by specifying the ARN of the prompt version and including a map of variables to values in the `promptVariables` field. You can append more messages to the prompt by using the `messages` field. If you use a prompt from Prompt management, you can't include the following fields in the request: `additionalModelRequestFields`, `inferenceConfig`, `system`, or `toolConfig`. Instead, these fields must be defined through Prompt management. For more information, see Test a prompt using Prompt management.

For information about the Converse API, see Use the Converse API. To use a guardrail, see Use a guardrail with the Converse API. To use a tool with a model, see Tool use (Function calling).

For example code, see Converse API examples.

This operation requires permission for the `bedrock:InvokeModel` action.

Important

To deny all inference access to resources that you specify in the modelId field, you need to deny access to the `bedrock:InvokeModel` and `bedrock:InvokeModelWithResponseStream` actions. Doing this also denies access to the resource through the base inference actions (InvokeModel and InvokeModelWithResponseStream). For more information see Deny access for inference on specific models.

For troubleshooting some of the common errors you might encounter when using the `Converse` API, see Troubleshooting Amazon Bedrock API Error Codes in the Amazon Bedrock User Guide

Request Syntax

```
POST /model/modelId/converse HTTP/1.1
Content-type: application/json

{
   "additionalModelRequestFields": JSON value,
   "additionalModelResponseFieldPaths": [ "string" ],
   "guardrailConfig": {
      "guardrailIdentifier": "string",
      "guardrailVersion": "string",
      "trace": "string"
   },
   "inferenceConfig": {
      "maxTokens": number,
      "stopSequences": [ "string" ],
      "temperature": number,
      "topP": number
   },
   "messages": [
      {
         "content": [
            { ... }
         ],
         "role": "string"
      }
   ],
   "outputConfig": {
      "textFormat": {
         "structure": { ... },
         "type": "string"
      }
   },
   "performanceConfig": {
      "latency": "string"
   },
   "promptVariables": {
      "string" : { ... }
   },
   "requestMetadata": {
```

```
      "string" : "string"
   },
   "serviceTier": {
      "type": "string"
   },
   "system": [
      { ... }
   ],
   "toolConfig": {
      "toolChoice": { ... },
      "tools": [
         { ... }
      ]
   }
}
```

URI Request Parameters

The request uses the following URI parameters.

modelId

Specifies the model or throughput with which to run inference, or the prompt resource to use in inference. The value depends on the resource that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see Amazon Bedrock base model IDs (on-demand throughput) in the Amazon Bedrock User Guide.
- If you use an Amazon Bedrock Marketplace model, specify the ID or ARN of the marketplace endpoint that you created. For more information about Amazon Bedrock Marketplace and setting up an endpoint, see Amazon Bedrock Marketplace in the Amazon Bedrock User Guide.
- If you use an inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for cross-region inference in the Amazon Bedrock User Guide.
- If you use a prompt created through Prompt management, specify the ARN of the prompt version. For more information, see Test a prompt using Prompt management.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see Run inference using a Provisioned Throughput in the Amazon Bedrock User Guide.

- If you use a custom model, specify the ARN of the custom model deployment (for on-demand inference) or the ARN of your provisioned model (for Provisioned Throughput). For more information, see Use a custom model in Amazon Bedrock in the Amazon Bedrock User Guide.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:custom-model-deployment/[a-z0-9]{12})|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_-]?)+)|([a-zA-Z0-9-:.]+)|(^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?))$|(^arn:aws:sagemaker:[a-z0-9-]+:[0-9]{12}:endpoint/[a-zA-Z0-9-]+$)|(^arn:aws(-[^:]+)?:bedrock:([0-9a-z-]{1,20}):([0-9]{12}):(default-)?prompt-router/[a-zA-Z0-9-:.]+$)`

Required: Yes

Request Body

The request accepts the following data in JSON format.

additionalModelRequestFields

Additional inference parameters that the model supports, beyond the base set of inference parameters that `Converse` and `ConverseStream` support in the `inferenceConfig` field. For more information, see Model parameters.

Type: JSON value

Required: No

additionalModelResponseFieldPaths

Additional model parameters field paths to return in the response. `Converse` and `ConverseStream` return the requested fields as a JSON Pointer object in the `additionalModelResponseFields` field. The following is example JSON for `additionalModelResponseFieldPaths`.

`[ "/stop_sequence" ]`

For information about the JSON Pointer syntax, see the Internet Engineering Task Force (IETF) documentation.

`Converse` and `ConverseStream` reject an empty JSON Pointer or incorrectly structured JSON Pointer with a `400` error code. if the JSON Pointer is valid, but the requested field is not in the model response, it is ignored by `Converse`.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: No

guardrailConfig

Configuration information for a guardrail that you want to use in the request. If you include `guardContent` blocks in the `content` field in the `messages` field, the guardrail operates only on those messages. If you include no `guardContent` blocks, the guardrail operates on all messages in the request body and in any included prompt resource.

Type: GuardrailConfiguration object

Required: No

inferenceConfig

Inference parameters to pass to the model. `Converse` and `ConverseStream` support a base set of inference parameters. If you need to pass additional parameters that the model supports, use the `additionalModelRequestFields` request field.

Type: InferenceConfiguration object

Required: No

messages

The messages that you want to send to the model.

Type: Array of Message objects

Required: No

outputConfig

Output configuration for a model response.

Type: OutputConfig object

Required: No

performanceConfig

Model performance settings for the request.

Type: PerformanceConfiguration object

Required: No

promptVariables

Contains a map of variables in a prompt from Prompt management to objects containing the values to fill in for them when running model invocation. This field is ignored if you don't specify a prompt resource in the `modelId` field.

Type: String to PromptVariableValues object map

Required: No

requestMetadata

Key-value pairs that you can use to filter invocation logs.

Type: String to string map

Map Entries: Maximum number of 16 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `[a-zA-Z0-9\s:_@$#=/+,-.]{1,256}`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s:_@$#=/+,-.]{0,256}`

Required: No

serviceTier

Specifies the processing tier configuration used for serving the request.

Type: ServiceTier object

Required: No

system

A prompt that provides instructions or context to the model about the task it should perform, or the persona it should adopt during the conversation.

Type: Array of SystemContentBlock objects

Required: No

toolConfig

Configuration information for the tools that the model can use when generating a response.

For information about models that support tool use, see Supported models and model features.

Type: ToolConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "additionalModelResponseFields": JSON value,
   "metrics": {
      "latencyMs": number
   },
   "output": { ... },
   "performanceConfig": {
      "latency": "string"
   },
   "serviceTier": {
      "type": "string"
   },
   "stopReason": "string",
   "trace": {
      "guardrail": {
         "actionReason": "string",
         "inputAssessment": {
```

```
        "string" : {
           "appliedGuardrailDetails": {
              "guardrailArn": "string",
              "guardrailId": "string",
              "guardrailOrigin": [ "string" ],
              "guardrailOwnership": "string",
              "guardrailVersion": "string"
           },
           "automatedReasoningPolicy": {
              "findings": [
                 { ... }
              ]
           },
           "contentPolicy": {
              "filters": [
                 {
                    "action": "string",
                    "confidence": "string",
                    "detected": boolean,
                    "filterStrength": "string",
                    "type": "string"
                 }
              ]
           },
           "contextualGroundingPolicy": {
              "filters": [
                 {
                    "action": "string",
                    "detected": boolean,
                    "score": number,
                    "threshold": number,
                    "type": "string"
                 }
              ]
           },
           "invocationMetrics": {
              "guardrailCoverage": {
                 "images": {
                    "guarded": number,
                    "total": number
                 },
                 "textCharacters": {
                    "guarded": number,
                    "total": number
```

```
            }
         },
         "guardrailProcessingLatency": number,
         "usage": {
            "automatedReasoningPolicies": number,
            "automatedReasoningPolicyUnits": number,
            "contentPolicyImageUnits": number,
            "contentPolicyUnits": number,
            "contextualGroundingPolicyUnits": number,
            "sensitiveInformationPolicyFreeUnits": number,
            "sensitiveInformationPolicyUnits": number,
            "topicPolicyUnits": number,
            "wordPolicyUnits": number
         }
      },
      "sensitiveInformationPolicy": {
         "piiEntities": [
            {
               "action": "string",
               "detected": boolean,
               "match": "string",
               "type": "string"
            }
         ],
         "regexes": [
            {
               "action": "string",
               "detected": boolean,
               "match": "string",
               "name": "string",
               "regex": "string"
            }
         ]
      },
      "topicPolicy": {
         "topics": [
            {
               "action": "string",
               "detected": boolean,
               "name": "string",
               "type": "string"
            }
         ]
      },
```

```
            "wordPolicy": {
               "customWords": [
                  {
                     "action": "string",
                     "detected": boolean,
                     "match": "string"
                  }
               ],
               "managedWordLists": [
                  {
                     "action": "string",
                     "detected": boolean,
                     "match": "string",
                     "type": "string"
                  }
               ]
            }
         }
      },
      "modelOutput": [ "string" ],
      "outputAssessments": {
         "string" : [
            {
               "appliedGuardrailDetails": {
                  "guardrailArn": "string",
                  "guardrailId": "string",
                  "guardrailOrigin": [ "string" ],
                  "guardrailOwnership": "string",
                  "guardrailVersion": "string"
               },
               "automatedReasoningPolicy": {
                  "findings": [
                     { ... }
                  ]
               },
               "contentPolicy": {
                  "filters": [
                     {
                        "action": "string",
                        "confidence": "string",
                        "detected": boolean,
                        "filterStrength": "string",
                        "type": "string"
                     }
```

```
            ]
         },
         "contextualGroundingPolicy": {
            "filters": [
               {
                  "action": "string",
                  "detected": boolean,
                  "score": number,
                  "threshold": number,
                  "type": "string"
               }
            ]
         },
         "invocationMetrics": {
            "guardrailCoverage": {
               "images": {
                  "guarded": number,
                  "total": number
               },
               "textCharacters": {
                  "guarded": number,
                  "total": number
               }
            },
            "guardrailProcessingLatency": number,
            "usage": {
               "automatedReasoningPolicies": number,
               "automatedReasoningPolicyUnits": number,
               "contentPolicyImageUnits": number,
               "contentPolicyUnits": number,
               "contextualGroundingPolicyUnits": number,
               "sensitiveInformationPolicyFreeUnits": number,
               "sensitiveInformationPolicyUnits": number,
               "topicPolicyUnits": number,
               "wordPolicyUnits": number
            }
         },
         "sensitiveInformationPolicy": {
            "piiEntities": [
               {
                  "action": "string",
                  "detected": boolean,
                  "match": "string",
                  "type": "string"
```

```
                        }
                    ],
                    "regexes": [
                        {
                            "action": "string",
                            "detected": boolean,
                            "match": "string",
                            "name": "string",
                            "regex": "string"
                        }
                    ]
                },
                "topicPolicy": {
                    "topics": [
                        {
                            "action": "string",
                            "detected": boolean,
                            "name": "string",
                            "type": "string"
                        }
                    ]
                },
                "wordPolicy": {
                    "customWords": [
                        {
                            "action": "string",
                            "detected": boolean,
                            "match": "string"
                        }
                    ],
                    "managedWordLists": [
                        {
                            "action": "string",
                            "detected": boolean,
                            "match": "string",
                            "type": "string"
                        }
                    ]
                }
            }
        ]
    }
},
"promptRouter": {
```

```
            "invokedModelId": "string"
         }
      },
      "usage": {
         "cacheDetails": [
            {
               "inputTokens": number,
               "ttl": "string"
            }
         ],
         "cacheReadInputTokens": number,
         "cacheWriteInputTokens": number,
         "inputTokens": number,
         "outputTokens": number,
         "totalTokens": number
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

additionalModelResponseFields

Additional fields in the response that are unique to the model.

Type: JSON value

metrics

Metrics for the call to `Converse`.

Type: ConverseMetrics object

output

The result from the call to `Converse`.

Type: ConverseOutput object

Note: This object is a Union. Only one member of this object can be specified or returned.

performanceConfig

Model performance settings for the request.

Type: PerformanceConfiguration object

serviceTier

Specifies the processing tier configuration used for serving the request.

Type: ServiceTier object

stopReason

The reason why the model stopped generating output.

Type: String

Valid Values: `end_turn | tool_use | max_tokens | stop_sequence | guardrail_intervened | content_filtered | malformed_model_output | malformed_tool_use | model_context_window_exceeded`

trace

A trace object that contains information about the Guardrail behavior.

Type: ConverseTrace object

usage

The total number of tokens used in the call to `Converse`. The total includes the tokens input to the model and the tokens generated by the model.

Type: TokenUsage object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

originalStatusCode

The original status code.

resourceName

The resource name.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

HTTP Status Code: 429

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

Examples

Send a message to a model

Send a messsage to Anthropic Claude Sonnet with `Converse`.

Sample Request

```
POST /model/anthropic.claude-3-sonnet-20240229-v1:0/converse HTTP/1.1
Content-type: application/json

{
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "text": "Write an article about impact of high inflation to GDP of
 a country"
                }
            ]
        }
    ],
    "system": [{"text" : "You are an economist with access to lots of data"}],
    "inferenceConfig": {
        "maxTokens": 1000,
        "temperature": 0.5
    }
```

```
}
```

Example response

Response for the above request.

Sample Request

```
HTTP/1.1 200
Content-type: application/json

{
    "output": {
        "message": {
            "content": [
                {
                    "text": "<text generated by the model>"
                }
            ],
            "role": "assistant"
        }
    },
    "stopReason": "end_turn",
    "usage": {
        "inputTokens": 30,
        "outputTokens": 628,
        "totalTokens": 658
    },
    "metrics": {
        "latencyMs": 1275
    }
}
```

Send a message with additional model fields

In the following example, the request passess a field (`top_k`) that the `Converse` field doesn't support. You pass the additional field in the `additionalModelRequestFields` field. The example also shows how to set the paths for the additional fields sent in the response from the model.

Sample Request

```
POST /model/anthropic.claude-3-sonnet-20240229-v1:0/converse HTTP/1.1
```

```
Content-type: application/json

{
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "text": "Provide general steps to debug a BSOD on a Windows
 laptop."
                }
            ]
        }
    ],
    "system": [{"text" : "You are a tech support expert who helps resolve technical
 issues. Signal 'SUCCESS' if you can resolve the issue, otherwise 'FAILURE'"}],
    "inferenceConfig": {
        "stopSequences": [ "SUCCESS", "FAILURE" ]
    },
    "additionalModelRequestFields": {
        "top_k": 200
    },
    "additionalModelResponseFieldPaths": [
        "/stop_sequence"
    ]
}
```

Example response

Response for the above example.

Sample Request

```
HTTP/1.1 200
Content-type: application/json

{
    "output": {
        "message": {
            "content": [
                {
                    "text": "<text generated by the model>"
                }
            ],
```

```
            "role": "assistant"
        }
    },
    "additionalModelResponseFields": {
        "stop_sequence": "SUCCESS"
    },
    "stopReason": "stop_sequence",
    "usage": {
        "inputTokens": 51,
        "outputTokens": 442,
        "totalTokens": 493
    },
    "metrics": {
        "latencyMs": 7944
    }
}
```

Use an inference profile in a conversation

The following request calls the US Anthropic Claude 3.5 Sonnet inference profile to route traffic to the us-east-1 and us-west-2 regions.

Sample Request

```
POST /model/us.anthropic.claude-3-5-sonnet-20240620-v1:0/converse HTTP/1.1

{
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "text": "Hello world"
                }
            ]
        }
    ]
}
```

Run inference on a prompt resource from Prompt management

Send the following request to run inference on version 1 of a prompt resource from Prompt management whose ID is PROMPT12345. Suppose the prompt contains a variable called

{{genre}}. This request would fill in the variable with the value pop. Check that you have bedrock:RenderPrompt permissions for the prompt resource. For more information, see Prerequisites for Prompt management.

Sample Request

```
POST /model/arn:aws:bedrock:us-west-2:123456789012:prompt/PROMPT12345:1/converse
 HTTP/1.1
Content-type: application/json

{
   "promptVariables": {
      "genre": {
         "text": "pop"
      }
   }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ConverseStream

Service: Amazon Bedrock Runtime

Sends messages to the specified Amazon Bedrock model and returns the response in a stream. `ConverseStream` provides a consistent API that works with all Amazon Bedrock models that support messages. This allows you to write code once and use it with different models. Should a model have unique inference parameters, you can also pass those unique parameters to the model.

To find out if a model supports streaming, call GetFoundationModel and check the `responseStreamingSupported` field in the response.

> **Note**
>
> The AWS CLI doesn't support streaming operations in Amazon Bedrock, including `ConverseStream`.

Amazon Bedrock doesn't store any text, images, or documents that you provide as content. The data is only used to generate the response.

You can submit a prompt by including it in the `messages` field, specifying the `modelId` of a foundation model or inference profile to run inference on it, and including any other fields that are relevant to your use case.

You can also submit a prompt from Prompt management by specifying the ARN of the prompt version and including a map of variables to values in the `promptVariables` field. You can append more messages to the prompt by using the `messages` field. If you use a prompt from Prompt management, you can't include the following fields in the request: `additionalModelRequestFields`, `inferenceConfig`, `system`, or `toolConfig`. Instead, these fields must be defined through Prompt management. For more information, see Test a prompt using Prompt management.

For information about the Converse API, see Use the Converse API. To use a guardrail, see Use a guardrail with the Converse API. To use a tool with a model, see Tool use (Function calling).

For example code, see Conversation streaming example.

This operation requires permission for the `bedrock:InvokeModelWithResponseStream` action.

Important

To deny all inference access to resources that you specify in the modelId field, you need to deny access to the `bedrock:InvokeModel` and `bedrock:InvokeModelWithResponseStream` actions. Doing this also denies access to the resource through the base inference actions (InvokeModel and InvokeModelWithResponseStream). For more information see Deny access for inference on specific models.

For troubleshooting some of the common errors you might encounter when using the `ConverseStream` API, see Troubleshooting Amazon Bedrock API Error Codes in the Amazon Bedrock User Guide

Request Syntax

```
POST /model/modelId/converse-stream HTTP/1.1
Content-type: application/json

{
   "additionalModelRequestFields": JSON value,
   "additionalModelResponseFieldPaths": [ "string" ],
   "guardrailConfig": {
      "guardrailIdentifier": "string",
      "guardrailVersion": "string",
      "streamProcessingMode": "string",
      "trace": "string"
   },
   "inferenceConfig": {
      "maxTokens": number,
      "stopSequences": [ "string" ],
      "temperature": number,
      "topP": number
   },
   "messages": [
      {
         "content": [
            { ... }
         ],
         "role": "string"
      }
```

```
    ],
    "outputConfig": {
       "textFormat": {
          "structure": { ... },
          "type": "string"
       }
    },
    "performanceConfig": {
       "latency": "string"
    },
    "promptVariables": {
       "string" : { ... }
    },
    "requestMetadata": {
       "string" : "string"
    },
    "serviceTier": {
       "type": "string"
    },
    "system": [
       { ... }
    ],
    "toolConfig": {
       "toolChoice": { ... },
       "tools": [
          { ... }
       ]
    }
}
```

URI Request Parameters

The request uses the following URI parameters.

modelId

Specifies the model or throughput with which to run inference, or the prompt resource to use in inference. The value depends on the resource that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see Amazon Bedrock base model IDs (on-demand throughput) in the Amazon Bedrock User Guide.

- If you use an Amazon Bedrock Marketplace model, specify the ID or ARN of the marketplace endpoint that you created. For more information about Amazon Bedrock Marketplace and setting up an endpoint, see Amazon Bedrock Marketplace in the Amazon Bedrock User Guide.
- If you use an inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for cross-region inference in the Amazon Bedrock User Guide.
- If you use a prompt created through Prompt management, specify the ARN of the prompt version. For more information, see Test a prompt using Prompt management.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see Run inference using a Provisioned Throughput in the Amazon Bedrock User Guide.
- If you use a custom model, specify the ARN of the custom model deployment (for on-demand inference) or the ARN of your provisioned model (for Provisioned Throughput). For more information, see Use a custom model in Amazon Bedrock in the Amazon Bedrock User Guide.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:custom-model-deployment/[a-z0-9]{12})|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_-]?)+)|([a-zA-Z0-9-:.]+)|(^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?))$|(^arn:aws:sagemaker:[a-z0-9-]+:[0-9]{12}:endpoint/[a-zA-Z0-9-]+$)|(^arn:aws(-[^:]+)?:bedrock:([0-9a-z-]{1,20}):([0-9]{12}):(default-)?prompt-router/[a-zA-Z0-9-:.]+$)`

Required: Yes

Request Body

The request accepts the following data in JSON format.

additionalModelRequestFields

Additional inference parameters that the model supports, beyond the base set of inference parameters that `Converse` and `ConverseStream` support in the `inferenceConfig` field. For more information, see Model parameters.

Type: JSON value

Required: No

additionalModelResponseFieldPaths

Additional model parameters field paths to return in the response. `Converse` and `ConverseStream` return the requested fields as a JSON Pointer object in the `additionalModelResponseFields` field. The following is example JSON for `additionalModelResponseFieldPaths`.

```
[ "/stop_sequence" ]
```

For information about the JSON Pointer syntax, see the Internet Engineering Task Force (IETF) documentation.

`Converse` and `ConverseStream` reject an empty JSON Pointer or incorrectly structured JSON Pointer with a `400` error code. if the JSON Pointer is valid, but the requested field is not in the model response, it is ignored by `Converse`.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: No

guardrailConfig

Configuration information for a guardrail that you want to use in the request. If you include `guardContent` blocks in the `content` field in the `messages` field, the guardrail operates only on those messages. If you include no `guardContent` blocks, the guardrail operates on all messages in the request body and in any included prompt resource.

Type: GuardrailStreamConfiguration object

Required: No

inferenceConfig

Inference parameters to pass to the model. `Converse` and `ConverseStream` support a base set of inference parameters. If you need to pass additional parameters that the model supports, use the `additionalModelRequestFields` request field.

Type: InferenceConfiguration object

Required: No

messages

The messages that you want to send to the model.

Type: Array of Message objects

Required: No

outputConfig

Output configuration for a model response.

Type: OutputConfig object

Required: No

performanceConfig

Model performance settings for the request.

Type: PerformanceConfiguration object

Required: No

promptVariables

Contains a map of variables in a prompt from Prompt management to objects containing the values to fill in for them when running model invocation. This field is ignored if you don't specify a prompt resource in the `modelId` field.

Type: String to PromptVariableValues object map

Required: No

requestMetadata

Key-value pairs that you can use to filter invocation logs.

Type: String to string map

Map Entries: Maximum number of 16 items.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `[a-zA-Z0-9\s:_@$#=/+,-.]{1,256}`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s:_@$#=/+,-.]{0,256}`

Required: No

serviceTier

Specifies the processing tier configuration used for serving the request.

Type: ServiceTier object

Required: No

system

A prompt that provides instructions or context to the model about the task it should perform, or the persona it should adopt during the conversation.

Type: Array of SystemContentBlock objects

Required: No

toolConfig

Configuration information for the tools that the model can use when generating a response.

For information about models that support streaming tool use, see Supported models and model features.

Type: ToolConfiguration object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "contentBlockDelta": {
      "contentBlockIndex": number,
      "delta": { ... }
   },
   "contentBlockStart": {
      "contentBlockIndex": number,
      "start": { ... }
   },
   "contentBlockStop": {
      "contentBlockIndex": number
   },
   "internalServerException": {
   },
   "messageStart": {
      "role": "string"
   },
   "messageStop": {
      "additionalModelResponseFields": JSON value,
      "stopReason": "string"
   },
   "metadata": {
      "metrics": {
         "latencyMs": number
      },
      "performanceConfig": {
         "latency": "string"
      },
      "serviceTier": {
         "type": "string"
      },
      "trace": {
         "guardrail": {
            "actionReason": "string",
            "inputAssessment": {
               "string" : {
                  "appliedGuardrailDetails": {
                     "guardrailArn": "string",
```

```
            "guardrailId": "string",
            "guardrailOrigin": [ "string" ],
            "guardrailOwnership": "string",
            "guardrailVersion": "string"
         },
         "automatedReasoningPolicy": {
            "findings": [
               { ... }
            ]
         },
         "contentPolicy": {
            "filters": [
               {
                  "action": "string",
                  "confidence": "string",
                  "detected": boolean,
                  "filterStrength": "string",
                  "type": "string"
               }
            ]
         },
         "contextualGroundingPolicy": {
            "filters": [
               {
                  "action": "string",
                  "detected": boolean,
                  "score": number,
                  "threshold": number,
                  "type": "string"
               }
            ]
         },
         "invocationMetrics": {
            "guardrailCoverage": {
               "images": {
                  "guarded": number,
                  "total": number
               },
               "textCharacters": {
                  "guarded": number,
                  "total": number
               }
            },
            "guardrailProcessingLatency": number,
```

```
      "usage": {
         "automatedReasoningPolicies": number,
         "automatedReasoningPolicyUnits": number,
         "contentPolicyImageUnits": number,
         "contentPolicyUnits": number,
         "contextualGroundingPolicyUnits": number,
         "sensitiveInformationPolicyFreeUnits": number,
         "sensitiveInformationPolicyUnits": number,
         "topicPolicyUnits": number,
         "wordPolicyUnits": number
      }
   },
   "sensitiveInformationPolicy": {
      "piiEntities": [
         {
            "action": "string",
            "detected": boolean,
            "match": "string",
            "type": "string"
         }
      ],
      "regexes": [
         {
            "action": "string",
            "detected": boolean,
            "match": "string",
            "name": "string",
            "regex": "string"
         }
      ]
   },
   "topicPolicy": {
      "topics": [
         {
            "action": "string",
            "detected": boolean,
            "name": "string",
            "type": "string"
         }
      ]
   },
   "wordPolicy": {
      "customWords": [
         {
```

```
                        "action": "string",
                        "detected": boolean,
                        "match": "string"
                     }
                  ],
                  "managedWordLists": [
                     {
                        "action": "string",
                        "detected": boolean,
                        "match": "string",
                        "type": "string"
                     }
                  ]
               }
            }
         },
         "modelOutput": [ "string" ],
         "outputAssessments": {
            "string" : [
               {
                  "appliedGuardrailDetails": {
                     "guardrailArn": "string",
                     "guardrailId": "string",
                     "guardrailOrigin": [ "string" ],
                     "guardrailOwnership": "string",
                     "guardrailVersion": "string"
                  },
                  "automatedReasoningPolicy": {
                     "findings": [
                        { ... }
                     ]
                  },
                  "contentPolicy": {
                     "filters": [
                        {
                           "action": "string",
                           "confidence": "string",
                           "detected": boolean,
                           "filterStrength": "string",
                           "type": "string"
                        }
                     ]
                  },
                  "contextualGroundingPolicy": {
```

```
            "filters": [
               {
                  "action": "string",
                  "detected": boolean,
                  "score": number,
                  "threshold": number,
                  "type": "string"
               }
            ]
         },
         "invocationMetrics": {
            "guardrailCoverage": {
               "images": {
                  "guarded": number,
                  "total": number
               },
               "textCharacters": {
                  "guarded": number,
                  "total": number
               }
            },
            "guardrailProcessingLatency": number,
            "usage": {
               "automatedReasoningPolicies": number,
               "automatedReasoningPolicyUnits": number,
               "contentPolicyImageUnits": number,
               "contentPolicyUnits": number,
               "contextualGroundingPolicyUnits": number,
               "sensitiveInformationPolicyFreeUnits": number,
               "sensitiveInformationPolicyUnits": number,
               "topicPolicyUnits": number,
               "wordPolicyUnits": number
            }
         },
         "sensitiveInformationPolicy": {
            "piiEntities": [
               {
                  "action": "string",
                  "detected": boolean,
                  "match": "string",
                  "type": "string"
               }
            ],
            "regexes": [
```

```
                        {
                           "action": "string",
                           "detected": boolean,
                           "match": "string",
                           "name": "string",
                           "regex": "string"
                        }
                     ]
                  },
                  "topicPolicy": {
                     "topics": [
                        {
                           "action": "string",
                           "detected": boolean,
                           "name": "string",
                           "type": "string"
                        }
                     ]
                  },
                  "wordPolicy": {
                     "customWords": [
                        {
                           "action": "string",
                           "detected": boolean,
                           "match": "string"
                        }
                     ],
                     "managedWordLists": [
                        {
                           "action": "string",
                           "detected": boolean,
                           "match": "string",
                           "type": "string"
                        }
                     ]
                  }
               }
            ]
         }
      },
      "promptRouter": {
         "invokedModelId": "string"
      }
   },
```

```
      "usage": {
         "cacheDetails": [
            {
               "inputTokens": number,
               "ttl": "string"
            }
         ],
         "cacheReadInputTokens": number,
         "cacheWriteInputTokens": number,
         "inputTokens": number,
         "outputTokens": number,
         "totalTokens": number
      }
   },
   "modelStreamErrorException": {
   },
   "serviceUnavailableException": {
   },
   "throttlingException": {
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

contentBlockDelta

The messages output content block delta.

Type: ContentBlockDeltaEvent object

contentBlockStart

Start information for a content block.

Type: ContentBlockStartEvent object

contentBlockStop

Stop information for a content block.

Type: ContentBlockStopEvent object

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

messageStart

Message start information.

Type: MessageStartEvent object

messageStop

Message stop information.

Type: MessageStopEvent object

metadata

Metadata for the converse output stream.

Type: ConverseStreamMetadataEvent object

modelStreamErrorException

A streaming error occurred. Retry your request.

Type: Exception

HTTP Status Code: 424

serviceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

Type: Exception

HTTP Status Code: 503

throttlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 429

validationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

originalStatusCode

The original status code.

resourceName

The resource name.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

HTTP Status Code: 429

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

Examples

Send a message to a model and stream the response.

Send a message to Anthropic Claude Sonnet with `ConverseStream` and stream the response.

Sample Request

```
POST /model/anthropic.claude-3-sonnet-20240229-v1:0/converse-stream HTTP/1.1
{
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "text": "Write an article about impact of high inflation to GDP of
 a country"
                }
            ]
        }
    ],
    "system": [{"text" : "You are an economist with access to lots of data"}],
    "inferenceConfig": {
        "maxTokens": 1000,
        "temperature": 0.5
    }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

CountTokens

Service: Amazon Bedrock Runtime

Returns the token count for a given inference request. This operation helps you estimate token usage before sending requests to foundation models by returning the token count that would be used if the same input were sent to the model in an inference request.

Token counting is model-specific because different models use different tokenization strategies. The token count returned by this operation will match the token count that would be charged if the same input were sent to the model in an `InvokeModel` or `Converse` request.

You can use this operation to:

- Estimate costs before sending inference requests.
- Optimize prompts to fit within token limits.
- Plan for token usage in your applications.

This operation accepts the same input formats as `InvokeModel` and `Converse`, allowing you to count tokens for both raw text inputs and structured conversation formats.

The following operations are related to `CountTokens`:

- InvokeModel - Sends inference requests to foundation models
- Converse - Sends conversation-based inference requests to foundation models

Request Syntax

```
POST /model/modelId/count-tokens HTTP/1.1
Content-type: application/json

{
   "input": { ... }
}
```

URI Request Parameters

The request uses the following URI parameters.

modelId

The unique identifier or ARN of the foundation model to use for token counting. Each model processes tokens differently, so the token count is specific to the model you specify.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z_\.\-/0-9:]+`

Required: Yes

Request Body

The request accepts the following data in JSON format.

input

The input for which to count tokens. The structure of this parameter depends on whether you're counting tokens for an `InvokeModel` or `Converse` request:

- For `InvokeModel` requests, provide the request body in the `invokeModel` field
- For `Converse` requests, provide the messages and system content in the `converse` field

The input format must be compatible with the model specified in the `modelId` parameter.

Type: CountTokensInput object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "inputTokens": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

inputTokens

The number of tokens in the provided input according to the specified model's tokenization rules. This count represents the number of input tokens that would be processed if the same input were sent to the model in an inference request. Use this value to estimate costs and ensure your inputs stay within model token limits.

Type: Integer

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

GetAsyncInvoke

Service: Amazon Bedrock Runtime

Retrieve information about an asynchronous invocation.

Request Syntax

```
GET /async-invoke/invocationArn HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

invocationArn

The invocation's ARN.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:async-invoke/[a-z0-9]{12}`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "clientRequestToken": "string",
   "endTime": "string",
   "failureMessage": "string",
   "invocationArn": "string",
   "lastModifiedTime": "string",
   "modelArn": "string",
   "outputDataConfig": { ... },
   "status": "string",
```

```
    "submitTime": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

clientRequestToken

The invocation's idempotency token.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[!-~]*`

endTime

When the invocation ended.

Type: Timestamp

failureMessage

An error message.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

invocationArn

The invocation's ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:async-invoke/[a-z0-9]{12}`

lastModifiedTime

The invocation's last modified time.

Type: Timestamp

modelArn

The invocation's model ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:[a-z0-9\-]+:bedrock:[a-z0-9\-]*:[0-9]*:(provisioned-model|foundation-model)/.+`

outputDataConfig

Output data settings.

Type: AsyncInvokeOutputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

status

The invocation's status.

Type: String

Valid Values: `InProgress | Completed | Failed`

submitTime

When the invocation request was submitted.

Type: Timestamp

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeGuardrailChecks

Service: Amazon Bedrock Runtime

Evaluates messages against inline guardrail checks. You specify the check configurations directly in the request, and Amazon Bedrock returns per-check results with severity or confidence scores.

Request Syntax

```
POST /guardrail-checks/invoke HTTP/1.1
Content-type: application/json

{
   "checks": {
      "contentFilter": {
         "categories": [
            {
               "category": "string"
            }
         ]
      },
      "promptAttack": {
         "categories": [
            {
               "category": "string"
            }
         ]
      },
      "sensitiveInformation": {
         "entities": [
            {
               "type": "string"
            }
         ]
      }
   },
   "messages": [
      {
         "content": [
            { ... }
         ],
         "role": "string"
      }
   ]
```

```
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

checks

The inline check configurations that specify which guardrail checks to run against the messages.

Type: GuardrailChecksConfig object

Required: Yes

messages

The messages to evaluate against the specified guardrail checks. Each message includes a role and one or more content blocks.

Type: Array of GuardrailChecksMessage objects

Array Members: Minimum number of 1 item.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "results": {
      "contentFilter": {
         "results": [
            {
               "category": "string",
               "severityScore": number
            }
         ]
      },
```

```
      "promptAttack": {
         "results": [
            {
               "category": "string",
               "severityScore": number
            }
         ]
      },
      "sensitiveInformation": {
         "results": [
            {
               "beginOffset": number,
               "confidenceScore": number,
               "contentIndex": number,
               "endOffset": number,
               "messageIndex": number,
               "type": "string"
            }
         ],
         "truncated": boolean
      }
   },
   "usage": {
      "contentFilter": {
         "textUnits": number
      },
      "promptAttack": {
         "textUnits": number
      },
      "sensitiveInformation": {
         "textUnits": number
      }
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

results

The per-check results containing findings from the guardrail evaluation.

Type: GuardrailChecksResults object

usage

The per-check text unit consumption for the guardrail evaluation.

Type: GuardrailChecksUsageResults object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeModel

Service: Amazon Bedrock Runtime

Invokes the specified Amazon Bedrock model to run inference using the prompt and inference parameters provided in the request body. You use model inference to generate text, images, and embeddings.

For example code, see Invoke model code examples.

This operation requires permission for the `bedrock:InvokeModel` action.

> **Important**
>
> To deny all inference access to resources that you specify in the modelId field, you need to deny access to the `bedrock:InvokeModel` and `bedrock:InvokeModelWithResponseStream` actions. Doing this also denies access to the resource through the Converse API actions (Converse and ConverseStream). For more information see Deny access for inference on specific models.

For troubleshooting some of the common errors you might encounter when using the `InvokeModel` API, see Troubleshooting Amazon Bedrock API Error Codes in the Amazon Bedrock User Guide

Request Syntax

```
POST /model/modelId/invoke HTTP/1.1
Content-Type: contentType
Accept: accept
X-Amzn-Bedrock-Trace: trace
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-PerformanceConfig-Latency: performanceConfigLatency
X-Amzn-Bedrock-Service-Tier: serviceTier
X-Amzn-Bedrock-Request-Metadata: requestMetadata

body
```

URI Request Parameters

The request uses the following URI parameters.

accept

The desired MIME type of the inference body in the response. The default value is `application/json`.

contentType

The MIME type of the input data in the request. You must specify `application/json`.

guardrailIdentifier

The unique identifier of the guardrail that you want to use. If you don't provide a value, no guardrail is applied to the invocation.

An error will be thrown in the following situations.

- You don't provide a guardrail identifier but you specify the `amazon-bedrock-guardrailConfig` field in the request body.
- You enable the guardrail but the `contentType` isn't `application/json`.
- You provide a guardrail identifier, but `guardrailVersion` isn't specified.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(|([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

guardrailVersion

The version number for the guardrail. The value can also be `DRAFT`.

Pattern: `(|([1-9][0-9]{0,7})|(DRAFT))`

modelId

Specifies the model or throughput with which to run inference, or the prompt resource to use in inference. The value depends on the resource that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see Amazon Bedrock base model IDs (on-demand throughput) in the Amazon Bedrock User Guide.
- If you use an Amazon Bedrock Marketplace model, specify the ID or ARN of the marketplace endpoint that you created. For more information about Amazon Bedrock Marketplace and setting up an endpoint, see Amazon Bedrock Marketplace in the Amazon Bedrock User Guide.

- If you use an inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for cross-region inference in the Amazon Bedrock User Guide.
- If you use a prompt created through Prompt management, specify the ARN of the prompt version. For more information, see Test a prompt using Prompt management.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see Run inference using a Provisioned Throughput in the Amazon Bedrock User Guide.
- If you use a custom model, specify the ARN of the custom model deployment (for on-demand inference) or the ARN of your provisioned model (for Provisioned Throughput). For more information, see Use a custom model in Amazon Bedrock in the Amazon Bedrock User Guide.
- If you use an imported model, specify the ARN of the imported model. You can get the model ARN from a successful call to CreateModelImportJob or from the Imported models page in the Amazon Bedrock console.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:custom-model-deployment/[a-z0-9]{12})|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_-]?)+)|([a-zA-Z0-9-:.]+)$|(^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?))$|(^arn:aws:sagemaker:[a-z0-9-]+:[0-9]{12}:endpoint/[a-zA-Z0-9-]+$)|(^arn:aws(-[^:]+)?:bedrock:([0-9a-z-]{1,20}):([0-9]{12}):(default-)?prompt-router/[a-zA-Z0-9-:.]+$)`

Required: Yes

performanceConfigLatency

Model performance settings for the request.

Valid Values: `standard | optimized`

requestMetadata

Key-value pairs that you can use to filter invocation logs.

Length Constraints: Minimum length of 0. Maximum length of 8500.

serviceTier

Specifies the processing tier type used for serving the request.

Valid Values: `priority | default | flex | reserved`

trace

Specifies whether to enable or disable the Bedrock trace. If enabled, you can see the full Bedrock trace.

Valid Values: `ENABLED | DISABLED | ENABLED_FULL`

Request Body

The request accepts the following binary data.

body

The prompt and inference parameters in the format specified in the `contentType` in the header. You must provide the body in JSON format. To see the format and content of the request and response bodies for different models, refer to Inference parameters. For more information, see Run inference in the Bedrock User Guide.

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Response Syntax

```
HTTP/1.1 200
Content-Type: contentType
X-Amzn-Bedrock-PerformanceConfig-Latency: performanceConfigLatency
X-Amzn-Bedrock-Service-Tier: serviceTier

body
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType

The MIME type of the inference result.

performanceConfigLatency

Model performance settings for the request.

Valid Values: `standard | optimized`

serviceTier

Specifies the processing tier type used for serving the request.

Valid Values: `priority | default | flex | reserved`

The response returns the following as the HTTP body.

body

Inference response from the model in the format specified in the `contentType` header. To see the format and content of the request and response bodies for different models, refer to Inference parameters.

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

originalStatusCode

The original status code.

resourceName

The resource name.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

HTTP Status Code: 429

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

HTTP Status Code: 400

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

Examples

Run inference on a text model

Send an invoke request to run inference on a Titan Text G1 - Express model. We set the `accept` parameter to accept any content type in the response.

Sample Request

```
POST https://bedrock-runtime.us-east-1.amazonaws.com/model/amazon.titan-text-express-v1/invoke

-H accept: */*
-H content-type: application/json

Payload
{"inputText": "Hello world"}
```

Run inference on an image model

In the following example, the request sets the `accept` parameter to `image/png`.

Sample Request

```
POST https://bedrock-runtime.us-east-1.amazonaws.com/model/stability.stable-diffusion-
xl-v1/invoke

-H accept: image/png
-H content-type: application/json

Payload
{"inputText": "Picture of a bird"}
```

Use a guardrail

This example shows how to use a guardrail with `InvokeModel`.

Sample Request

```
POST /model/modelId/invoke HTTP/1.1
Accept: accept
Content-Type: contentType
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-GuardrailTrace: guardrailTrace
X-Amzn-Bedrock-Trace: trace

body

// body
{
    "amazon-bedrock-guardrailConfig": {
        "tagSuffix": "string"
    }
}
```

Example response

This is an example response from `InvokeModel` when using a guardrail.

Sample Request

```
HTTP/1.1 200
Content-Type: contentType
```

```
body

// body
{
  "amazon-bedrock-guardrailAction": "INTERVENED | NONE",
  "amazon-bedrock-trace": {
    "guardrails": {
      //  Detailed guardrail trace
    }
  }
}
```

Use an inference profile in model invocation

The following request calls the US Anthropic Claude 3.5 Sonnet inference profile to route traffic to the us-east-1 and us-west-2 regions.

Sample Request

```
POST /model/us.anthropic.claude-3-5-sonnet-20240620-v1:0/invoke HTTP/1.1

{
    "anthropic_version": "bedrock-2023-05-31",
    "max_tokens": 1024,
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": "Hello world"
                }
            ]
        }
    ]
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeModelWithBidirectionalStream

Service: Amazon Bedrock Runtime

Invoke the specified Amazon Bedrock model to run inference using the bidirectional stream. The response is returned in a stream that remains open for 8 minutes. A single session can contain multiple prompts and responses from the model. The prompts to the model are provided as audio files and the model's responses are spoken back to the user and transcribed.

It is possible for users to interrupt the model's response with a new prompt, which will halt the response speech. The model will retain contextual awareness of the conversation while pivoting to respond to the new prompt.

This operation requires permission for the `bedrock:InvokeModel` action.

Note that to log `InvokeModelWithBidirectionalStream` API calls in CloudTrail, you need to configure advanced event selectors to record data events for the `AWS::Bedrock::Model` resource type and `AWS:Bedrock::AsyncInvoke`. To learn more, see Monitor Amazon Bedrock API calls using CloudTrail.

> **Important**
>
> To deny all inference access to resources that you specify in the modelId field, you need to deny access to the `bedrock:InvokeModel` and `bedrock:InvokeModelWithResponseStream` actions. Doing this also denies access to the resource through the Converse API actions (Converse and ConverseStream). For more information see Deny access for inference on specific models.

For more information, see Using the Amazon Nova Sonic Speech-to-Speech model.

Request Syntax

```
POST /model/modelId/invoke-with-bidirectional-stream HTTP/1.1
Content-type: application/json

{
   "chunk": {
      "bytes": blob
   }
```

```
}
```

URI Request Parameters

The request uses the following URI parameters.

modelId

The model ID or ARN of the model ID to use. Currently, only `amazon.nova-sonic-v1:0` is supported.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:custom-model-deployment/[a-z0-9]{12})|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_-]?)+)|([a-zA-Z0-9-:.]+)$|(^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?))$|(^arn:aws:sagemaker:[a-z0-9-]+:[0-9]{12}:endpoint/[a-zA-Z0-9-]+$)|(^arn:aws(-[^:]+)?:bedrock:([0-9a-z-]{1,20}):([0-9]{12}):(default-)?prompt-router/[a-zA-Z0-9-:.]+$)`

Required: Yes

Request Body

The request accepts the following data in JSON format.

chunk

The audio chunk that is used as input for the invocation step.

Type: BidirectionalInputPayloadPart object

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "chunk": {
      "bytes": blob
   },
   "internalServerException": {
   },
   "modelStreamErrorException": {
   },
   "modelTimeoutException": {
   },
   "serviceUnavailableException": {
   },
   "throttlingException": {
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

chunk

The speech chunk that was provided as output from the invocation step.

Type: BidirectionalOutputPayloadPart object

internalServerException

The request encountered an unknown internal error.

Type: Exception

HTTP Status Code: 500

modelStreamErrorException

The request encountered an error with the model stream.

Type: Exception

HTTP Status Code: 424

modelTimeoutException

The connection was closed because a request was not received within the timeout period.

Type: Exception

HTTP Status Code: 408

serviceUnavailableException

The request has failed due to a temporary failure of the server.

Type: Exception

HTTP Status Code: 503

throttlingException

The request was denied due to request throttling.

Type: Exception

HTTP Status Code: 429

validationException

The input fails to satisfy the constraints specified by an AWS service.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

originalStatusCode

The original status code.

resourceName

The resource name.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

HTTP Status Code: 429

ModelStreamErrorException

An error occurred while streaming the response. Retry your request.

originalMessage

The original message.

originalStatusCode

The original status code.

HTTP Status Code: 424

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

HTTP Status Code: 400

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

InvokeModelWithResponseStream

Service: Amazon Bedrock Runtime

Invoke the specified Amazon Bedrock model to run inference using the prompt and inference parameters provided in the request body. The response is returned in a stream.

To see if a model supports streaming, call GetFoundationModel and check the `responseStreamingSupported` field in the response.

> **Note**
>
> The AWS CLI doesn't support streaming operations in Amazon Bedrock, including `InvokeModelWithResponseStream`.

For example code, see Invoke model with streaming code example.

This operation requires permissions to perform the `bedrock:InvokeModelWithResponseStream` action.

> **Important**
>
> To deny all inference access to resources that you specify in the modelId field, you need to deny access to the `bedrock:InvokeModel` and `bedrock:InvokeModelWithResponseStream` actions. Doing this also denies access to the resource through the Converse API actions (Converse and ConverseStream). For more information see Deny access for inference on specific models.

For troubleshooting some of the common errors you might encounter when using the `InvokeModelWithResponseStream` API, see Troubleshooting Amazon Bedrock API Error Codes in the Amazon Bedrock User Guide

Request Syntax

```
POST /model/modelId/invoke-with-response-stream HTTP/1.1
Content-Type: contentType
X-Amzn-Bedrock-Accept: accept
X-Amzn-Bedrock-Trace: trace
```

```
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-PerformanceConfig-Latency: performanceConfigLatency
X-Amzn-Bedrock-Service-Tier: serviceTier
X-Amzn-Bedrock-Request-Metadata: requestMetadata

body
```

URI Request Parameters

The request uses the following URI parameters.

accept

The desired MIME type of the inference body in the response. The default value is `application/json`.

contentType

The MIME type of the input data in the request. You must specify `application/json`.

guardrailIdentifier

The unique identifier of the guardrail that you want to use. If you don't provide a value, no guardrail is applied to the invocation.

An error is thrown in the following situations.

- You don't provide a guardrail identifier but you specify the `amazon-bedrock-guardrailConfig` field in the request body.
- You enable the guardrail but the `contentType` isn't `application/json`.
- You provide a guardrail identifier, but `guardrailVersion` isn't specified.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(|([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

guardrailVersion

The version number for the guardrail. The value can also be `DRAFT`.

Pattern: `(|([1-9][0-9]{0,7})|(DRAFT))`

modelId

Specifies the model or throughput with which to run inference, or the prompt resource to use in inference. The value depends on the resource that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs for base models, see Amazon Bedrock base model IDs (on-demand throughput) in the Amazon Bedrock User Guide.
- If you use an Amazon Bedrock Marketplace model, specify the ID or ARN of the marketplace endpoint that you created. For more information about Amazon Bedrock Marketplace and setting up an endpoint, see Amazon Bedrock Marketplace in the Amazon Bedrock User Guide.
- If you use an inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for cross-region inference in the Amazon Bedrock User Guide.
- If you use a prompt created through Prompt management, specify the ARN of the prompt version. For more information, see Test a prompt using Prompt management.
- If you use a provisioned model, specify the ARN of the Provisioned Throughput. For more information, see Run inference using a Provisioned Throughput in the Amazon Bedrock User Guide.
- If you use a custom model, specify the ARN of the custom model deployment (for on-demand inference) or the ARN of your provisioned model (for Provisioned Throughput). For more information, see Use a custom model in Amazon Bedrock in the Amazon Bedrock User Guide.
- If you use an imported model, specify the ARN of the imported model. You can get the model ARN from a successful call to CreateModelImportJob or from the Imported models page in the Amazon Bedrock console.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:custom-model-deployment/[a-z0-9]{12})|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|(([0-9a-zA-Z][_-]?)+)|([a-zA-Z0-9-:.]+)$|(^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?))$|`

`(^arn:aws:sagemaker:[a-z0-9-]+:[0-9]{12}:endpoint/[a-zA-Z0-9-]+$)|(^arn:aws(-[^:]+)?:bedrock:([0-9a-z-]{1,20}):([0-9]{12}):(default-)?prompt-router/[a-zA-Z0-9-:.]+$)`

Required: Yes

performanceConfigLatency

Model performance settings for the request.

Valid Values: `standard | optimized`

requestMetadata

Key-value pairs that you can use to filter invocation logs.

Length Constraints: Minimum length of 0. Maximum length of 8500.

serviceTier

Specifies the processing tier type used for serving the request.

Valid Values: `priority | default | flex | reserved`

trace

Specifies whether to enable or disable the Bedrock trace. If enabled, you can see the full Bedrock trace.

Valid Values: `ENABLED | DISABLED | ENABLED_FULL`

Request Body

The request accepts the following binary data.

body

The prompt and inference parameters in the format specified in the `contentType` in the header. You must provide the body in JSON format. To see the format and content of the request and response bodies for different models, refer to Inference parameters. For more information, see Run inference in the Bedrock User Guide.

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Response Syntax

```
HTTP/1.1 200
X-Amzn-Bedrock-Content-Type: contentType
X-Amzn-Bedrock-PerformanceConfig-Latency: performanceConfigLatency
X-Amzn-Bedrock-Service-Tier: serviceTier
Content-type: application/json

{
   "chunk": {
      "bytes": blob
   },
   "internalServerException": {
   },
   "modelStreamErrorException": {
   },
   "modelTimeoutException": {
   },
   "serviceUnavailableException": {
   },
   "throttlingException": {
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType

The MIME type of the inference result.

performanceConfigLatency

Model performance settings for the request.

Valid Values: `standard | optimized`

serviceTier

Specifies the processing tier type used for serving the request.

Valid Values: `priority | default | flex | reserved`

The following data is returned in JSON format by the service.

chunk

Content included in the response.

Type: PayloadPart object

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

modelStreamErrorException

An error occurred while streaming the response. Retry your request.

Type: Exception

HTTP Status Code: 424

modelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

Type: Exception

HTTP Status Code: 408

serviceUnavailableException

The service isn't available. Try again later.

Type: Exception

HTTP Status Code: 503

throttlingException

Your request was throttled because of service-wide limitations. Resubmit your request later or in a different region. You can also purchase Provisioned Throughput to increase the rate or number of tokens you can process.

Type: Exception

HTTP Status Code: 429

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

originalStatusCode

The original status code.

resourceName

The resource name.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

HTTP Status Code: 429

ModelStreamErrorException

An error occurred while streaming the response. Retry your request.

originalMessage

The original message.

originalStatusCode

The original status code.

HTTP Status Code: 424

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

HTTP Status Code: 400

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

Examples

Run inference with streaming on a text model

For streaming, you can set `x-amzn-bedrock-accept-type` in the header to contain the desired content type of the response. In this example, we set it to accept any content type. The default value is `application/json`.

Sample Request

```
POST https://bedrock-runtime.us-east-1.amazonaws.com/model/amazon.titan-text-express-v1/invoke-with-response-stream

-H accept: application/vnd.amazon.eventstream
-H content-type: application/json
-H x-amzn-bedrock-accept: */*

Payload
{"inputText": "Hello world"}
```

Example response

For streaming, the content type in the response is always set to `application/vnd.amazon.eventstream`. The response includes an additional header (x-amzn-bedrock-content-type), which contains the actual content type of the response.

Sample Request

```
-H content-type: application/vnd.amazon.eventstream
-H x-amzn-bedrock-content-type: application/json

Payload (stream events)
<response chunk>
```

Use a guardrail

This examples show how to use a guardrail with `InvokeModelWithResponseStream`.

```
POST /model/modelId/invoke-with-response-stream HTTP/1.1
X-Amzn-Bedrock-Accept: accept
Content-Type: contentType
X-Amzn-Bedrock-GuardrailIdentifier: guardrailIdentifier
X-Amzn-Bedrock-GuardrailVersion: guardrailVersion
X-Amzn-Bedrock-GuardrailTrace: guardrailTrace
X-Amzn-Bedrock-Trace: trace

body

// body
{
    "amazon-bedrock-guardrailConfig": {
        "tagSuffix": "string",
        "streamProcessingMode": "string"
    }
}
```

Example response

This examples shows the response from a call to `InvokeModelWithResponseStream` when using a guardrail.

```
HTTP/1.1 200
X-Amzn-Bedrock-Content-Type: contentType
Content-type: application/json

// chunk 1
{
  "completion": "...",
```

```
   "amazon-bedrock-guardrailAction": "INTERVENED | NONE"
}

// chunk 2
{
  "completion": "...",
  "amazon-bedrock-guardrailAction": "INTERVENED | NONE"
}

// last chunk
{
  "completion": "...",
   "amazon-bedrock-guardrailAction": "INTERVENED | NONE",
  "amazon-bedrock-trace": {
    "guardrail": {
    ... // Detailed guardrail trace
    }
  }
}
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

ListAsyncInvokes

Service: Amazon Bedrock Runtime

Lists asynchronous invocations.

Request Syntax

```
GET /async-invoke?
maxResults=maxResults&nextToken=nextToken&sortBy=sortBy&sortOrder=sortOrder&statusEquals=status
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

maxResults

The maximum number of invocations to return in one page of results.

Valid Range: Minimum value of 1. Maximum value of 1000.

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `\S*`

sortBy

How to sort the response.

Valid Values: `SubmissionTime`

sortOrder

The sorting order for the response.

Valid Values: `Ascending | Descending`

statusEquals

Filter invocations by status.

Valid Values: `InProgress | Completed | Failed`

submitTimeAfter

Include invocations submitted after this time.

submitTimeBefore

Include invocations submitted before this time.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "asyncInvokeSummaries": [
      {
         "clientRequestToken": "string",
         "endTime": "string",
         "failureMessage": "string",
         "invocationArn": "string",
         "lastModifiedTime": "string",
         "modelArn": "string",
         "outputDataConfig": { ... },
         "status": "string",
         "submitTime": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

asyncInvokeSummaries

A list of invocation summaries.

Type: Array of AsyncInvokeSummary objects

nextToken

Specify the pagination token from a previous request to retrieve the next page of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

StartAsyncInvoke

Service: Amazon Bedrock Runtime

Starts an asynchronous invocation.

This operation requires permission for the `bedrock:InvokeModel` action.

> **Important**
>
> To deny all inference access to resources that you specify in the modelId field, you need to deny access to the `bedrock:InvokeModel` and `bedrock:InvokeModelWithResponseStream` actions. Doing this also denies access to the resource through the Converse API actions (Converse and ConverseStream). For more information see Deny access for inference on specific models.

Request Syntax

```
POST /async-invoke HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "modelId": "string",
   "modelInput": JSON value,
   "outputDataConfig": { ... },
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken

Specify idempotency token to ensure that requests are not duplicated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[!-~]*`

Required: No

modelId

The model to invoke.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z_\.\-/0-9:]+`

Required: Yes

modelInput

Input to send to the model.

Type: JSON value

Required: Yes

outputDataConfig

Where to store the output.

Type: AsyncInvokeOutputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

tags

Tags to apply to the invocation.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "invocationArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

invocationArn

The ARN of the invocation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:async-invoke/[a-z0-9]{12}`

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. For troubleshooting this error, see InternalFailure in the Amazon Bedrock User Guide

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide

HTTP Status Code: 404

ServiceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

HTTP Status Code: 400

ServiceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

HTTP Status Code: 503

ThrottlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide

HTTP Status Code: 429

ValidationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby V3

Data Types

The following data types are supported by Amazon Bedrock:

- AccountEnforcedGuardrailInferenceInputConfiguration
- AccountEnforcedGuardrailOutputConfiguration
- AdvancedPromptOptimizationInputConfig
- AdvancedPromptOptimizationJobSummary
- AdvancedPromptOptimizationOutputConfig
- AgreementAvailability
- AutomatedEvaluationConfig
- AutomatedEvaluationCustomMetricConfig
- AutomatedEvaluationCustomMetricSource
- AutomatedReasoningCheckFinding
- AutomatedReasoningCheckImpossibleFinding
- AutomatedReasoningCheckInputTextReference
- AutomatedReasoningCheckInvalidFinding
- AutomatedReasoningCheckLogicWarning
- AutomatedReasoningCheckNoTranslationsFinding
- AutomatedReasoningCheckRule
- AutomatedReasoningCheckSatisfiableFinding
- AutomatedReasoningCheckScenario
- AutomatedReasoningCheckTooComplexFinding
- AutomatedReasoningCheckTranslation
- AutomatedReasoningCheckTranslationAmbiguousFinding
- AutomatedReasoningCheckTranslationOption
- AutomatedReasoningCheckValidFinding
- AutomatedReasoningLogicStatement
- AutomatedReasoningPolicyAddRuleAnnotation
- AutomatedReasoningPolicyAddRuleFromNaturalLanguageAnnotation
- AutomatedReasoningPolicyAddRuleMutation

- AutomatedReasoningPolicyAddTypeAnnotation
- AutomatedReasoningPolicyAddTypeMutation
- AutomatedReasoningPolicyAddTypeValue
- AutomatedReasoningPolicyAddVariableAnnotation
- AutomatedReasoningPolicyAddVariableMutation
- AutomatedReasoningPolicyAnnotatedChunk
- AutomatedReasoningPolicyAnnotatedContent
- AutomatedReasoningPolicyAnnotatedLine
- AutomatedReasoningPolicyAnnotation
- AutomatedReasoningPolicyAtomicStatement
- AutomatedReasoningPolicyBuildLog
- AutomatedReasoningPolicyBuildLogEntry
- AutomatedReasoningPolicyBuildResultAssetManifest
- AutomatedReasoningPolicyBuildResultAssetManifestEntry
- AutomatedReasoningPolicyBuildResultAssets
- AutomatedReasoningPolicyBuildStep
- AutomatedReasoningPolicyBuildStepContext
- AutomatedReasoningPolicyBuildStepMessage
- AutomatedReasoningPolicyBuildWorkflowDocument
- AutomatedReasoningPolicyBuildWorkflowRepairContent
- AutomatedReasoningPolicyBuildWorkflowSource
- AutomatedReasoningPolicyBuildWorkflowSummary
- AutomatedReasoningPolicyDefinition
- AutomatedReasoningPolicyDefinitionElement
- AutomatedReasoningPolicyDefinitionQualityReport
- AutomatedReasoningPolicyDefinitionRule
- AutomatedReasoningPolicyDefinitionType
- AutomatedReasoningPolicyDefinitionTypeValue
- AutomatedReasoningPolicyDefinitionTypeValuePair
- AutomatedReasoningPolicyDefinitionVariable

- AutomatedReasoningPolicyDeleteRuleAnnotation
- AutomatedReasoningPolicyDeleteRuleMutation
- AutomatedReasoningPolicyDeleteTypeAnnotation
- AutomatedReasoningPolicyDeleteTypeMutation
- AutomatedReasoningPolicyDeleteTypeValue
- AutomatedReasoningPolicyDeleteVariableAnnotation
- AutomatedReasoningPolicyDeleteVariableMutation
- AutomatedReasoningPolicyDisjointRuleSet
- AutomatedReasoningPolicyFidelityReport
- AutomatedReasoningPolicyGeneratedTestCase
- AutomatedReasoningPolicyGeneratedTestCases
- AutomatedReasoningPolicyGenerateFidelityReportContent
- AutomatedReasoningPolicyIngestContentAnnotation
- AutomatedReasoningPolicyIterativeRefinementContent
- AutomatedReasoningPolicyMutation
- AutomatedReasoningPolicyPlanning
- AutomatedReasoningPolicyReportSourceDocument
- AutomatedReasoningPolicyRuleReport
- AutomatedReasoningPolicyScenario
- AutomatedReasoningPolicyScenarios
- AutomatedReasoningPolicySourceDocument
- AutomatedReasoningPolicyStatementLocation
- AutomatedReasoningPolicyStatementReference
- AutomatedReasoningPolicySummary
- AutomatedReasoningPolicyTestCase
- AutomatedReasoningPolicyTestResult
- AutomatedReasoningPolicyTypeValueAnnotation
- AutomatedReasoningPolicyUpdateFromRuleFeedbackAnnotation
- AutomatedReasoningPolicyUpdateFromScenarioFeedbackAnnotation
- AutomatedReasoningPolicyUpdateRuleAnnotation

- AutomatedReasoningPolicyUpdateRuleMutation
- AutomatedReasoningPolicyUpdateTypeAnnotation
- AutomatedReasoningPolicyUpdateTypeMutation
- AutomatedReasoningPolicyUpdateTypeValue
- AutomatedReasoningPolicyUpdateVariableAnnotation
- AutomatedReasoningPolicyUpdateVariableMutation
- AutomatedReasoningPolicyVariableReport
- AutomatedReasoningPolicyWorkflowTypeContent
- BatchDeleteAdvancedPromptOptimizationJobError
- BatchDeleteAdvancedPromptOptimizationJobItem
- BatchDeleteEvaluationJobError
- BatchDeleteEvaluationJobItem
- BedrockEvaluatorModel
- ByteContentDoc
- CloudWatchConfig
- CustomizationConfig
- CustomMetricBedrockEvaluatorModel
- CustomMetricDefinition
- CustomMetricEvaluatorModelConfig
- CustomModelDataSource
- CustomModelDeploymentSummary
- CustomModelDeploymentUpdateDetails
- CustomModelSummary
- CustomModelUnits
- DataProcessingDetails
- DimensionalPriceRate
- DistillationConfig
- EndpointConfig
- EvaluationBedrockModel
- EvaluationConfig

- EvaluationDataset
- EvaluationDatasetLocation
- EvaluationDatasetMetricConfig
- EvaluationInferenceConfig
- EvaluationInferenceConfigSummary
- EvaluationModelConfig
- EvaluationModelConfigSummary
- EvaluationOutputDataConfig
- EvaluationPrecomputedInferenceSource
- EvaluationPrecomputedRagSourceConfig
- EvaluationPrecomputedRetrieveAndGenerateSourceConfig
- EvaluationPrecomputedRetrieveSourceConfig
- EvaluationRagConfigSummary
- EvaluationSummary
- EvaluatorModelConfig
- ExternalSource
- ExternalSourcesGenerationConfiguration
- ExternalSourcesRetrieveAndGenerateConfiguration
- FieldForReranking
- FilterAttribute
- FoundationModelDetails
- FoundationModelLifecycle
- FoundationModelSummary
- GenerationConfiguration
- GraderConfig
- GuardrailAutomatedReasoningPolicy
- GuardrailAutomatedReasoningPolicyConfig
- GuardrailConfiguration
- GuardrailContentFilter
- GuardrailContentFilterConfig

- GuardrailContentFiltersTier
- GuardrailContentFiltersTierConfig
- GuardrailContentPolicy
- GuardrailContentPolicyConfig
- GuardrailContextualGroundingFilter
- GuardrailContextualGroundingFilterConfig
- GuardrailContextualGroundingPolicy
- GuardrailContextualGroundingPolicyConfig
- GuardrailCrossRegionConfig
- GuardrailCrossRegionDetails
- GuardrailManagedWords
- GuardrailManagedWordsConfig
- GuardrailPiiEntity
- GuardrailPiiEntityConfig
- GuardrailRegex
- GuardrailRegexConfig
- GuardrailSensitiveInformationPolicy
- GuardrailSensitiveInformationPolicyConfig
- GuardrailSummary
- GuardrailTopic
- GuardrailTopicConfig
- GuardrailTopicPolicy
- GuardrailTopicPolicyConfig
- GuardrailTopicsTier
- GuardrailTopicsTierConfig
- GuardrailWord
- GuardrailWordConfig
- GuardrailWordPolicy
- GuardrailWordPolicyConfig
- HumanEvaluationConfig

- HumanEvaluationCustomMetric
- HumanWorkflowConfig
- ImplicitFilterConfiguration
- ImportedModelSummary
- InferenceConfiguration
- InferenceProfileModel
- InferenceProfileModelSource
- InferenceProfileSummary
- InvocationLogsConfig
- InvocationLogSource
- KbInferenceConfig
- KnowledgeBaseConfig
- KnowledgeBaseRetrievalConfiguration
- KnowledgeBaseRetrieveAndGenerateConfiguration
- KnowledgeBaseVectorSearchConfiguration
- LambdaGraderConfig
- LegalTerm
- LoggingConfig
- MarketplaceModelEndpoint
- MarketplaceModelEndpointSummary
- MetadataAttributeSchema
- MetadataConfigurationForReranking
- ModelConfiguration
- ModelCopyJobSummary
- ModelCustomizationJobSummary
- ModelDataSource
- ModelEnforcement
- ModelImportJobSummary
- ModelInvocationJobInputDataConfig
- ModelInvocationJobOutputDataConfig

- ModelInvocationJobS3InputDataConfig
- ModelInvocationJobS3OutputDataConfig
- ModelInvocationJobSummary
- ModelPackageArnDataSource
- Offer
- OrchestrationConfiguration
- OutputDataConfig
- PerformanceConfiguration
- PricingTerm
- PromptRouterSummary
- PromptRouterTargetModel
- PromptTemplate
- ProvisionedModelSummary
- QueryTransformationConfiguration
- RAGConfig
- RatingScaleItem
- RatingScaleItemValue
- RequestMetadataBaseFilters
- RequestMetadataFilters
- RerankingMetadataSelectiveModeConfiguration
- RetrievalFilter
- RetrieveAndGenerateConfiguration
- RetrieveConfig
- RFTConfig
- RFTHyperParameters
- RoutingCriteria
- S3Config
- S3DataSource
- S3ObjectDoc
- SageMakerEndpoint

- SelectiveContentGuarding
- StatusDetails
- SupportTerm
- Tag
- TeacherModelConfig
- TermDetails
- TextInferenceConfig
- TrainingDataConfig
- TrainingDetails
- TrainingMetrics
- ValidationDataConfig
- ValidationDetails
- Validator
- ValidatorMetric
- ValidityTerm
- VectorSearchBedrockRerankingConfiguration
- VectorSearchBedrockRerankingModelConfiguration
- VectorSearchRerankingConfiguration
- VpcConfig

The following data types are supported by Agents for Amazon Bedrock:

- ActionGroupExecutor
- ActionGroupSummary
- Agent
- AgentActionGroup
- AgentAlias
- AgentAliasHistoryEvent
- AgentAliasRoutingConfigurationListItem
- AgentAliasSummary
- AgentCollaborator

- AgentCollaboratorSummary
- AgentDescriptor
- AgentFlowNodeConfiguration
- AgentKnowledgeBase
- AgentKnowledgeBaseSummary
- AgentSummary
- AgentVersion
- AgentVersionSummary
- AnyToolChoice
- APISchema
- AudioConfiguration
- AudioExtractionConfiguration
- AudioSegmentationConfiguration
- AutoToolChoice
- BedrockDataAutomationConfiguration
- BedrockEmbeddingModelConfiguration
- BedrockFoundationModelConfiguration
- BedrockFoundationModelContextEnrichmentConfiguration
- ByteContentDoc
- CachePointBlock
- ChatPromptTemplateConfiguration
- ChunkingConfiguration
- CollectorFlowNodeConfiguration
- ConditionFlowNodeConfiguration
- ConfluenceCrawlerConfiguration
- ConfluenceDataSourceConfiguration
- ConfluenceSourceConfiguration
- ContentBlock
- ContextEnrichmentConfiguration
- CrawlFilterConfiguration

- CuratedQuery
- CustomContent
- CustomDocumentIdentifier
- CustomOrchestration
- CustomS3Location
- CustomTransformationConfiguration
- CyclicConnectionFlowValidationDetails
- DailySchedule
- DataSource
- DataSourceConfiguration
- DataSourceSummary
- DayOfMonth
- DeletionProtectionConfiguration
- DocumentAccessControlEntry
- DocumentContent
- DocumentIdentifier
- DocumentMetadata
- DuplicateConditionExpressionFlowValidationDetails
- DuplicateConnectionsFlowValidationDetails
- EmbeddingModelConfiguration
- EnrichmentStrategyConfiguration
- FieldForReranking
- FixedSizeChunkingConfiguration
- FlowAliasConcurrencyConfiguration
- FlowAliasRoutingConfigurationListItem
- FlowAliasSummary
- FlowCondition
- FlowConditionalConnectionConfiguration
- FlowConnection
- FlowConnectionConfiguration

- FlowDataConnectionConfiguration
- FlowDefinition
- FlowNode
- FlowNodeConfiguration
- FlowNodeInput
- FlowNodeOutput
- FlowSummary
- FlowValidation
- FlowValidationDetails
- FlowVersionSummary
- Function
- FunctionSchema
- GuardrailConfiguration
- HierarchicalChunkingConfiguration
- HierarchicalChunkingLevelConfiguration
- ImageExtractionConfiguration
- IncompatibleConnectionDataTypeFlowValidationDetails
- InferenceConfiguration
- IngestionJob
- IngestionJobFilter
- IngestionJobSortBy
- IngestionJobStatistics
- IngestionJobSummary
- InlineCodeFlowNodeConfiguration
- InlineContent
- InputFlowNodeConfiguration
- IntermediateStorage
- InvalidLoopBoundaryFlowValidationDetails
- IteratorFlowNodeConfiguration
- KendraKnowledgeBaseConfiguration

- KnowledgeBase
- KnowledgeBaseConfiguration
- KnowledgeBaseDocument
- KnowledgeBaseDocumentDetail
- KnowledgeBaseFlowNodeConfiguration
- KnowledgeBaseOrchestrationConfiguration
- KnowledgeBasePromptTemplate
- KnowledgeBaseSummary
- LambdaFunctionFlowNodeConfiguration
- LastDayOfMonth
- LexFlowNodeConfiguration
- LoopControllerFlowNodeConfiguration
- LoopFlowNodeConfiguration
- LoopIncompatibleNodeTypeFlowValidationDetails
- LoopInputFlowNodeConfiguration
- MalformedConditionExpressionFlowValidationDetails
- MalformedNodeInputExpressionFlowValidationDetails
- ManagedKnowledgeBaseConfiguration
- ManagedKnowledgeBaseConnectorConfiguration
- MediaExtractionConfiguration
- MemoryConfiguration
- Message
- MetadataAttribute
- MetadataAttributeValue
- MetadataConfigurationForReranking
- MismatchedNodeInputTypeFlowValidationDetails
- MismatchedNodeOutputTypeFlowValidationDetails
- MissingConnectionConfigurationFlowValidationDetails
- MissingDefaultConditionFlowValidationDetails
- MissingEndingNodesFlowValidationDetails

- MissingLoopControllerNodeFlowValidationDetails
- MissingLoopInputNodeFlowValidationDetails
- MissingNodeConfigurationFlowValidationDetails
- MissingNodeInputFlowValidationDetails
- MissingNodeOutputFlowValidationDetails
- MissingStartingNodesFlowValidationDetails
- MongoDbAtlasConfiguration
- MongoDbAtlasFieldMapping
- MonthlySchedule
- MultipleLoopControllerNodesFlowValidationDetails
- MultipleLoopInputNodesFlowValidationDetails
- MultipleNodeInputConnectionsFlowValidationDetails
- NeptuneAnalyticsConfiguration
- NeptuneAnalyticsFieldMapping
- OpenSearchManagedClusterConfiguration
- OpenSearchManagedClusterFieldMapping
- OpenSearchServerlessConfiguration
- OpenSearchServerlessFieldMapping
- OrchestrationExecutor
- OutputFlowNodeConfiguration
- ParameterDetail
- ParsingConfiguration
- ParsingPrompt
- PatternObjectFilter
- PatternObjectFilterConfiguration
- PerformanceConfiguration
- PineconeConfiguration
- PineconeFieldMapping
- PromptAgentResource
- PromptConfiguration

- PromptFlowNodeConfiguration
- PromptFlowNodeInlineConfiguration
- PromptFlowNodeResourceConfiguration
- PromptFlowNodeSourceConfiguration
- PromptGenAiResource
- PromptInferenceConfiguration
- PromptInputVariable
- PromptMetadataEntry
- PromptModelInferenceConfiguration
- PromptOverrideConfiguration
- PromptSummary
- PromptTemplateConfiguration
- PromptVariant
- QueryGenerationColumn
- QueryGenerationConfiguration
- QueryGenerationContext
- QueryGenerationTable
- RdsConfiguration
- RdsFieldMapping
- RedisEnterpriseCloudConfiguration
- RedisEnterpriseCloudFieldMapping
- RedshiftConfiguration
- RedshiftProvisionedAuthConfiguration
- RedshiftProvisionedConfiguration
- RedshiftQueryEngineAwsDataCatalogStorageConfiguration
- RedshiftQueryEngineConfiguration
- RedshiftQueryEngineRedshiftStorageConfiguration
- RedshiftQueryEngineStorageConfiguration
- RedshiftServerlessAuthConfiguration
- RedshiftServerlessConfiguration

- RerankingMetadataSelectiveModeConfiguration
- RetrievalFlowNodeConfiguration
- RetrievalFlowNodeS3Configuration
- RetrievalFlowNodeServiceConfiguration
- S3Content
- S3DataSourceConfiguration
- S3Identifier
- S3Location
- S3VectorsConfiguration
- SalesforceCrawlerConfiguration
- SalesforceDataSourceConfiguration
- SalesforceSourceConfiguration
- SeedUrl
- SemanticChunkingConfiguration
- ServerSideEncryptionConfiguration
- SessionSummaryConfiguration
- SharePointCrawlerConfiguration
- SharePointDataSourceConfiguration
- SharePointSourceConfiguration
- SpecificToolChoice
- SqlKnowledgeBaseConfiguration
- StorageConfiguration
- StorageFlowNodeConfiguration
- StorageFlowNodeS3Configuration
- StorageFlowNodeServiceConfiguration
- SupplementalDataStorageConfiguration
- SupplementalDataStorageLocation
- SyncSchedule
- SystemContentBlock
- TextContentDoc

- TextPromptTemplateConfiguration
- Tool
- ToolChoice
- ToolConfiguration
- ToolInputSchema
- ToolSpecification
- Transformation
- TransformationFunction
- TransformationLambdaConfiguration
- UnfulfilledNodeInputFlowValidationDetails
- UnknownConnectionConditionFlowValidationDetails
- UnknownConnectionSourceFlowValidationDetails
- UnknownConnectionSourceOutputFlowValidationDetails
- UnknownConnectionTargetFlowValidationDetails
- UnknownConnectionTargetInputFlowValidationDetails
- UnknownNodeInputFlowValidationDetails
- UnknownNodeOutputFlowValidationDetails
- UnreachableNodeFlowValidationDetails
- UnsatisfiedConnectionConditionsFlowValidationDetails
- UnspecifiedFlowValidationDetails
- UrlConfiguration
- ValidationExceptionField
- VectorIngestionConfiguration
- VectorKnowledgeBaseConfiguration
- VectorSearchBedrockRerankingConfiguration
- VectorSearchBedrockRerankingModelConfiguration
- VectorSearchRerankingConfiguration
- VideoConfiguration
- VideoExtractionConfiguration
- VideoSegmentationConfiguration

- VpcConfiguration
- VpcConfigurationSummary
- WebCrawlerConfiguration
- WebCrawlerLimits
- WebDataSourceConfiguration
- WebSourceConfiguration
- WeeklySchedule

The following data types are supported by Agents for Amazon Bedrock Runtime:

- ActionGroupExecutor
- ActionGroupInvocationInput
- ActionGroupInvocationOutput
- AgentActionGroup
- AgentCollaboratorInputPayload
- AgentCollaboratorInvocationInput
- AgentCollaboratorInvocationOutput
- AgentCollaboratorOutputPayload
- AgenticRetrieveAction
- AgenticRetrieveActionDetails
- AgenticRetrieveBedrockGuardrailConfiguration
- AgenticRetrieveBedrockRerankingConfiguration
- AgenticRetrieveBedrockRerankingModelConfiguration
- AgenticRetrieveCitation
- AgenticRetrieveCitationReference
- AgenticRetrieveConfiguration
- AgenticRetrieveFailure
- AgenticRetrieveFullDocExpansionDetails
- AgenticRetrieveGeneratedResponse
- AgenticRetrieveGuardrailWarning
- AgenticRetrieveMemoryConfiguration

- AgenticRetrieveMemoryMetadataFilter
- AgenticRetrieveMemoryMetadataFilterLeft
- AgenticRetrieveMemoryMetadataFilterRight
- AgenticRetrieveMemoryMetadataValue
- AgenticRetrieveMemoryRetrievalConfig
- AgenticRetrieveMemoryRetrieveDetails
- AgenticRetrieveMemorySessionBinding
- AgenticRetrieveMessage
- AgenticRetrieveMessageContent
- AgenticRetrievePolicyConfiguration
- AgenticRetriever
- AgenticRetrieveRerankingConfiguration
- AgenticRetrieveResponseEvent
- AgenticRetrieveResultEvent
- AgenticRetrieveResultItem
- AgenticRetrieveSourceMetadata
- AgenticRetrieveSourceRetriever
- AgenticRetrieveStreamResponseOutput
- AgenticRetrieveTraceEvent
- AgenticRetrieveTraceEventAttributes
- AgenticRetrieveTraceResultItem
- AgenticRetrieveWarning
- AgenticRetrieveWarningMessage
- AnalyzePromptEvent
- ApiInvocationInput
- ApiParameter
- ApiRequestBody
- ApiResult
- APISchema
- Attribution

- AudioSegment
- BedrockFoundationModelConfiguration
- BedrockFoundationModelModelConfiguration
- BedrockModelConfigurations
- BedrockRerankingConfiguration
- BedrockRerankingModelConfiguration
- BedrockSessionContentBlock
- ByteContentDoc
- ByteContentFile
- Caller
- Citation
- CitationEvent
- CodeInterpreterInvocationInput
- CodeInterpreterInvocationOutput
- Collaborator
- CollaboratorConfiguration
- ConditionResultEvent
- ContentBlock
- ContentBody
- ConversationHistory
- CustomOrchestration
- CustomOrchestrationTrace
- CustomOrchestrationTraceEvent
- DocumentAcl
- DocumentAclCondition
- DocumentAclGroup
- DocumentAclMembership
- DocumentAclUser
- ExternalSource
- ExternalSourcesGenerationConfiguration

- ExternalSourcesRetrieveAndGenerateConfiguration
- FailureTrace
- FieldForReranking
- FilePart
- FileSource
- FilterAttribute
- FinalResponse
- FlowCompletionEvent
- FlowExecutionContent
- FlowExecutionError
- FlowExecutionEvent
- FlowExecutionInputEvent
- FlowExecutionOutputEvent
- FlowExecutionSummary
- FlowFailureEvent
- FlowInput
- FlowInputContent
- FlowInputField
- FlowMultiTurnInputContent
- FlowMultiTurnInputRequestEvent
- FlowOutputContent
- FlowOutputEvent
- FlowOutputField
- FlowResponseStream
- FlowTrace
- FlowTraceCondition
- FlowTraceConditionNodeResultEvent
- FlowTraceDependencyEvent
- FlowTraceEvent
- FlowTraceNodeActionEvent

- FlowTraceNodeInputContent
- FlowTraceNodeInputEvent
- FlowTraceNodeInputExecutionChainItem
- FlowTraceNodeInputField
- FlowTraceNodeInputSource
- FlowTraceNodeOutputContent
- FlowTraceNodeOutputEvent
- FlowTraceNodeOutputField
- FlowTraceNodeOutputNext
- FoundationModelConfiguration
- FunctionDefinition
- FunctionInvocationInput
- FunctionParameter
- FunctionResult
- FunctionSchema
- GeneratedQuery
- GeneratedResponsePart
- GenerationConfiguration
- GuardrailAssessment
- GuardrailConfiguration
- GuardrailConfigurationWithArn
- GuardrailContentFilter
- GuardrailContentPolicyAssessment
- GuardrailCustomWord
- GuardrailEvent
- GuardrailManagedWord
- GuardrailPiiEntityFilter
- GuardrailRegexFilter
- GuardrailSensitiveInformationPolicyAssessment
- GuardrailTopic

- GuardrailTopicPolicyAssessment
- GuardrailTrace
- GuardrailWordPolicyAssessment
- ImageBlock
- ImageInput
- ImageInputSource
- ImageSource
- ImplicitFilterConfiguration
- InferenceConfig
- InferenceConfiguration
- InlineAgentFilePart
- InlineAgentPayloadPart
- InlineAgentResponseStream
- InlineAgentReturnControlPayload
- InlineAgentTracePart
- InlineBedrockModelConfigurations
- InlineSessionState
- InputFile
- InputImage
- InputPrompt
- InvocationInput
- InvocationInputMember
- InvocationResultMember
- InvocationStep
- InvocationStepPayload
- InvocationStepSummary
- InvocationSummary
- KnowledgeBase
- KnowledgeBaseConfiguration
- KnowledgeBaseLookupInput

- KnowledgeBaseLookupOutput
- KnowledgeBaseQuery
- KnowledgeBaseRetrievalConfiguration
- KnowledgeBaseRetrievalResult
- KnowledgeBaseRetrieveAndGenerateConfiguration
- KnowledgeBaseRetrieverConfiguration
- KnowledgeBaseVectorSearchConfiguration
- ManagedSearchBedrockRerankingConfiguration
- ManagedSearchBedrockRerankingModelConfiguration
- ManagedSearchConfiguration
- ManagedSearchRerankingConfiguration
- Memory
- MemorySessionSummary
- Message
- Metadata
- MetadataAttributeSchema
- MetadataConfigurationForReranking
- ModelInvocationInput
- ModelPerformanceConfiguration
- NodeActionEvent
- NodeDependencyEvent
- NodeExecutionContent
- NodeFailureEvent
- NodeInputEvent
- NodeInputExecutionChainItem
- NodeInputField
- NodeInputSource
- NodeOutputEvent
- NodeOutputField
- NodeOutputNext

- NodeTraceElements
- Observation
- OptimizedPrompt
- OptimizedPromptEvent
- OptimizedPromptStream
- OrchestrationConfiguration
- OrchestrationExecutor
- OrchestrationModelInvocationOutput
- OrchestrationTrace
- OutputFile
- Parameter
- ParameterDetail
- PayloadPart
- PerformanceConfiguration
- PostProcessingModelInvocationOutput
- PostProcessingParsedResponse
- PostProcessingTrace
- PreProcessingModelInvocationOutput
- PreProcessingParsedResponse
- PreProcessingTrace
- PromptConfiguration
- PromptCreationConfigurations
- PromptOverrideConfiguration
- PromptTemplate
- PropertyParameters
- QueryGenerationInput
- QueryTransformationConfiguration
- Rationale
- RawResponse
- ReasoningContentBlock

- ReasoningTextBlock
- RepromptResponse
- RequestBody
- RerankDocument
- RerankingConfiguration
- RerankingMetadataSelectiveModeConfiguration
- RerankQuery
- RerankResult
- RerankSource
- RerankTextDocument
- ResponseStream
- RetrievalContent
- RetrievalFilter
- RetrievalOverrides
- RetrievalResultConfluenceLocation
- RetrievalResultContent
- RetrievalResultContentColumn
- RetrievalResultCustomDocumentLocation
- RetrievalResultGoogleDriveLocation
- RetrievalResultKendraDocumentLocation
- RetrievalResultLocation
- RetrievalResultOneDriveLocation
- RetrievalResultS3Location
- RetrievalResultSalesforceLocation
- RetrievalResultSharePointLocation
- RetrievalResultSqlLocation
- RetrievalResultWebLocation
- RetrieveAndGenerateConfiguration
- RetrieveAndGenerateInput
- RetrieveAndGenerateOutput

- RetrieveAndGenerateOutputEvent
- RetrieveAndGenerateSessionConfiguration
- RetrieveAndGenerateStreamResponseOutput
- RetrievedReference
- RetrieverConfiguration
- ReturnControlPayload
- ReturnControlResults
- RoutingClassifierModelInvocationOutput
- RoutingClassifierTrace
- S3Identifier
- S3Location
- S3ObjectDoc
- S3ObjectFile
- SatisfiedCondition
- SessionState
- SessionSummary
- Span
- StreamingConfigurations
- TextInferenceConfig
- TextPrompt
- TextResponsePart
- TextToSqlConfiguration
- TextToSqlKnowledgeBaseConfiguration
- Trace
- TraceElements
- TracePart
- TransformationConfiguration
- Usage
- UserContext
- VectorSearchBedrockRerankingConfiguration

- VectorSearchBedrockRerankingModelConfiguration
- VectorSearchRerankingConfiguration
- VideoSegment

The following data types are supported by Data Automation for Amazon Bedrock:

- AudioExtractionCategory
- AudioExtractionCategoryTypeConfiguration
- AudioLanguageConfiguration
- AudioOverrideConfiguration
- AudioStandardExtraction
- AudioStandardGenerativeField
- AudioStandardOutputConfiguration
- Blueprint
- BlueprintFilter
- BlueprintItem
- BlueprintOptimizationObject
- BlueprintOptimizationOutputConfiguration
- BlueprintOptimizationSample
- BlueprintSummary
- ChannelLabelingConfiguration
- CustomOutputConfiguration
- DataAutomationLibrary
- DataAutomationLibraryConfiguration
- DataAutomationLibraryEntitySummary
- DataAutomationLibraryFilter
- DataAutomationLibraryIngestionJob
- DataAutomationLibraryIngestionJobSummary
- DataAutomationLibraryItem
- DataAutomationLibrarySummary
- DataAutomationProject

- DataAutomationProjectFilter
- DataAutomationProjectSummary
- DeleteEntitiesInfo
- DocumentBoundingBox
- DocumentCustomOutputConfiguration
- DocumentExtractionGranularity
- DocumentOutputAdditionalFileFormat
- DocumentOutputFormat
- DocumentOutputTextFormat
- DocumentOverrideConfiguration
- DocumentStandardExtraction
- DocumentStandardGenerativeField
- DocumentStandardOutputConfiguration
- EncryptionConfiguration
- EntityDetails
- EntityTypeInfo
- EventBridgeConfiguration
- ImageBoundingBox
- ImageExtractionCategory
- ImageOverrideConfiguration
- ImageStandardExtraction
- ImageStandardGenerativeField
- ImageStandardOutputConfiguration
- InlinePayload
- InputConfiguration
- ModalityProcessingConfiguration
- ModalityRoutingConfiguration
- NotificationConfiguration
- OutputConfiguration
- OverrideConfiguration

- Phrase
- PIIEntitiesConfiguration
- S3Object
- SensitiveDataConfiguration
- SpeakerLabelingConfiguration
- SplitterConfiguration
- StandardOutputConfiguration
- Tag
- TranscriptConfiguration
- UpsertEntityInfo
- ValidationExceptionField
- VideoBoundingBox
- VideoExtractionCategory
- VideoOverrideConfiguration
- VideoStandardExtraction
- VideoStandardGenerativeField
- VideoStandardOutputConfiguration
- VocabularyEntity
- VocabularyEntityInfo
- VocabularyEntitySummary

The following data types are supported by Runtime for Amazon Bedrock Data Automation:

- AssetProcessingConfiguration
- Blueprint
- DataAutomationConfiguration
- EncryptionConfiguration
- EventBridgeConfiguration
- InputConfiguration
- NotificationConfiguration
- OutputConfiguration

- OutputSegment
- SyncInputConfiguration
- Tag
- TimestampSegment
- VideoAssetProcessingConfiguration
- VideoSegmentConfiguration

The following data types are supported by Amazon Bedrock Runtime:

- AnyToolChoice
- AppliedGuardrailDetails
- AsyncInvokeOutputDataConfig
- AsyncInvokeS3OutputDataConfig
- AsyncInvokeSummary
- AudioBlock
- AudioSource
- AutoToolChoice
- BidirectionalInputPayloadPart
- BidirectionalOutputPayloadPart
- CacheDetail
- CachePointBlock
- Citation
- CitationGeneratedContent
- CitationLocation
- CitationsConfig
- CitationsContentBlock
- CitationsDelta
- CitationSourceContent
- CitationSourceContentDelta
- ContentBlock
- ContentBlockDelta

- ContentBlockDeltaEvent
- ContentBlockStart
- ContentBlockStartEvent
- ContentBlockStopEvent
- ConverseMetrics
- ConverseOutput
- ConverseStreamMetadataEvent
- ConverseStreamMetrics
- ConverseStreamOutput
- ConverseStreamTrace
- ConverseTokensRequest
- ConverseTrace
- CountTokensInput
- DocumentBlock
- DocumentCharLocation
- DocumentChunkLocation
- DocumentContentBlock
- DocumentPageLocation
- DocumentSource
- ErrorBlock
- GuardrailAssessment
- GuardrailAutomatedReasoningFinding
- GuardrailAutomatedReasoningImpossibleFinding
- GuardrailAutomatedReasoningInputTextReference
- GuardrailAutomatedReasoningInvalidFinding
- GuardrailAutomatedReasoningLogicWarning
- GuardrailAutomatedReasoningNoTranslationsFinding
- GuardrailAutomatedReasoningPolicyAssessment
- GuardrailAutomatedReasoningRule
- GuardrailAutomatedReasoningSatisfiableFinding

- GuardrailAutomatedReasoningScenario
- GuardrailAutomatedReasoningStatement
- GuardrailAutomatedReasoningTooComplexFinding
- GuardrailAutomatedReasoningTranslation
- GuardrailAutomatedReasoningTranslationAmbiguousFinding
- GuardrailAutomatedReasoningTranslationOption
- GuardrailAutomatedReasoningValidFinding
- GuardrailChecksConfig
- GuardrailChecksContentBlock
- GuardrailChecksContentFilterCategoryConfig
- GuardrailChecksContentFilterConfig
- GuardrailChecksContentFilterResult
- GuardrailChecksContentFilterResultEntry
- GuardrailChecksContentFilterUsage
- GuardrailChecksMessage
- GuardrailChecksPromptAttackCategoryConfig
- GuardrailChecksPromptAttackConfig
- GuardrailChecksPromptAttackResult
- GuardrailChecksPromptAttackResultEntry
- GuardrailChecksPromptAttackUsage
- GuardrailChecksResults
- GuardrailChecksSensitiveInformationConfig
- GuardrailChecksSensitiveInformationEntityConfig
- GuardrailChecksSensitiveInformationResult
- GuardrailChecksSensitiveInformationResultEntry
- GuardrailChecksSensitiveInformationUsage
- GuardrailChecksUsageResults
- GuardrailConfiguration
- GuardrailContentBlock
- GuardrailContentFilter

- GuardrailContentPolicyAssessment
- GuardrailContextualGroundingFilter
- GuardrailContextualGroundingPolicyAssessment
- GuardrailConverseContentBlock
- GuardrailConverseImageBlock
- GuardrailConverseImageSource
- GuardrailConverseTextBlock
- GuardrailCoverage
- GuardrailCustomWord
- GuardrailImageBlock
- GuardrailImageCoverage
- GuardrailImageSource
- GuardrailInvocationMetrics
- GuardrailManagedWord
- GuardrailOutputContent
- GuardrailPiiEntityFilter
- GuardrailRegexFilter
- GuardrailSensitiveInformationPolicyAssessment
- GuardrailStreamConfiguration
- GuardrailTextBlock
- GuardrailTextCharactersCoverage
- GuardrailTopic
- GuardrailTopicPolicyAssessment
- GuardrailTraceAssessment
- GuardrailUsage
- GuardrailWordPolicyAssessment
- ImageBlock
- ImageBlockDelta
- ImageBlockStart
- ImageSource

- InferenceConfiguration
- InvokeModelTokensRequest
- InvokeModelWithBidirectionalStreamInput
- InvokeModelWithBidirectionalStreamOutput
- JsonSchemaDefinition
- Message
- MessageStartEvent
- MessageStopEvent
- OutputConfig
- OutputFormat
- OutputFormatStructure
- PayloadPart
- PerformanceConfiguration
- PromptRouterTrace
- PromptVariableValues
- ReasoningContentBlock
- ReasoningContentBlockDelta
- ReasoningTextBlock
- ResponseStream
- S3Location
- SearchResultBlock
- SearchResultContentBlock
- SearchResultLocation
- ServiceTier
- SpecificToolChoice
- SystemContentBlock
- SystemTool
- Tag
- TokenUsage
- Tool

- ToolChoice
- ToolConfiguration
- ToolInputSchema
- ToolResultBlock
- ToolResultBlockDelta
- ToolResultBlockStart
- ToolResultContentBlock
- ToolSpecification
- ToolUseBlock
- ToolUseBlockDelta
- ToolUseBlockStart
- VideoBlock
- VideoSource
- WebLocation

Amazon Bedrock

The following data types are supported by Amazon Bedrock:

- AccountEnforcedGuardrailInferenceInputConfiguration
- AccountEnforcedGuardrailOutputConfiguration
- AdvancedPromptOptimizationInputConfig
- AdvancedPromptOptimizationJobSummary
- AdvancedPromptOptimizationOutputConfig
- AgreementAvailability
- AutomatedEvaluationConfig
- AutomatedEvaluationCustomMetricConfig
- AutomatedEvaluationCustomMetricSource
- AutomatedReasoningCheckFinding
- AutomatedReasoningCheckImpossibleFinding
- AutomatedReasoningCheckInputTextReference

- AutomatedReasoningCheckInvalidFinding
- AutomatedReasoningCheckLogicWarning
- AutomatedReasoningCheckNoTranslationsFinding
- AutomatedReasoningCheckRule
- AutomatedReasoningCheckSatisfiableFinding
- AutomatedReasoningCheckScenario
- AutomatedReasoningCheckTooComplexFinding
- AutomatedReasoningCheckTranslation
- AutomatedReasoningCheckTranslationAmbiguousFinding
- AutomatedReasoningCheckTranslationOption
- AutomatedReasoningCheckValidFinding
- AutomatedReasoningLogicStatement
- AutomatedReasoningPolicyAddRuleAnnotation
- AutomatedReasoningPolicyAddRuleFromNaturalLanguageAnnotation
- AutomatedReasoningPolicyAddRuleMutation
- AutomatedReasoningPolicyAddTypeAnnotation
- AutomatedReasoningPolicyAddTypeMutation
- AutomatedReasoningPolicyAddTypeValue
- AutomatedReasoningPolicyAddVariableAnnotation
- AutomatedReasoningPolicyAddVariableMutation
- AutomatedReasoningPolicyAnnotatedChunk
- AutomatedReasoningPolicyAnnotatedContent
- AutomatedReasoningPolicyAnnotatedLine
- AutomatedReasoningPolicyAnnotation
- AutomatedReasoningPolicyAtomicStatement
- AutomatedReasoningPolicyBuildLog
- AutomatedReasoningPolicyBuildLogEntry
- AutomatedReasoningPolicyBuildResultAssetManifest
- AutomatedReasoningPolicyBuildResultAssetManifestEntry
- AutomatedReasoningPolicyBuildResultAssets

- AutomatedReasoningPolicyBuildStep
- AutomatedReasoningPolicyBuildStepContext
- AutomatedReasoningPolicyBuildStepMessage
- AutomatedReasoningPolicyBuildWorkflowDocument
- AutomatedReasoningPolicyBuildWorkflowRepairContent
- AutomatedReasoningPolicyBuildWorkflowSource
- AutomatedReasoningPolicyBuildWorkflowSummary
- AutomatedReasoningPolicyDefinition
- AutomatedReasoningPolicyDefinitionElement
- AutomatedReasoningPolicyDefinitionQualityReport
- AutomatedReasoningPolicyDefinitionRule
- AutomatedReasoningPolicyDefinitionType
- AutomatedReasoningPolicyDefinitionTypeValue
- AutomatedReasoningPolicyDefinitionTypeValuePair
- AutomatedReasoningPolicyDefinitionVariable
- AutomatedReasoningPolicyDeleteRuleAnnotation
- AutomatedReasoningPolicyDeleteRuleMutation
- AutomatedReasoningPolicyDeleteTypeAnnotation
- AutomatedReasoningPolicyDeleteTypeMutation
- AutomatedReasoningPolicyDeleteTypeValue
- AutomatedReasoningPolicyDeleteVariableAnnotation
- AutomatedReasoningPolicyDeleteVariableMutation
- AutomatedReasoningPolicyDisjointRuleSet
- AutomatedReasoningPolicyFidelityReport
- AutomatedReasoningPolicyGeneratedTestCase
- AutomatedReasoningPolicyGeneratedTestCases
- AutomatedReasoningPolicyGenerateFidelityReportContent
- AutomatedReasoningPolicyIngestContentAnnotation
- AutomatedReasoningPolicyIterativeRefinementContent
- AutomatedReasoningPolicyMutation

- AutomatedReasoningPolicyPlanning
- AutomatedReasoningPolicyReportSourceDocument
- AutomatedReasoningPolicyRuleReport
- AutomatedReasoningPolicyScenario
- AutomatedReasoningPolicyScenarios
- AutomatedReasoningPolicySourceDocument
- AutomatedReasoningPolicyStatementLocation
- AutomatedReasoningPolicyStatementReference
- AutomatedReasoningPolicySummary
- AutomatedReasoningPolicyTestCase
- AutomatedReasoningPolicyTestResult
- AutomatedReasoningPolicyTypeValueAnnotation
- AutomatedReasoningPolicyUpdateFromRuleFeedbackAnnotation
- AutomatedReasoningPolicyUpdateFromScenarioFeedbackAnnotation
- AutomatedReasoningPolicyUpdateRuleAnnotation
- AutomatedReasoningPolicyUpdateRuleMutation
- AutomatedReasoningPolicyUpdateTypeAnnotation
- AutomatedReasoningPolicyUpdateTypeMutation
- AutomatedReasoningPolicyUpdateTypeValue
- AutomatedReasoningPolicyUpdateVariableAnnotation
- AutomatedReasoningPolicyUpdateVariableMutation
- AutomatedReasoningPolicyVariableReport
- AutomatedReasoningPolicyWorkflowTypeContent
- BatchDeleteAdvancedPromptOptimizationJobError
- BatchDeleteAdvancedPromptOptimizationJobItem
- BatchDeleteEvaluationJobError
- BatchDeleteEvaluationJobItem
- BedrockEvaluatorModel
- ByteContentDoc
- CloudWatchConfig

- CustomizationConfig
- CustomMetricBedrockEvaluatorModel
- CustomMetricDefinition
- CustomMetricEvaluatorModelConfig
- CustomModelDataSource
- CustomModelDeploymentSummary
- CustomModelDeploymentUpdateDetails
- CustomModelSummary
- CustomModelUnits
- DataProcessingDetails
- DimensionalPriceRate
- DistillationConfig
- EndpointConfig
- EvaluationBedrockModel
- EvaluationConfig
- EvaluationDataset
- EvaluationDatasetLocation
- EvaluationDatasetMetricConfig
- EvaluationInferenceConfig
- EvaluationInferenceConfigSummary
- EvaluationModelConfig
- EvaluationModelConfigSummary
- EvaluationOutputDataConfig
- EvaluationPrecomputedInferenceSource
- EvaluationPrecomputedRagSourceConfig
- EvaluationPrecomputedRetrieveAndGenerateSourceConfig
- EvaluationPrecomputedRetrieveSourceConfig
- EvaluationRagConfigSummary
- EvaluationSummary
- EvaluatorModelConfig

- ExternalSource
- ExternalSourcesGenerationConfiguration
- ExternalSourcesRetrieveAndGenerateConfiguration
- FieldForReranking
- FilterAttribute
- FoundationModelDetails
- FoundationModelLifecycle
- FoundationModelSummary
- GenerationConfiguration
- GraderConfig
- GuardrailAutomatedReasoningPolicy
- GuardrailAutomatedReasoningPolicyConfig
- GuardrailConfiguration
- GuardrailContentFilter
- GuardrailContentFilterConfig
- GuardrailContentFiltersTier
- GuardrailContentFiltersTierConfig
- GuardrailContentPolicy
- GuardrailContentPolicyConfig
- GuardrailContextualGroundingFilter
- GuardrailContextualGroundingFilterConfig
- GuardrailContextualGroundingPolicy
- GuardrailContextualGroundingPolicyConfig
- GuardrailCrossRegionConfig
- GuardrailCrossRegionDetails
- GuardrailManagedWords
- GuardrailManagedWordsConfig
- GuardrailPiiEntity
- GuardrailPiiEntityConfig
- GuardrailRegex

- GuardrailRegexConfig
- GuardrailSensitiveInformationPolicy
- GuardrailSensitiveInformationPolicyConfig
- GuardrailSummary
- GuardrailTopic
- GuardrailTopicConfig
- GuardrailTopicPolicy
- GuardrailTopicPolicyConfig
- GuardrailTopicsTier
- GuardrailTopicsTierConfig
- GuardrailWord
- GuardrailWordConfig
- GuardrailWordPolicy
- GuardrailWordPolicyConfig
- HumanEvaluationConfig
- HumanEvaluationCustomMetric
- HumanWorkflowConfig
- ImplicitFilterConfiguration
- ImportedModelSummary
- InferenceConfiguration
- InferenceProfileModel
- InferenceProfileModelSource
- InferenceProfileSummary
- InvocationLogsConfig
- InvocationLogSource
- KbInferenceConfig
- KnowledgeBaseConfig
- KnowledgeBaseRetrievalConfiguration
- KnowledgeBaseRetrieveAndGenerateConfiguration
- KnowledgeBaseVectorSearchConfiguration

- LambdaGraderConfig
- LegalTerm
- LoggingConfig
- MarketplaceModelEndpoint
- MarketplaceModelEndpointSummary
- MetadataAttributeSchema
- MetadataConfigurationForReranking
- ModelConfiguration
- ModelCopyJobSummary
- ModelCustomizationJobSummary
- ModelDataSource
- ModelEnforcement
- ModelImportJobSummary
- ModelInvocationJobInputDataConfig
- ModelInvocationJobOutputDataConfig
- ModelInvocationJobS3InputDataConfig
- ModelInvocationJobS3OutputDataConfig
- ModelInvocationJobSummary
- ModelPackageArnDataSource
- Offer
- OrchestrationConfiguration
- OutputDataConfig
- PerformanceConfiguration
- PricingTerm
- PromptRouterSummary
- PromptRouterTargetModel
- PromptTemplate
- ProvisionedModelSummary
- QueryTransformationConfiguration
- RAGConfig

- RatingScaleItem
- RatingScaleItemValue
- RequestMetadataBaseFilters
- RequestMetadataFilters
- RerankingMetadataSelectiveModeConfiguration
- RetrievalFilter
- RetrieveAndGenerateConfiguration
- RetrieveConfig
- RFTConfig
- RFTHyperParameters
- RoutingCriteria
- S3Config
- S3DataSource
- S3ObjectDoc
- SageMakerEndpoint
- SelectiveContentGuarding
- StatusDetails
- SupportTerm
- Tag
- TeacherModelConfig
- TermDetails
- TextInferenceConfig
- TrainingDataConfig
- TrainingDetails
- TrainingMetrics
- ValidationDataConfig
- ValidationDetails
- Validator
- ValidatorMetric
- ValidityTerm

- VectorSearchBedrockRerankingConfiguration
- VectorSearchBedrockRerankingModelConfiguration
- VectorSearchRerankingConfiguration
- VpcConfig

AccountEnforcedGuardrailInferenceInputConfiguration

Service: Amazon Bedrock

Account-level enforced guardrail input configuration.

Contents

guardrailIdentifier

Identifier for the guardrail, could be the ID or the ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

guardrailVersion

Numerical guardrail version.

Type: String

Pattern: `[1-9][0-9]{0,7}`

Required: Yes

modelEnforcement

Model-specific information for the enforced guardrail configuration. If not present, the configuration is enforced on all models

Type: ModelEnforcement object

Required: No

selectiveContentGuarding

Selective content guarding controls for enforced guardrails.

Type: SelectiveContentGuarding object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AccountEnforcedGuardrailOutputConfiguration

Service: Amazon Bedrock

Account enforced guardrail output configuration.

Contents

configId

Unique ID for the account enforced configuration.

Type: String

Pattern: `[a-z0-9]+`

Required: No

createdAt

Timestamp.

Type: Timestamp

Required: No

createdBy

The ARN of the role used to update the configuration.

Type: String

Required: No

guardrailArn

ARN representation for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+`

Required: No

guardrailId

Unique ID for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

Required: No

guardrailVersion

Numerical guardrail version.

Type: String

Pattern: `[1-9][0-9]{0,7}`

Required: No

inputTags

This member has been deprecated.

Whether to honor or ignore input tags at runtime.

Type: String

Valid Values: `HONOR | IGNORE`

Required: No

modelEnforcement

Model-specific information for the enforced guardrail configuration.

Type: ModelEnforcement object

Required: No

owner

Configuration owner type.

Type: String

Valid Values: ACCOUNT

Required: No

selectiveContentGuarding

Selective content guarding controls for enforced guardrails.

Type: SelectiveContentGuarding object

Required: No

updatedAt

Timestamp.

Type: Timestamp

Required: No

updatedBy

The ARN of the role used to update the configuration.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdvancedPromptOptimizationInputConfig

Service: Amazon Bedrock

The input data configuration for an advanced prompt optimization job.

Contents

s3Uri

The S3 URI of the input JSONL file.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdvancedPromptOptimizationJobSummary

Service: Amazon Bedrock

A summary of an advanced prompt optimization job.

Contents

creationTime

The time at which the job was created.

Type: Timestamp

Required: Yes

jobArn

The ARN of the job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/[a-z0-9]{12}`

Required: Yes

jobName

The name of the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9.+-]*`

Required: Yes

jobStatus

The status of the job.

Type: String

Valid Values: `InProgress | Completed | Failed | PartiallyCompleted | Stopping | Stopped | Deleting`

Required: Yes

lastModifiedTime

The time at which the job was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AdvancedPromptOptimizationOutputConfig

Service: Amazon Bedrock

The output data configuration for an advanced prompt optimization job.

Contents

s3Uri

The S3 URI prefix for the output location.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgreementAvailability

Service: Amazon Bedrock

Information about the agreement availability

Contents

status

Status of the agreement.

Type: String

Valid Values: `AVAILABLE | PENDING | NOT_AVAILABLE | ERROR`

Required: Yes

errorMessage

Error message.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedEvaluationConfig

Service: Amazon Bedrock

The configuration details of an automated evaluation job. The `EvaluationDatasetMetricConfig` object is used to specify the prompt datasets, task type, and metric names.

Contents

datasetMetricConfigs

Configuration details of the prompt datasets and metrics you want to use for your evaluation job.

Type: Array of EvaluationDatasetMetricConfig objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

customMetricConfig

Defines the configuration of custom metrics to be used in an evaluation job.

Type: AutomatedEvaluationCustomMetricConfig object

Required: No

evaluatorModelConfig

Contains the evaluator model configuration details. `EvaluatorModelConfig` is required for evaluation jobs that use a knowledge base or in model evaluation job that use a model as judge. This model computes all evaluation related metrics.

Type: EvaluatorModelConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedEvaluationCustomMetricConfig

Service: Amazon Bedrock

Defines the configuration of custom metrics to be used in an evaluation job. To learn more about using custom metrics in Amazon Bedrock evaluation jobs, see Create a prompt for a custom metrics (LLM-as-a-judge model evaluations) and Create a prompt for a custom metrics (RAG evaluations).

Contents

customMetrics

Defines a list of custom metrics to be used in an Amazon Bedrock evaluation job.

Type: Array of AutomatedEvaluationCustomMetricSource objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: Yes

evaluatorModelConfig

Configuration of the evaluator model you want to use to evaluate custom metrics in an Amazon Bedrock evaluation job.

Type: CustomMetricEvaluatorModelConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedEvaluationCustomMetricSource

Service: Amazon Bedrock

An array item definining a single custom metric for use in an Amazon Bedrock evaluation job.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

customMetricDefinition

The definition of a custom metric for use in an Amazon Bedrock evaluation job.

Type: CustomMetricDefinition object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckFinding

Service: Amazon Bedrock

Represents the result of an Automated Reasoning validation check, indicating whether the content is logically valid, invalid, or falls into other categories based on the policy rules.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

impossible

Indicates that Automated Reasoning cannot make a statement about the claims. This can happen if the premises are logically incorrect, or if there is a conflict within the Automated Reasoning policy itself.

Type: AutomatedReasoningCheckImpossibleFinding object

Required: No

invalid

Indicates that the claims are false. The claims are not implied by the premises and Automated Reasoning policy. Furthermore, there exist different claims that are consistent with the premises and Automated Reasoning policy.

Type: AutomatedReasoningCheckInvalidFinding object

Required: No

noTranslations

Identifies that some or all of the input prompt wasn't translated into logic. This can happen if the input isn't relevant to the Automated Reasoning policy, or if the policy doesn't have variables to model relevant input.

Type: AutomatedReasoningCheckNoTranslationsFinding object

Required: No

satisfiable

Indicates that the claims can be true or false. It depends on what assumptions are made for the claim to be implied from the premises and Automated Reasoning policy rules. In this situation, different assumptions can make input claims false and alternative claims true.

Type: AutomatedReasoningCheckSatisfiableFinding object

Required: No

tooComplex

Indicates that the input contains too much information for Automated Reasoning to process within its latency limits.

Type: AutomatedReasoningCheckTooComplexFinding object

Required: No

translationAmbiguous

Indicates that an ambiguity was detected in the translation, making it unsound to continue with validity checking. Additional context or follow-up questions might be needed to get translation to succeed.

Type: AutomatedReasoningCheckTranslationAmbiguousFinding object

Required: No

valid

Indicates that the claims are true. The claims are implied by the premises and the Automated Reasoning policy. Given the Automated Reasoning policy and premises, it is not possible for these claims to be false.

Type: AutomatedReasoningCheckValidFinding object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckImpossibleFinding

Service: Amazon Bedrock

Indicates that no valid claims can be made due to logical contradictions in the premises or rules.

Contents

contradictingRules

The automated reasoning policy rules that contradict the claims and/or premises in the input.

Type: Array of AutomatedReasoningCheckRule objects

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: AutomatedReasoningCheckLogicWarning object

Required: No

translation

The logical translation of the input that this finding evaluates.

Type: AutomatedReasoningCheckTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckInputTextReference

Service: Amazon Bedrock

References a portion of the original input text that corresponds to logical elements.

Contents

text

The specific text from the original input that this reference points to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckInvalidFinding

Service: Amazon Bedrock

Indicates that the claims are logically false and contradictory to the established rules or premises.

Contents

contradictingRules

The automated reasoning policy rules that contradict the claims in the input.

Type: Array of AutomatedReasoningCheckRule objects

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: AutomatedReasoningCheckLogicWarning object

Required: No

translation

The logical translation of the input that this finding invalidates.

Type: AutomatedReasoningCheckTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckLogicWarning

Service: Amazon Bedrock

Identifies logical issues in the translated statements that exist independent of any policy rules, such as statements that are always true or always false.

Contents

claims

The logical statements that are validated while assuming the policy and premises.

Type: Array of AutomatedReasoningLogicStatement objects

Required: No

premises

The logical statements that serve as premises under which the claims are validated.

Type: Array of AutomatedReasoningLogicStatement objects

Required: No

type

The category of the detected logical issue, such as statements that are always true or always false.

Type: String

Valid Values: `ALWAYS_TRUE | ALWAYS_FALSE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AutomatedReasoningCheckNoTranslationsFinding

Service: Amazon Bedrock

Indicates that no relevant logical information could be extracted from the input for validation.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckRule

Service: Amazon Bedrock

References a specific automated reasoning policy rule that was applied during evaluation.

Contents

id

The unique identifier of the automated reasoning rule.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: No

policyVersionArn

The ARN of the automated reasoning policy version that contains this rule.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckSatisfiableFinding

Service: Amazon Bedrock

Indicates that the claims could be either true or false depending on additional assumptions not provided in the input.

Contents

claimsFalseScenario

An example scenario demonstrating how the claims could be logically false.

Type: AutomatedReasoningCheckScenario object

Required: No

claimsTrueScenario

An example scenario demonstrating how the claims could be logically true.

Type: AutomatedReasoningCheckScenario object

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: AutomatedReasoningCheckLogicWarning object

Required: No

translation

The logical translation of the input that this finding evaluates.

Type: AutomatedReasoningCheckTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckScenario

Service: Amazon Bedrock

Represents a logical scenario where claims can be evaluated as true or false, containing specific logical assignments.

Contents

statements

List of logical assignments and statements that define this scenario.

Type: Array of AutomatedReasoningLogicStatement objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckTooComplexFinding

Service: Amazon Bedrock

Indicates that the input exceeds the processing capacity due to the volume or complexity of the logical information.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckTranslation

Service: Amazon Bedrock

Contains the logical translation of natural language input into formal logical statements, including premises, claims, and confidence scores.

Contents

claims

The logical statements that are being validated against the premises and policy rules.

Type: Array of AutomatedReasoningLogicStatement objects

Required: Yes

confidence

A confidence score between 0 and 1 indicating how certain the system is about the logical translation.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: Yes

premises

The logical statements that serve as the foundation or assumptions for the claims.

Type: Array of AutomatedReasoningLogicStatement objects

Required: No

untranslatedClaims

References to portions of the original input text that correspond to the claims but could not be fully translated.

Type: Array of AutomatedReasoningCheckInputTextReference objects

Required: No

untranslatedPremises

References to portions of the original input text that correspond to the premises but could not be fully translated.

Type: Array of AutomatedReasoningCheckInputTextReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckTranslationAmbiguousFinding

Service: Amazon Bedrock

Indicates that the input has multiple valid logical interpretations, requiring additional context or clarification.

Contents

differenceScenarios

Scenarios showing how the different translation options differ in meaning.

Type: Array of AutomatedReasoningCheckScenario objects

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Required: No

options

Different logical interpretations that were detected during translation of the input.

Type: Array of AutomatedReasoningCheckTranslationOption objects

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckTranslationOption

Service: Amazon Bedrock

Represents one possible logical interpretation of ambiguous input content.

Contents

translations

Different logical interpretations that were detected during translation of the input.

Type: Array of AutomatedReasoningCheckTranslation objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningCheckValidFinding

Service: Amazon Bedrock

Indicates that the claims are definitively true and logically implied by the premises, with no possible alternative interpretations.

Contents

claimsTrueScenario

An example scenario demonstrating how the claims are logically true.

Type: AutomatedReasoningCheckScenario object

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: AutomatedReasoningCheckLogicWarning object

Required: No

supportingRules

The automated reasoning policy rules that support why this result is considered valid.

Type: Array of AutomatedReasoningCheckRule objects

Required: No

translation

The logical translation of the input that this finding validates.

Type: AutomatedReasoningCheckTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningLogicStatement

Service: Amazon Bedrock

Represents a logical statement that can be expressed both in formal logic notation and natural language, providing dual representations for better understanding and validation.

Contents

logic

The formal logic representation of the statement using mathematical notation and logical operators.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: Yes

naturalLanguage

The natural language representation of the logical statement, providing a human-readable interpretation of the formal logic.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddRuleAnnotation

Service: Amazon Bedrock

An annotation for adding a new rule to an Automated Reasoning policy using a formal logical expression.

Contents

expression

The formal logical expression that defines the rule, using mathematical notation and referencing policy variables and types.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[\s\S]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddRuleFromNaturalLanguageAnnotation

Service: Amazon Bedrock

An annotation for adding a new rule to the policy by converting a natural language description into a formal logical expression.

Contents

naturalLanguage

The natural language description of the rule that should be converted into a formal logical expression.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddRuleMutation

Service: Amazon Bedrock

A mutation operation that adds a new rule to the policy definition during the build process.

Contents

rule

The rule definition that specifies the formal logical expression and metadata for the new rule being added to the policy.

Type: AutomatedReasoningPolicyDefinitionRule object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddTypeAnnotation

Service: Amazon Bedrock

An annotation for adding a new custom type to an Automated Reasoning policy, defining a set of possible values for variables.

Contents

description

A description of what the custom type represents and how it should be used in the policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

name

The name of the new custom type. This name will be used to reference the type in variable definitions and rules.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

values

The list of possible values that variables of this type can take, each with its own description and identifier.

Type: Array of AutomatedReasoningPolicyDefinitionTypeValue objects

Array Members: Minimum number of 1 item. Maximum number of 150 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddTypeMutation

Service: Amazon Bedrock

A mutation operation that adds a new custom type to the policy definition during the build process.

Contents

type

The type definition that specifies the name, description, and possible values for the new custom type being added to the policy.

Type: AutomatedReasoningPolicyDefinitionType object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddTypeValue

Service: Amazon Bedrock

Represents a single value that can be added to an existing custom type in the policy.

Contents

value

The identifier or name of the new value to add to the type.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

description

A description of what this new type value represents and when it should be used.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddVariableAnnotation

Service: Amazon Bedrock

An annotation for adding a new variable to an Automated Reasoning policy, which can be used in rule expressions.

Contents

description

A description of what the variable represents and how it should be used in rules.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

name

The name of the new variable. This name will be used to reference the variable in rule expressions.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

type

The type of the variable, which can be a built-in type (like string or number) or a custom type defined in the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAddVariableMutation

Service: Amazon Bedrock

A mutation operation that adds a new variable to the policy definition during the build process.

Contents

variable

The variable definition that specifies the name, type, and description for the new variable being added to the policy.

Type: AutomatedReasoningPolicyDefinitionVariable object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAnnotatedChunk

Service: Amazon Bedrock

Represents a portion of a source document with line number annotations. Chunks help organize document content for easier navigation and reference.

Contents

content

The lines of text contained within this chunk, each annotated with its line number.

Type: Array of AutomatedReasoningPolicyAnnotatedContent objects

Required: Yes

pageNumber

The page number where this chunk begins, if the document is divided into pages.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAnnotatedContent

Service: Amazon Bedrock

Represents a content element within an annotated chunk. This union type allows for different types of content elements to be included in document chunks, such as individual lines of text with their line numbers.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

line

An annotated line of text from the source document, including both the line number and the text content.

Type: AutomatedReasoningPolicyAnnotatedLine object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAnnotatedLine

Service: Amazon Bedrock

Represents a single line of text from a source document, annotated with its line number for precise referencing.

Contents

lineNumber

The line number of this text within the source document.

Type: Integer

Required: No

lineText

The actual text content of this line from the source document.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyAnnotation

Service: Amazon Bedrock

Contains the various operations that can be performed on an Automated Reasoning policy, including adding, updating, and deleting rules, variables, and types.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

addRule

An operation to add a new logical rule to the policy using formal mathematical expressions.

Type: AutomatedReasoningPolicyAddRuleAnnotation object

Required: No

addRuleFromNaturalLanguage

An operation to add a new rule by converting natural language descriptions into formal logical expressions.

Type: AutomatedReasoningPolicyAddRuleFromNaturalLanguageAnnotation object

Required: No

addType

An operation to add a new custom type to the policy, defining a set of possible values for policy variables.

Type: AutomatedReasoningPolicyAddTypeAnnotation object

Required: No

addVariable

An operation to add a new variable to the policy, which can be used in rule expressions to represent dynamic values.

Type: AutomatedReasoningPolicyAddVariableAnnotation object

Required: No

deleteRule

An operation to remove a rule from the policy.

Type: AutomatedReasoningPolicyDeleteRuleAnnotation object

Required: No

deleteType

An operation to remove a custom type from the policy. The type must not be referenced by any variables or rules.

Type: AutomatedReasoningPolicyDeleteTypeAnnotation object

Required: No

deleteVariable

An operation to remove a variable from the policy. The variable must not be referenced by any rules.

Type: AutomatedReasoningPolicyDeleteVariableAnnotation object

Required: No

ingestContent

An operation to process and incorporate new content into the policy, extracting additional rules and concepts.

Type: AutomatedReasoningPolicyIngestContentAnnotation object

Required: No

updateFromRulesFeedback

An operation to update the policy based on feedback about how specific rules performed during testing or validation.

Type: AutomatedReasoningPolicyUpdateFromRuleFeedbackAnnotation object

Required: No

updateFromScenarioFeedback

An operation to update the policy based on feedback about how it performed on specific test scenarios.

Type: AutomatedReasoningPolicyUpdateFromScenarioFeedbackAnnotation object

Required: No

updateRule

An operation to modify an existing rule in the policy, such as changing its logical expression or conditions.

Type: AutomatedReasoningPolicyUpdateRuleAnnotation object

Required: No

updateType

An operation to modify an existing custom type in the policy, such as changing its name, description, or allowed values.

Type: AutomatedReasoningPolicyUpdateTypeAnnotation object

Required: No

updateVariable

An operation to modify an existing variable in the policy, such as changing its name, type, or description.

Type: AutomatedReasoningPolicyUpdateVariableAnnotation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AutomatedReasoningPolicyAtomicStatement

Service: Amazon Bedrock

Represents a single, indivisible statement extracted from a source document. Atomic statements are the fundamental units used to ground policy rules and variables to their source material.

Contents

id

A unique identifier for this atomic statement within the fidelity report.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8.

Pattern: `[a-zA-Z0-9]*`

Required: Yes

location

Information about where this statement appears in the source document, including line numbers.

Type: AutomatedReasoningPolicyStatementLocation object

Required: Yes

text

The actual text content of the atomic statement as extracted from the source document.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildLog

Service: Amazon Bedrock

Contains detailed logging information about the policy build process, including steps taken, decisions made, and any issues encountered.

Contents

entries

A list of log entries documenting each step in the policy build process, including timestamps, status, and detailed messages.

Type: Array of AutomatedReasoningPolicyBuildLogEntry objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildLogEntry

Service: Amazon Bedrock

Represents a single entry in the policy build log, containing information about a specific step or event in the build process.

Contents

annotation

The annotation or operation that was being processed when this log entry was created.

Type: AutomatedReasoningPolicyAnnotation object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

buildSteps

Detailed information about the specific build steps that were executed, including any sub-operations or transformations.

Type: Array of AutomatedReasoningPolicyBuildStep objects

Required: Yes

status

The status of the build step (e.g., SUCCESS, FAILED, IN_PROGRESS).

Type: String

Valid Values: `APPLIED | FAILED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildResultAssetManifest

Service: Amazon Bedrock

A catalog of all artifacts produced by a build workflow, providing a comprehensive list of available assets including their types and identifiers.

Contents

entries

The list of asset entries in the manifest, each describing an available artifact that can be retrieved.

Type: Array of AutomatedReasoningPolicyBuildResultAssetManifestEntry objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildResultAssetManifestEntry

Service: Amazon Bedrock

Represents a single entry in the asset manifest, describing one artifact produced by the build workflow.

Contents

assetType

The type of asset (e.g., BUILD_LOG, QUALITY_REPORT, POLICY_DEFINITION, GENERATED_TEST_CASES, POLICY_SCENARIOS, FIDELITY_REPORT, ASSET_MANIFEST, SOURCE_DOCUMENT).

Type: String

Valid Values: `BUILD_LOG | QUALITY_REPORT | POLICY_DEFINITION | GENERATED_TEST_CASES | POLICY_SCENARIOS | FIDELITY_REPORT | ASSET_MANIFEST | SOURCE_DOCUMENT`

Required: Yes

assetId

A unique identifier for the asset, if applicable. Use this ID when requesting specific assets through the API.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[0-9a-fA-F\-]+`

Required: No

assetName

A human-readable name for the asset, if applicable. This helps identify specific documents or reports within the workflow results.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildResultAssets

Service: Amazon Bedrock

Contains the various assets generated during a policy build workflow, including logs, quality reports, test cases, and the final policy definition.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

assetManifest

A manifest listing all available artifacts produced by the build workflow. This provides a catalog of all assets that can be retrieved, including their types, names, and identifiers.

Type: AutomatedReasoningPolicyBuildResultAssetManifest object

Required: No

buildLog

The complete build log containing detailed information about each step in the policy generation process.

Type: AutomatedReasoningPolicyBuildLog object

Required: No

document

A source document that was used as input during the build workflow. This allows you to retrieve the original documents that were processed to generate the policy.

Type: AutomatedReasoningPolicySourceDocument object

Required: No

fidelityReport

A comprehensive fidelity report that measures how accurately the generated policy represents the source documents. The report includes coverage and accuracy scores, along with detailed grounding information for rules and variables.

Type: AutomatedReasoningPolicyFidelityReport object

Required: No

generatedTestCases

A comprehensive test suite generated by the build workflow, providing validation capabilities for automated reasoning policies.

Type: AutomatedReasoningPolicyGeneratedTestCases object

Required: No

policyDefinition

The complete policy definition generated by the build workflow, containing all rules, variables, and custom types extracted from the source documents.

Type: AutomatedReasoningPolicyDefinition object

Required: No

policyScenarios

An entity encompassing all the policy scenarios generated by the build workflow, which can be used to validate an Automated Reasoning policy.

Type: AutomatedReasoningPolicyScenarios object

Required: No

qualityReport

A comprehensive report analyzing the quality of the generated policy, including metrics about rule coverage, potential conflicts, and unused elements.

Type: AutomatedReasoningPolicyDefinitionQualityReport object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildStep

Service: Amazon Bedrock

Represents a single step in the policy build process, containing context about what was being processed and any messages or results.

Contents

context

Contextual information about what was being processed during this build step, such as the type of operation or the source material being analyzed.

Type: AutomatedReasoningPolicyBuildStepContext object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

messages

A list of messages generated during this build step, including informational messages, warnings, and error details.

Type: Array of AutomatedReasoningPolicyBuildStepMessage objects

Required: Yes

priorElement

Reference to the previous element or step in the build process, helping to trace the sequence of operations.

Type: AutomatedReasoningPolicyDefinitionElement object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildStepContext

Service: Amazon Bedrock

Provides context about what type of operation was being performed during a build step.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

mutation

Indicates that this build step involved modifying the policy structure, such as adding or updating rules, variables, or types.

Type: AutomatedReasoningPolicyMutation object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

planning

Indicates that this build step was part of the planning phase, where the system determines what operations to perform.

Type: AutomatedReasoningPolicyPlanning object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildStepMessage

Service: Amazon Bedrock

Represents a message generated during a build step, providing information about what happened or any issues encountered.

Contents

message

The content of the message, describing what occurred during the build step.

Type: String

Required: Yes

messageType

The type of message (e.g., INFO, WARNING, ERROR) indicating its severity and purpose.

Type: String

Valid Values: `INFO | WARNING | ERROR`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildWorkflowDocument

Service: Amazon Bedrock

Represents a source document used in the policy build workflow, containing the content and metadata needed for policy generation.

Contents

document

The actual content of the source document that will be analyzed to extract policy rules and concepts.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 5000000.

Required: Yes

documentContentType

The MIME type of the document content (e.g., text/plain, application/pdf, text/markdown).

Type: String

Valid Values: `pdf | txt`

Required: Yes

documentName

A descriptive name for the document that helps identify its purpose and content.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

documentDescription

A detailed description of the document's content and how it should be used in the policy generation process.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildWorkflowRepairContent

Service: Amazon Bedrock

Contains content and instructions for repairing or improving an existing Automated Reasoning policy.

Contents

annotations

Specific annotations or modifications to apply during the policy repair process, such as rule corrections or variable updates.

Type: Array of AutomatedReasoningPolicyAnnotation objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildWorkflowSource

Service: Amazon Bedrock

Defines the source content for a policy build workflow, which can include documents, repair instructions, or other input materials.

Contents

policyDefinition

An existing policy definition that serves as the starting point for the build workflow, typically used in policy repair or update scenarios.

Type: AutomatedReasoningPolicyDefinition object

Required: No

workflowContent

The actual content to be processed in the build workflow, such as documents to analyze or repair instructions to apply.

Type: AutomatedReasoningPolicyWorkflowTypeContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyBuildWorkflowSummary

Service: Amazon Bedrock

Provides a summary of a policy build workflow, including its current status, timing information, and key identifiers.

Contents

buildWorkflowId

The unique identifier of the build workflow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 36.

Pattern: `[a-f0-9]{8}-?[a-f0-9]{4}-?4[a-f0-9]{3}-?[89ab][a-f0-9]{3}-?[a-f0-9]{12}`

Required: Yes

buildWorkflowType

The type of build workflow (e.g., DOCUMENT_INGESTION, POLICY_REPAIR).

Type: String

Valid Values: `INGEST_CONTENT | REFINE_POLICY | IMPORT_POLICY | GENERATE_FIDELITY_REPORT | GENERATE_POLICY_SCENARIOS | RESOLVE_POLICY_AMBIGUITIES | ITERATIVELY_REFINE_POLICY`

Required: Yes

createdAt

The timestamp when the build workflow was created.

Type: Timestamp

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy associated with this build workflow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

status

The current status of the build workflow (e.g., RUNNING, COMPLETED, FAILED, CANCELLED).

Type: String

Valid Values: `SCHEDULED | CANCEL_REQUESTED | PREPROCESSING | BUILDING | TESTING | COMPLETED | FAILED | CANCELLED`

Required: Yes

updatedAt

The timestamp when the build workflow was last updated.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinition

Service: Amazon Bedrock

Contains the formal logic rules, variables, and custom variable types that define an Automated Reasoning policy. The policy definition specifies the constraints used to validate foundation model responses for accuracy and logical consistency.

Contents

rules

The formal logic rules extracted from the source document. Rules define the logical constraints that determine whether model responses are valid, invalid, or satisfiable.

Type: Array of AutomatedReasoningPolicyDefinitionRule objects

Array Members: Minimum number of 0 items. Maximum number of 1500 items.

Required: No

types

The custom user-defined vairable types used in the policy. Types are enum-based variable types that provide additional context beyond the predefined variable types.

Type: Array of AutomatedReasoningPolicyDefinitionType objects

Array Members: Minimum number of 0 items. Maximum number of 150 items.

Required: No

variables

The variables that represent concepts in the policy. Variables can have values assigned when translating natural language into formal logic. Their descriptions are crucial for accurate translation.

Type: Array of AutomatedReasoningPolicyDefinitionVariable objects

Array Members: Minimum number of 0 items. Maximum number of 600 items.

Required: No

version

The version of the policy definition format.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionElement

Service: Amazon Bedrock

Represents a single element in an Automated Reasoning policy definition, such as a rule, variable, or type definition.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

policyDefinitionRule

A rule element within the policy definition that contains a formal logical expression used for validation.

Type: AutomatedReasoningPolicyDefinitionRule object

Required: No

policyDefinitionType

A custom type element within the policy definition that defines a set of possible values for variables.

Type: AutomatedReasoningPolicyDefinitionType object

Required: No

policyDefinitionVariable

A variable element within the policy definition that represents a concept used in logical expressions and rules.

Type: AutomatedReasoningPolicyDefinitionVariable object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionQualityReport

Service: Amazon Bedrock

Provides a comprehensive analysis of the quality and completeness of an Automated Reasoning policy definition, highlighting potential issues and optimization opportunities.

Contents

conflictingRules

A list of rules that may conflict with each other, potentially leading to inconsistent policy behavior.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 500 items.

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

disjointRuleSets

Groups of rules that operate on completely separate sets of variables, indicating the policy may be addressing multiple unrelated concerns.

Type: Array of AutomatedReasoningPolicyDisjointRuleSet objects

Required: Yes

ruleCount

The total number of rules defined in the policy.

Type: Integer

Required: Yes

typeCount

The total number of custom types defined in the policy.

Type: Integer

Required: Yes

unusedTypes

A list of custom types that are defined but not referenced by any variables or rules, suggesting they may be unnecessary.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 150 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

unusedTypeValues

A list of type values that are defined but never used in any rules, indicating potential cleanup opportunities.

Type: Array of AutomatedReasoningPolicyDefinitionTypeValuePair objects

Array Members: Minimum number of 1 item. Maximum number of 22500 items.

Required: Yes

unusedVariables

A list of variables that are defined but not referenced by any rules, suggesting they may be unnecessary.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 600 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

variableCount

The total number of variables defined in the policy.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionRule

Service: Amazon Bedrock

Represents a formal logic rule in an Automated Reasoning policy. For example, rules can be expressed as if-then statements that define logical constraints.

Contents

expression

The formal logic expression of the rule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[\s\S]+`

Required: Yes

id

The unique identifier of the rule within the policy.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

alternateExpression

The human-readable form of the rule expression, often in natural language or simplified notation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[\s\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionType

Service: Amazon Bedrock

Represents a custom user-defined viarble type in an Automated Reasoning policy. Types are enum-based and provide additional context beyond predefined variable types.

Contents

name

The name of the custom type.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

values

The possible values for this enum-based type, each with its own description.

Type: Array of AutomatedReasoningPolicyDefinitionTypeValue objects

Array Members: Minimum number of 1 item. Maximum number of 150 items.

Required: Yes

description

The description of what the custom type represents.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionTypeValue

Service: Amazon Bedrock

Represents a single value within a custom type definition, including its identifier and description.

Contents

value

The actual value or identifier for this type value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

description

A human-readable description explaining what this type value represents and when it should be used.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionTypeValuePair

Service: Amazon Bedrock

Associates a type name with a specific value name, used for referencing type values in rules and other policy elements.

Contents

typeName

The name of the custom type that contains the referenced value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

valueName

The name of the specific value within the type.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDefinitionVariable

Service: Amazon Bedrock

Represents a variable in an Automated Reasoning policy. Variables represent concepts that can have values assigned during natural language translation.

Contents

description

The description of the variable that explains what it represents and how users might refer to it. Clear and comprehensive descriptions are essential for accurate natural language translation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

name

The name of the variable. Use descriptive names that clearly indicate the concept being represented.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

type

The data type of the variable. Valid types include bool, int, real, enum, and custom types that you can provide.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteRuleAnnotation

Service: Amazon Bedrock

An annotation for removing a rule from an Automated Reasoning policy.

Contents

ruleId

The unique identifier of the rule to delete from the policy.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteRuleMutation

Service: Amazon Bedrock

A mutation operation that removes a rule from the policy definition during the build process.

Contents

id

The unique identifier of the rule to delete.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteTypeAnnotation

Service: Amazon Bedrock

An annotation for removing a custom type from an Automated Reasoning policy.

Contents

name

The name of the custom type to delete from the policy. The type must not be referenced by any variables or rules.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteTypeMutation

Service: Amazon Bedrock

A mutation operation that removes a custom type from the policy definition during the build process.

Contents

name

The name of the custom type to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteTypeValue

Service: Amazon Bedrock

Represents a value to be removed from an existing custom type in the policy.

Contents

value

The identifier or name of the value to remove from the type.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteVariableAnnotation

Service: Amazon Bedrock

An annotation for removing a variable from an Automated Reasoning policy.

Contents

name

The name of the variable to delete from the policy. The variable must not be referenced by any rules.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDeleteVariableMutation

Service: Amazon Bedrock

A mutation operation that removes a variable from the policy definition during the build process.

Contents

name

The name of the variable to delete.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyDisjointRuleSet

Service: Amazon Bedrock

Represents a set of rules that operate on completely separate variables, indicating they address different concerns or domains within the policy.

Contents

rules

The list of rules that form this disjoint set, all operating on the same set of variables.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 500 items.

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

variables

The set of variables that are used by the rules in this disjoint set.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 600 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyFidelityReport

Service: Amazon Bedrock

A comprehensive analysis report that measures how accurately a generated policy represents the source documents. The report includes coverage and accuracy scores, detailed grounding information linking policy elements to source statements, and annotated document content.

Contents

accuracyScore

A score from 0.0 to 1.0 indicating how accurate the policy rules are relative to the source documents. A higher score means the policy rules more faithfully represent the source material.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: Yes

coverageScore

A score from 0.0 to 1.0 indicating how well the policy covers the statements in the source documents. A higher score means more of the source content is represented in the policy.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: Yes

documentSources

A list of source documents with their content broken down into atomic statements and annotated with line numbers for precise referencing.

Type: Array of AutomatedReasoningPolicyReportSourceDocument objects

Required: Yes

ruleReports

A mapping from rule identifiers to detailed fidelity reports for each rule, showing which source statements ground each rule and how accurate it is.

Type: String to AutomatedReasoningPolicyRuleReport object map

Key Length Constraints: Fixed length of 12.

Key Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

variableReports

A mapping from variable names to detailed fidelity reports for each variable, showing which source statements ground each variable and how accurate it is.

Type: String to AutomatedReasoningPolicyVariableReport object map

Key Length Constraints: Minimum length of 1. Maximum length of 64.

Key Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyGeneratedTestCase

Service: Amazon Bedrock

Represents a generated test case, consisting of query content, guard content, and expected results.

Contents

expectedAggregatedFindingsResult

The expected results of the generated test case. Possible values include:

- `VALID` - The claims are true. The claims are implied by the premises and the Automated Reasoning policy. Given the Automated Reasoning policy and premises, it is not possible for these claims to be false. In other words, there are no alternative answers that are true that contradict the claims.
- `INVALID` - The claims are false. The claims are not implied by the premises and Automated Reasoning policy. Furthermore, there exists different claims that are consistent with the premises and Automated Reasoning policy.
- `SATISFIABLE` - The claims can be true or false. It depends on what assumptions are made for the claim to be implied from the premises and Automated Reasoning policy rules. In this situation, different assumptions can make input claims false and alternative claims true.
- `IMPOSSIBLE` - Automated Reasoning can't make a statement about the claims. This can happen if the premises are logically incorrect, or if there is a conflict within the Automated Reasoning policy itself.

Type: String

Valid Values: `VALID | INVALID | SATISFIABLE | IMPOSSIBLE | TRANSLATION_AMBIGUOUS | TOO_COMPLEX | NO_TRANSLATION`

Required: Yes

guardContent

The output content that's validated by the Automated Reasoning policy. This represents the foundation model response that will be checked for accuracy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

queryContent

The input query or prompt that generated the content. This provides context for the validation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyGeneratedTestCases

Service: Amazon Bedrock

Contains a comprehensive test suite generated by the build workflow, providing validation capabilities for automated reasoning policies.

Contents

generatedTestCases

Represents a collection of generated test cases.

Type: Array of AutomatedReasoningPolicyGeneratedTestCase objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyGenerateFidelityReportContent

Service: Amazon Bedrock

Configuration for generating a fidelity report, which can either analyze new documents or update an existing fidelity report with a new policy definition.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

documents

Source documents to analyze for generating a new fidelity report. The documents will be processed to create atomic statements and grounding information.

Type: Array of AutomatedReasoningPolicyBuildWorkflowDocument objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyIngestContentAnnotation

Service: Amazon Bedrock

An annotation for processing and incorporating new content into an Automated Reasoning policy.

Contents

content

The new content to be analyzed and incorporated into the policy, such as additional documents or rule descriptions.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4096.

Pattern: `[\s\S]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyIterativeRefinementContent

Service: Amazon Bedrock

Configuration for an iterative policy refinement workflow, including source documents to process and optional feedback to guide the refinement.

Contents

documents

Source documents used for iterative policy refinement. These documents provide context for refining the policy definition.

Type: Array of AutomatedReasoningPolicyBuildWorkflowDocument objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

feedback

Optional feedback to guide the iterative refinement workflow. Provide specific instructions or constraints for policy refinement.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyMutation

Service: Amazon Bedrock

A container for various mutation operations that can be applied to an Automated Reasoning policy, including adding, updating, and deleting policy elements.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

addRule

A mutation to add a new rule to the policy.

Type: AutomatedReasoningPolicyAddRuleMutation object

Required: No

addType

A mutation to add a new custom type to the policy.

Type: AutomatedReasoningPolicyAddTypeMutation object

Required: No

addVariable

A mutation to add a new variable to the policy.

Type: AutomatedReasoningPolicyAddVariableMutation object

Required: No

deleteRule

A mutation to remove a rule from the policy.

Type: AutomatedReasoningPolicyDeleteRuleMutation object

Required: No

deleteType

A mutation to remove a custom type from the policy.

Type: AutomatedReasoningPolicyDeleteTypeMutation object

Required: No

deleteVariable

A mutation to remove a variable from the policy.

Type: AutomatedReasoningPolicyDeleteVariableMutation object

Required: No

updateRule

A mutation to modify an existing rule in the policy.

Type: AutomatedReasoningPolicyUpdateRuleMutation object

Required: No

updateType

A mutation to modify an existing custom type in the policy.

Type: AutomatedReasoningPolicyUpdateTypeMutation object

Required: No

updateVariable

A mutation to modify an existing variable in the policy.

Type: AutomatedReasoningPolicyUpdateVariableMutation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyPlanning

Service: Amazon Bedrock

Represents the planning phase of policy build workflow, where the system analyzes source content and determines what operations to perform.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyReportSourceDocument

Service: Amazon Bedrock

Represents a source document that was analyzed during fidelity report generation, including the document's metadata and its content broken down into atomic statements.

Contents

atomicStatements

The list of atomic statements extracted from this document, representing the fundamental units of meaning used for grounding.

Type: Array of AutomatedReasoningPolicyAtomicStatement objects

Required: Yes

documentContent

The document's content organized into annotated chunks with line number information for precise referencing.

Type: Array of AutomatedReasoningPolicyAnnotatedChunk objects

Required: Yes

documentHash

A SHA-256 hash of the document content, used for verification and ensuring the document hasn't changed since analysis.

Type: String

Length Constraints: Fixed length of 64.

Required: Yes

documentId

A unique identifier for this document within the fidelity report.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8.

Pattern: `[a-zA-Z0-9]*`

Required: Yes

documentName

The name of the source document that was analyzed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyRuleReport

Service: Amazon Bedrock

Provides detailed fidelity analysis for a specific policy rule, including which source document statements support it and how accurate the rule is.

Contents

rule

The identifier of the policy rule being analyzed in this report.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

accuracyJustification

A textual explanation of the accuracy score, describing why the rule received this particular accuracy rating.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

accuracyScore

A score from 0.0 to 1.0 indicating how accurately this rule represents the source material.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: No

groundingJustifications

Explanations describing how the source statements support and justify this specific rule.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

groundingStatements

References to statements from the source documents that provide the basis or justification for this rule.

Type: Array of AutomatedReasoningPolicyStatementReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyScenario

Service: Amazon Bedrock

Represents a test scenario used to validate an Automated Reasoning policy, including the test conditions and expected outcomes.

Contents

alternateExpression

An alternative way to express the same test scenario, used for validation and comparison purposes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

expectedResult

The expected outcome when this scenario is evaluated against the policy (e.g., PASS, FAIL, VIOLATION).

Type: String

Valid Values: `VALID | INVALID | SATISFIABLE | IMPOSSIBLE | TRANSLATION_AMBIGUOUS | TOO_COMPLEX | NO_TRANSLATION`

Required: Yes

expression

The logical expression or condition that defines this test scenario.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

ruleIds

The list of rule identifiers that are expected to be triggered or evaluated by this test scenario.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyScenarios

Service: Amazon Bedrock

Contains a comprehensive entity encompassing all the scenarios generated by the build workflow, which can be used to validate an Automated Reasoning policy.

Contents

policyScenarios

Represents a collection of generated policy scenarios.

Type: Array of AutomatedReasoningPolicyScenario objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicySourceDocument

Service: Amazon Bedrock

Represents a source document that was processed during a build workflow. Contains the document content, metadata, and a hash for verification.

Contents

document

The raw content of the source document as a binary blob.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 5000000.

Required: Yes

documentContentType

The MIME type of the document (e.g., application/pdf, text/plain).

Type: String

Valid Values: `pdf | txt`

Required: Yes

documentHash

A SHA-256 hash of the document content, used for verification and integrity checking.

Type: String

Length Constraints: Fixed length of 64.

Required: Yes

documentName

The name of the source document for identification purposes.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

documentDescription

An optional description providing context about the document's content and purpose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 4000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyStatementLocation

Service: Amazon Bedrock

Describes the location of a statement within a source document using line numbers.

Contents

lines

The line numbers in the source document where this statement appears.

Type: Array of integers

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyStatementReference

Service: Amazon Bedrock

References a specific atomic statement within a source document, used to link policy elements back to their source material.

Contents

documentId

The unique identifier of the document containing the referenced statement.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8.

Pattern: `[a-zA-Z0-9]*`

Required: Yes

statementId

The unique identifier of the specific atomic statement being referenced.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 8.

Pattern: `[a-zA-Z0-9]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicySummary

Service: Amazon Bedrock

Contains summary information about an Automated Reasoning policy, including metadata and timestamps.

Contents

createdAt

The timestamp when the policy was created.

Type: Timestamp

Required: Yes

name

The name of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[0-9a-zA-Z-_ ]+`

Required: Yes

policyArn

The Amazon Resource Name (ARN) of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

policyId

The unique identifier of the policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[a-z0-9]{12}`

Required: Yes

updatedAt

The timestamp when the policy was last updated.

Type: Timestamp

Required: Yes

version

The version of the policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `([1-9][0-9]{0,11})`

Required: Yes

description

The description of the policy.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyTestCase

Service: Amazon Bedrock

Represents a test for validating an Automated Reasoning policy. tests contain sample inputs and expected outcomes to verify policy behavior.

Contents

createdAt

The timestamp when the test was created.

Type: Timestamp

Required: Yes

guardContent

The output content to be validated by the policy, typically representing a foundation model response.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

testCaseId

The unique identifier of the test.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[0-9A-Z]{12}`

Required: Yes

updatedAt

The timestamp when the test was last updated.

Type: Timestamp

Required: Yes

confidenceThreshold

The minimum confidence level for logic validation. Content meeting this threshold is considered high-confidence and can be validated.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

expectedAggregatedFindingsResult

The expected result of the Automated Reasoning check for this test.

Type: String

Valid Values: `VALID | INVALID | SATISFIABLE | IMPOSSIBLE | TRANSLATION_AMBIGUOUS | TOO_COMPLEX | NO_TRANSLATION`

Required: No

queryContent

The input query or prompt that generated the content. This provides context for the validation.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyTestResult

Service: Amazon Bedrock

Contains the results of testing an Automated Reasoning policy against various scenarios and validation checks.

Contents

policyArn

The Amazon Resource Name (ARN) of the Automated Reasoning policy that was tested.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

testCase

The test case that was executed, including the input content, expected results, and configuration parameters used during validation.

Type: AutomatedReasoningPolicyTestCase object

Required: Yes

testRunStatus

The overall status of the test run (e.g., COMPLETED, FAILED, IN_PROGRESS).

Type: String

Valid Values: `NOT_STARTED | SCHEDULED | IN_PROGRESS | COMPLETED | FAILED`

Required: Yes

updatedAt

The timestamp when the test results were last updated.

Type: Timestamp

Required: Yes

aggregatedTestFindingsResult

A summary of all test findings, aggregated to provide an overall assessment of policy quality and correctness.

Type: String

Valid Values: `VALID | INVALID | SATISFIABLE | IMPOSSIBLE | TRANSLATION_AMBIGUOUS | TOO_COMPLEX | NO_TRANSLATION`

Required: No

testFindings

Detailed findings from the test run, including any issues, violations, or unexpected behaviors discovered.

Type: Array of AutomatedReasoningCheckFinding objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

testRunResult

The overall result of the test run, indicating whether the policy passed or failed validation.

Type: String

Valid Values: `PASSED | FAILED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyTypeValueAnnotation

Service: Amazon Bedrock

An annotation for managing values within custom types, including adding, updating, or removing specific type values.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

addTypeValue

An operation to add a new value to an existing custom type.

Type: AutomatedReasoningPolicyAddTypeValue object

Required: No

deleteTypeValue

An operation to remove a value from an existing custom type.

Type: AutomatedReasoningPolicyDeleteTypeValue object

Required: No

updateTypeValue

An operation to modify an existing value within a custom type.

Type: AutomatedReasoningPolicyUpdateTypeValue object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateFromRuleFeedbackAnnotation

Service: Amazon Bedrock

An annotation for updating the policy based on feedback about how specific rules performed during testing or real-world usage.

Contents

feedback

The feedback information about rule performance, including suggestions for improvements or corrections.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

ruleIds

The list of rule identifiers that the feedback applies to.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateFromScenarioFeedbackAnnotation

Service: Amazon Bedrock

An annotation for updating the policy based on feedback about how it performed on specific test scenarios.

Contents

scenarioExpression

The logical expression that defines the test scenario that generated this feedback.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: Yes

feedback

The feedback information about scenario performance, including any issues or improvements identified.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

ruleIds

The list of rule identifiers that were involved in the scenario being evaluated.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateRuleAnnotation

Service: Amazon Bedrock

An annotation for modifying an existing rule in an Automated Reasoning policy.

Contents

expression

The new formal logical expression for the rule, replacing the previous expression.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[\s\S]+`

Required: Yes

ruleId

The unique identifier of the rule to update.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[A-Z][0-9A-Z]{11}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateRuleMutation

Service: Amazon Bedrock

A mutation operation that modifies an existing rule in the policy definition during the build process.

Contents

rule

The updated rule definition containing the modified formal logical expression and any changed metadata for the existing rule.

Type: AutomatedReasoningPolicyDefinitionRule object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateTypeAnnotation

Service: Amazon Bedrock

An annotation for modifying an existing custom type in an Automated Reasoning policy.

Contents

name

The current name of the custom type to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

values

The updated list of values for the custom type, which can include additions, modifications, or removals.

Type: Array of AutomatedReasoningPolicyTypeValueAnnotation objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: Yes

description

The new description for the custom type, replacing the previous description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

newName

The new name for the custom type, if you want to rename it. If not provided, the name remains unchanged.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateTypeMutation

Service: Amazon Bedrock

A mutation operation that modifies an existing custom type in the policy definition during the build process.

Contents

type

The updated type definition containing the modified name, description, or values for the existing custom type.

Type: AutomatedReasoningPolicyDefinitionType object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateTypeValue

Service: Amazon Bedrock

Represents a modification to a value within an existing custom type.

Contents

value

The current identifier or name of the type value to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

description

The new description for the type value, replacing the previous description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

newValue

The new identifier or name for the type value, if you want to rename it.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateVariableAnnotation

Service: Amazon Bedrock

An annotation for modifying an existing variable in an Automated Reasoning policy.

Contents

name

The current name of the variable to update.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

description

The new description for the variable, replacing the previous description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: No

newName

The new name for the variable, if you want to rename it. If not provided, the name remains unchanged.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyUpdateVariableMutation

Service: Amazon Bedrock

A mutation operation that modifies an existing variable in the policy definition during the build process.

Contents

variable

The updated variable definition containing the modified name, type, or description for the existing variable.

Type: AutomatedReasoningPolicyDefinitionVariable object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyVariableReport

Service: Amazon Bedrock

Provides detailed fidelity analysis for a specific policy variable, including which source document statements support it and how accurate the variable definition is.

Contents

policyVariable

The name of the policy variable being analyzed in this report.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `\p{L}[\p{L}\p{M}0-9_]*( [\p{L}\p{M}0-9_]+)*`

Required: Yes

accuracyJustification

A textual explanation of the accuracy score, describing why the variable received this particular accuracy rating.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

accuracyScore

A score from 0.0 to 1.0 indicating how accurately this variable represents concepts from the source material.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: No

groundingJustifications

Explanations describing how the source statements support and justify this specific variable definition.

Type: Array of strings

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

groundingStatements

References to statements from the source documents that provide the basis or justification for this variable.

Type: Array of AutomatedReasoningPolicyStatementReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutomatedReasoningPolicyWorkflowTypeContent

Service: Amazon Bedrock

Defines the content and configuration for different types of policy build workflows.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

documents

The list of documents to be processed in a document ingestion workflow.

Type: Array of AutomatedReasoningPolicyBuildWorkflowDocument objects

Array Members: Fixed number of 1 item.

Required: No

generateFidelityReportContent

The content configuration for generating a fidelity report workflow. This can include source documents to analyze or an existing fidelity report to update with a new policy definition.

Type: AutomatedReasoningPolicyGenerateFidelityReportContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

iterativeRefinementContent

Content configuration to start an iterative policy refinement workflow that uses generative AI to automatically make changes to the policy based on test results and the optional feedback provided.

Type: AutomatedReasoningPolicyIterativeRefinementContent object

Required: No

policyRepairAssets

The assets and instructions needed for a policy repair workflow, including repair annotations and guidance.

Type: AutomatedReasoningPolicyBuildWorkflowRepairContent object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDeleteAdvancedPromptOptimizationJobError

Service: Amazon Bedrock

Information about an error deleting a job.

Contents

code

The error code.

Type: String

Required: Yes

jobIdentifier

The identifier of the job that could not be deleted.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/)?[a-z0-9]{12})`

Required: Yes

message

A message describing the error.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

BatchDeleteAdvancedPromptOptimizationJobItem

Service: Amazon Bedrock

Information about a successfully deleted job.

Contents

jobIdentifier

The identifier of the deleted job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `((arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:advanced-prompt-optimization-job/)?[a-z0-9]{12})`

Required: Yes

jobStatus

The status of the deleted job.

Type: String

Valid Values: `InProgress | Completed | Failed | PartiallyCompleted | Stopping | Stopped | Deleting`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BatchDeleteEvaluationJobError

Service: Amazon Bedrock

A JSON array that provides the status of the evaluation jobs being deleted.

Contents

code

A HTTP status code of the evaluation job being deleted.

Type: String

Required: Yes

jobIdentifier

The ARN of the evaluation job being deleted.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})`

Required: Yes

message

A status message about the evaluation job deletion.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

BatchDeleteEvaluationJobItem

Service: Amazon Bedrock

An evaluation job for deletion, and it's current status.

Contents

jobIdentifier

The Amazon Resource Name (ARN) of the evaluation job for deletion.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12})`

Required: Yes

jobStatus

The status of the evaluation job for deletion.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockEvaluatorModel

Service: Amazon Bedrock

The evaluator model used in knowledge base evaluation job or in model evaluation job that use a model as judge. This model computes all evaluation related metrics.

Contents

modelIdentifier

The Amazon Resource Name (ARN) of the evaluator model used used in knowledge base evaluation job or in model evaluation job that use a model as judge.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))))$|(^[a-z0-9-]+[.][a-z0-9-]+([.][a-z0-9-]+)*(:[a-z0-9-]+)?$)|^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ByteContentDoc

Service: Amazon Bedrock

Contains the document contained in the wrapper object, along with its attributes/fields.

Contents

contentType

The MIME type of the document contained in the wrapper object.

Type: String

Pattern: `.*[a-z]{1,20}/.{1,20}.*`

Required: Yes

data

The byte value of the file to upload, encoded as a Base-64 string.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 10485760.

Required: Yes

identifier

The file name of the document contained in the wrapper object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

CloudWatchConfig

Service: Amazon Bedrock

CloudWatch logging configuration.

Contents

logGroupName

The log group name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Required: Yes

roleArn

The role Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

largeDataDeliveryS3Config

S3 configuration for delivering a large amount of data.

Type: S3Config object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomizationConfig

Service: Amazon Bedrock

A model customization configuration

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

distillationConfig

The Distillation configuration for the custom model.

Type: DistillationConfig object

Required: No

rftConfig

Configuration settings for reinforcement fine-tuning (RFT) model customization, including grader configuration and hyperparameters.

Type: RFTConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomMetricBedrockEvaluatorModel

Service: Amazon Bedrock

Defines the model you want to evaluate custom metrics in an Amazon Bedrock evaluation job.

Contents

modelIdentifier

The Amazon Resource Name (ARN) of the evaluator model for custom metrics. For a list of supported evaluator models, see Evaluate model performance using another LLM as a judge and Evaluate the performance of RAG sources using Amazon Bedrock evaluations.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))))$|(^[a-z0-9-]+[.][a-z0-9-]+([.][a-z0-9-]+)*(:[a-z0-9-]+)?$)|^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomMetricDefinition

Service: Amazon Bedrock

The definition of a custom metric for use in an Amazon Bedrock evaluation job. A custom metric definition includes a metric name, prompt (instructions) and optionally, a rating scale. Your prompt must include a task description and input variables. The required input variables are different for model-as-a-judge and RAG evaluations.

For more information about how to define a custom metric in Amazon Bedrock, see Create a prompt for a custom metrics (LLM-as-a-judge model evaluations) and Create a prompt for a custom metrics (RAG evaluations).

Contents

instructions

The prompt for a custom metric that instructs the evaluator model how to rate the model or RAG source under evaluation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5000.

Required: Yes

name

The name for a custom metric. Names must be unique in your AWS region.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[0-9a-zA-Z-_.]+`

Required: Yes

ratingScale

Defines the rating scale to be used for a custom metric. We recommend that you always define a ratings scale when creating a custom metric. If you don't define a scale, Amazon Bedrock won't be able to visually display the results of the evaluation in the console or calculate average

values of numerical scores. For more information on specifying a rating scale, see Specifying an output schema (rating scale).

Type: Array of RatingScaleItem objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomMetricEvaluatorModelConfig

Service: Amazon Bedrock

Configuration of the evaluator model you want to use to evaluate custom metrics in an Amazon Bedrock evaluation job.

Contents

bedrockEvaluatorModels

Defines the model you want to evaluate custom metrics in an Amazon Bedrock evaluation job.

Type: Array of CustomMetricBedrockEvaluatorModel objects

Array Members: Fixed number of 1 item.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomModelDataSource

Service: Amazon Bedrock

The data source for a custom model. This is a union type that supports the following member:

- `modelPackageArnDataSource` — Specifies a SageMaker AI model package as the data source.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

modelPackageArnDataSource

A SageMaker AI model package ARN as the data source for the custom model. When you specify a model package ARN, Amazon Bedrock resolves the model package to retrieve the model artifacts.

Type: ModelPackageArnDataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomModelDeploymentSummary

Service: Amazon Bedrock

Contains summary information about a custom model deployment, including its ARN, name, status, and associated custom model.

Contents

createdAt

The date and time when the custom model deployment was created.

Type: Timestamp

Required: Yes

customModelDeploymentArn

The Amazon Resource Name (ARN) of the custom model deployment.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model-deployment/[a-z0-9]{12}`

Required: Yes

customModelDeploymentName

The name of the custom model deployment.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the custom model associated with this deployment.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

status

The status of the custom model deployment. Possible values are `CREATING`, `ACTIVE`, and `FAILED`.

Type: String

Valid Values: `Creating | Active | Failed`

Required: Yes

failureMessage

If the deployment status is `FAILED`, this field contains a message describing the failure reason.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

lastUpdatedAt

The date and time when the custom model deployment was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomModelDeploymentUpdateDetails

Service: Amazon Bedrock

Details about an update to a custom model deployment, including the new custom model resource ARN and current update status.

Contents

modelArn

ARN of the new custom model being deployed as part of the update.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

updateStatus

Current status of the deployment update.

Type: String

Valid Values: `Updating | UpdateCompleted | UpdateFailed`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomModelSummary

Service: Amazon Bedrock

Summary information for a custom model.

Contents

baseModelArn

The base model Amazon Resource Name (ARN).

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

baseModelName

The base model name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63})`

Required: Yes

creationTime

Creation time of the model.

Type: Timestamp

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

Required: Yes

modelName

The name of the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: Yes

customizationType

Specifies whether to carry out continued pre-training of a model or whether to fine-tune it. For more information, see Custom models.

Type: String

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION | REINFORCEMENT_FINE_TUNING | IMPORTED`

Required: No

modelStatus

The current status of the custom model. Possible values include:

- `Creating` - The model is being created and validated.
- `Active` - The model has been successfully created and is ready for use.
- `Failed` - The model creation process failed.

Type: String

Valid Values: `Active | Creating | Failed`

Required: No

ownerAccountId

The unique identifier of the account that owns the model.

Type: String

Pattern: `[0-9]{12}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomModelUnits

Service: Amazon Bedrock

A `CustomModelUnit` (CMU) is an abstract view of the hardware utilization that Amazon Bedrock needs to host a single copy of your custom model. A model copy represents a single instance of your imported model that is ready to serve inference requests. Amazon Bedrock determines the number of custom model units that a model copy needs when you import the custom model.

You can use `CustomModelUnits` to estimate the cost of running your custom model. For more information, see Calculate the cost of running a custom model.

Contents

customModelUnitsPerModelCopy

The number of custom model units used to host a model copy.

Type: Integer

Required: No

customModelUnitsVersion

The version of the custom model unit. Use to determine the billing rate for the custom model unit.

Type: String

Pattern: `v\d+.\d+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataProcessingDetails

Service: Amazon Bedrock

For a Distillation job, the status details for the data processing sub-task of the job.

Contents

creationTime

The start time of the data processing sub-task of the job.

Type: Timestamp

Required: No

lastModifiedTime

The latest update to the data processing sub-task of the job.

Type: Timestamp

Required: No

status

The status of the data processing sub-task of the job.

Type: String

Valid Values: `InProgress | Completed | Stopping | Stopped | Failed | NotStarted`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DimensionalPriceRate

Service: Amazon Bedrock

Dimensional price rate.

Contents

description

Description of the price rate.

Type: String

Required: No

dimension

Dimension for the price rate.

Type: String

Required: No

price

Single-dimensional rate information.

Type: String

Required: No

unit

Unit associated with the price.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DistillationConfig

Service: Amazon Bedrock

Settings for distilling a foundation model into a smaller and more efficient model.

Contents

teacherModelConfig

The teacher model configuration.

Type: TeacherModelConfig object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EndpointConfig

Service: Amazon Bedrock

Specifies the configuration for the endpoint.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

sageMaker

The configuration specific to Amazon SageMaker for the endpoint.

Type: SageMakerEndpoint object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationBedrockModel

Service: Amazon Bedrock

Contains the ARN of the Amazon Bedrock model or inference profile specified in your evaluation job. Each Amazon Bedrock model supports different `inferenceParams`. To learn more about supported inference parameters for Amazon Bedrock models, see Inference parameters for foundation models.

The `inferenceParams` are specified using JSON. To successfully insert JSON as string make sure that all quotations are properly escaped. For example, `"temperature":"0.25"` key value pair would need to be formatted as `\"temperature\":\"0.25\"` to successfully accepted in the request.

Contents

modelIdentifier

The ARN of the Amazon Bedrock model or inference profile specified.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:application-inference-profile/[a-z0-9]{12})|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))|([0-9]{12}:(default-prompt-router|prompt-router)/[a-zA-Z0-9-:.]+)))|(([a-z]{2,4}[.]{1})([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|arn:aws(-[^:]+)?:sagemaker:[a-z0-9-]{1,20}:[0-9]{12}:endpoint/[a-z0-9-]{1,63}`

Required: Yes

inferenceParams

Each Amazon Bedrock support different inference parameters that change how the model behaves during inference.

For code examples that demonstrate how to use the `inferenceParams` parameter, see Create a model evaluation job in Amazon Bedrock.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

performanceConfig

Specifies performance settings for the model or inference profile.

Type: PerformanceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationConfig

Service: Amazon Bedrock

The configuration details of either an automated or human-based evaluation job.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

automated

Contains the configuration details of an automated evaluation job that computes metrics.

Type: AutomatedEvaluationConfig object

Required: No

human

Contains the configuration details of an evaluation job that uses human workers.

Type: HumanEvaluationConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationDataset

Service: Amazon Bedrock

Used to specify the name of a built-in prompt dataset and optionally, the Amazon S3 bucket where a custom prompt dataset is saved.

Contents

name

Used to specify supported built-in prompt datasets. Valid values are `Builtin.Bold`, `Builtin.BoolQ`, `Builtin.NaturalQuestions`, `Builtin.Gigaword`, `Builtin.RealToxicityPrompts`, `Builtin.TriviaQA`, `Builtin.T-Rex`, `Builtin.WomensEcommerceClothingReviews` and `Builtin.Wikitext2`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[0-9a-zA-Z-_.]+`

Required: Yes

datasetLocation

For custom prompt datasets, you must specify the location in Amazon S3 where the prompt dataset is saved.

Type: EvaluationDatasetLocation object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

EvaluationDatasetLocation

Service: Amazon Bedrock

The location in Amazon S3 where your prompt dataset is stored.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3Uri

The S3 URI of the S3 bucket specified in the job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationDatasetMetricConfig

Service: Amazon Bedrock

Defines the prompt datasets, built-in metric names and custom metric names, and the task type.

Contents

dataset

Specifies the prompt dataset.

Type: EvaluationDataset object

Required: Yes

metricNames

The names of the metrics you want to use for your evaluation job.

For knowledge base evaluation jobs that evaluate retrieval only, valid values are "`Builtin.ContextRelevance`", "`Builtin.ContextCoverage`".

For knowledge base evaluation jobs that evaluate retrieval with response generation, valid values are "`Builtin.Correctness`", "`Builtin.Completeness`", "`Builtin.Helpfulness`", "`Builtin.LogicalCoherence`", "`Builtin.Faithfulness`", "`Builtin.Harmfulness`", "`Builtin.Stereotyping`", "`Builtin.Refusal`".

For automated model evaluation jobs, valid values are "`Builtin.Accuracy`", "`Builtin.Robustness`", and "`Builtin.Toxicity`". In model evaluation jobs that use a LLM as judge you can specify "`Builtin.Correctness`", "`Builtin.Completeness`", "`Builtin.Faithfulness`", "`Builtin.Helpfulness`", "`Builtin.Coherence`", "`Builtin.Relevance`", "`Builtin.FollowingInstructions`", "`Builtin.ProfessionalStyleAndTone`", You can also specify the following responsible AI related metrics only for model evaluation job that use a LLM as judge "`Builtin.Harmfulness`", "`Builtin.Stereotyping`", and "`Builtin.Refusal`".

For human-based model evaluation jobs, the list of strings must match the `name` parameter specified in `HumanEvaluationCustomMetric`.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[0-9a-zA-Z-_.]+`

Required: Yes

taskType

The the type of task you want to evaluate for your evaluation job. This applies only to model evaluation jobs and is ignored for knowledge base evaluation jobs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[A-Za-z0-9]+`

Valid Values: `Summarization | Classification | QuestionAndAnswer | Generation | Custom`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationInferenceConfig

Service: Amazon Bedrock

The configuration details of the inference model for an evaluation job.

For automated model evaluation jobs, only a single model is supported.

For human-based model evaluation jobs, your annotator can compare the responses for up to two different models.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

models

Specifies the inference models.

Type: Array of EvaluationModelConfig objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: No

ragConfigs

Contains the configuration details of the inference for a knowledge base evaluation job, including either the retrieval only configuration or the retrieval with response generation configuration.

Type: Array of RAGConfig objects

Array Members: Fixed number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationInferenceConfigSummary

Service: Amazon Bedrock

Identifies the models, Knowledge Bases, or other RAG sources evaluated in a model or Knowledge Base evaluation job.

Contents

modelConfigSummary

A summary of the models used in an Amazon Bedrock model evaluation job. These resources can be models in Amazon Bedrock or models outside of Amazon Bedrock that you use to generate your own inference response data.

Type: EvaluationModelConfigSummary object

Required: No

ragConfigSummary

A summary of the RAG resources used in an Amazon Bedrock Knowledge Base evaluation job. These resources can be Knowledge Bases in Amazon Bedrock or RAG sources outside of Amazon Bedrock that you use to generate your own inference response data.

Type: EvaluationRagConfigSummary object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationModelConfig

Service: Amazon Bedrock

Defines the models used in the model evaluation job.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bedrockModel

Defines the Amazon Bedrock model or inference profile and inference parameters you want used.

Type: EvaluationBedrockModel object

Required: No

precomputedInferenceSource

Defines the model used to generate inference response data for a model evaluation job where you provide your own inference response data.

Type: EvaluationPrecomputedInferenceSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationModelConfigSummary

Service: Amazon Bedrock

A summary of the models used in an Amazon Bedrock model evaluation job. These resources can be models in Amazon Bedrock or models outside of Amazon Bedrock that you use to generate your own inference response data.

Contents

bedrockModelIdentifiers

The Amazon Resource Names (ARNs) of the models used for the evaluation job.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:application-inference-profile/[a-z0-9]{12})|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))|([0-9]{12}:(default-prompt-router|prompt-router)/[a-zA-Z0-9-:.]+)))|(([a-z]{2,4}[.]{1})([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|arn:aws(-[^:]+)?:sagemaker:[a-z0-9-]{1,20}:[0-9]{12}:endpoint/[a-z0-9-]{1,63}`

Required: No

precomputedInferenceSourceIdentifiers

A label that identifies the models used for a model evaluation job where you provide your own inference response data.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9._-]){0,255}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationOutputDataConfig

Service: Amazon Bedrock

The Amazon S3 location where the results of your evaluation job are saved.

Contents

s3Uri

The Amazon S3 URI where the results of the evaluation job are saved.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationPrecomputedInferenceSource

Service: Amazon Bedrock

A summary of a model used for a model evaluation job where you provide your own inference response data.

Contents

inferenceSourceIdentifier

A label that identifies a model used in a model evaluation job where you provide your own inference response data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9._-]){0,255}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationPrecomputedRagSourceConfig

Service: Amazon Bedrock

A summary of a RAG source used for a Knowledge Base evaluation job where you provide your own inference response data.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

retrieveAndGenerateSourceConfig

A summary of a RAG source used for a retrieve-and-generate Knowledge Base evaluation job where you provide your own inference response data.

Type: EvaluationPrecomputedRetrieveAndGenerateSourceConfig object

Required: No

retrieveSourceConfig

A summary of a RAG source used for a retrieve-only Knowledge Base evaluation job where you provide your own inference response data.

Type: EvaluationPrecomputedRetrieveSourceConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationPrecomputedRetrieveAndGenerateSourceConfig

Service: Amazon Bedrock

A summary of a RAG source used for a retrieve-and-generate Knowledge Base evaluation job where you provide your own inference response data.

Contents

ragSourceIdentifier

A label that identifies the RAG source used for a retrieve-and-generate Knowledge Base evaluation job where you provide your own inference response data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9._-]){0,255}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationPrecomputedRetrieveSourceConfig

Service: Amazon Bedrock

A summary of a RAG source used for a retrieve-only Knowledge Base evaluation job where you provide your own inference response data.

Contents

ragSourceIdentifier

A label that identifies the RAG source used for a retrieve-only Knowledge Base evaluation job where you provide your own inference response data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9._-]){0,255}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationRagConfigSummary

Service: Amazon Bedrock

A summary of the RAG resources used in an Amazon Bedrock Knowledge Base evaluation job. These resources can be Knowledge Bases in Amazon Bedrock or RAG sources outside of Amazon Bedrock that you use to generate your own inference response data.

Contents

bedrockKnowledgeBaseIdentifiers

The Amazon Resource Names (ARNs) of the Knowledge Base resources used for a Knowledge Base evaluation job where Amazon Bedrock invokes the Knowledge Base for you.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: No

precomputedRagSourceIdentifiers

A label that identifies the RAG sources used for a Knowledge Base evaluation job where you provide your own inference response data.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9._-]){0,255}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluationSummary

Service: Amazon Bedrock

Summary information of an evaluation job.

Contents

creationTime

The time the evaluation job was created.

Type: Timestamp

Required: Yes

evaluationTaskTypes

The type of task for model evaluation.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[A-Za-z0-9]+`

Valid Values: `Summarization | Classification | QuestionAndAnswer | Generation | Custom`

Required: Yes

jobArn

The Amazon Resource Name (ARN) of the evaluation job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:evaluation-job/[a-z0-9]{12}`

Required: Yes

jobName

The name for the evaluation job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-z0-9](-*[a-z0-9]){0,62}`

Required: Yes

jobType

Specifies whether the evaluation job is automated or human-based.

Type: String

Valid Values: `Human | Automated`

Required: Yes

status

The current status of the evaluation job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped | Deleting`

Required: Yes

applicationType

Specifies whether the evaluation job is for evaluating a model or evaluating a knowledge base (retrieval and response generation).

Type: String

Valid Values: `ModelEvaluation | RagEvaluation`

Required: No

customMetricsEvaluatorModelIdentifiers

The Amazon Resource Names (ARNs) of the models used to compute custom metrics in an Amazon Bedrock evaluation job.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))))$|(^[a-z0-9-]+[.][a-z0-9-]+([.][a-z0-9-]+)*(:[a-z0-9-]+)?$)|^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: No

evaluatorModelIdentifiers

The Amazon Resource Names (ARNs) of the models used to compute the metrics for a knowledge base evaluation job.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))))$|(^[a-z0-9-]+[.][a-z0-9-]+([.][a-z0-9-]+)*(:[a-z0-9-]+)?$)|^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: No

inferenceConfigSummary

Identifies the models, Knowledge Bases, or other RAG sources evaluated in a model or Knowledge Base evaluation job.

Type: EvaluationInferenceConfigSummary object

Required: No

modelIdentifiers

This member has been deprecated.

The Amazon Resource Names (ARNs) of the model(s) used for the evaluation job.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|([0-9]{12}:provisioned-model/[a-z0-9]{12})|([0-9]{12}:imported-model/[a-z0-9]{12})|([0-9]{12}:application-inference-profile/[a-z0-9]{12})|([0-9]{12}:inference-profile/(([a-z-]{2,8}.)[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))|([0-9]{12}:(default-prompt-router|prompt-router)/[a-zA-Z0-9-:.]+)))|(([a-z]{2,4}[.]{1})([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63}))|arn:aws(-[^:]+)?:sagemaker:[a-z0-9-]{1,20}:[0-9]{12}:endpoint/[a-z0-9-]{1,63}`

Required: No

ragIdentifiers

This member has been deprecated.

The Amazon Resource Names (ARNs) of the knowledge base resources used for a knowledge base evaluation job.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EvaluatorModelConfig

Service: Amazon Bedrock

Specifies the model configuration for the evaluator model. `EvaluatorModelConfig` is required for evaluation jobs that use a knowledge base or in model evaluation job that use a model as judge. This model computes all evaluation related metrics.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bedrockEvaluatorModels

The evaluator model used in knowledge base evaluation job or in model evaluation job that use a model as judge. This model computes all evaluation related metrics.

Type: Array of BedrockEvaluatorModel objects

Array Members: Fixed number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExternalSource

Service: Amazon Bedrock

The unique external source of the content contained in the wrapper object.

Contents

sourceType

The source type of the external source wrapper object.

Type: String

Valid Values: `S3 | BYTE_CONTENT`

Required: Yes

byteContent

The identifier, content type, and data of the external source wrapper object.

Type: ByteContentDoc object

Required: No

s3Location

The S3 location of the external source wrapper object.

Type: S3ObjectDoc object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExternalSourcesGenerationConfiguration

Service: Amazon Bedrock

The response generation configuration of the external source wrapper object.

Contents

additionalModelRequestFields

Additional model parameters and their corresponding values not included in the text inference configuration for an external source. Takes in custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

guardrailConfiguration

Configuration details for the guardrail.

Type: GuardrailConfiguration object

Required: No

kbInferenceConfig

Configuration details for inference when using `RetrieveAndGenerate` to generate responses while using an external source.

Type: KbInferenceConfig object

Required: No

promptTemplate

Contains the template for the prompt for the external source wrapper object.

Type: PromptTemplate object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExternalSourcesRetrieveAndGenerateConfiguration

Service: Amazon Bedrock

The configuration of the external source wrapper object in the `retrieveAndGenerate` function.

Contents

modelArn

The Amazon Resource Name (ARN) of the foundation model or inference profile used to generate responses.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

sources

The document for the external source wrapper object in the `retrieveAndGenerate` function.

Type: Array of ExternalSource objects

Array Members: Fixed number of 1 item.

Required: Yes

generationConfiguration

Contains configurations details for response generation based on retrieved text chunks.

Type: ExternalSourcesGenerationConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FieldForReranking

Service: Amazon Bedrock

Specifies a field to be used during the reranking process in a Knowledge Base vector search. This structure identifies metadata fields that should be considered when reordering search results to improve relevance.

Contents

fieldName

The name of the metadata field to be used during the reranking process.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FilterAttribute

Service: Amazon Bedrock

Specifies the name of the metadata attribute/field to apply filters. You must match the name of the attribute/field in your data source/document metadata.

Contents

key

The name of metadata attribute/field, which must match the name in your data source/document metadata.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

value

The value of the metadata attribute/field.

Type: JSON value

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FoundationModelDetails

Service: Amazon Bedrock

Information about a foundation model.

Contents

modelArn

The model Amazon Resource Name (ARN).

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

Required: Yes

modelId

The model identifier.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

customizationsSupported

The customization that the model supports.

Type: Array of strings

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION`

Required: No

inferenceTypesSupported

The inference types that the model supports.

Type: Array of strings

Valid Values: `ON_DEMAND | PROVISIONED`

Required: No

inputModalities

The input modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

modelLifecycle

Contains details about whether a model version is available or deprecated

Type: FoundationModelLifecycle object

Required: No

modelName

The model name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `.*`

Required: No

outputModalities

The output modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

providerName

The model's provider name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `.*`

Required: No

responseStreamingSupported

Indicates whether the model supports streaming.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FoundationModelLifecycle

Service: Amazon Bedrock

Details about whether a model version is available or deprecated.

Contents

status

Specifies whether a model version is available (`ACTIVE`) or deprecated (`LEGACY`.

Type: String

Valid Values: `ACTIVE | LEGACY`

Required: Yes

endOfLifeTime

Time when model is no longer available for use in all AWS regions

Type: Timestamp

Required: No

legacyTime

Time when the model enters legacy state

Type: Timestamp

Required: No

publicExtendedAccessTime

Public extended accesss portion of the legacy period, where users should expect higher pricing

Type: Timestamp

Required: No

startOfLifeTime

Launch time when the model first becomes available

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FoundationModelSummary

Service: Amazon Bedrock

Summary information for a foundation model.

Contents

modelArn

The Amazon Resource Name (ARN) of the foundation model.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

Required: Yes

modelId

The model ID of the foundation model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)`

Required: Yes

customizationsSupported

Whether the model supports fine-tuning or continual pre-training.

Type: Array of strings

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION`

Required: No

inferenceTypesSupported

The inference types that the model supports.

Type: Array of strings

Valid Values: `ON_DEMAND | PROVISIONED`

Required: No

inputModalities

The input modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

modelLifecycle

Contains details about whether a model version is available or deprecated.

Type: FoundationModelLifecycle object

Required: No

modelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `.*`

Required: No

outputModalities

The output modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

providerName

The model's provider name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `.*`

Required: No

responseStreamingSupported

Indicates whether the model supports streaming.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GenerationConfiguration

Service: Amazon Bedrock

The configuration details for response generation based on retrieved text chunks.

Contents

additionalModelRequestFields

Additional model parameters and corresponding values not included in the `textInferenceConfig` structure for a knowledge base. This allows you to provide custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

guardrailConfiguration

Contains configuration details for the guardrail.

Type: GuardrailConfiguration object

Required: No

kbInferenceConfig

Contains configuration details for inference for knowledge base retrieval and response generation.

Type: KbInferenceConfig object

Required: No

promptTemplate

Contains the template for the prompt that's sent to the model for response generation.

Type: PromptTemplate object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GraderConfig

Service: Amazon Bedrock

Configuration for the grader used in reinforcement fine-tuning to evaluate model responses and provide reward signals.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

lambdaGrader

Configuration for using an AWS Lambda function as the grader for evaluating model responses and provide reward signals in reinforcement fine-tuning.

Type: LambdaGraderConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningPolicy

Service: Amazon Bedrock

Represents the configuration of Automated Reasoning policies within a Amazon Bedrock Guardrail, including the policies to apply and confidence thresholds.

Contents

policies

The list of Automated Reasoning policy ARNs that should be applied as part of this guardrail configuration.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

confidenceThreshold

The minimum confidence level required for Automated Reasoning policy violations to trigger guardrail actions. Values range from 0.0 to 1.0.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

GuardrailAutomatedReasoningPolicyConfig

Service: Amazon Bedrock

Configuration settings for integrating Automated Reasoning policies with Amazon Bedrock Guardrails.

Contents

policies

The list of Automated Reasoning policy ARNs to include in the guardrail configuration.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: Yes

confidenceThreshold

The confidence threshold for triggering guardrail actions based on Automated Reasoning policy violations.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

GuardrailConfiguration

Service: Amazon Bedrock

The configuration details for the guardrail.

Contents

guardrailId

The unique identifier for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

Required: Yes

guardrailVersion

The version of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(([1-9][0-9]{0,7})|(DRAFT))`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentFilter

Service: Amazon Bedrock

Contains filter strengths for harmful content. Guardrails support the following content filters to detect and filter harmful user inputs and FM-generated outputs.

- **Hate** – Describes language or a statement that discriminates, criticizes, insults, denounces, or dehumanizes a person or group on the basis of an identity (such as race, ethnicity, gender, religion, sexual orientation, ability, and national origin).
- **Insults** – Describes language or a statement that includes demeaning, humiliating, mocking, insulting, or belittling language. This type of language is also labeled as bullying.
- **Sexual** – Describes language or a statement that indicates sexual interest, activity, or arousal using direct or indirect references to body parts, physical traits, or sex.
- **Violence** – Describes language or a statement that includes glorification of or threats to inflict physical pain, hurt, or injury toward a person, group or thing.

Content filtering depends on the confidence classification of user inputs and FM responses across each of the four harmful categories. All input and output statements are classified into one of four confidence levels (NONE, LOW, MEDIUM, HIGH) for each harmful category. For example, if a statement is classified as *Hate* with HIGH confidence, the likelihood of the statement representing hateful content is high. A single statement can be classified across multiple categories with varying confidence levels. For example, a single statement can be classified as *Hate* with HIGH confidence, *Insults* with LOW confidence, *Sexual* with NONE confidence, and *Violence* with MEDIUM confidence.

For more information, see Guardrails content filters.

This data type is used in the following API operations:

- GetGuardrail response body

Contents

inputStrength

The strength of the content filter to apply to prompts. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: Yes

outputStrength

The strength of the content filter to apply to model responses. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: Yes

type

The harmful category that the content filter is applied to.

Type: String

Valid Values: `SEXUAL | VIOLENCE | HATE | INSULTS | MISCONDUCT | PROMPT_ATTACK`

Required: Yes

inputAction

The action to take when harmful content is detected in the input. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

inputEnabled

Indicates whether guardrail evaluation is enabled on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

inputModalities

The input modalities selected for the guardrail content filter.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `TEXT | IMAGE`

Required: No

outputAction

The action to take when harmful content is detected in the output. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

outputEnabled

Indicates whether guardrail evaluation is enabled on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputModalities

The output modalities selected for the guardrail content filter.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `TEXT | IMAGE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentFilterConfig

Service: Amazon Bedrock

Contains filter strengths for harmful content. Guardrails support the following content filters to detect and filter harmful user inputs and FM-generated outputs.

- **Hate** – Describes language or a statement that discriminates, criticizes, insults, denounces, or dehumanizes a person or group on the basis of an identity (such as race, ethnicity, gender, religion, sexual orientation, ability, and national origin).
- **Insults** – Describes language or a statement that includes demeaning, humiliating, mocking, insulting, or belittling language. This type of language is also labeled as bullying.
- **Sexual** – Describes language or a statement that indicates sexual interest, activity, or arousal using direct or indirect references to body parts, physical traits, or sex.
- **Violence** – Describes language or a statement that includes glorification of or threats to inflict physical pain, hurt, or injury toward a person, group or thing.

Content filtering depends on the confidence classification of user inputs and FM responses across each of the four harmful categories. All input and output statements are classified into one of four confidence levels (NONE, LOW, MEDIUM, HIGH) for each harmful category. For example, if a statement is classified as *Hate* with HIGH confidence, the likelihood of the statement representing hateful content is high. A single statement can be classified across multiple categories with varying confidence levels. For example, a single statement can be classified as *Hate* with HIGH confidence, *Insults* with LOW confidence, *Sexual* with NONE confidence, and *Violence* with MEDIUM confidence.

For more information, see Guardrails content filters.

Contents

inputStrength

The strength of the content filter to apply to prompts. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: Yes

outputStrength

The strength of the content filter to apply to model responses. As you increase the filter strength, the likelihood of filtering harmful content increases and the probability of seeing harmful content in your application reduces.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: Yes

type

The harmful category that the content filter is applied to.

Type: String

Valid Values: `SEXUAL | VIOLENCE | HATE | INSULTS | MISCONDUCT | PROMPT_ATTACK`

Required: Yes

inputAction

Specifies the action to take when harmful content is detected. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

inputEnabled

Specifies whether to enable guardrail evaluation on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

inputModalities

The input modalities selected for the guardrail content filter configuration.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `TEXT | IMAGE`

Required: No

outputAction

Specifies the action to take when harmful content is detected in the output. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

outputEnabled

Specifies whether to enable guardrail evaluation on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputModalities

The output modalities selected for the guardrail content filter configuration.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `TEXT | IMAGE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentFiltersTier

Service: Amazon Bedrock

The tier that your guardrail uses for content filters.

Contents

tierName

The tier that your guardrail uses for content filters. Valid values include:

- `CLASSIC` tier – Provides established guardrails functionality supporting English, French, and Spanish languages.
- `STANDARD` tier – Provides a more robust solution than the `CLASSIC` tier and has more comprehensive language support. This tier requires that your guardrail use cross-Region inference.

Type: String

Valid Values: `CLASSIC | STANDARD`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentFiltersTierConfig

Service: Amazon Bedrock

The tier that your guardrail uses for content filters. Consider using a tier that balances performance, accuracy, and compatibility with your existing generative AI workflows.

Contents

tierName

The tier that your guardrail uses for content filters. Valid values include:

- `CLASSIC` tier – Provides established guardrails functionality supporting English, French, and Spanish languages.
- `STANDARD` tier – Provides a more robust solution than the `CLASSIC` tier and has more comprehensive language support. This tier requires that your guardrail use cross-Region inference.

Type: String

Valid Values: `CLASSIC | STANDARD`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentPolicy

Service: Amazon Bedrock

Contains details about how to handle harmful content.

This data type is used in the following API operations:

- GetGuardrail response body

Contents

filters

Contains the type of the content filter and how strongly it should apply to prompts and model responses.

Type: Array of GuardrailContentFilter objects

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Required: No

tier

The tier that your guardrail uses for content filters.

Type: GuardrailContentFiltersTier object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentPolicyConfig

Service: Amazon Bedrock

Contains details about how to handle harmful content.

Contents

filtersConfig

Contains the type of the content filter and how strongly it should apply to prompts and model responses.

Type: Array of GuardrailContentFilterConfig objects

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Required: Yes

tierConfig

The tier that your guardrail uses for content filters.

Type: GuardrailContentFiltersTierConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContextualGroundingFilter

Service: Amazon Bedrock

The details for the guardrails contextual grounding filter.

Contents

threshold

The threshold details for the guardrails contextual grounding filter.

Type: Double

Valid Range: Minimum value of 0.

Required: Yes

type

The filter type details for the guardrails contextual grounding filter.

Type: String

Valid Values: `GROUNDING | RELEVANCE`

Required: Yes

action

The action to take when content fails the contextual grounding evaluation. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

enabled

Indicates whether contextual grounding is enabled for evaluation. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContextualGroundingFilterConfig

Service: Amazon Bedrock

The filter configuration details for the guardrails contextual grounding filter.

Contents

threshold

The threshold details for the guardrails contextual grounding filter.

Type: Double

Valid Range: Minimum value of 0.

Required: Yes

type

The filter details for the guardrails contextual grounding filter.

Type: String

Valid Values: `GROUNDING | RELEVANCE`

Required: Yes

action

Specifies the action to take when content fails the contextual grounding evaluation. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

enabled

Specifies whether to enable contextual grounding evaluation. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContextualGroundingPolicy

Service: Amazon Bedrock

The details for the guardrails contextual grounding policy.

Contents

filters

The filter details for the guardrails contextual grounding policy.

Type: Array of GuardrailContextualGroundingFilter objects

Array Members: Minimum number of 1 item.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContextualGroundingPolicyConfig

Service: Amazon Bedrock

The policy configuration details for the guardrails contextual grounding policy.

Contents

filtersConfig

The filter configuration details for the guardrails contextual grounding policy.

Type: Array of GuardrailContextualGroundingFilterConfig objects

Array Members: Minimum number of 1 item.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailCrossRegionConfig

Service: Amazon Bedrock

The system-defined guardrail profile that you're using with your guardrail. Guardrail profiles define the destination AWS Regions where guardrail inference requests can be automatically routed. Using guardrail profiles helps maintain guardrail performance and reliability when demand increases.

For more information, see the Amazon Bedrock User Guide.

Contents

guardrailProfileIdentifier

The ID or Amazon Resource Name (ARN) of the guardrail profile that your guardrail is using. Guardrail profile availability depends on your current AWS Region. For more information, see the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 15. Maximum length of 2048.

Pattern: `[a-z0-9-]+[.]{1}guardrail[.]{1}v[0-9:]+|arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail-profile/[a-z0-9-]+[.]{1}guardrail[.]{1}v[0-9:]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailCrossRegionDetails

Service: Amazon Bedrock

Contains details about the system-defined guardrail profile that you're using with your guardrail for cross-Region inference.

For more information, see the Amazon Bedrock User Guide.

Contents

guardrailProfileArn

The Amazon Resource Name (ARN) of the guardrail profile that you're using with your guardrail.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail-profile/[a-z0-9-]+[.]{1}guardrail[.]{1}v[0-9:]+`

Required: No

guardrailProfileId

The ID of the guardrail profile that your guardrail is using. Profile availability depends on your current AWS Region. For more information, see the Amazon Bedrock User Guide.

Type: String

Length Constraints: Minimum length of 15. Maximum length of 30.

Pattern: `[a-z0-9-]+[.]{1}guardrail[.]{1}v[0-9:]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailManagedWords

Service: Amazon Bedrock

The managed word list that was configured for the guardrail. (This is a list of words that are pre-defined and managed by guardrails only.)

Contents

type

ManagedWords$type The managed word type that was configured for the guardrail. (For now, we only offer profanity word list)

Type: String

Valid Values: `PROFANITY`

Required: Yes

inputAction

The action to take when harmful content is detected in the input. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

inputEnabled

Indicates whether guardrail evaluation is enabled on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

The action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

outputEnabled

Indicates whether guardrail evaluation is enabled on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailManagedWordsConfig

Service: Amazon Bedrock

The managed word list to configure for the guardrail.

Contents

type

The managed word type to configure for the guardrail.

Type: String

Valid Values: `PROFANITY`

Required: Yes

inputAction

Specifies the action to take when harmful content is detected in the input. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `NONE` – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

inputEnabled

Specifies whether to enable guardrail evaluation on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

Specifies the action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

outputEnabled

Specifies whether to enable guardrail evaluation on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailPiiEntity

Service: Amazon Bedrock

The PII entity configured for the guardrail.

Contents

action

The configured guardrail action when PII entity is detected.

Type: String

Valid Values: `BLOCK | ANONYMIZE | NONE`

Required: Yes

type

The type of PII entity. For example, Social Security Number.

Type: String

Valid Values: `ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER`

Required: Yes

inputAction

The action to take when harmful content is detected in the input. Supported values include:

- `BLOCK` – Block the content and replace it with blocked messaging.
- `ANONYMIZE` – Mask the content and replace it with identifier tags.

- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

inputEnabled

Indicates whether guardrail evaluation is enabled on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

The action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- ANONYMIZE – Mask the content and replace it with identifier tags.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

outputEnabled

Indicates whether guardrail evaluation is enabled on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailPiiEntityConfig

Service: Amazon Bedrock

The PII entity to configure for the guardrail.

Contents

action

Configure guardrail action when the PII entity is detected.

Type: String

Valid Values: `BLOCK | ANONYMIZE | NONE`

Required: Yes

type

Configure guardrail type when the PII entity is detected.

The following PIIs are used to block or mask sensitive information:

- **General**
  - **ADDRESS**

    A physical address, such as "100 Main Street, Anytown, USA" or "Suite #12, Building 123". An address can include information such as the street, building, location, city, state, country, county, zip code, precinct, and neighborhood.
  - **AGE**

    An individual's age, including the quantity and unit of time. For example, in the phrase "I am 40 years old," Guardrails recognizes "40 years" as an age.
  - **NAME**

    An individual's name. This entity type does not include titles, such as Dr., Mr., Mrs., or Miss. guardrails doesn't apply this entity type to names that are part of organizations or addresses. For example, guardrails recognizes the "John Doe Organization" as an organization, and it recognizes "Jane Doe Street" as an address.
  - **EMAIL**

    An email address, such as *marymajor@email.com*.

- **PHONE**

  A phone number. This entity type also includes fax and pager numbers.

- **USERNAME**

  A user name that identifies an account, such as a login name, screen name, nick name, or handle.

- **PASSWORD**

  An alphanumeric string that is used as a password, such as "*very20special#pass*".

- **DRIVER_ID**

  The number assigned to a driver's license, which is an official document permitting an individual to operate one or more motorized vehicles on a public road. A driver's license number consists of alphanumeric characters.

- **LICENSE_PLATE**

  A license plate for a vehicle is issued by the state or country where the vehicle is registered. The format for passenger vehicles is typically five to eight digits, consisting of upper-case letters and numbers. The format varies depending on the location of the issuing state or country.

- **VEHICLE_IDENTIFICATION_NUMBER**

  A Vehicle Identification Number (VIN) uniquely identifies a vehicle. VIN content and format are defined in the *ISO 3779* specification. Each country has specific codes and formats for VINs.

- **Finance**

  - **CREDIT_DEBIT_CARD_CVV**

    A three-digit card verification code (CVV) that is present on VISA, MasterCard, and Discover credit and debit cards. For American Express credit or debit cards, the CVV is a four-digit numeric code.

  - **CREDIT_DEBIT_CARD_EXPIRY**

    The expiration date for a credit or debit card. This number is usually four digits long and is often formatted as *month/year* or *MM/YY*. Guardrails recognizes expiration dates such as *01/21*, *01/2021*, and *Jan 2021*.

- **CREDIT_DEBIT_CARD_NUMBER**

  The number for a credit or debit card. These numbers can vary from 13 to 16 digits in length. However, Amazon Comprehend also recognizes credit or debit card numbers when only the last four digits are present.

- **PIN**

  A four-digit personal identification number (PIN) with which you can access your bank account.

- **INTERNATIONAL_BANK_ACCOUNT_NUMBER**

  An International Bank Account Number has specific formats in each country. For more information, see www.iban.com/structure.

- **SWIFT_CODE**

  A SWIFT code is a standard format of Bank Identifier Code (BIC) used to specify a particular bank or branch. Banks use these codes for money transfers such as international wire transfers.

  SWIFT codes consist of eight or 11 characters. The 11-digit codes refer to specific branches, while eight-digit codes (or 11-digit codes ending in 'XXX') refer to the head or primary office.

- **IT**
  - **IP_ADDRESS**

    An IPv4 address, such as *198.51.100.0*.

  - **MAC_ADDRESS**

    A *media access control* (MAC) address is a unique identifier assigned to a network interface controller (NIC).

  - **URL**

    A web address, such as *www.example.com*.

  - **AWS_ACCESS_KEY**

    A unique identifier that's associated with a secret access key; you use the access key ID and secret access key to sign programmatic AWS requests cryptographically.

- **AWS_SECRET_KEY**

  A unique identifier that's associated with an access key. You use the access key ID and secret access key to sign programmatic AWS requests cryptographically.

- **USA specific**
  - **US_BANK_ACCOUNT_NUMBER**

    A US bank account number, which is typically 10 to 12 digits long.

  - **US_BANK_ROUTING_NUMBER**

    A US bank account routing number. These are typically nine digits long,

  - **US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER**

    A US Individual Taxpayer Identification Number (ITIN) is a nine-digit number that starts with a "9" and contain a "7" or "8" as the fourth digit. An ITIN can be formatted with a space or a dash after the third and forth digits.

  - **US_PASSPORT_NUMBER**

    A US passport number. Passport numbers range from six to nine alphanumeric characters.

  - **US_SOCIAL_SECURITY_NUMBER**

    A US Social Security Number (SSN) is a nine-digit number that is issued to US citizens, permanent residents, and temporary working residents.

- **Canada specific**
  - **CA_HEALTH_NUMBER**

    A Canadian Health Service Number is a 10-digit unique identifier, required for individuals to access healthcare benefits.

  - **CA_SOCIAL_INSURANCE_NUMBER**

    A Canadian Social Insurance Number (SIN) is a nine-digit unique identifier, required for individuals to access government programs and benefits.

    The SIN is formatted as three groups of three digits, such as *123-456-789*. A SIN can be validated through a simple check-digit process called the Luhn algorithm.

- **UK Specific**
  - **UK_NATIONAL_HEALTH_SERVICE_NUMBER**

A UK National Health Service Number is a 10-17 digit number, such as *485 777 3456*. The current system formats the 10-digit number with spaces after the third and sixth digits. The final digit is an error-detecting checksum.

- **UK_NATIONAL_INSURANCE_NUMBER**

  A UK National Insurance Number (NINO) provides individuals with access to National Insurance (social security) benefits. It is also used for some purposes in the UK tax system.

  The number is nine digits long and starts with two letters, followed by six numbers and one letter. A NINO can be formatted with a space or a dash after the two letters and after the second, forth, and sixth digits.

- **UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER**

  A UK Unique Taxpayer Reference (UTR) is a 10-digit number that identifies a taxpayer or a business.

- **Custom**
  - **Regex filter** - You can use a regular expressions to define patterns for a guardrail to recognize and act upon such as serial number, booking ID etc..

Type: String

Valid Values: `ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER`

Required: Yes

inputAction

Specifies the action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- ANONYMIZE – Mask the content and replace it with identifier tags.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

inputEnabled

Specifies whether to enable guardrail evaluation on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

Specifies the action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- ANONYMIZE – Mask the content and replace it with identifier tags.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

outputEnabled

Specifies whether to enable guardrail evaluation on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailRegex

Service: Amazon Bedrock

The regular expression configured for the guardrail.

Contents

action

The action taken when a match to the regular expression is detected.

Type: String

Valid Values: `BLOCK | ANONYMIZE | NONE`

Required: Yes

name

The name of the regular expression for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

pattern

The pattern of the regular expression configured for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

description

The description of the regular expression for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

inputAction

The action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

inputEnabled

Indicates whether guardrail evaluation is enabled on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

The action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

outputEnabled

Indicates whether guardrail evaluation is enabled on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailRegexConfig

Service: Amazon Bedrock

The regular expression to configure for the guardrail.

Contents

action

The guardrail action to configure when matching regular expression is detected.

Type: String

Valid Values: `BLOCK | ANONYMIZE | NONE`

Required: Yes

name

The name of the regular expression to configure for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

pattern

The regular expression pattern to configure for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: Yes

description

The description of the regular expression to configure for the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

inputAction

Specifies the action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

inputEnabled

Specifies whether to enable guardrail evaluation on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

Specifies the action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | ANONYMIZE | NONE

Required: No

outputEnabled

Specifies whether to enable guardrail evaluation on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailSensitiveInformationPolicy

Service: Amazon Bedrock

Contains details about PII entities and regular expressions configured for the guardrail.

Contents

piiEntities

The list of PII entities configured for the guardrail.

Type: Array of GuardrailPiiEntity objects

Array Members: Minimum number of 1 item.

Required: No

regexes

The list of regular expressions configured for the guardrail.

Type: Array of GuardrailRegex objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailSensitiveInformationPolicyConfig

Service: Amazon Bedrock

Contains details about PII entities and regular expressions to configure for the guardrail.

Contents

piiEntitiesConfig

A list of PII entities to configure to the guardrail.

Type: Array of GuardrailPiiEntityConfig objects

Array Members: Minimum number of 1 item.

Required: No

regexesConfig

A list of regular expressions to configure to the guardrail.

Type: Array of GuardrailRegexConfig objects

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailSummary

Service: Amazon Bedrock

Contains details about a guardrail.

This data type is used in the following API operations:

- ListGuardrails response body

Contents

arn

The ARN of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+`

Required: Yes

createdAt

The date and time at which the guardrail was created.

Type: Timestamp

Required: Yes

id

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

Required: Yes

name

The name of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Pattern: `[0-9a-zA-Z-_]+`

Required: Yes

status

The status of the guardrail.

Type: String

Valid Values: `CREATING | UPDATING | VERSIONING | READY | FAILED | DELETING`

Required: Yes

updatedAt

The date and time at which the guardrail was last updated.

Type: Timestamp

Required: Yes

version

The version of the guardrail.

Type: String

Pattern: `(([1-9][0-9]{0,7})|(DRAFT))`

Required: Yes

crossRegionDetails

Details about the system-defined guardrail profile that you're using with your guardrail, including the guardrail profile ID and Amazon Resource Name (ARN).

Type: GuardrailCrossRegionDetails object

Required: No

description

A description of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopic

Service: Amazon Bedrock

Details about topics for the guardrail to identify and deny.

This data type is used in the following API operations:

- GetGuardrail response body

Contents

definition

A definition of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

name

The name of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[0-9a-zA-Z-_ !?.]+`

Required: Yes

examples

A list of prompts, each of which is an example of a prompt that can be categorized as belonging to the topic.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

inputAction

The action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

inputEnabled

Indicates whether guardrail evaluation is enabled on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

The action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

outputEnabled

Indicates whether guardrail evaluation is enabled on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

type

Specifies to deny the topic.

Type: String

Valid Values: DENY

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicConfig

Service: Amazon Bedrock

Details about topics for the guardrail to identify and deny.

Contents

definition

A definition of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

name

The name of the topic to deny.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[0-9a-zA-Z-_ !?.]+`

Required: Yes

type

Specifies to deny the topic.

Type: String

Valid Values: DENY

Required: Yes

examples

A list of prompts, each of which is an example of a prompt that can be categorized as belonging to the topic.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

inputAction

Specifies the action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

inputEnabled

Specifies whether to enable guardrail evaluation on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

Specifies the action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

outputEnabled

Specifies whether to enable guardrail evaluation on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicPolicy

Service: Amazon Bedrock

Contains details about topics that the guardrail should identify and deny.

This data type is used in the following API operations:

- GetGuardrail response body

Contents

topics

A list of policies related to topics that the guardrail should deny.

Type: Array of GuardrailTopic objects

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Required: Yes

tier

The tier that your guardrail uses for denied topic filters.

Type: GuardrailTopicsTier object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicPolicyConfig

Service: Amazon Bedrock

Contains details about topics that the guardrail should identify and deny.

Contents

topicsConfig

A list of policies related to topics that the guardrail should deny.

Type: Array of GuardrailTopicConfig objects

Array Members: Minimum number of 1 item. Maximum number of 30 items.

Required: Yes

tierConfig

The tier that your guardrail uses for denied topic filters.

Type: GuardrailTopicsTierConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicsTier

Service: Amazon Bedrock

The tier that your guardrail uses for denied topic filters.

Contents

tierName

The tier that your guardrail uses for denied topic filters. Valid values include:

- `CLASSIC` tier – Provides established guardrails functionality supporting English, French, and Spanish languages.
- `STANDARD` tier – Provides a more robust solution than the `CLASSIC` tier and has more comprehensive language support. This tier requires that your guardrail use cross-Region inference.

Type: String

Valid Values: `CLASSIC | STANDARD`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicsTierConfig

Service: Amazon Bedrock

The tier that your guardrail uses for denied topic filters. Consider using a tier that balances performance, accuracy, and compatibility with your existing generative AI workflows.

Contents

tierName

The tier that your guardrail uses for denied topic filters. Valid values include:

- `CLASSIC` tier – Provides established guardrails functionality supporting English, French, and Spanish languages.
- `STANDARD` tier – Provides a more robust solution than the `CLASSIC` tier and has more comprehensive language support. This tier requires that your guardrail use cross-Region inference.

Type: String

Valid Values: `CLASSIC | STANDARD`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailWord

Service: Amazon Bedrock

A word configured for the guardrail.

Contents

text

Text of the word configured for the guardrail to block.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

inputAction

The action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

inputEnabled

Indicates whether guardrail evaluation is enabled on the input. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

The action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: `BLOCK | NONE`

Required: No

outputEnabled

Indicates whether guardrail evaluation is enabled on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailWordConfig

Service: Amazon Bedrock

A word to configure for the guardrail.

Contents

text

Text of the word configured for the guardrail to block.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

inputAction

Specifies the action to take when harmful content is detected in the input. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

inputEnabled

Specifies whether to enable guardrail evaluation on the intput. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

outputAction

Specifies the action to take when harmful content is detected in the output. Supported values include:

- BLOCK – Block the content and replace it with blocked messaging.
- NONE – Take no action but return detection information in the trace response.

Type: String

Valid Values: BLOCK | NONE

Required: No

outputEnabled

Specifies whether to enable guardrail evaluation on the output. When disabled, you aren't charged for the evaluation. The evaluation doesn't appear in the response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailWordPolicy

Service: Amazon Bedrock

Contains details about the word policy configured for the guardrail.

Contents

managedWordLists

A list of managed words configured for the guardrail.

Type: Array of GuardrailManagedWords objects

Required: No

words

A list of words configured for the guardrail.

Type: Array of GuardrailWord objects

Array Members: Minimum number of 1 item. Maximum number of 10000 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailWordPolicyConfig

Service: Amazon Bedrock

Contains details about the word policy to configured for the guardrail.

Contents

managedWordListsConfig

A list of managed words to configure for the guardrail.

Type: Array of GuardrailManagedWordsConfig objects

Required: No

wordsConfig

A list of words to configure for the guardrail.

Type: Array of GuardrailWordConfig objects

Array Members: Minimum number of 1 item. Maximum number of 10000 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanEvaluationConfig

Service: Amazon Bedrock

Specifies the custom metrics, how tasks will be rated, the flow definition ARN, and your custom prompt datasets. Model evaluation jobs use human workers *only* support the use of custom prompt datasets. To learn more about custom prompt datasets and the required format, see Custom prompt datasets.

When you create custom metrics in `HumanEvaluationCustomMetric` you must specify the metric's `name`. The list of `names` specified in the `HumanEvaluationCustomMetric` array, must match the `metricNames` array of strings specified in `EvaluationDatasetMetricConfig`. For example, if in the `HumanEvaluationCustomMetric` array your specified the names `"accuracy", "toxicity", "readability"` as custom metrics *then* the `metricNames` array would need to look like the following `["accuracy", "toxicity", "readability"]` in `EvaluationDatasetMetricConfig`.

Contents

datasetMetricConfigs

Use to specify the metrics, task, and prompt dataset to be used in your model evaluation job.

Type: Array of EvaluationDatasetMetricConfig objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

customMetrics

A `HumanEvaluationCustomMetric` object. It contains the names the metrics, how the metrics are to be evaluated, an optional description.

Type: Array of HumanEvaluationCustomMetric objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

humanWorkflowConfig

The parameters of the human workflow.

Type: HumanWorkflowConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanEvaluationCustomMetric

Service: Amazon Bedrock

In a model evaluation job that uses human workers you must define the name of the metric, and how you want that metric rated `ratingMethod`, and an optional description of the metric.

Contents

name

The name of the metric. Your human evaluators will see this name in the evaluation UI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[0-9a-zA-Z-_.]+`

Required: Yes

ratingMethod

Choose how you want your human workers to evaluation your model. Valid values for rating methods are `ThumbsUpDown`, `IndividualLikertScale`,`ComparisonLikertScale`, `ComparisonChoice`, and `ComparisonRank`

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[0-9a-zA-Z-_]+`

Required: Yes

description

An optional description of the metric. Use this parameter to provide more details about the metric.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `.+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HumanWorkflowConfig

Service: Amazon Bedrock

Contains `SageMakerFlowDefinition` object. The object is used to specify the prompt dataset, task type, rating method and metric names.

Contents

flowDefinitionArn

The Amazon Resource Number (ARN) for the flow definition

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `arn:aws(-[^:]+)?:sagemaker:[a-z0-9-]{1,20}:[0-9]{12}:flow-definition/.*`

Required: Yes

instructions

Instructions for the flow definition

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5000.

Pattern: `[\S\s]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImplicitFilterConfiguration

Service: Amazon Bedrock

Configuration for implicit filtering in Knowledge Base vector searches. Implicit filtering allows you to automatically filter search results based on metadata attributes without requiring explicit filter expressions in each query.

Contents

metadataAttributes

A list of metadata attribute schemas that define the structure and properties of metadata fields used for implicit filtering. Each attribute defines a key, type, and optional description.

Type: Array of MetadataAttributeSchema objects

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the foundation model used for implicit filtering. This model processes the query to extract relevant filtering criteria.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImportedModelSummary

Service: Amazon Bedrock

Information about the imported model.

Contents

creationTime

Creation time of the imported model.

Type: Timestamp

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the imported model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:imported-model/[a-z0-9]{12}`

Required: Yes

modelName

Name of the imported model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Required: Yes

instructSupported

Specifies if the imported model supports converse.

Type: Boolean

Required: No

modelArchitecture

The architecture of the imported model.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceConfiguration

Service: Amazon Bedrock

Inference parameters for a model.

Contents

maxTokens

The maximum number of tokens to generate.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

stopSequences

Stop sequences that end generation.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2500 items.

Length Constraints: Minimum length of 1.

Required: No

temperature

The temperature for sampling.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topP

The top-p value for nucleus sampling.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceProfileModel

Service: Amazon Bedrock

Contains information about a model.

Contents

modelArn

The Amazon Resource Name (ARN) of the model.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceProfileModelSource

Service: Amazon Bedrock

Contains information about the model or system-defined inference profile that is the source for an inference profile..

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

copyFrom

The ARN of the model or system-defined inference profile that is the source for the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|foundation-model)/[a-zA-Z0-9-:.]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceProfileSummary

Service: Amazon Bedrock

Contains information about an inference profile.

Contents

inferenceProfileArn

The Amazon Resource Name (ARN) of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+`

Required: Yes

inferenceProfileId

The unique identifier of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9-:.]+`

Required: Yes

inferenceProfileName

The name of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `([0-9a-zA-Z][ _-]?)+`

Required: Yes

models

A list of information about each model in the inference profile.

Type: Array of InferenceProfileModel objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

status

The status of the inference profile. `ACTIVE` means that the inference profile is ready to be used.

Type: String

Valid Values: `ACTIVE`

Required: Yes

type

The type of the inference profile. The following types are possible:

- `SYSTEM_DEFINED` – The inference profile is defined by Amazon Bedrock. You can route inference requests across regions with these inference profiles.
- `APPLICATION` – The inference profile was created by a user. This type of inference profile can track metrics and costs when invoking the model in it. The inference profile may route requests to one or multiple regions.

Type: String

Valid Values: `SYSTEM_DEFINED | APPLICATION`

Required: Yes

createdAt

The time at which the inference profile was created.

Type: Timestamp

Required: No

description

The description of the inference profile.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `([0-9a-zA-Z:.][ _-]?)+`

Required: No

updatedAt

The time at which the inference profile was last updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationLogsConfig

Service: Amazon Bedrock

Settings for using invocation logs to customize a model.

Contents

invocationLogSource

The source of the invocation logs.

Type: InvocationLogSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

requestMetadataFilters

Rules for filtering invocation logs based on request metadata.

Type: RequestMetadataFilters object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

usePromptResponse

Whether to use the model's response for training, or just the prompt. The default value is `False`.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

InvocationLogSource

Service: Amazon Bedrock

A storage location for invocation logs.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3Uri

The URI of an invocation log in a bucket.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KbInferenceConfig

Service: Amazon Bedrock

Contains configuration details of the inference for knowledge base retrieval and response generation.

Contents

textInferenceConfig

Contains configuration details for text generation using a language model via the `RetrieveAndGenerate` function.

Type: TextInferenceConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseConfig

Service: Amazon Bedrock

The configuration details for retrieving information from a knowledge base and generating responses.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

retrieveAndGenerateConfig

Contains configuration details for retrieving information from a knowledge base and generating responses.

Type: RetrieveAndGenerateConfiguration object

Required: No

retrieveConfig

Contains configuration details for retrieving information from a knowledge base.

Type: RetrieveConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseRetrievalConfiguration

Service: Amazon Bedrock

Contains configuration details for retrieving information from a knowledge base.

Contents

vectorSearchConfiguration

Contains configuration details for returning the results from the vector search.

Type: KnowledgeBaseVectorSearchConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseRetrieveAndGenerateConfiguration

Service: Amazon Bedrock

Contains configuration details for retrieving information from a knowledge base and generating responses.

Contents

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the foundation model or inference profile used to generate responses.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

generationConfiguration

Contains configurations details for response generation based on retrieved text chunks.

Type: GenerationConfiguration object

Required: No

orchestrationConfiguration

Contains configuration details for the model to process the prompt prior to retrieval and response generation.

Type: OrchestrationConfiguration object

Required: No

retrievalConfiguration

Contains configuration details for retrieving text chunks.

Type: KnowledgeBaseRetrievalConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseVectorSearchConfiguration

Service: Amazon Bedrock

The configuration details for returning the results from the knowledge base vector search.

Contents

filter

Specifies the filters to use on the metadata fields in the knowledge base data sources before returning results.

Type: RetrievalFilter object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

implicitFilterConfiguration

Configuration for implicit filtering in Knowledge Base vector searches. This allows the system to automatically apply filters based on the query context without requiring explicit filter expressions.

Type: ImplicitFilterConfiguration object

Required: No

numberOfResults

The number of text chunks to retrieve; the number of results to return.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

overrideSearchType

By default, Amazon Bedrock decides a search strategy for you. If you're using an Amazon OpenSearch Serverless vector store that contains a filterable text field, you can specify whether to query the knowledge base with a `HYBRID` search using both vector embeddings and raw text,

or SEMANTIC search using only vector embeddings. For other vector store configurations, only SEMANTIC search is available.

Type: String

Valid Values: HYBRID | SEMANTIC

Required: No

rerankingConfiguration

Configuration for reranking search results in Knowledge Base vector searches. Reranking improves search relevance by reordering initial vector search results using more sophisticated relevance models.

Type: VectorSearchRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LambdaGraderConfig

Service: Amazon Bedrock

Configuration for using an AWS Lambda function to grade model responses during reinforcement fine-tuning training.

Contents

lambdaArn

ARN of the AWS Lambda function that will evaluate model responses and return reward scores for RFT training.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LegalTerm

Service: Amazon Bedrock

The legal term of the agreement.

Contents

url

URL to the legal term document.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LoggingConfig

Service: Amazon Bedrock

Configuration fields for invocation logging.

Contents

audioDataDeliveryEnabled

Set to include audio data in the log delivery.

Type: Boolean

Required: No

cloudWatchConfig

CloudWatch logging configuration.

Type: CloudWatchConfig object

Required: No

embeddingDataDeliveryEnabled

Set to include embeddings data in the log delivery.

Type: Boolean

Required: No

imageDataDeliveryEnabled

Set to include image data in the log delivery.

Type: Boolean

Required: No

s3Config

S3 configuration for storing log data.

Type: S3Config object

Required: No

textDataDeliveryEnabled

Set to include text data in the log delivery.

Type: Boolean

Required: No

videoDataDeliveryEnabled

Set to include video data in the log delivery.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MarketplaceModelEndpoint

Service: Amazon Bedrock

Contains details about an endpoint for a model from Amazon Bedrock Marketplace.

Contents

createdAt

The timestamp when the endpoint was registered.

Type: Timestamp

Required: Yes

endpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

endpointConfig

The configuration of the endpoint, including the number and type of instances used.

Type: EndpointConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

endpointStatus

The current status of the endpoint (e.g., Creating, InService, Updating, Failed).

Type: String

Required: Yes

modelSourceIdentifier

The ARN of the model from Amazon Bedrock Marketplace that is deployed on this endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*arn:aws:sagemaker:.*:hub-content/SageMakerPublicHub/Model/.*`

Required: Yes

updatedAt

The timestamp when the endpoint was last updated.

Type: Timestamp

Required: Yes

endpointStatusMessage

Additional information about the endpoint status, if available.

Type: String

Required: No

status

The overall status of the endpoint in Amazon Bedrock Marketplace (e.g., ACTIVE, INACTIVE).

Type: String

Valid Values: `REGISTERED | INCOMPATIBLE_ENDPOINT`

Required: No

statusMessage

Additional information about the overall status, if available.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MarketplaceModelEndpointSummary

Service: Amazon Bedrock

Provides a summary of an endpoint for a model from Amazon Bedrock Marketplace.

Contents

createdAt

The timestamp when the endpoint was created.

Type: Timestamp

Required: Yes

endpointArn

The Amazon Resource Name (ARN) of the endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

modelSourceIdentifier

The ARN of the model from Amazon Bedrock Marketplace that is deployed on this endpoint.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*arn:aws:sagemaker:.*:hub-content/SageMakerPublicHub/Model/.*`

Required: Yes

updatedAt

The timestamp when the endpoint was last updated.

Type: Timestamp

Required: Yes

status

The overall status of the endpoint in Amazon Bedrock Marketplace.

Type: String

Valid Values: `REGISTERED | INCOMPATIBLE_ENDPOINT`

Required: No

statusMessage

Additional information about the overall status, if available.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataAttributeSchema

Service: Amazon Bedrock

Defines the schema for a metadata attribute used in Knowledge Base vector searches. Metadata attributes provide additional context for documents and can be used for filtering and reranking search results.

Contents

description

An optional description of the metadata attribute that provides additional context about its purpose and usage.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

key

The unique identifier for the metadata attribute. This key is used to reference the attribute in filter expressions and reranking configurations.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[\s\S]+`

Required: Yes

type

The data type of the metadata attribute. The type determines how the attribute can be used in filter expressions and reranking.

Type: String

Valid Values: `STRING | NUMBER | BOOLEAN | STRING_LIST`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataConfigurationForReranking

Service: Amazon Bedrock

Configuration for how metadata should be used during the reranking process in Knowledge Base vector searches. This determines which metadata fields are included or excluded when reordering search results.

Contents

selectionMode

The mode for selecting which metadata fields to include in the reranking process. Valid values are ALL (use all available metadata fields) or SELECTIVE (use only specified fields).

Type: String

Valid Values: `SELECTIVE | ALL`

Required: Yes

selectiveModeConfiguration

Configuration for selective mode, which allows you to explicitly include or exclude specific metadata fields during reranking. This is only used when selectionMode is set to SELECTIVE.

Type: RerankingMetadataSelectiveModeConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelConfiguration

Service: Amazon Bedrock

Configuration for a model used in advanced prompt optimization.

Contents

modelId

The model to use for optimization. The value depends on the resource that you use:

- If you use a base model, specify the model ID or its ARN. For a list of model IDs, see Models at a glance in the Amazon Bedrock User Guide.
- If you use a cross-Region (system-defined) inference profile, specify the inference profile ID or its ARN. For a list of inference profile IDs, see Supported Regions and models for inference profiles in the Amazon Bedrock User Guide.
- If you use an application inference profile, specify its full ARN, including the account ID and Region.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|))|(arn:aws(|-us-gov|-cn|-iso|-iso-b|-iso-f):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:application-inference-profile/[a-zA-Z0-9-:.]+)|(arn:aws(|-us-gov|-cn|-iso|-iso-b|-iso-f):bedrock:[a-z0-9-]{1,20}:([0-9]{12})?:inference-profile/[a-zA-Z0-9-:.]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.:]?[a-z0-9-]{1,63})))`

Required: Yes

additionalModelRequestFields

Additional model-specific request fields.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

inferenceConfig

The inference configuration for the model.

Type: InferenceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCopyJobSummary

Service: Amazon Bedrock

Contains details about each model copy job.

This data type is used in the following API operations:

- ListModelCopyJobs response

Contents

creationTime

The time that the model copy job was created.

Type: Timestamp

Required: Yes

jobArn

The Amazon Resoource Name (ARN) of the model copy job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-copy-job/[a-z0-9]{12}`

Required: Yes

sourceAccountId

The unique identifier of the account that the model being copied originated from.

Type: String

Pattern: `[0-9]{12}`

Required: Yes

sourceModelArn

The Amazon Resource Name (ARN) of the original model being copied.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

status

The status of the model copy job.

Type: String

Valid Values: `InProgress | Completed | Failed`

Required: Yes

targetModelArn

The Amazon Resource Name (ARN) of the copied model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

Required: Yes

failureMessage

If a model fails to be copied, a message describing why the job failed is included here.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

sourceModelName

The name of the original model being copied.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: No

targetModelKmsKeyArn

The Amazon Resource Name (ARN) of the KMS key used to encrypt the copied model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

targetModelName

The name of the copied model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: No

targetModelTags

Tags associated with the copied model.

Type: Array of Tag objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCustomizationJobSummary

Service: Amazon Bedrock

Information about one customization job

Contents

baseModelArn

Amazon Resource Name (ARN) of the base model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

creationTime

Creation time of the custom model.

Type: Timestamp

Required: Yes

jobArn

Amazon Resource Name (ARN) of the customization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}`

Required: Yes

jobName

Name of the customization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

Required: Yes

status

Status of the customization job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

customizationType

Specifies whether to carry out continued pre-training of a model or whether to fine-tune it. For more information, see Custom models.

Type: String

Valid Values: `FINE_TUNING | CONTINUED_PRE_TRAINING | DISTILLATION | REINFORCEMENT_FINE_TUNING | IMPORTED`

Required: No

customModelArn

Amazon Resource Name (ARN) of the custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/(imported|[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-`

`z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})/[a-z0-9]{12}`

Required: No

customModelName

Name of the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?){1,63}`

Required: No

endTime

Time that the customization job ended.

Type: Timestamp

Required: No

lastModifiedTime

Time that the customization job was last modified.

Type: Timestamp

Required: No

statusDetails

Details about the status of the data processing sub-task of the job.

Type: StatusDetails object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelDataSource

Service: Amazon Bedrock

The data source of the model to import.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3DataSource

The Amazon S3 data source of the model to import.

Type: S3DataSource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelEnforcement

Service: Amazon Bedrock

Model-specific information for the enforced guardrail configuration.

Contents

excludedModels

Models to exclude from enforcement of the guardrail.

Type: Array of strings

Array Members: Minimum number of 0 items.

Pattern: `([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}){0,1}`

Required: Yes

includedModels

Models to enforce the guardrail on.

Type: Array of strings

Array Members: Minimum number of 1 item.

Pattern: `(ALL|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}){0,1})`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelImportJobSummary

Service: Amazon Bedrock

Information about the import job.

Contents

creationTime

The time import job was created.

Type: Timestamp

Required: Yes

jobArn

The Amazon Resource Name (ARN) of the import job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-import-job/[a-z0-9]{12}`

Required: Yes

jobName

The name of the import job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*`

Required: Yes

status

The status of the imported job.

Type: String

Valid Values: `InProgress | Completed | Failed`

Required: Yes

endTime

The time when import job ended.

Type: Timestamp

Required: No

importedModelArn

The Amazon resource Name (ARN) of the imported model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:imported-model/[a-z0-9]{12}`

Required: No

importedModelName

The name of the imported model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Required: No

lastModifiedTime

The time when the import job was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInvocationJobInputDataConfig

Service: Amazon Bedrock

Details about the location of the input to the batch inference job.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3InputDataConfig

Contains the configuration of the S3 location of the input data.

Type: ModelInvocationJobS3InputDataConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInvocationJobOutputDataConfig

Service: Amazon Bedrock

Contains the configuration of the S3 location of the output data.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3OutputDataConfig

Contains the configuration of the S3 location of the output data.

Type: ModelInvocationJobS3OutputDataConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInvocationJobS3InputDataConfig

Service: Amazon Bedrock

Contains the configuration of the S3 location of the input data.

Contents

s3Uri

The S3 location of the input data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

s3BucketOwner

The ID of the AWS account that owns the S3 bucket containing the input data.

Type: String

Pattern: `[0-9]{12}`

Required: No

s3InputFormat

The format of the input data.

Type: String

Valid Values: `JSONL`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInvocationJobS3OutputDataConfig

Service: Amazon Bedrock

Contains the configuration of the S3 location of the output data.

Contents

s3Uri

The S3 location of the output data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

s3BucketOwner

The ID of the AWS account that owns the S3 bucket containing the output data.

Type: String

Pattern: `[0-9]{12}`

Required: No

s3EncryptionKeyId

The unique identifier of the key that encrypts the S3 location of the output data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInvocationJobSummary

Service: Amazon Bedrock

A summary of a batch inference job.

Contents

inputDataConfig

Details about the location of the input to the batch inference job.

Type: ModelInvocationJobInputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

jobArn

The Amazon Resource Name (ARN) of the batch inference job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-invocation-job/[a-z0-9]{12})`

Required: Yes

jobName

The name of the batch inference job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `[a-zA-Z0-9]{1,63}(-*[a-zA-Z0-9\+\-\.]){0,63}`

Required: Yes

modelId

The unique identifier of the foundation model used for model inference.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-:]{1,63}/[a-z0-9]{12}$)|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$)|([0-9]{12}:(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+$)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)`

Required: Yes

outputDataConfig

Details about the location of the output of the batch inference job.

Type: ModelInvocationJobOutputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the service role with permissions to carry out and manage batch inference. You can use the console to create a default service role or follow the steps at Create a service role for batch inference.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

submitTime

The time at which the batch inference job was submitted.

Type: Timestamp

Required: Yes

clientRequestToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9]{1,256}(-*[a-zA-Z0-9]){0,256}`

Required: No

endTime

The time at which the batch inference job ended.

Type: Timestamp

Required: No

errorRecordCount

The number of records that failed to process in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

Required: No

jobExpirationTime

The time at which the batch inference job times or timed out.

Type: Timestamp

Required: No

lastModifiedTime

The time at which the batch inference job was last modified.

Type: Timestamp

Required: No

message

If the batch inference job failed, this field contains a message describing why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

modelInvocationType

The invocation endpoint for ModelInvocationJob

Type: String

Valid Values: `InvokeModel | Converse`

Required: No

processedRecordCount

The number of records that have been processed in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

Required: No

status

The status of the batch inference job.

The following statuses are possible:

- Submitted – This job has been submitted to a queue for validation.
- Validating – This job is being validated for the requirements described in Format and upload your batch inference data. The criteria include the following:
  - Your IAM service role has access to the Amazon S3 buckets containing your files.
  - Your files are .jsonl files and each individual record is a JSON object in the correct format. Note that validation doesn't check if the `modelInput` value matches the request body for the model.

  - Your files fulfill the requirements for file size and number of records. For more information, see Quotas for Amazon Bedrock.
- Scheduled – This job has been validated and is now in a queue. The job will automatically start when it reaches its turn.
- Expired – This job timed out because it was scheduled but didn't begin before the set timeout duration. Submit a new job request.
- InProgress – This job has begun. You can start viewing the results in the output S3 location.
- Completed – This job has successfully completed. View the output files in the output S3 location.
- PartiallyCompleted – This job has partially completed. Not all of your records could be processed in time. View the output files in the output S3 location.
- Failed – This job has failed. Check the failure message for any further details. For further assistance, reach out to the Support Center.
- Stopped – This job was stopped by a user.
- Stopping – This job is being stopped by a user.

Type: String

Valid Values: `Submitted | InProgress | Completed | Failed | Stopping | Stopped | PartiallyCompleted | Expired | Validating | Scheduled`

Required: No

successRecordCount

The number of records that were successfully processed in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

Required: No

timeoutDurationInHours

The number of hours after which the batch inference job was set to time out.

Type: Integer

Valid Range: Minimum value of 24. Maximum value of 168.

Required: No

totalRecordCount

The total number of records in the batch inference job.

Type: Long

Valid Range: Minimum value of 0.

Required: No

vpcConfig

The configuration of the Virtual Private Cloud (VPC) for the data in the batch inference job. For more information, see Protect batch inference jobs using a VPC.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPackageArnDataSource

Service: Amazon Bedrock

Contains the Amazon Resource Name (ARN) of a SageMaker AI model package to use as the data source for a custom model.

Contents

modelPackageArn

The Amazon Resource Name (ARN) of the SageMaker AI model package. The ARN must be for a model package of `restricted` type.

To use a model package ARN, you must have the `sagemaker:DescribeModelPackage` and `sagemaker:AccessModelPackageData` permissions on the model package resource.

Type: String

Pattern: `arn:aws[a-z\-]*:sagemaker:[a-z0-9\-]{9,16}:[0-9]{12}:model-package/[\S]{1,2048}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Offer

Service: Amazon Bedrock

An offer dictates usage terms for the model.

Contents

offerToken

Offer token.

Type: String

Required: Yes

termDetails

Details about the terms of the offer.

Type: TermDetails object

Required: Yes

offerId

Offer Id for a model offer.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OrchestrationConfiguration

Service: Amazon Bedrock

The configuration details for the model to process the prompt prior to retrieval and response generation.

Contents

queryTransformationConfiguration

Contains configuration details for transforming the prompt.

Type: QueryTransformationConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputDataConfig

Service: Amazon Bedrock

S3 Location of the output data.

Contents

s3Uri

The S3 URI where the output data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PerformanceConfiguration

Service: Amazon Bedrock

Contains performance settings for a model.

Contents

latency

Specifies whether to use the latency-optimized or standard version of a model or inference profile.

Type: String

Valid Values: `standard | optimized`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PricingTerm

Service: Amazon Bedrock

Describes the usage-based pricing term.

Contents

rateCard

Describes a usage price for each dimension.

Type: Array of DimensionalPriceRate objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptRouterSummary

Service: Amazon Bedrock

Details about a prompt router.

Contents

fallbackModel

The router's fallback model.

Type: PromptRouterTargetModel object

Required: Yes

models

The router's models.

Type: Array of PromptRouterTargetModel objects

Required: Yes

promptRouterArn

The router's ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:default-prompt-router/[a-zA-Z0-9-:.]+`

Required: Yes

promptRouterName

The router's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `([0-9a-zA-Z][ _-]?)+`

Required: Yes

routingCriteria

The router's routing criteria.

Type: RoutingCriteria object

Required: Yes

status

The router's status.

Type: String

Valid Values: `AVAILABLE`

Required: Yes

type

The summary's type.

Type: String

Valid Values: `custom | default`

Required: Yes

createdAt

When the router was created.

Type: Timestamp

Required: No

description

The router's description.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `([0-9a-zA-Z:.][ _-]?)+`

Required: No

updatedAt

When the router was updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptRouterTargetModel

Service: Amazon Bedrock

The target model for a prompt router.

Contents

modelArn

The target model's ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})|(^arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{0,20}):(|[0-9]{12}):(inference-profile|application-inference-profile)/[a-zA-Z0-9-:.]+)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptTemplate

Service: Amazon Bedrock

The template for the prompt that's sent to the model for response generation.

Contents

textPromptTemplate

The template for the prompt that's sent to the model for response generation. You can include prompt placeholders, which become replaced before the prompt is sent to the model to provide instructions and context to the model. In addition, you can include XML tags to delineate meaningful sections of the prompt template.

For more information, see Knowledge base prompt template and Use XML tags with Anthropic Claude models.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProvisionedModelSummary

Service: Amazon Bedrock

A summary of information about a Provisioned Throughput.

This data type is used in the following API operations:

- ListProvisionedThroughputs response

Contents

creationTime

The time that the Provisioned Throughput was created.

Type: Timestamp

Required: Yes

desiredModelArn

The Amazon Resource Name (ARN) of the model requested to be associated to this Provisioned Throughput. This value differs from the `modelArn` if updating hasn't completed.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

desiredModelUnits

The number of model units that was requested to be allocated to the Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

foundationModelArn

The Amazon Resource Name (ARN) of the base model for which the Provisioned Throughput was created, or of the base model that the custom model for which the Provisioned Throughput was created was customized.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}`

Required: Yes

lastModifiedTime

The time that the Provisioned Throughput was last modified.

Type: Timestamp

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the model associated with the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/((imported)|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}))(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))`

Required: Yes

modelUnits

The number of model units allocated to the Provisioned Throughput.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

provisionedModelArn

The Amazon Resource Name (ARN) of the Provisioned Throughput.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}`

Required: Yes

provisionedModelName

The name of the Provisioned Throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `([0-9a-zA-Z][_-]?)+`

Required: Yes

status

The status of the Provisioned Throughput.

Type: String

Valid Values: `Creating | InService | Updating | Failed`

Required: Yes

commitmentDuration

The duration for which the Provisioned Throughput was committed.

Type: String

Valid Values: `OneMonth | SixMonths`

Required: No

commitmentExpirationTime

The timestamp for when the commitment term of the Provisioned Throughput expires.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryTransformationConfiguration

Service: Amazon Bedrock

The configuration details for transforming the prompt.

Contents

type

The type of transformation to apply to the prompt.

Type: String

Valid Values: `QUERY_DECOMPOSITION`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RAGConfig

Service: Amazon Bedrock

Contains configuration details for retrieval of information and response generation.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

knowledgeBaseConfig

Contains configuration details for knowledge base retrieval and response generation.

Type: KnowledgeBaseConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

precomputedRagSourceConfig

Contains configuration details about the RAG source used to generate inference response data for a Knowledge Base evaluation job.

Type: EvaluationPrecomputedRagSourceConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

RatingScaleItem

Service: Amazon Bedrock

Defines the value and corresponding definition for one rating in a custom metric rating scale.

Contents

definition

Defines the definition for one rating in a custom metric rating scale.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

value

Defines the value for one rating in a custom metric rating scale.

Type: RatingScaleItemValue object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RatingScaleItemValue

Service: Amazon Bedrock

Defines the value for one rating in a custom metric rating scale.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

floatValue

A floating point number representing the value for a rating in a custom metric rating scale.

Type: Float

Required: No

stringValue

A string representing the value for a rating in a custom metric rating scale.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RequestMetadataBaseFilters

Service: Amazon Bedrock

A mapping of a metadata key to a value that it should or should not equal.

Contents

equals

Include results where the key equals the value.

Type: String to string map

Map Entries: Maximum number of 1 item.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `[a-zA-Z0-9\s._:/=+$@-]{1,256}`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+$@-]{0,256}`

Required: No

notEquals

Include results where the key does not equal the value.

Type: String to string map

Map Entries: Maximum number of 1 item.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `[a-zA-Z0-9\s._:/=+$@-]{1,256}`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+$@-]{0,256}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RequestMetadataFilters

Service: Amazon Bedrock

Rules for filtering invocation logs. A filter can be a mapping of a metadata key to a value that it should or should not equal (a base filter), or a list of base filters that are all applied with `AND` or `OR` logical operators

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

andAll

Include results where all of the based filters match.

Type: Array of RequestMetadataBaseFilters objects

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Required: No

equals

Include results where the key equals the value.

Type: String to string map

Map Entries: Maximum number of 1 item.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `[a-zA-Z0-9\s._:/=+$@-]{1,256}`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+$@-]{0,256}`

Required: No

notEquals

Include results where the key does not equal the value.

Type: String to string map

Map Entries: Maximum number of 1 item.

Key Length Constraints: Minimum length of 1. Maximum length of 256.

Key Pattern: `[a-zA-Z0-9\s._:/=+$@-]{1,256}`

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Value Pattern: `[a-zA-Z0-9\s._:/=+$@-]{0,256}`

Required: No

orAll

Include results where any of the base filters match.

Type: Array of RequestMetadataBaseFilters objects

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankingMetadataSelectiveModeConfiguration

Service: Amazon Bedrock

Configuration for selectively including or excluding metadata fields during the reranking process. This allows you to control which metadata attributes are considered when reordering search results.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

fieldsToExclude

A list of metadata field names to explicitly exclude from the reranking process. All metadata fields except these will be considered when reordering search results. This parameter cannot be used together with fieldsToInclude.

Type: Array of FieldForReranking objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

fieldsToInclude

A list of metadata field names to explicitly include in the reranking process. Only these fields will be considered when reordering search results. This parameter cannot be used together with fieldsToExclude.

Type: Array of FieldForReranking objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalFilter

Service: Amazon Bedrock

Specifies the filters to use on the metadata attributes/fields in the knowledge base data sources before returning results.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

andAll

Knowledge base data sources are returned if their metadata attributes fulfill all the filter conditions inside this list.

Type: Array of RetrievalFilter objects

Array Members: Minimum number of 2 items.

Required: No

equals

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value matches the value in this object.

The following example would return data sources with an animal attribute whose value is 'cat': `"equals": { "key": "animal", "value": "cat" }`

Type: FilterAttribute object

Required: No

greaterThan

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is greater than the value in this object.

The following example would return data sources with an year attribute whose value is greater than '1989': `"greaterThan": { "key": "year", "value": 1989 }`

Type: FilterAttribute object

Required: No

greaterThanOrEquals

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is greater than or equal to the value in this object.

The following example would return data sources with an year attribute whose value is greater than or equal to '1989': `"greaterThanOrEquals": { "key": "year", "value": 1989 }`

Type: FilterAttribute object

Required: No

in

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is in the list specified in the value in this object.

The following example would return data sources with an animal attribute that is either 'cat' or 'dog': `"in": { "key": "animal", "value": ["cat", "dog"] }`

Type: FilterAttribute object

Required: No

lessThan

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is less than the value in this object.

The following example would return data sources with an year attribute whose value is less than to '1989': `"lessThan": { "key": "year", "value": 1989 }`

Type: FilterAttribute object

Required: No

lessThanOrEquals

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is less than or equal to the value in this object.

The following example would return data sources with an year attribute whose value is less than or equal to '1989': `"lessThanOrEquals": { "key": "year", "value": 1989 }`

Type: FilterAttribute object

Required: No

listContains

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is a list that contains the value as one of its members.

The following example would return data sources with an animals attribute that is a list containing a cat member (for example, `["dog", "cat"]`): `"listContains": { "key": "animals", "value": "cat" }`

Type: FilterAttribute object

Required: No

notEquals

Knowledge base data sources that contain a metadata attribute whose name matches the key and whose value doesn't match the value in this object are returned.

The following example would return data sources that don't contain an animal attribute whose value is 'cat': `"notEquals": { "key": "animal", "value": "cat" }`

Type: FilterAttribute object

Required: No

notIn

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value isn't in the list specified in the value in this object.

The following example would return data sources whose animal attribute is neither 'cat' nor 'dog': `"notIn": { "key": "animal", "value": ["cat", "dog"] }`

Type: FilterAttribute object

Required: No

orAll

Knowledge base data sources are returned if their metadata attributes fulfill at least one of the filter conditions inside this list.

Type: Array of RetrievalFilter objects

Array Members: Minimum number of 2 items.

Required: No

startsWith

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value starts with the value in this object. This filter is currently only supported for Amazon OpenSearch Serverless vector stores.

The following example would return data sources with an animal attribute starts with 'ca' (for example, 'cat' or 'camel'). `"startsWith": { "key": "animal", "value": "ca" }`

Type: FilterAttribute object

Required: No

stringContains

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the key and whose value is one of the following:

A string that contains the value as a substring. The following example would return data sources with an animal attribute that contains the substring at (for example, 'cat'): `"stringContains": { "key": "animal", "value": "at" }`

A list with a member that contains the value as a substring. The following example would return data sources with an animals attribute that is a list containing a member that contains the substring at (for example, `["dog", "cat"]`): `"stringContains": { "key": "animals", "value": "at" }`

Type: FilterAttribute object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateConfiguration

Service: Amazon Bedrock

Contains configuration details for a knowledge base retrieval and response generation.

Contents

type

The type of resource that contains your data for retrieving information and generating responses.

If you choose to use `EXTERNAL_SOURCES`, then currently only Claude 3 Sonnet models for knowledge bases are supported.

Type: String

Valid Values: `KNOWLEDGE_BASE | EXTERNAL_SOURCES`

Required: Yes

externalSourcesConfiguration

The configuration for the external source wrapper object in the `retrieveAndGenerate` function.

Type: ExternalSourcesRetrieveAndGenerateConfiguration object

Required: No

knowledgeBaseConfiguration

Contains configuration details for the knowledge base retrieval and response generation.

Type: KnowledgeBaseRetrieveAndGenerateConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveConfig

Service: Amazon Bedrock

The configuration details for retrieving information from a knowledge base.

Contents

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

knowledgeBaseRetrievalConfiguration

Contains configuration details for knowledge base retrieval.

Type: KnowledgeBaseRetrievalConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RFTConfig

Service: Amazon Bedrock

Configuration settings for reinforcement fine-tuning (RFT), including grader configuration and training hyperparameters.

Contents

graderConfig

Configuration for the grader that evaluates model responses and provides reward signals during RFT training.

Type: GraderConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

hyperParameters

Hyperparameters that control the reinforcement fine-tuning training process, including learning rate, batch size, and epoch count.

Type: RFTHyperParameters object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RFTHyperParameters

Service: Amazon Bedrock

Hyperparameters for controlling the reinforcement fine-tuning training process, including learning settings and evaluation intervals.

Contents

batchSize

Number of training samples processed in each batch during reinforcement fine-tuning (RFT) training. Larger batches may improve training stability.

Type: Integer

Valid Range: Minimum value of 16. Maximum value of 512.

Required: No

epochCount

Number of training epochs to run during reinforcement fine-tuning. Higher values may improve performance but increase training time.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 50.

Required: No

evalInterval

Interval between evaluation runs during RFT training, measured in training steps. More frequent evaluation provides better monitoring.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

inferenceMaxTokens

Maximum number of tokens the model can generate in response to each prompt during RFT training.

Type: Integer

Required: No

learningRate

Learning rate for the reinforcement fine-tuning. Controls how quickly the model adapts to reward signals.

Type: Float

Valid Range: Minimum value of 1.0e-07. Maximum value of 0.001.

Required: No

maxPromptLength

Maximum length of input prompts during RFT training, measured in tokens. Longer prompts allow more context but increase memory usage and training-time.

Type: Integer

Required: No

reasoningEffort

Level of reasoning effort applied during RFT training. Higher values may improve response quality but increase training time.

Type: String

Valid Values: `low | medium | high`

Required: No

trainingSamplePerPrompt

Number of response samples generated per prompt during RFT training. More samples provide better reward signal estimation.

Type: Integer

Valid Range: Minimum value of 2. Maximum value of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RoutingCriteria

Service: Amazon Bedrock

Routing criteria for a prompt router.

Contents

responseQualityDifference

The criteria's response quality difference.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 100.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Config

Service: Amazon Bedrock

S3 configuration for storing log data.

Contents

bucketName

S3 bucket name.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Required: Yes

keyPrefix

S3 prefix.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3DataSource

Service: Amazon Bedrock

The Amazon S3 data source of the model to import. For the CreateCustomModel API operation, you must specify the Amazon S3 URI for the Amazon-managed Amazon S3 bucket containing your model artifacts. SageMaker AI creates this bucket when you run your first SageMaker AI training job.

Contents

s3Uri

The URI of the Amazon S3 data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3ObjectDoc

Service: Amazon Bedrock

The unique wrapper object of the document from the S3 location.

Contents

uri

The S3 URI location for the wrapper object of the document.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]/.{1,1024}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SageMakerEndpoint

Service: Amazon Bedrock

Specifies the configuration for a Amazon SageMaker endpoint.

Contents

executionRole

The ARN of the IAM role that Amazon SageMaker can assume to access model artifacts and docker image for deployment on Amazon EC2 compute instances or for batch transform jobs.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

initialInstanceCount

The number of Amazon EC2 compute instances to deploy for initial endpoint creation.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

instanceType

The Amazon EC2 compute instance type to deploy for hosting the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Required: Yes

kmsEncryptionKey

The AWS KMS key that Amazon SageMaker uses to encrypt data on the storage volume attached to the Amazon EC2 compute instance that hosts the endpoint.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)))|([a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+)`

Required: No

vpc

The VPC configuration for the endpoint.

Type: VpcConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SelectiveContentGuarding

Service: Amazon Bedrock

Selective content guarding controls for enforced guardrails.

Contents

messages

Selective guarding mode for user messages.

Type: String

Valid Values: `SELECTIVE | COMPREHENSIVE`

Required: No

system

Selective guarding mode for system prompts.

Type: String

Valid Values: `SELECTIVE | COMPREHENSIVE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StatusDetails

Service: Amazon Bedrock

The status details for sub-tasks of a customization job. Possible statuses for each sub-task include the following:

- NotStarted
- InProgress
- Completed
- Stopping
- Stopped
- Failed

Contents

dataProcessingDetails

For a Distillation job, the status details for the data processing sub-task of the customization job.

Type: DataProcessingDetails object

Required: No

trainingDetails

The status details for the training sub-task of the job.

Type: TrainingDetails object

Required: No

validationDetails

The status details for the validation sub-task of the job.

Type: ValidationDetails object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SupportTerm

Service: Amazon Bedrock

Describes a support term.

Contents

refundPolicyDescription

Describes the refund policy.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tag

Service: Amazon Bedrock

Definition of the key/value pair for a tag.

Contents

key

Key for the tag.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

value

Value for the tag.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TeacherModelConfig

Service: Amazon Bedrock

Details about a teacher model used for model customization.

Contents

teacherModelIdentifier

The identifier of the teacher model.

Type: String

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:((:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$)|([0-9]{12}:inference-profile/[a-zA-Z0-9-:.]+$)))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})`

Required: Yes

maxResponseLengthForInference

The maximum number of tokens requested when the customization job invokes the teacher model.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TermDetails

Service: Amazon Bedrock

Describes the usage terms of an offer.

Contents

legalTerm

Describes the legal terms.

Type: LegalTerm object

Required: Yes

supportTerm

Describes the support terms.

Type: SupportTerm object

Required: Yes

usageBasedPricingTerm

Describes the pricing terms.

Type: PricingTerm object

Required: Yes

validityTerm

Describes the validity terms.

Type: ValidityTerm object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextInferenceConfig

Service: Amazon Bedrock

The configuration details for text generation using a language model via the `RetrieveAndGenerate` function.

Contents

maxTokens

The maximum number of tokens to generate in the output text. Do not use the minimum of 0 or the maximum of 65536. The limit values described here are arbitrary values, for actual values consult the limits defined by your specific model.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 65536.

Required: No

stopSequences

A list of sequences of characters that, if generated, will cause the model to stop generating further tokens. Do not use a minimum length of 1 or a maximum length of 1000. The limit values described here are arbitrary values, for actual values consult the limits defined by your specific model.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

temperature

Controls the random-ness of text generated by the language model, influencing how much the model sticks to the most predictable next words versus exploring more surprising options. A lower temperature value (e.g. 0.2 or 0.3) makes model outputs more deterministic or predictable, while a higher temperature (e.g. 0.8 or 0.9) makes the outputs more creative or unpredictable.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topP

A probability distribution threshold which controls what the model considers for the set of possible next tokens. The model will only consider the top p% of the probability distribution when generating the next token.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingDataConfig

Service: Amazon Bedrock

S3 Location of the training data.

Contents

invocationLogsConfig

Settings for using invocation logs to customize a model.

Type: InvocationLogsConfig object

Required: No

s3Uri

The S3 URI where the training data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingDetails

Service: Amazon Bedrock

The status details for the training sub-task of the customization job.

Contents

creationTime

The start time of the training sub-task of the job.

Type: Timestamp

Required: No

lastModifiedTime

The latest update to the training sub-task of the job.

Type: Timestamp

Required: No

status

The status of the training sub-task of the job.

Type: String

Valid Values: `InProgress | Completed | Stopping | Stopped | Failed | NotStarted`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingMetrics

Service: Amazon Bedrock

Metrics associated with the custom job.

Contents

trainingLoss

Loss metric associated with the custom job.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidationDataConfig

Service: Amazon Bedrock

Array of up to 10 validators.

Contents

validators

Information about the validators.

Type: Array of Validator objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidationDetails

Service: Amazon Bedrock

The status details for the validation sub-task of the customization job.

Contents

creationTime

The start time of the validation sub-task of the job.

Type: Timestamp

Required: No

lastModifiedTime

The latest update to the validation sub-task of the job.

Type: Timestamp

Required: No

status

The status of the validation sub-task of the job.

Type: String

Valid Values: `InProgress | Completed | Stopping | Stopped | Failed | NotStarted`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Validator

Service: Amazon Bedrock

Information about a validator.

Contents

s3Uri

The S3 URI where the validation data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][-.a-z0-9]{1,61}[a-z0-9](?:/[-!_*'().a-z0-9A-Z]+(?:/[-!_*'().a-z0-9A-Z]+)*)?/?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidatorMetric

Service: Amazon Bedrock

The metric for the validator.

Contents

validationLoss

The validation loss associated with this validator.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidityTerm

Service: Amazon Bedrock

Describes the validity terms.

Contents

agreementDuration

Describes the agreement duration.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchBedrockRerankingConfiguration

Service: Amazon Bedrock

Configuration for using Amazon Bedrock foundation models to rerank Knowledge Base vector search results. This enables more sophisticated relevance ranking using large language models.

Contents

modelConfiguration

Configuration for the Amazon Bedrock foundation model used for reranking. This includes the model ARN and any additional request fields required by the model.

Type: VectorSearchBedrockRerankingModelConfiguration object

Required: Yes

metadataConfiguration

Configuration for how document metadata should be used during the reranking process. This determines which metadata fields are included when reordering search results.

Type: MetadataConfigurationForReranking object

Required: No

numberOfRerankedResults

The maximum number of results to rerank. This limits how many of the initial vector search results will be processed by the reranking model. A smaller number improves performance but may exclude potentially relevant results.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchBedrockRerankingModelConfiguration

Service: Amazon Bedrock

Configuration for the Amazon Bedrock foundation model used for reranking vector search results. This specifies which model to use and any additional parameters required by the model.

Contents

modelArn

The Amazon Resource Name (ARN) of the foundation model to use for reranking. This model processes the query and search results to determine a more relevant ordering.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/(.*))?`

Required: Yes

additionalModelRequestFields

A list of additional fields to include in the model request during reranking. These fields provide extra context or configuration options specific to the selected foundation model.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchRerankingConfiguration

Service: Amazon Bedrock

Configuration for reranking vector search results to improve relevance. Reranking applies additional relevance models to reorder the initial vector search results based on more sophisticated criteria.

Contents

type

The type of reranking to apply to vector search results. Currently, the only supported value is BEDROCK, which uses Amazon Bedrock foundation models for reranking.

Type: String

Valid Values: `BEDROCK_RERANKING_MODEL`

Required: Yes

bedrockRerankingConfiguration

Configuration for using Amazon Bedrock foundation models to rerank search results. This is required when the reranking type is set to BEDROCK.

Type: VectorSearchBedrockRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VpcConfig

Service: Amazon Bedrock

The configuration of a virtual private cloud (VPC). For more information, see Protect your data using Amazon Virtual Private Cloud and AWS PrivateLink.

Contents

securityGroupIds

An array of IDs for each security group in the VPC to use.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

subnetIds

An array of IDs for each subnet in the VPC to use.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `[-0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

Agents for Amazon Bedrock

The following data types are supported by Agents for Amazon Bedrock:

- ActionGroupExecutor
- ActionGroupSummary
- Agent
- AgentActionGroup
- AgentAlias
- AgentAliasHistoryEvent
- AgentAliasRoutingConfigurationListItem
- AgentAliasSummary
- AgentCollaborator
- AgentCollaboratorSummary
- AgentDescriptor
- AgentFlowNodeConfiguration
- AgentKnowledgeBase
- AgentKnowledgeBaseSummary
- AgentSummary
- AgentVersion
- AgentVersionSummary
- AnyToolChoice
- APISchema
- AudioConfiguration
- AudioExtractionConfiguration
- AudioSegmentationConfiguration
- AutoToolChoice
- BedrockDataAutomationConfiguration

- BedrockEmbeddingModelConfiguration
- BedrockFoundationModelConfiguration
- BedrockFoundationModelContextEnrichmentConfiguration
- ByteContentDoc
- CachePointBlock
- ChatPromptTemplateConfiguration
- ChunkingConfiguration
- CollectorFlowNodeConfiguration
- ConditionFlowNodeConfiguration
- ConfluenceCrawlerConfiguration
- ConfluenceDataSourceConfiguration
- ConfluenceSourceConfiguration
- ContentBlock
- ContextEnrichmentConfiguration
- CrawlFilterConfiguration
- CuratedQuery
- CustomContent
- CustomDocumentIdentifier
- CustomOrchestration
- CustomS3Location
- CustomTransformationConfiguration
- CyclicConnectionFlowValidationDetails
- DailySchedule
- DataSource
- DataSourceConfiguration
- DataSourceSummary
- DayOfMonth
- DeletionProtectionConfiguration
- DocumentAccessControlEntry
- DocumentContent

- DocumentIdentifier
- DocumentMetadata
- DuplicateConditionExpressionFlowValidationDetails
- DuplicateConnectionsFlowValidationDetails
- EmbeddingModelConfiguration
- EnrichmentStrategyConfiguration
- FieldForReranking
- FixedSizeChunkingConfiguration
- FlowAliasConcurrencyConfiguration
- FlowAliasRoutingConfigurationListItem
- FlowAliasSummary
- FlowCondition
- FlowConditionalConnectionConfiguration
- FlowConnection
- FlowConnectionConfiguration
- FlowDataConnectionConfiguration
- FlowDefinition
- FlowNode
- FlowNodeConfiguration
- FlowNodeInput
- FlowNodeOutput
- FlowSummary
- FlowValidation
- FlowValidationDetails
- FlowVersionSummary
- Function
- FunctionSchema
- GuardrailConfiguration
- HierarchicalChunkingConfiguration
- HierarchicalChunkingLevelConfiguration

- ImageExtractionConfiguration
- IncompatibleConnectionDataTypeFlowValidationDetails
- InferenceConfiguration
- IngestionJob
- IngestionJobFilter
- IngestionJobSortBy
- IngestionJobStatistics
- IngestionJobSummary
- InlineCodeFlowNodeConfiguration
- InlineContent
- InputFlowNodeConfiguration
- IntermediateStorage
- InvalidLoopBoundaryFlowValidationDetails
- IteratorFlowNodeConfiguration
- KendraKnowledgeBaseConfiguration
- KnowledgeBase
- KnowledgeBaseConfiguration
- KnowledgeBaseDocument
- KnowledgeBaseDocumentDetail
- KnowledgeBaseFlowNodeConfiguration
- KnowledgeBaseOrchestrationConfiguration
- KnowledgeBasePromptTemplate
- KnowledgeBaseSummary
- LambdaFunctionFlowNodeConfiguration
- LastDayOfMonth
- LexFlowNodeConfiguration
- LoopControllerFlowNodeConfiguration
- LoopFlowNodeConfiguration
- LoopIncompatibleNodeTypeFlowValidationDetails
- LoopInputFlowNodeConfiguration

- MalformedConditionExpressionFlowValidationDetails
- MalformedNodeInputExpressionFlowValidationDetails
- ManagedKnowledgeBaseConfiguration
- ManagedKnowledgeBaseConnectorConfiguration
- MediaExtractionConfiguration
- MemoryConfiguration
- Message
- MetadataAttribute
- MetadataAttributeValue
- MetadataConfigurationForReranking
- MismatchedNodeInputTypeFlowValidationDetails
- MismatchedNodeOutputTypeFlowValidationDetails
- MissingConnectionConfigurationFlowValidationDetails
- MissingDefaultConditionFlowValidationDetails
- MissingEndingNodesFlowValidationDetails
- MissingLoopControllerNodeFlowValidationDetails
- MissingLoopInputNodeFlowValidationDetails
- MissingNodeConfigurationFlowValidationDetails
- MissingNodeInputFlowValidationDetails
- MissingNodeOutputFlowValidationDetails
- MissingStartingNodesFlowValidationDetails
- MongoDbAtlasConfiguration
- MongoDbAtlasFieldMapping
- MonthlySchedule
- MultipleLoopControllerNodesFlowValidationDetails
- MultipleLoopInputNodesFlowValidationDetails
- MultipleNodeInputConnectionsFlowValidationDetails
- NeptuneAnalyticsConfiguration
- NeptuneAnalyticsFieldMapping
- OpenSearchManagedClusterConfiguration

- OpenSearchManagedClusterFieldMapping
- OpenSearchServerlessConfiguration
- OpenSearchServerlessFieldMapping
- OrchestrationExecutor
- OutputFlowNodeConfiguration
- ParameterDetail
- ParsingConfiguration
- ParsingPrompt
- PatternObjectFilter
- PatternObjectFilterConfiguration
- PerformanceConfiguration
- PineconeConfiguration
- PineconeFieldMapping
- PromptAgentResource
- PromptConfiguration
- PromptFlowNodeConfiguration
- PromptFlowNodeInlineConfiguration
- PromptFlowNodeResourceConfiguration
- PromptFlowNodeSourceConfiguration
- PromptGenAiResource
- PromptInferenceConfiguration
- PromptInputVariable
- PromptMetadataEntry
- PromptModelInferenceConfiguration
- PromptOverrideConfiguration
- PromptSummary
- PromptTemplateConfiguration
- PromptVariant
- QueryGenerationColumn
- QueryGenerationConfiguration

- QueryGenerationContext
- QueryGenerationTable
- RdsConfiguration
- RdsFieldMapping
- RedisEnterpriseCloudConfiguration
- RedisEnterpriseCloudFieldMapping
- RedshiftConfiguration
- RedshiftProvisionedAuthConfiguration
- RedshiftProvisionedConfiguration
- RedshiftQueryEngineAwsDataCatalogStorageConfiguration
- RedshiftQueryEngineConfiguration
- RedshiftQueryEngineRedshiftStorageConfiguration
- RedshiftQueryEngineStorageConfiguration
- RedshiftServerlessAuthConfiguration
- RedshiftServerlessConfiguration
- RerankingMetadataSelectiveModeConfiguration
- RetrievalFlowNodeConfiguration
- RetrievalFlowNodeS3Configuration
- RetrievalFlowNodeServiceConfiguration
- S3Content
- S3DataSourceConfiguration
- S3Identifier
- S3Location
- S3VectorsConfiguration
- SalesforceCrawlerConfiguration
- SalesforceDataSourceConfiguration
- SalesforceSourceConfiguration
- SeedUrl
- SemanticChunkingConfiguration
- ServerSideEncryptionConfiguration

- SessionSummaryConfiguration
- SharePointCrawlerConfiguration
- SharePointDataSourceConfiguration
- SharePointSourceConfiguration
- SpecificToolChoice
- SqlKnowledgeBaseConfiguration
- StorageConfiguration
- StorageFlowNodeConfiguration
- StorageFlowNodeS3Configuration
- StorageFlowNodeServiceConfiguration
- SupplementalDataStorageConfiguration
- SupplementalDataStorageLocation
- SyncSchedule
- SystemContentBlock
- TextContentDoc
- TextPromptTemplateConfiguration
- Tool
- ToolChoice
- ToolConfiguration
- ToolInputSchema
- ToolSpecification
- Transformation
- TransformationFunction
- TransformationLambdaConfiguration
- UnfulfilledNodeInputFlowValidationDetails
- UnknownConnectionConditionFlowValidationDetails
- UnknownConnectionSourceFlowValidationDetails
- UnknownConnectionSourceOutputFlowValidationDetails
- UnknownConnectionTargetFlowValidationDetails
- UnknownConnectionTargetInputFlowValidationDetails

- UnknownNodeInputFlowValidationDetails
- UnknownNodeOutputFlowValidationDetails
- UnreachableNodeFlowValidationDetails
- UnsatisfiedConnectionConditionsFlowValidationDetails
- UnspecifiedFlowValidationDetails
- UrlConfiguration
- ValidationExceptionField
- VectorIngestionConfiguration
- VectorKnowledgeBaseConfiguration
- VectorSearchBedrockRerankingConfiguration
- VectorSearchBedrockRerankingModelConfiguration
- VectorSearchRerankingConfiguration
- VideoConfiguration
- VideoExtractionConfiguration
- VideoSegmentationConfiguration
- VpcConfiguration
- VpcConfigurationSummary
- WebCrawlerConfiguration
- WebCrawlerLimits
- WebDataSourceConfiguration
- WebSourceConfiguration
- WeeklySchedule

ActionGroupExecutor

Service: Agents for Amazon Bedrock

Contains details about the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

customControl

To return the action group invocation results directly in the `InvokeAgent` response, specify `RETURN_CONTROL`.

Type: String

Valid Values: `RETURN_CONTROL`

Required: No

lambda

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ActionGroupSummary

Service: Agents for Amazon Bedrock

Contains details about an action group.

Contents

actionGroupId

The unique identifier of the action group.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

actionGroupName

The name of the action group.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

updatedAt

The time at which the action group was last updated.

Type: Timestamp

Required: Yes

description

The description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Agent

Service: Agents for Amazon Bedrock

Contains details about an agent.

Contents

agentArn

The Amazon Resource Name (ARN) of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent/[0-9a-zA-Z]{10}`

Required: Yes

agentId

The unique identifier of the agent.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentName

The name of the agent.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

agentStatus

The status of the agent and whether it is ready for use. The following statuses are possible:

- CREATING – The agent is being created.
- PREPARING – The agent is being prepared.
- PREPARED – The agent is prepared and ready to be invoked.
- NOT_PREPARED – The agent has been created but not yet prepared.
- FAILED – The agent API operation failed.
- UPDATING – The agent is being updated.
- DELETING – The agent is being deleted.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

Required: Yes

agentVersion

The version of the agent.

Type: String

Length Constraints: Fixed length of 5.

Pattern: `DRAFT`

Required: Yes

createdAt

The time at which the agent was created.

Type: Timestamp

Required: Yes

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 5400.

Required: Yes

updatedAt

The time at which the agent was last updated.

Type: Timestamp

Required: Yes

agentCollaboration

The agent's collaboration settings.

Type: String

Valid Values: `SUPERVISOR | SUPERVISOR_ROUTER | DISABLED`

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key that encrypts the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

customOrchestration

Contains custom orchestration configurations for the agent.

Type: CustomOrchestration object

Required: No

description

The description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

Contains reasons that the agent-related API that you invoked failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

foundationModel

The foundation model used for orchestration by the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

guardrailConfiguration

Details about the guardrail associated with the agent.

Type: GuardrailConfiguration object

Required: No

instruction

Instructions that tell the agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 4000.

Required: No

memoryConfiguration

Contains memory configuration for the agent.

Type: MemoryConfiguration object

Required: No

orchestrationType

Specifies the orchestration strategy for the agent.

Type: String

Valid Values: `DEFAULT | CUSTOM_ORCHESTRATION`

Required: No

preparedAt

The time at which the agent was last prepared.

Type: Timestamp

Required: No

promptOverrideConfiguration

Contains configurations to override prompt templates in different parts of an agent sequence. For more information, see Advanced prompts.

Type: PromptOverrideConfiguration object

Required: No

recommendedActions

Contains recommended actions to take for the agent-related API that you invoked to succeed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentActionGroup

Service: Agents for Amazon Bedrock

Contains details about an action group.

Contents

actionGroupId

The unique identifier of the action group.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

actionGroupName

The name of the action group.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

actionGroupState

Specifies whether the action group is available for the agent to invoke or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

agentId

The unique identifier of the agent to which the action group belongs.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent to which the action group belongs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

createdAt

The time at which the action group was created.

Type: Timestamp

Required: Yes

updatedAt

The time at which the action group was last updated.

Type: Timestamp

Required: Yes

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Type: ActionGroupExecutor object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see Action group OpenAPI schemas.

Type: APISchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

The description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Defines functions that each define parameters that the agent needs to invoke from the user. Each function represents an action in an action group.

Type: FunctionSchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionGroupSignatureParams

The configuration settings for a computer use action.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Type: String to string map

Key Length Constraints: Minimum length of 0. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 100.

Required: No

parentActionSignature

If this field is set as `AMAZON.UserInput`, the agent can request the user for additional information when trying to complete a task. The `description`, `apiSchema`, and `actionGroupExecutor` fields must be blank for this action group.

During orchestration, if the agent determines that it needs to invoke an API in an action group, but doesn't have enough information to complete the API request, it will invoke this action group instead and return an Observation reprompting the user for more information.

Type: String

Valid Values: `AMAZON.UserInput | AMAZON.CodeInterpreter | ANTHROPIC.Computer | ANTHROPIC.Bash | ANTHROPIC.TextEditor`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentAlias

Service: Agents for Amazon Bedrock

Contains details about an alias of an agent.

Contents

agentAliasArn

The Amazon Resource Name (ARN) of the alias of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: Yes

agentAliasId

The unique identifier of the alias of the agent.

Type: String

Length Constraints: Fixed length of 10.

Pattern: `(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

agentAliasName

The name of the alias of the agent.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentAliasStatus

The status of the alias of the agent and whether it is ready for use. The following statuses are possible:

- CREATING – The agent alias is being created.
- PREPARED – The agent alias is finished being created or updated and is ready to be invoked.
- FAILED – The agent alias API operation failed.
- UPDATING – The agent alias is being updated.
- DELETING – The agent alias is being deleted.
- DISSOCIATED - The agent alias has no version associated with it.

Type: String

Valid Values: `CREATING | PREPARED | FAILED | UPDATING | DELETING | DISSOCIATED`

Required: Yes

agentId

The unique identifier of the agent.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

createdAt

The time at which the alias of the agent was created.

Type: Timestamp

Required: Yes

routingConfiguration

Contains details about the routing configuration of the alias.

Type: Array of AgentAliasRoutingConfigurationListItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: Yes

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Required: Yes

agentAliasHistoryEvents

Contains details about the history of the alias.

Type: Array of AgentAliasHistoryEvent objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

aliasInvocationState

The invocation state for the agent alias. If the agent alias is running, the value is `ACCEPT_INVOCATIONS`. If the agent alias is paused, the value is `REJECT_INVOCATIONS`. Use the `UpdateAgentAlias` operation to change the invocation state.

Type: String

Valid Values: `ACCEPT_INVOCATIONS | REJECT_INVOCATIONS`

Required: No

clientToken

A unique, case-sensitive identifier to ensure that the API request completes no more than one time. If this token matches a previous request, Amazon Bedrock ignores the request, but does not return an error. For more information, see Ensuring idempotency.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

description

The description of the alias of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

Information on the failure of Provisioned Throughput assigned to an agent alias.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentAliasHistoryEvent

Service: Agents for Amazon Bedrock

Contains details about the history of the alias.

Contents

endDate

The date that the alias stopped being associated to the version in the `routingConfiguration` object

Type: Timestamp

Required: No

routingConfiguration

Contains details about the version of the agent with which the alias is associated.

Type: Array of AgentAliasRoutingConfigurationListItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

startDate

The date that the alias began being associated to the version in the `routingConfiguration` object.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AgentAliasRoutingConfigurationListItem

Service: Agents for Amazon Bedrock

Contains details about the routing configuration of the alias.

Contents

agentVersion

The version of the agent with which the alias is associated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: No

provisionedThroughput

Information on the Provisioned Throughput assigned to an agent alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(([0-9a-zA-Z][_-]?){1,63})|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentAliasSummary

Service: Agents for Amazon Bedrock

Contains details about an alias of an agent.

Contents

agentAliasId

Contains details about

Type: String

Length Constraints: Fixed length of 10.

Pattern: `(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

agentAliasName

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentAliasStatus

The status of the alias.

Type: String

Valid Values: `CREATING | PREPARED | FAILED | UPDATING | DELETING | DISSOCIATED`

Required: Yes

createdAt

The time at which the alias of the agent was created.

Type: Timestamp

Required: Yes

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Required: Yes

aliasInvocationState

The invocation state for the agent alias. If the agent alias is running, the value is `ACCEPT_INVOCATIONS`. If the agent alias is paused, the value is `REJECT_INVOCATIONS`. Use the `UpdateAgentAlias` operation to change the invocation state.

Type: String

Valid Values: `ACCEPT_INVOCATIONS | REJECT_INVOCATIONS`

Required: No

description

The description of the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

routingConfiguration

Contains details about the version of the agent with which the alias is associated.

Type: Array of AgentAliasRoutingConfigurationListItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentCollaborator

Service: Agents for Amazon Bedrock

An agent collaborator.

Contents

agentDescriptor

The collaborator's agent descriptor.

Type: AgentDescriptor object

Required: Yes

agentId

The collaborator's agent ID.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The collaborator's agent version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

collaborationInstruction

The collaborator's instructions.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: Yes

collaboratorId

The collaborator's collaborator ID.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

collaboratorName

The collaborator's collaborator name.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

createdAt

When the collaborator was created.

Type: Timestamp

Required: Yes

lastUpdatedAt

When the collaborator was updated.

Type: Timestamp

Required: Yes

clientToken

The collaborator's client token.

Type: String

Length Constraints: Minimum length of 33. Maximum length of 256.

Pattern: `[a-zA-Z0-9](-*[a-zA-Z0-9]){0,256}`

Required: No

relayConversationHistory

The collaborator's relay conversation history.

Type: String

Valid Values: `TO_COLLABORATOR | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentCollaboratorSummary

Service: Agents for Amazon Bedrock

An agent collaborator summary.

Contents

agentDescriptor

The collaborator's agent descriptor.

Type: AgentDescriptor object

Required: Yes

agentId

The collaborator's agent ID.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The collaborator's agent version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

collaborationInstruction

The collaborator's collaboration instruction.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: Yes

collaboratorId

The collaborator's ID.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

collaboratorName

The collaborator's name.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

createdAt

When the collaborator was created.

Type: Timestamp

Required: Yes

lastUpdatedAt

When the collaborator was last updated.

Type: Timestamp

Required: Yes

relayConversationHistory

The collaborator's relay conversation history.

Type: String

Valid Values: `TO_COLLABORATOR | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentDescriptor

Service: Agents for Amazon Bedrock

An agent descriptor.

Contents

aliasArn

The agent's alias ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Defines an agent node in your flow. You specify the agent to invoke at this point in the flow. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

agentAliasArn

The Amazon Resource Name (ARN) of the alias of the agent to invoke.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `$|^arn:aws(-cn|-us-gov|-eusc|-iso(-[b-f])?)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentKnowledgeBase

Service: Agents for Amazon Bedrock

Contains details about a knowledge base that is associated with an agent.

Contents

agentId

The unique identifier of the agent with which the knowledge base is associated.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentVersion

The version of the agent with which the knowledge base is associated.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

createdAt

The time at which the association between the agent and the knowledge base was created.

Type: Timestamp

Required: Yes

description

The description of the association between the agent and the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

knowledgeBaseId

The unique identifier of the association between the agent and the knowledge base.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseState

Specifies whether to use the knowledge base or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

updatedAt

The time at which the association between the agent and the knowledge base was last updated.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentKnowledgeBaseSummary

Service: Agents for Amazon Bedrock

Contains details about a knowledge base associated with an agent.

Contents

knowledgeBaseId

The unique identifier of the knowledge base associated with an agent.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseState

Specifies whether the agent uses the knowledge base or not when sending an InvokeAgent request.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

updatedAt

The time at which the knowledge base associated with an agent was last updated.

Type: Timestamp

Required: Yes

description

The description of the knowledge base associated with an agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentSummary

Service: Agents for Amazon Bedrock

Contains details about an agent.

Contents

agentId

The unique identifier of the agent.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentName

The name of the agent.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentStatus

The status of the agent.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

Required: Yes

updatedAt

The time at which the agent was last updated.

Type: Timestamp

Required: Yes

description

The description of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

guardrailConfiguration

Details about the guardrail associated with the agent.

Type: GuardrailConfiguration object

Required: No

latestAgentVersion

The latest version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentVersion

Service: Agents for Amazon Bedrock

Contains details about a version of an agent.

Contents

agentArn

The Amazon Resource Name (ARN) of the agent that the version belongs to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent/[0-9a-zA-Z]{10}`

Required: Yes

agentId

The unique identifier of the agent that the version belongs to.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

agentName

The name of the agent that the version belongs to.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentResourceRoleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

agentStatus

The status of the agent that the version belongs to.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

Required: Yes

createdAt

The time at which the version was created.

Type: Timestamp

Required: Yes

idleSessionTTLInSeconds

The number of seconds for which Amazon Bedrock keeps information about a user's conversation with the agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 5400.

Required: Yes

updatedAt

The time at which the version was last updated.

Type: Timestamp

Required: Yes

version

The version number.

Type: String

Pattern: `[0-9]{1,5}`

Required: Yes

agentCollaboration

The agent's collaboration settings.

Type: String

Valid Values: `SUPERVISOR | SUPERVISOR_ROUTER | DISABLED`

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key that encrypts the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

description

The description of the version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

A list of reasons that the API operation on the version failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

foundationModel

The foundation model that the version invokes.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

guardrailConfiguration

Details about the guardrail associated with the agent.

Type: GuardrailConfiguration object

Required: No

instruction

The instructions provided to the agent.

Type: String

Length Constraints: Minimum length of 40. Maximum length of 4000.

Required: No

memoryConfiguration

Contains details of the memory configuration on the version of the agent.

Type: MemoryConfiguration object

Required: No

promptOverrideConfiguration

Contains configurations to override prompt templates in different parts of an agent sequence. For more information, see Advanced prompts.

Type: PromptOverrideConfiguration object

Required: No

recommendedActions

A list of recommended actions to take for the failed API operation on the version to succeed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentVersionSummary

Service: Agents for Amazon Bedrock

Contains details about a version of an agent.

Contents

agentName

The name of the agent to which the version belongs.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentStatus

The status of the agent to which the version belongs.

Type: String

Valid Values: `CREATING | PREPARING | PREPARED | NOT_PREPARED | DELETING | FAILED | VERSIONING | UPDATING`

Required: Yes

agentVersion

The version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

createdAt

The time at which the version was created.

Type: Timestamp

Required: Yes

updatedAt

The time at which the version was last updated.

Type: Timestamp

Required: Yes

description

The description of the version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

guardrailConfiguration

Details about the guardrail associated with the agent.

Type: GuardrailConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AnyToolChoice

Service: Agents for Amazon Bedrock

Defines tools, at least one of which must be requested by the model. No text is generated but the results of tool use are sent back to the model to help generate a response. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

APISchema

Service: Agents for Amazon Bedrock

Contains details about the OpenAPI schema for the action group. For more information, see Action group OpenAPI schemas. You can either include the schema directly in the `payload` field or you can upload it to an S3 bucket and specify the S3 bucket location in the `s3` field.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

payload

The JSON or YAML-formatted payload defining the OpenAPI schema for the action group. For more information, see Action group OpenAPI schemas.

Type: String

Required: No

s3

Contains details about the S3 object containing the OpenAPI schema for the action group. For more information, see Action group OpenAPI schemas.

Type: S3Identifier object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AudioConfiguration

Service: Agents for Amazon Bedrock

Configuration settings for processing audio content in multimodal knowledge bases.

Contents

segmentationConfiguration

Configuration for segmenting audio content during processing.

Type: AudioSegmentationConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioExtractionConfiguration

Service: Agents for Amazon Bedrock

Configuration for audio extraction.

Contents

audioExtractionStatus

Whether audio extraction is enabled or disabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioSegmentationConfiguration

Service: Agents for Amazon Bedrock

Configuration for segmenting audio content during multimodal knowledge base ingestion. Determines how audio files are divided into chunks for processing.

Contents

fixedLengthDuration

The duration in seconds for each audio segment. Audio files will be divided into chunks of this length for processing.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 30.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AutoToolChoice

Service: Agents for Amazon Bedrock

Defines tools. The model automatically decides whether to call a tool or to generate text instead. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockDataAutomationConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for using Amazon Bedrock Data Automation as the parser for ingesting your data sources.

Contents

parsingModality

Specifies whether to enable parsing of multimodal data, including both text and/or images.

Type: String

Valid Values: `MULTIMODAL`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockEmbeddingModelConfiguration

Service: Agents for Amazon Bedrock

The vector configuration details for the Bedrock embeddings model.

Contents

audio

Configuration settings for processing audio content in multimodal knowledge bases.

> **Important**
>
> This field is deprecated. Use `modelConfiguration` instead.

Type: Array of AudioConfiguration objects

Array Members: Fixed number of 1 item.

Required: No

dimensions

The dimensions details for the vector configuration used on the Bedrock embeddings model.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 4096.

Required: No

embeddingDataType

The data type for the vectors when using a model to convert text into vector embeddings. The model must support the specified data type for vector embeddings. Floating-point (float32) is the default data type, and is supported by most models for vector embeddings. See Supported embeddings models for information on the available models and their vector data types.

Type: String

Valid Values: `FLOAT32 | BINARY`

Required: No

modelConfiguration

Model-specific configuration for the embedding model, provided as a JSON object. Use this field to specify settings that apply to the embedding model that you selected, such as how audio and video files are divided into segments.

The fields that this object accepts depend on the embedding model. For the settings that each model accepts, see the documentation for that model.

For an example of a CreateKnowledgeBase request that uses this field to configure a multimodal embedding model, see the Examples section of CreateKnowledgeBase.

Type: JSON value

Required: No

video

Configuration settings for processing video content in multimodal knowledge bases.

> **Important**
>
> This field is deprecated. Use `modelConfiguration` instead.

Type: Array of VideoConfiguration objects

Array Members: Fixed number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockFoundationModelConfiguration

Service: Agents for Amazon Bedrock

Settings for a foundation model used to parse documents for a data source.

Contents

modelArn

The ARN of the foundation model to use for parsing.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

parsingModality

Specifies whether to enable parsing of multimodal data, including both text and/or images.

Type: String

Valid Values: `MULTIMODAL`

Required: No

parsingPrompt

Instructions for interpreting the contents of a document.

Type: ParsingPrompt object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockFoundationModelContextEnrichmentConfiguration

Service: Agents for Amazon Bedrock

Context enrichment configuration is used to provide additional context to the RAG application using Amazon Bedrock foundation models.

Contents

enrichmentStrategyConfiguration

The enrichment stategy used to provide additional context. For example, Neptune GraphRAG uses Amazon Bedrock foundation models to perform chunk entity extraction.

Type: EnrichmentStrategyConfiguration object

Required: Yes

modelArn

The Amazon Resource Name (ARN) of the model used to create vector embeddings for the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ByteContentDoc

Service: Agents for Amazon Bedrock

Contains information about content defined inline in bytes.

Contents

data

The base64-encoded string of the content.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 5242880.

Required: Yes

mimeType

The MIME type of the content. For a list of MIME types, see Media Types. The following MIME types are supported:

- text/plain
- text/html
- text/csv
- text/vtt
- message/rfc822
- application/xhtml+xml
- application/pdf
- application/msword
- application/vnd.ms-word.document.macroenabled.12
- application/vnd.ms-word.template.macroenabled.12
- application/vnd.ms-excel
- application/vnd.ms-excel.addin.macroenabled.12
- application/vnd.ms-excel.sheet.macroenabled.12
- application/vnd.ms-excel.template.macroenabled.12
- application/vnd.ms-excel.sheet.binary.macroenabled.12
- application/vnd.ms-spreadsheetml

- application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
- application/vnd.openxmlformats-officedocument.spreadsheetml.template
- application/vnd.openxmlformats-officedocument.wordprocessingml.document
- application/vnd.openxmlformats-officedocument.wordprocessingml.template

Type: String

Pattern: `.*[a-z]{1,20}/.{1,20}.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CachePointBlock

Service: Agents for Amazon Bedrock

Indicates where a cache checkpoint is located. All information before this checkpoint is cached to be accessed on subsequent requests.

Contents

type

Indicates that the CachePointBlock is of the default type

Type: String

Valid Values: `default`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ChatPromptTemplateConfiguration

Service: Agents for Amazon Bedrock

Contains configurations to use a prompt in a conversational format. For more information, see Create a prompt using Prompt management.

Contents

messages

Contains messages in the chat for the prompt.

Type: Array of Message objects

Required: Yes

inputVariables

An array of the variables in the prompt template.

Type: Array of PromptInputVariable objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

system

Contains system prompts to provide context to the model or to describe how it should behave.

Type: Array of SystemContentBlock objects

Required: No

toolConfiguration

Configuration information for the tools that the model can use when generating a response.

Type: ToolConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ChunkingConfiguration

Service: Agents for Amazon Bedrock

Details about how to chunk the documents in the data source. A *chunk* refers to an excerpt from a data source that is returned when the knowledge base that it belongs to is queried.

Contents

chunkingStrategy

Knowledge base can split your source data into chunks. A *chunk* refers to an excerpt from a data source that is returned when the knowledge base that it belongs to is queried. You have the following options for chunking your data. If you opt for `NONE`, then you may want to pre-process your files by splitting them up such that each file corresponds to a chunk.

- `FIXED_SIZE` – Amazon Bedrock splits your source data into chunks of the approximate size that you set in the `fixedSizeChunkingConfiguration`.
- `HIERARCHICAL` – Split documents into layers of chunks where the first layer contains large chunks, and the second layer contains smaller chunks derived from the first layer.
- `SEMANTIC` – Split documents into chunks based on groups of similar content derived with natural language processing.
- `NONE` – Amazon Bedrock treats each file as one chunk. If you choose this option, you may want to pre-process your documents by splitting them into separate files.

Type: String

Valid Values: `FIXED_SIZE | NONE | HIERARCHICAL | SEMANTIC`

Required: Yes

fixedSizeChunkingConfiguration

Configurations for when you choose fixed-size chunking. If you set the `chunkingStrategy` as `NONE`, exclude this field.

Type: FixedSizeChunkingConfiguration object

Required: No

hierarchicalChunkingConfiguration

Settings for hierarchical document chunking for a data source. Hierarchical chunking splits documents into layers of chunks where the first layer contains large chunks, and the second layer contains smaller chunks derived from the first layer.

Type: HierarchicalChunkingConfiguration object

Required: No

semanticChunkingConfiguration

Settings for semantic document chunking for a data source. Semantic chunking splits a document into into smaller documents based on groups of similar content derived from the text with natural language processing.

Type: SemanticChunkingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CollectorFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Defines a collector node in your flow. This node takes an iteration of inputs and consolidates them into an array in the output. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConditionFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Defines a condition node in your flow. You can specify conditions that determine which node comes next in the flow. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

conditions

An array of conditions. Each member contains the name of a condition and an expression that defines the condition.

Type: Array of FlowCondition objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConfluenceCrawlerConfiguration

Service: Agents for Amazon Bedrock

The configuration of the Confluence content. For example, configuring specific types of Confluence content.

Contents

filterConfiguration

The configuration of filtering the Confluence content. For example, configuring regular expression patterns to include or exclude certain content.

Type: CrawlFilterConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConfluenceDataSourceConfiguration

Service: Agents for Amazon Bedrock

The configuration information to connect to Confluence as your data source for self-managed knowledge bases.

Contents

sourceConfiguration

The endpoint information to connect to your Confluence data source.

Type: ConfluenceSourceConfiguration object

Required: Yes

crawlerConfiguration

The configuration of the Confluence content. For example, configuring specific types of Confluence content.

Type: ConfluenceCrawlerConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConfluenceSourceConfiguration

Service: Agents for Amazon Bedrock

The endpoint information to connect to your Confluence data source.

Contents

authType

The supported authentication type to authenticate and connect to your Confluence instance.

Type: String

Valid Values: `BASIC | OAUTH2_CLIENT_CREDENTIALS`

Required: Yes

credentialsSecretArn

The Amazon Resource Name of an AWS Secrets Manager secret that stores your authentication credentials for your Confluence instance URL. For more information on the key-value pairs that must be included in your secret, depending on your authentication type, see Confluence connection configuration.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

hostType

The supported host type, whether online/cloud or server/on-premises.

Type: String

Valid Values: SAAS

Required: Yes

hostUrl

The Confluence host URL or instance URL.

Type: String

Pattern: `https://[A-Za-z0-9][^\s]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlock

Service: Agents for Amazon Bedrock

Contains the content for the message you pass to, or receive from a model. For more information, see Create a prompt using Prompt management.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

cachePoint

Creates a cache checkpoint within a message.

Type: CachePointBlock object

Required: No

text

The text in the message.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContextEnrichmentConfiguration

Service: Agents for Amazon Bedrock

Context enrichment configuration is used to provide additional context to the RAG application.

Contents

type

The method used for context enrichment. It must be Amazon Bedrock foundation models.

Type: String

Valid Values: `BEDROCK_FOUNDATION_MODEL`

Required: Yes

bedrockFoundationModelConfiguration

The configuration of the Amazon Bedrock foundation model used for context enrichment.

Type: BedrockFoundationModelContextEnrichmentConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CrawlFilterConfiguration

Service: Agents for Amazon Bedrock

The configuration of filtering the data source content. For example, configuring regular expression patterns to include or exclude certain content.

Contents

type

The type of filtering that you want to apply to certain objects or content of the data source. For example, the PATTERN type is regular expression patterns you can apply to filter your content.

Type: String

Valid Values: PATTERN

Required: Yes

patternObjectFilter

The configuration of filtering certain objects or content types of the data source.

Type: PatternObjectFilterConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CuratedQuery

Service: Agents for Amazon Bedrock

Contains configurations for a query, each of which defines information about example queries to help the query engine generate appropriate SQL queries.

Contents

naturalLanguage

An example natural language query.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: Yes

sql

The SQL equivalent of the natural language query.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomContent

Service: Agents for Amazon Bedrock

Contains information about the content to ingest into a knowledge base connected to a custom data source. Choose a `sourceType` and include the field that corresponds to it.

Contents

customDocumentIdentifier

A unique identifier for the document.

Type: CustomDocumentIdentifier object

Required: Yes

sourceType

The source of the data to ingest.

Type: String

Valid Values: `IN_LINE | S3_LOCATION`

Required: Yes

inlineContent

Contains information about content defined inline to ingest into a knowledge base.

Type: InlineContent object

Required: No

s3Location

Contains information about the Amazon S3 location of the file from which to ingest data.

Type: CustomS3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomDocumentIdentifier

Service: Agents for Amazon Bedrock

Contains information about the identifier of the document to ingest into a custom data source.

Contents

id

The identifier of the document to ingest into a custom data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomOrchestration

Service: Agents for Amazon Bedrock

Details of custom orchestration.

Contents

executor

The structure of the executor invoking the actions in custom orchestration.

Type: OrchestrationExecutor object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomS3Location

Service: Agents for Amazon Bedrock

Contains information about the Amazon S3 location of the file containing the content to ingest into a knowledge base connected to a custom data source.

Contents

uri

The S3 URI of the file containing the content to ingest.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]/.{1,1024}`

Required: Yes

bucketOwnerAccountId

The identifier of the AWS account that owns the S3 bucket containing the content to ingest.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[0-9]{12}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomTransformationConfiguration

Service: Agents for Amazon Bedrock

Settings for customizing steps in the data source content ingestion pipeline.

You can configure the data source to process documents with a Lambda function after they are parsed and converted into chunks. When you add a post-chunking transformation, the service stores chunked documents in an S3 bucket and invokes a Lambda function to process them.

To process chunked documents with a Lambda function, define an S3 bucket path for input and output objects, and a transformation that specifies the Lambda function to invoke. You can use the Lambda function to customize how chunks are split, and the metadata for each chunk.

Contents

intermediateStorage

An S3 bucket path for input and output objects.

Type: IntermediateStorage object

Required: Yes

transformations

A Lambda function that processes documents.

Type: Array of Transformation objects

Array Members: Fixed number of 1 item.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CyclicConnectionFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a cyclic connection detected in the flow.

Contents

connection

The name of the connection that causes the cycle in the flow.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DailySchedule

Service: Agents for Amazon Bedrock

A daily sync. The run time is system-chosen (off-peak) and not configurable.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataSource

Service: Agents for Amazon Bedrock

Contains details about a data source.

Contents

createdAt

The time at which the data source was created.

Type: Timestamp

Required: Yes

dataSourceConfiguration

The connection configuration for the data source.

Type: DataSourceConfiguration object

Required: Yes

dataSourceId

The unique identifier of the data source.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source belongs.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

name

The name of the data source.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

status

The status of the data source. The following statuses are possible:

- Available – The data source has been created and is ready for ingestion into the knowledge base.
- Deleting – The data source is being deleted.

Type: String

Valid Values: `AVAILABLE | DELETING | DELETE_UNSUCCESSFUL | CREATING | UPDATING | FAILED`

Required: Yes

updatedAt

The time at which the data source was last updated.

Type: Timestamp

Required: Yes

dataDeletionPolicy

The data deletion policy for the data source.

Type: String

Valid Values: `RETAIN | DELETE`

Required: No

description

The description of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

The detailed reasons on the failure to delete a data source.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

serverSideEncryptionConfiguration

Contains details about the configuration of the server-side encryption.

Type: ServerSideEncryptionConfiguration object

Required: No

vectorIngestionConfiguration

Contains details about how to ingest the documents in the data source.

Type: VectorIngestionConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataSourceConfiguration

Service: Agents for Amazon Bedrock

The connection configuration for the data source.

Contents

type

The type of data source.

Type: String

Valid Values: `S3 | WEB | CONFLUENCE | SALESFORCE | SHAREPOINT | CUSTOM | REDSHIFT_METADATA | MANAGED_KNOWLEDGE_BASE_CONNECTOR`

Required: Yes

confluenceConfiguration

The configuration information to connect to Confluence as your data source for self-managed knowledge bases.

> **Note**
>
> To configure this data source for managed knowledge bases, use managedKnowledgeBaseConnectorConfiguration. Confluence data source connector for self-managed knowledge bases is in preview release and is subject to change.

Type: ConfluenceDataSourceConfiguration object

Required: No

managedKnowledgeBaseConnectorConfiguration

Contains the configuration for a data source that connects a managed knowledge base to a supported data source connector. Specify this object when the data source type is `MANAGED_KNOWLEDGE_BASE_CONNECTOR`.

Type: ManagedKnowledgeBaseConnectorConfiguration object

Required: No

s3Configuration

The configuration information to connect to Amazon S3 as your data source for self-managed knowledge bases. To configure this data source for managed knowledge bases, use managedKnowledgeBaseConnectorConfiguration.

Type: S3DataSourceConfiguration object

Required: No

salesforceConfiguration

The configuration information to connect to Salesforce as your data source.

> **Note**
>
> Salesforce data source connector for self-managed knowledge bases is in preview release and is subject to change.

Type: SalesforceDataSourceConfiguration object

Required: No

sharePointConfiguration

The configuration information to connect to SharePoint as your data source for self-managed knowledge bases.

> **Note**
>
> To configure this data source for managed knowledge bases, use managedKnowledgeBaseConnectorConfiguration. SharePoint data source connector for self-managed knowledge bases is in preview release and is subject to change.

Type: SharePointDataSourceConfiguration object

Required: No

webConfiguration

The configuration of web URLs to crawl for your data source. You should be authorized to crawl the URLs.

> **Note**
>
> To configure this data source for managed knowledge bases, use managedKnowledgeBaseConnectorConfiguration. Web crawler data source connector for self-managed knowledge bases is in preview release and is subject to change.

Type: WebDataSourceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataSourceSummary

Service: Agents for Amazon Bedrock

Contains details about a data source.

Contents

dataSourceId

The unique identifier of the data source.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base to which the data source belongs.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

name

The name of the data source.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

status

The status of the data source.

Type: String

Valid Values: `AVAILABLE | DELETING | DELETE_UNSUCCESSFUL | CREATING | UPDATING | FAILED`

Required: Yes

updatedAt

The time at which the data source was last updated.

Type: Timestamp

Required: Yes

description

The description of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DayOfMonth

Service: Agents for Amazon Bedrock

The day of the month on which a monthly sync runs. Specify exactly one of `dayNumber` or `lastDayOfMonth`.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

dayNumber

A specific day of the month, from 1 to 28. Values are capped at 28, so a monthly sync runs in every month, including February.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 28.

Required: No

lastDayOfMonth

Set this option to run the monthly sync on the last calendar day of each month.

Type: LastDayOfMonth object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

DeletionProtectionConfiguration

Service: Agents for Amazon Bedrock

Configuration for deletion protection.

Contents

deletionProtectionStatus

Enable or disable deletion protection for the connector.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

deletionProtectionThreshold

The threshold is the maximum percentage of documents that a sync job can delete from your index. If a sync would delete more than this percentage, the sync skips its delete phase, leaving your indexed documents in place. Not supported for the Custom connector.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentAccessControlEntry

Service: Agents for Amazon Bedrock

An access control entry specifying a principal and their access level.

Contents

access

Whether to allow or deny access.

Type: String

Valid Values: `ALLOW | DENY`

Required: Yes

name

The user identifier.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Required: Yes

type

The type of principal.

Type: String

Valid Values: `USER`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentContent

Service: Agents for Amazon Bedrock

Contains information about the content of a document. Choose a `dataSourceType` and include the field that corresponds to it.

Contents

dataSourceType

The type of data source that is connected to the knowledge base to which to ingest this document.

Type: String

Valid Values: `CUSTOM | S3`

Required: Yes

custom

Contains information about the content to ingest into a knowledge base connected to a custom data source.

Type: CustomContent object

Required: No

s3

Contains information about the content to ingest into a knowledge base connected to an Amazon S3 data source

Type: S3Content object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentIdentifier

Service: Agents for Amazon Bedrock

Contains information that identifies the document.

Contents

dataSourceType

The type of data source connected to the knowledge base that contains the document.

Type: String

Valid Values: `CUSTOM | S3`

Required: Yes

custom

Contains information that identifies the document in a custom data source.

Type: CustomDocumentIdentifier object

Required: No

s3

Contains information that identifies the document in an S3 data source.

Type: S3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentMetadata

Service: Agents for Amazon Bedrock

Contains information about the metadata associate with the content to ingest into a knowledge base. Choose a `type` and include the field that corresponds to it.

Contents

type

The type of the source source from which to add metadata.

Type: String

Valid Values: `IN_LINE_ATTRIBUTE | S3_LOCATION`

Required: Yes

accessControlList

Access control list for the document. Used when metadata type is IN_LINE_ATTRIBUTE.

Type: Array of DocumentAccessControlEntry objects

Array Members: Minimum number of 1 item.

Required: No

inlineAttributes

An array of objects, each of which defines a metadata attribute to associate with the content to ingest. You define the attributes inline.

Type: Array of MetadataAttribute objects

Array Members: Minimum number of 1 item. Maximum number of 50 items.

Required: No

s3Location

The Amazon S3 location of the file containing metadata to associate with the content to ingest.

Type: CustomS3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DuplicateConditionExpressionFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about duplicate condition expressions found in a condition node.

Contents

expression

The duplicated condition expression.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

node

The name of the node containing the duplicate condition expressions.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DuplicateConnectionsFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about duplicate connections found between two nodes in the flow.

Contents

source

The name of the source node where the duplicate connection starts.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

target

The name of the target node where the duplicate connection ends.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EmbeddingModelConfiguration

Service: Agents for Amazon Bedrock

The configuration details for the embeddings model.

Contents

bedrockEmbeddingModelConfiguration

The vector configuration details on the Bedrock embeddings model.

Type: BedrockEmbeddingModelConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EnrichmentStrategyConfiguration

Service: Agents for Amazon Bedrock

The strategy used for performing context enrichment.

Contents

method

The method used for the context enrichment strategy.

Type: String

Valid Values: `CHUNK_ENTITY_EXTRACTION`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FieldForReranking

Service: Agents for Amazon Bedrock

Specifies a metadata field to include or exclude during the reranking process.

Contents

fieldName

The name of the metadata field to include or exclude during reranking.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FixedSizeChunkingConfiguration

Service: Agents for Amazon Bedrock

Configurations for when you choose fixed-size chunking. If you set the `chunkingStrategy` as `NONE`, exclude this field.

Contents

maxTokens

The maximum number of tokens to include in a chunk.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 8192.

Required: Yes

overlapPercentage

The percentage of overlap between adjacent chunks of a data source.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 99.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowAliasConcurrencyConfiguration

Service: Agents for Amazon Bedrock

Determines how multiple nodes in a flow can run in parallel. Running nodes concurrently can improve your flow's performance.

Contents

type

The type of concurrency to use for parallel node execution. Specify one of the following options:

- `Automatic` - Amazon Bedrock determines which nodes can be executed in parallel based on the flow definition and its dependencies.
- `Manual` - You specify which nodes can be executed in parallel.

Type: String

Valid Values: `Automatic | Manual`

Required: Yes

maxConcurrency

The maximum number of nodes that can be executed concurrently in the flow.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowAliasRoutingConfigurationListItem

Service: Agents for Amazon Bedrock

Contains information about a version that the alias maps to.

Contents

flowVersion

The version that the alias maps to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowAliasSummary

Service: Agents for Amazon Bedrock

Contains information about an alias of a flow.

This data type is used in the following API operations:

- ListFlowAliases response

Contents

arn

The Amazon Resource Name (ARN) of the alias.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/(TSTALIASID|[0-9a-zA-Z]{10})`

Required: Yes

createdAt

The time at which the alias was created.

Type: Timestamp

Required: Yes

flowId

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

id

The unique identifier of the alias of the flow.

Type: String

Pattern: `(TSTALIASID|[0-9a-zA-Z]{10})`

Required: Yes

name

The name of the alias.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

routingConfiguration

A list of configurations about the versions that the alias maps to. Currently, you can only specify one.

Type: Array of FlowAliasRoutingConfigurationListItem objects

Array Members: Fixed number of 1 item.

Required: Yes

updatedAt

The time at which the alias was last updated.

Type: Timestamp

Required: Yes

concurrencyConfiguration

The configuration that specifies how nodes in the flow are executed concurrently.

Type: FlowAliasConcurrencyConfiguration object

Required: No

description

A description of the alias.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowCondition

Service: Agents for Amazon Bedrock

Defines a condition in the condition node.

Contents

name

A name for the condition that you can reference.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

expression

Defines the condition. You must refer to at least one of the inputs in the condition. For more information, expand the Condition node section in Node types in prompt flows.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowConditionalConnectionConfiguration

Service: Agents for Amazon Bedrock

The configuration of a connection between a condition node and another node.

Contents

condition

The condition that triggers this connection. For more information about how to write conditions, see the **Condition** node type in the Node types topic in the Amazon Bedrock User Guide.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowConnection

Service: Agents for Amazon Bedrock

Contains information about a connection between two nodes in the flow.

Contents

name

A name for the connection that you can reference.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

source

The node that the connection starts at.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

target

The node that the connection ends at.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

type

Whether the source node that the connection begins from is a condition node (`Conditional`) or not (`Data`).

Type: String

Valid Values: `Data | Conditional`

Required: Yes

configuration

The configuration of the connection.

Type: FlowConnectionConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowConnectionConfiguration

Service: Agents for Amazon Bedrock

The configuration of the connection.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

conditional

The configuration of a connection originating from a Condition node.

Type: FlowConditionalConnectionConfiguration object

Required: No

data

The configuration of a connection originating from a node that isn't a Condition node.

Type: FlowDataConnectionConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowDataConnectionConfiguration

Service: Agents for Amazon Bedrock

The configuration of a connection originating from a node that isn't a Condition node.

Contents

sourceOutput

The name of the output in the source node that the connection begins from.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

targetInput

The name of the input in the target node that the connection ends at.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowDefinition

Service: Agents for Amazon Bedrock

The definition of the nodes and connections between nodes in the flow.

Contents

connections

An array of connection definitions in the flow.

Type: Array of FlowConnection objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

nodes

An array of node definitions in the flow.

Type: Array of FlowNode objects

Array Members: Minimum number of 0 items. Maximum number of 40 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowNode

Service: Agents for Amazon Bedrock

Contains configurations about a node in the flow.

Contents

name

A name for the node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

type

The type of node. This value must match the name of the key that you provide in the configuration you provide in the `FlowNodeConfiguration` field.

Type: String

Valid Values: `Input | Output | KnowledgeBase | Condition | Lex | Prompt | LambdaFunction | Storage | Agent | Retrieval | Iterator | Collector | InlineCode | Loop | LoopInput | LoopController`

Required: Yes

configuration

Contains configurations for the node.

Type: FlowNodeConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

inputs

An array of objects, each of which contains information about an input into the node.

Type: Array of FlowNodeInput objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

outputs

A list of objects, each of which contains information about an output from the node.

Type: Array of FlowNodeOutput objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a node in your flow. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

agent

Contains configurations for an agent node in your flow. Invokes an alias of an agent and returns the response.

Type: AgentFlowNodeConfiguration object

Required: No

collector

Contains configurations for a collector node in your flow. Collects an iteration of inputs and consolidates them into an array of outputs.

Type: CollectorFlowNodeConfiguration object

Required: No

condition

Contains configurations for a condition node in your flow. Defines conditions that lead to different branches of the flow.

Type: ConditionFlowNodeConfiguration object

Required: No

inlineCode

Contains configurations for an inline code node in your flow. Inline code nodes let you write and execute code directly within your flow, enabling data transformations, custom logic, and integrations without needing an external Lambda function.

Type: InlineCodeFlowNodeConfiguration object

Required: No

input

Contains configurations for an input flow node in your flow. The first node in the flow. `inputs` can't be specified for this node.

Type: InputFlowNodeConfiguration object

Required: No

iterator

Contains configurations for an iterator node in your flow. Takes an input that is an array and iteratively sends each item of the array as an output to the following node. The size of the array is also returned in the output.

The output flow node at the end of the flow iteration will return a response for each member of the array. To return only one response, you can include a collector node downstream from the iterator node.

Type: IteratorFlowNodeConfiguration object

Required: No

knowledgeBase

Contains configurations for a knowledge base node in your flow. Queries a knowledge base and returns the retrieved results or generated response.

Type: KnowledgeBaseFlowNodeConfiguration object

Required: No

lambdaFunction

Contains configurations for a Lambda function node in your flow. Invokes an AWS Lambda function.

Type: LambdaFunctionFlowNodeConfiguration object

Required: No

lex

Contains configurations for a Lex node in your flow. Invokes an Amazon Lex bot to identify the intent of the input and return the intent as the output.

Type: LexFlowNodeConfiguration object

Required: No

loop

Contains configurations for a DoWhile loop in your flow.

Type: LoopFlowNodeConfiguration object

Required: No

loopController

Contains controller node configurations for a DoWhile loop in your flow.

Type: LoopControllerFlowNodeConfiguration object

Required: No

loopInput

Contains input node configurations for a DoWhile loop in your flow.

Type: LoopInputFlowNodeConfiguration object

Required: No

output

Contains configurations for an output flow node in your flow. The last node in the flow. `outputs` can't be specified for this node.

Type: OutputFlowNodeConfiguration object

Required: No

prompt

Contains configurations for a prompt node in your flow. Runs a prompt and generates the model response as the output. You can use a prompt from Prompt management or you can configure one in this node.

Type: PromptFlowNodeConfiguration object

Required: No

retrieval

Contains configurations for a retrieval node in your flow. Retrieves data from an Amazon S3 location and returns it as the output.

Type: RetrievalFlowNodeConfiguration object

Required: No

storage

Contains configurations for a storage node in your flow. Stores an input in an Amazon S3 location.

Type: StorageFlowNodeConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowNodeInput

Service: Agents for Amazon Bedrock

Contains configurations for an input in an Amazon Bedrock Flows node.

Contents

expression

An expression that formats the input for the node. For an explanation of how to create expressions, see Expressions in Prompt flows in Amazon Bedrock.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

name

Specifies a name for the input that you can reference.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

type

Specifies the data type of the input. If the input doesn't match this type at runtime, a validation error will be thrown.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: Yes

category

Specifies how input data flows between iterations in a DoWhile loop.

- `LoopCondition` - Controls whether the loop continues by evaluating condition expressions against the input data. Use this category to define the condition that determines if the loop should continue.

- `ReturnValueToLoopStart` - Defines data to pass back to the start of the loop's next iteration. Use this category for variables that you want to update for each loop iteration.
- `ExitLoop` - Defines the value that's available once the loop ends. Use this category to expose loop results to nodes outside the loop.

Type: String

Valid Values: `LoopCondition | ReturnValueToLoopStart | ExitLoop`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowNodeOutput

Service: Agents for Amazon Bedrock

Contains configurations for an output from a node.

Contents

name

A name for the output that you can reference.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

type

The data type of the output. If the output doesn't match this type at runtime, a validation error will be thrown.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowSummary

Service: Agents for Amazon Bedrock

Contains the definition of a flow.

Contents

arn

The Amazon Resource Name (ARN) of the flow.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

Required: Yes

createdAt

The time at which the flow was created.

Type: Timestamp

Required: Yes

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

name

The name of the flow.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

status

The status of the flow. The following statuses are possible:

- NotPrepared – The flow has been created or updated, but hasn't been prepared. If you just created the flow, you can't test it. If you updated the flow, the DRAFT version won't contain the latest changes for testing. Send a PrepareFlow request to package the latest changes into the DRAFT version.
- Preparing – The flow is being prepared so that the DRAFT version contains the latest changes for testing.
- Prepared – The flow is prepared and the DRAFT version contains the latest changes for testing.
- Failed – The last API operation that you invoked on the flow failed. Send a GetFlow request and check the error message in the `validations` field.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

Required: Yes

updatedAt

The time at which the flow was last updated.

Type: Timestamp

Required: Yes

version

The latest version of the flow.

Type: String

Length Constraints: Fixed length of 5.

Pattern: DRAFT

Required: Yes

description

A description of the flow.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowValidation

Service: Agents for Amazon Bedrock

Contains information about validation of the flow.

This data type is used in the following API operations:

- GetFlow response
- GetFlowVersion response

Contents

message

A message describing the validation error.

Type: String

Pattern: `[\s\S]+`

Required: Yes

severity

The severity of the issue described in the message.

Type: String

Valid Values: `Warning | Error`

Required: Yes

details

Specific details about the validation issue encountered in the flow.

Type: FlowValidationDetails object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

type

The type of validation issue encountered in the flow.

Type: String

Valid Values: `CyclicConnection | DuplicateConnections | DuplicateConditionExpression | UnreachableNode | UnknownConnectionSource | UnknownConnectionSourceOutput | UnknownConnectionTarget | UnknownConnectionTargetInput | UnknownConnectionCondition | MalformedConditionExpression | MalformedNodeInputExpression | MismatchedNodeInputType | MismatchedNodeOutputType | IncompatibleConnectionDataType | MissingConnectionConfiguration | MissingDefaultCondition | MissingEndingNodes | MissingNodeConfiguration | MissingNodeInput | MissingNodeOutput | MissingStartingNodes | MultipleNodeInputConnections | UnfulfilledNodeInput | UnsatisfiedConnectionConditions | Unspecified | UnknownNodeInput | UnknownNodeOutput | MissingLoopInputNode | MissingLoopControllerNode | MultipleLoopInputNodes | MultipleLoopControllerNodes | LoopIncompatibleNodeType | InvalidLoopBoundary`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowValidationDetails

Service: Agents for Amazon Bedrock

A union type containing various possible validation issues in the flow.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

cyclicConnection

Details about a cyclic connection in the flow.

Type: CyclicConnectionFlowValidationDetails object

Required: No

duplicateConditionExpression

Details about duplicate condition expressions in a node.

Type: DuplicateConditionExpressionFlowValidationDetails object

Required: No

duplicateConnections

Details about duplicate connections between nodes.

Type: DuplicateConnectionsFlowValidationDetails object

Required: No

incompatibleConnectionDataType

Details about incompatible data types in a connection.

Type: IncompatibleConnectionDataTypeFlowValidationDetails object

Required: No

invalidLoopBoundary

Details about a flow that includes connections that violate loop boundary rules.

Type: InvalidLoopBoundaryFlowValidationDetails object

Required: No

loopIncompatibleNodeType

Details about a flow that includes incompatible node types in a DoWhile loop.

Type: LoopIncompatibleNodeTypeFlowValidationDetails object

Required: No

malformedConditionExpression

Details about a malformed condition expression in a node.

Type: MalformedConditionExpressionFlowValidationDetails object

Required: No

malformedNodeInputExpression

Details about a malformed input expression in a node.

Type: MalformedNodeInputExpressionFlowValidationDetails object

Required: No

mismatchedNodeInputType

Details about mismatched input data types in a node.

Type: MismatchedNodeInputTypeFlowValidationDetails object

Required: No

mismatchedNodeOutputType

Details about mismatched output data types in a node.

Type: MismatchedNodeOutputTypeFlowValidationDetails object

Required: No

missingConnectionConfiguration

Details about missing configuration for a connection.

Type: MissingConnectionConfigurationFlowValidationDetails object

Required: No

missingDefaultCondition

Details about a missing default condition in a conditional node.

Type: MissingDefaultConditionFlowValidationDetails object

Required: No

missingEndingNodes

Details about missing ending nodes in the flow.

Type: MissingEndingNodesFlowValidationDetails object

Required: No

missingLoopControllerNode

Details about a flow that's missing a required `LoopController` node in a DoWhile loop.

Type: MissingLoopControllerNodeFlowValidationDetails object

Required: No

missingLoopInputNode

Details about a flow that's missing a required `LoopInput` node in a DoWhile loop.

Type: MissingLoopInputNodeFlowValidationDetails object

Required: No

missingNodeConfiguration

Details about missing configuration for a node.

Type: MissingNodeConfigurationFlowValidationDetails object

Required: No

missingNodeInput

Details about a missing required input in a node.

Type: MissingNodeInputFlowValidationDetails object

Required: No

missingNodeOutput

Details about a missing required output in a node.

Type: MissingNodeOutputFlowValidationDetails object

Required: No

missingStartingNodes

Details about missing starting nodes in the flow.

Type: MissingStartingNodesFlowValidationDetails object

Required: No

multipleLoopControllerNodes

Details about a flow that contains multiple `LoopController` nodes in a DoWhile loop.

Type: MultipleLoopControllerNodesFlowValidationDetails object

Required: No

multipleLoopInputNodes

Details about a flow that contains multiple `LoopInput` nodes in a DoWhile loop.

Type: MultipleLoopInputNodesFlowValidationDetails object

Required: No

multipleNodeInputConnections

Details about multiple connections to a single node input.

Type: MultipleNodeInputConnectionsFlowValidationDetails object

Required: No

unfulfilledNodeInput

Details about an unfulfilled node input with no valid connections.

Type: UnfulfilledNodeInputFlowValidationDetails object

Required: No

unknownConnectionCondition

Details about an unknown condition for a connection.

Type: UnknownConnectionConditionFlowValidationDetails object

Required: No

unknownConnectionSource

Details about an unknown source node for a connection.

Type: UnknownConnectionSourceFlowValidationDetails object

Required: No

unknownConnectionSourceOutput

Details about an unknown source output for a connection.

Type: UnknownConnectionSourceOutputFlowValidationDetails object

Required: No

unknownConnectionTarget

Details about an unknown target node for a connection.

Type: UnknownConnectionTargetFlowValidationDetails object

Required: No

unknownConnectionTargetInput

Details about an unknown target input for a connection.

Type: UnknownConnectionTargetInputFlowValidationDetails object

Required: No

unknownNodeInput

Details about an unknown input for a node.

Type: UnknownNodeInputFlowValidationDetails object

Required: No

unknownNodeOutput

Details about an unknown output for a node.

Type: UnknownNodeOutputFlowValidationDetails object

Required: No

unreachableNode

Details about an unreachable node in the flow.

Type: UnreachableNodeFlowValidationDetails object

Required: No

unsatisfiedConnectionConditions

Details about unsatisfied conditions for a connection.

Type: UnsatisfiedConnectionConditionsFlowValidationDetails object

Required: No

unspecified

Details about an unspecified validation.

Type: UnspecifiedFlowValidationDetails object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowVersionSummary

Service: Agents for Amazon Bedrock

Contains information about a version of a flow.

This data type is used in the following API operations:

- ListFlowVersions response

Contents

arn

The Amazon Resource Name (ARN) of the flow that the version belongs to.

Type: String

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}`

Required: Yes

createdAt

The time at the version was created.

Type: Timestamp

Required: Yes

id

The unique identifier of the flow.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

status

The status of the flow.

Type: String

Valid Values: `Failed | Prepared | Preparing | NotPrepared`

Required: Yes

version

The version of the flow.

Type: String

Pattern: `[0-9]{1,5}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Function

Service: Agents for Amazon Bedrock

Defines parameters that the agent needs to invoke from the user to complete the function. Corresponds to an action in an action group.

This data type is used in the following API operations:

- CreateAgentActionGroup request
- CreateAgentActionGroup response
- UpdateAgentActionGroup request
- UpdateAgentActionGroup response
- GetAgentActionGroup response

Contents

name

A name for the function.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

description

A description of the function and its purpose.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1200.

Required: No

parameters

The parameters that the agent elicits from the user to fulfill the function.

Type: String to ParameterDetail object map

Key Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

requireConfirmation

Contains information if user confirmation is required to invoke the function.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FunctionSchema

Service: Agents for Amazon Bedrock

Defines functions that each define parameters that the agent needs to invoke from the user. Each function represents an action in an action group.

This data type is used in the following API operations:

- CreateAgentActionGroup request
- CreateAgentActionGroup response
- UpdateAgentActionGroup request
- UpdateAgentActionGroup response
- GetAgentActionGroup response

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

functions

A list of functions that each define an action in the action group.

Type: Array of Function objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

GuardrailConfiguration

Service: Agents for Amazon Bedrock

Details about a guardrail associated with a resource.

Contents

guardrailIdentifier

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: No

guardrailVersion

The version of the guardrail.

Type: String

Pattern: `(([0-9]{1,8})|(DRAFT))`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

HierarchicalChunkingConfiguration

Service: Agents for Amazon Bedrock

Settings for hierarchical document chunking for a data source. Hierarchical chunking splits documents into layers of chunks where the first layer contains large chunks, and the second layer contains smaller chunks derived from the first layer.

You configure the number of tokens to overlap, or repeat across adjacent chunks. For example, if you set overlap tokens to 60, the last 60 tokens in the first chunk are also included at the beginning of the second chunk. For each layer, you must also configure the maximum number of tokens in a chunk.

Contents

levelConfigurations

Token settings for each layer.

Type: Array of HierarchicalChunkingLevelConfiguration objects

Array Members: Fixed number of 2 items.

Required: Yes

overlapTokens

The number of tokens to repeat across chunks in the same layer.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

HierarchicalChunkingLevelConfiguration

Service: Agents for Amazon Bedrock

Token settings for a layer in a hierarchical chunking configuration.

Contents

maxTokens

The maximum number of tokens that a chunk can contain in this layer.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 8192.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageExtractionConfiguration

Service: Agents for Amazon Bedrock

Configuration for image extraction.

Contents

imageExtractionStatus

Whether image extraction is enabled or disabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IncompatibleConnectionDataTypeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about incompatible data types in a connection between nodes.

Contents

connection

The name of the connection with incompatible data types.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceConfiguration

Service: Agents for Amazon Bedrock

Contains inference parameters to use when the agent invokes a foundation model in the part of the agent sequence defined by the `promptType`. For more information, see Inference parameters for foundation models.

Contents

maximumLength

The maximum number of tokens to allow in the generated response.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

stopSequences

A list of stop sequences. A stop sequence is a sequence of characters that causes the model to stop generating the response.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

temperature

The likelihood of the model selecting higher-probability options while generating a response. A lower value makes the model more likely to choose higher-probability options, while a higher value makes the model more likely to choose lower-probability options.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topK

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for `topK` is the number of most-likely

candidates from which the model chooses the next token in the sequence. For example, if you set `topK` to 50, the model selects the next token from among the top 50 most likely choices.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 500.

Required: No

topP

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for `Top P` determines the number of most-likely candidates from which the model chooses the next token in the sequence. For example, if you set `topP` to 0.8, the model only selects the next token from the top 80% of the probability distribution of next tokens.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IngestionJob

Service: Agents for Amazon Bedrock

Contains details about a data ingestion job. Data sources are ingested into a knowledge base so that Large Language Models (LLMs) can use your data.

This data type is used in the following API operations:

- StartIngestionJob response
- GetIngestionJob response
- ListIngestionJob response

Contents

dataSourceId

The unique identifier of the data source for the data ingestion job.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

ingestionJobId

The unique identifier of the data ingestion job.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge for the data ingestion job.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

startedAt

The time the data ingestion job started.

If you stop a data ingestion job, the `startedAt` time is the time the job was started before the job was stopped.

Type: Timestamp

Required: Yes

status

The status of the data ingestion job.

Type: String

Valid Values: `STARTING | IN_PROGRESS | COMPLETE | FAILED | STOPPING | STOPPED`

Required: Yes

updatedAt

The time the data ingestion job was last updated.

If you stop a data ingestion job, the `updatedAt` time is the time the job was stopped.

Type: Timestamp

Required: Yes

description

The description of the data ingestion job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

A list of reasons that the data ingestion job failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

statistics

Contains statistics about the data ingestion job.

Type: IngestionJobStatistics object

Required: No

textReadyAt

The time at which all text content in the data ingestion job finished extraction and became available to query.

This time isn't returned until text extraction is complete for all the documents in the job.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IngestionJobFilter

Service: Agents for Amazon Bedrock

The definition of a filter to filter the data.

Contents

attribute

The name of field or attribute to apply the filter.

Type: String

Valid Values: STATUS

Required: Yes

operator

The operation to apply to the field or attribute.

Type: String

Valid Values: EQ

Required: Yes

values

A list of values that belong to the field or attribute.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Length Constraints: Minimum length of 0. Maximum length of 100.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IngestionJobSortBy

Service: Agents for Amazon Bedrock

The parameters of sorting the data.

Contents

attribute

The name of field or attribute to apply sorting of data.

Type: String

Valid Values: `STATUS | STARTED_AT`

Required: Yes

order

The order for sorting the data.

Type: String

Valid Values: `ASCENDING | DESCENDING`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IngestionJobStatistics

Service: Agents for Amazon Bedrock

Contains the statistics for the data ingestion job.

Contents

numberOfDocumentsDeleted

The number of source documents that were deleted.

Type: Long

Required: No

numberOfDocumentsFailed

The number of source documents that failed to be ingested.

Type: Long

Required: No

numberOfDocumentsScanned

The total number of source documents that were scanned. Includes new, updated, and unchanged documents.

Type: Long

Required: No

numberOfDocumentsSkipped

The number of source documents that were skipped during ingestion.

Type: Long

Required: No

numberOfMetadataDocumentsModified

The number of metadata files that were updated or deleted.

Type: Long

Required: No

numberOfMetadataDocumentsScanned

The total number of metadata files that were scanned. Includes new, updated, and unchanged files.

Type: Long

Required: No

numberOfModifiedDocumentsIndexed

The number of modified source documents in the data source that were successfully indexed.

Type: Long

Required: No

numberOfNewDocumentsIndexed

The number of new source documents in the data source that were successfully indexed.

Type: Long

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IngestionJobSummary

Service: Agents for Amazon Bedrock

Contains details about a data ingestion job.

Contents

dataSourceId

The unique identifier of the data source for the data ingestion job.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

ingestionJobId

The unique identifier of the data ingestion job.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base for the data ingestion job.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

startedAt

The time the data ingestion job started.

Type: Timestamp

Required: Yes

status

The status of the data ingestion job.

Type: String

Valid Values: `STARTING | IN_PROGRESS | COMPLETE | FAILED | STOPPING | STOPPED`

Required: Yes

updatedAt

The time the data ingestion job was last updated.

Type: Timestamp

Required: Yes

description

The description of the data ingestion job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

statistics

Contains statistics for the data ingestion job.

Type: IngestionJobStatistics object

Required: No

textReadyAt

The time at which all text content in the data ingestion job finished extraction and became available to query.

This time isn't returned until text extraction is complete for all the documents in the job.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineCodeFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for an inline code node in your flow. Inline code nodes let you write and execute code directly within your flow, enabling data transformations, custom logic, and integrations without needing an external Lambda function.

Contents

code

The code that's executed in your inline code node. The code can access input data from previous nodes in the flow, perform operations on that data, and produce output that can be used by other nodes in your flow.

The code must be valid in the programming `language` that you specify.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 5000000.

Required: Yes

language

The programming language used by your inline code node.

The code must be valid in the programming `language` that you specify. Currently, only Python 3 (`Python_3`) is supported.

Type: String

Valid Values: `Python_3`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineContent

Service: Agents for Amazon Bedrock

Contains information about content defined inline to ingest into a data source. Choose a `type` and include the field that corresponds to it.

Contents

type

The type of inline content to define.

Type: String

Valid Values: `BYTE | TEXT`

Required: Yes

byteContent

Contains information about content defined inline in bytes.

Type: ByteContentDoc object

Required: No

textContent

Contains information about content defined inline in text.

Type: TextContentDoc object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for the input flow node for a flow. This node takes the input from flow invocation and passes it to the next node in the data type that you specify.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

IntermediateStorage

Service: Agents for Amazon Bedrock

A location for storing content from data sources temporarily as it is processed by custom components in the ingestion pipeline.

Contents

s3Location

An S3 bucket path.

Type: S3Location object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvalidLoopBoundaryFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a flow that contains connections that violate loop boundary rules.

Contents

connection

The name of the connection that violates loop boundary rules.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

source

The source node of the connection that violates DoWhile loop boundary rules.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

target

The target node of the connection that violates DoWhile loop boundary rules.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

IteratorFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for an iterator node in a flow. Takes an input that is an array and iteratively sends each item of the array as an output to the following node. The size of the array is also returned in the output.

The output flow node at the end of the flow iteration will return a response for each member of the array. To return only one response, you can include a collector node downstream from the iterator node.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KendraKnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Settings for an Amazon Kendra knowledge base.

Contents

kendraIndexArn

The ARN of the Amazon Kendra index.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):kendra:[a-z0-9-]{1,20}:([0-9]{12}|):index/([a-zA-Z0-9][a-zA-Z0-9-]{35}|[a-zA-Z0-9][a-zA-Z0-9-]{35}-[a-zA-Z0-9][a-zA-Z0-9-]{35})`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBase

Service: Agents for Amazon Bedrock

Contains information about a knowledge base.

Contents

createdAt

The time the knowledge base was created.

Type: Timestamp

Required: Yes

knowledgeBaseArn

The Amazon Resource Name (ARN) of the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

Required: Yes

knowledgeBaseConfiguration

Contains details about the embeddings configuration of the knowledge base.

Type: KnowledgeBaseConfiguration object

Required: Yes

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

name

The name of the knowledge base.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

roleArn

The Amazon Resource Name (ARN) of the IAM role with permissions to invoke API operations on the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+`

Required: Yes

status

The status of the knowledge base. The following statuses are possible:

- CREATING – The knowledge base is being created.
- ACTIVE – The knowledge base is ready to be queried.
- DELETING – The knowledge base is being deleted.
- UPDATING – The knowledge base is being updated.
- FAILED – The knowledge base API operation failed.

Type: String

Valid Values: `CREATING | ACTIVE | DELETING | UPDATING | FAILED | DELETE_UNSUCCESSFUL | UPDATE_UNSUCCESSFUL`

Required: Yes

updatedAt

The time the knowledge base was last updated.

Type: Timestamp

Required: Yes

description

The description of the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

failureReasons

A list of reasons that the API operation on the knowledge base failed.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2048 items.

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

storageConfiguration

Contains details about the storage configuration of the knowledge base.

Type: StorageConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Contains details about the vector embeddings configuration of the knowledge base.

Contents

type

The type of data that the data source is converted into for the knowledge base. Choose `MANAGED` to create a managed knowledge base.

Type: String

Valid Values: `VECTOR | KENDRA | SQL | MANAGED`

Required: Yes

kendraKnowledgeBaseConfiguration

Settings for an Amazon Kendra knowledge base.

Type: KendraKnowledgeBaseConfiguration object

Required: No

managedKnowledgeBaseConfiguration

Contains configuration details for a knowledge base that uses a vector store fully managed by Amazon Bedrock. Specify this object when the knowledge base type is MANAGED.

Type: ManagedKnowledgeBaseConfiguration object

Required: No

sqlKnowledgeBaseConfiguration

Specifies configurations for a knowledge base connected to an SQL database.

Type: SqlKnowledgeBaseConfiguration object

Required: No

vectorKnowledgeBaseConfiguration

Contains details about the model that's used to convert the data source into vector embeddings.

Type: VectorKnowledgeBaseConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseDocument

Service: Agents for Amazon Bedrock

Contains information about a document to ingest into a knowledge base and metadata to associate with it.

Contents

content

Contains the content of the document.

Type: DocumentContent object

Required: Yes

metadata

Contains the metadata to associate with the document.

Type: DocumentMetadata object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseDocumentDetail

Service: Agents for Amazon Bedrock

Contains the details for a document that was ingested or deleted.

Contents

dataSourceId

The identifier of the data source connected to the knowledge base that the document was ingested into or deleted from.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

identifier

Contains information that identifies the document.

Type: DocumentIdentifier object

Required: Yes

knowledgeBaseId

The identifier of the knowledge base that the document was ingested into or deleted from.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

status

The ingestion status of the document. The following statuses are possible:

- STARTING – You submitted the ingestion job containing the document.
- PENDING – The document is waiting to be ingested.
- IN_PROGRESS – The document is being ingested.
- INDEXED – The document was successfully indexed.

- PARTIALLY_INDEXED – The document was partially indexed.
- METADATA_PARTIALLY_INDEXED – You submitted metadata for an existing document and it was partially indexed.
- METADATA_UPDATE_FAILED – You submitted a metadata update for an existing document but it failed.
- FAILED – The document failed to be ingested.
- NOT_FOUND – The document wasn't found.
- IGNORED – The document was ignored during ingestion.
- DELETING – You submitted the delete job containing the document.
- DELETE_IN_PROGRESS – The document is being deleted.

Type: String

Valid Values: `INDEXED | PARTIALLY_INDEXED | PENDING | FAILED | METADATA_PARTIALLY_INDEXED | METADATA_UPDATE_FAILED | IGNORED | NOT_FOUND | STARTING | IN_PROGRESS | DELETING | DELETE_IN_PROGRESS`

Required: Yes

statusReason

The reason for the status. Appears alongside the status `IGNORED`.

Type: String

Required: No

updatedAt

The date and time at which the document was last updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a knowledge base node in a flow. This node takes a query as the input and returns, as the output, the retrieved responses directly (as an array) or a response generated based on the retrieved responses. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

knowledgeBaseId

The unique identifier of the knowledge base to query.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `$|^[0-9a-zA-Z]+`

Required: Yes

guardrailConfiguration

Contains configurations for a guardrail to apply during query and response generation for the knowledge base in this configuration.

Type: GuardrailConfiguration object

Required: No

inferenceConfiguration

Contains inference configurations for the prompt.

Type: PromptInferenceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

modelId

The unique identifier of the model or inference profile to use to generate a response from the query results. Omit this field if you want to return the retrieved results as an array.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

numberOfResults

The number of results to retrieve from the knowledge base.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

orchestrationConfiguration

The configuration for orchestrating the retrieval and generation process in the knowledge base node.

Type: KnowledgeBaseOrchestrationConfiguration object

Required: No

promptTemplate

A custom prompt template to use with the knowledge base for generating responses.

Type: KnowledgeBasePromptTemplate object

Required: No

rerankingConfiguration

The configuration for reranking the retrieved results from the knowledge base to improve relevance.

Type: VectorSearchRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseOrchestrationConfiguration

Service: Agents for Amazon Bedrock

Configures how the knowledge base orchestrates the retrieval and generation process, allowing for customization of prompts, inference parameters, and performance settings.

Contents

additionalModelRequestFields

The additional model-specific request parameters as key-value pairs to be included in the request to the foundation model.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

inferenceConfig

Contains inference configurations for the prompt.

Type: PromptInferenceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

performanceConfig

The performance configuration options for the knowledge base retrieval and generation process.

Type: PerformanceConfiguration object

Required: No

promptTemplate

A custom prompt template for orchestrating the retrieval and generation process.

Type: KnowledgeBasePromptTemplate object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBasePromptTemplate

Service: Agents for Amazon Bedrock

Defines a custom prompt template for orchestrating the retrieval and generation process.

Contents

textPromptTemplate

The text of the prompt template.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseSummary

Service: Agents for Amazon Bedrock

Contains details about a knowledge base.

Contents

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

name

The name of the knowledge base.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

status

The status of the knowledge base.

Type: String

Valid Values: `CREATING | ACTIVE | DELETING | UPDATING | FAILED | DELETE_UNSUCCESSFUL | UPDATE_UNSUCCESSFUL`

Required: Yes

updatedAt

The time the knowledge base was last updated.

Type: Timestamp

Required: Yes

description

The description of the knowledge base.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LambdaFunctionFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a Lambda function node in the flow. You specify the Lambda function to invoke and the inputs into the function. The output is the response that is defined in the Lambda function. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

lambdaArn

The Amazon Resource Name (ARN) of the Lambda function to invoke.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `$|^arn:aws(-cn|-us-gov|-eusc|-iso(-[b-f])?)?:lambda:([a-z]{2,}-){2,}\d:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LastDayOfMonth

Service: Agents for Amazon Bedrock

The option to run the monthly sync on the last calendar day of each month.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LexFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a Lex node in the flow. You specify a Amazon Lex bot to invoke. This node takes an utterance as the input and returns as the output the intent identified by the Amazon Lex bot. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

botAliasArn

The Amazon Resource Name (ARN) of the Amazon Lex bot alias to invoke.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 78.

Pattern: `$|^arn:aws(-cn|-us-gov|-eusc|-iso(-[b-f])?)?:lex:([a-z]{2,}-){2,}\d:\d{12}:bot-alias/[0-9a-zA-Z]+/[0-9a-zA-Z]+`

Required: Yes

localeId

The Region to invoke the Amazon Lex bot in.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LoopControllerFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for the controller node of a DoWhile loop in the flow.

Contents

continueCondition

Specifies the condition that determines when the flow exits the DoWhile loop. The loop executes until this condition evaluates to true.

Type: FlowCondition object

Required: Yes

maxIterations

Specifies the maximum number of times the DoWhile loop can iterate before the flow exits the loop.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LoopFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for the nodes of a DoWhile loop in your flow.

A DoWhile loop is made up of the following nodes:

- `Loop` - The container node that holds the loop's flow definition. This node encompasses the entire loop structure.
- `LoopInput` - The entry point node for the loop. This node receives inputs from nodes outside the loop and from previous loop iterations.
- Body nodes - The processing nodes that execute within each loop iteration. These can be nodes for handling data in your flow, such as a prompt or Lambda function nodes. Some node types aren't supported inside a DoWhile loop body. For more information, see LoopIncompatibleNodeTypeFlowValidationDetails.
- `LoopController` - The node that evaluates whether the loop should continue or exit based on a condition.

These nodes work together to create a loop that runs at least once and continues until a specified condition is met or a maximum number of iterations is reached.

Contents

definition

The definition of the DoWhile loop nodes and connections between nodes in the flow.

Type: FlowDefinition object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

LoopIncompatibleNodeTypeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a flow that contains an incompatible node in a DoWhile loop.

Contents

incompatibleNodeName

The node that's incompatible in the DoWhile loop.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

incompatibleNodeType

The node type of the incompatible node in the DoWhile loop. Some node types, like a condition node, aren't allowed in a DoWhile loop.

Type: String

Valid Values: `Input | Condition | Iterator | Collector`

Required: Yes

node

The Loop container node that contains an incompatible node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LoopInputFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for the input node of a DoWhile loop in the flow.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MalformedConditionExpressionFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a malformed condition expression in a node.

Contents

cause

The error message describing why the condition expression is malformed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

condition

The name of the malformed condition.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the node containing the malformed condition expression.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

MalformedNodeInputExpressionFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a malformed input expression in a node.

Contents

cause

The error message describing why the input expression is malformed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: Yes

input

The name of the input with the malformed expression.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the node containing the malformed input expression.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedKnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Configurations for a managed knowledge base.

Contents

embeddingModelArn

The ARN for the embeddings model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

embeddingModelConfiguration

The configuration details for the embeddings model. Not required when choosing the MANAGED embeddingModelType.

Type: EmbeddingModelConfiguration object

Required: No

embeddingModelType

Choose CUSTOM to provide your own Bedrock embedding model ARN. Choose MANAGED to use a service-managed embedding model.

Type: String

Valid Values: `CUSTOM | MANAGED`

Required: No

serverSideEncryptionConfiguration

Contains the configuration for server-side encryption for your managed knowledge base.

Type: ServerSideEncryptionConfiguration object

Required: No

supplementalDataStorageConfiguration

Use this object to specify the Amazon S3 location that the knowledge base uses to process and ingest multimodal content. This field is required when you use a native multimodal embedding model.

Type: SupplementalDataStorageConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedKnowledgeBaseConnectorConfiguration

Service: Agents for Amazon Bedrock

Configuration for managed knowledge base connector data sources.

Contents

connectorParameters

Connector-specific parameters. For more information, see Connect a data source.

Type: JSON value

Required: No

deletionProtectionConfiguration

A safeguard against accidental bulk deletion of indexed content.

Type: DeletionProtectionConfiguration object

Required: No

mediaExtractionConfiguration

Configuration for extracting media (images, audio, video) from data source files.

Type: MediaExtractionConfiguration object

Required: No

syncSchedule

The recurring schedule on which the connector automatically syncs this data source. If not specified, the data source is not synced automatically and you start each sync yourself. Not supported for the Custom connector.

Type: SyncSchedule object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MediaExtractionConfiguration

Service: Agents for Amazon Bedrock

Configuration for media extraction settings.

Contents

audioExtractionConfiguration

Configuration for audio extraction.

Type: AudioExtractionConfiguration object

Required: No

imageExtractionConfiguration

Configuration for image extraction.

Type: ImageExtractionConfiguration object

Required: No

videoExtractionConfiguration

Configuration for video extraction.

Type: VideoExtractionConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MemoryConfiguration

Service: Agents for Amazon Bedrock

Details of the memory configuration.

Contents

enabledMemoryTypes

The type of memory that is stored.

Type: Array of strings

Array Members: Fixed number of 1 item.

Valid Values: `SESSION_SUMMARY`

Required: Yes

sessionSummaryConfiguration

Contains the configuration for SESSION_SUMMARY memory type enabled for the agent.

Type: SessionSummaryConfiguration object

Required: No

storageDays

The number of days the agent is configured to retain the conversational context.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 365.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

Message

Service: Agents for Amazon Bedrock

A message input or response from a model. For more information, see Create a prompt using Prompt management.

Contents

content

The content in the message.

Type: Array of ContentBlock objects

Required: Yes

role

The role that the message belongs to.

Type: String

Valid Values: `user | assistant`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataAttribute

Service: Agents for Amazon Bedrock

Contains information about a metadata attribute.

Contents

key

The key of the metadata attribute.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

value

Contains the value of the metadata attribute.

Type: MetadataAttributeValue object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataAttributeValue

Service: Agents for Amazon Bedrock

Contains the value of the metadata attribute. Choose a `type` and include the field that corresponds to it.

Contents

type

The type of the metadata attribute.

Type: String

Valid Values: `BOOLEAN | NUMBER | STRING | STRING_LIST`

Required: Yes

booleanValue

The value of the Boolean metadata attribute.

Type: Boolean

Required: No

numberValue

The value of the numeric metadata attribute.

Type: Double

Required: No

stringListValue

An array of strings that define the value of the metadata attribute.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: No

stringValue

The value of the string metadata attribute.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataConfigurationForReranking

Service: Agents for Amazon Bedrock

Specifies how metadata fields should be handled during the reranking process.

Contents

selectionMode

The mode for selecting metadata fields for reranking.

Type: String

Valid Values: `SELECTIVE | ALL`

Required: Yes

selectiveModeConfiguration

The configuration for selective metadata field inclusion or exclusion during reranking.

Type: RerankingMetadataSelectiveModeConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MismatchedNodeInputTypeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about mismatched input data types in a node.

Contents

expectedType

The expected data type for the node input.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: Yes

input

The name of the input with the mismatched data type.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the node containing the input with the mismatched data type.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

MismatchedNodeOutputTypeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about mismatched output data types in a node.

Contents

expectedType

The expected data type for the node output.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: Yes

node

The name of the node containing the output with the mismatched data type.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

output

The name of the output with the mismatched data type.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingConnectionConfigurationFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a connection missing required configuration.

Contents

connection

The name of the connection missing configuration.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingDefaultConditionFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a missing default condition in a conditional node.

Contents

node

The name of the node missing the default condition.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingEndingNodesFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about missing ending nodes (such as FlowOutputNode) in the flow.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingLoopControllerNodeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a flow that's missing a required `LoopController` node in a DoWhile loop.

Contents

loopNode

The DoWhile loop in a flow that's missing a required `LoopController` node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingLoopInputNodeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a flow that's missing a required `LoopInput` node in a DoWhile loop.

Contents

loopNode

The DoWhile loop in a flow that's missing a required `LoopInput` node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingNodeConfigurationFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a node missing a required configuration.

Contents

node

The name of the node missing a required configuration.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingNodeInputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a missing required input in a node.

Contents

input

The name of the missing input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the node missing the required input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingNodeOutputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a missing required output in a node.

Contents

node

The name of the node missing the required output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

output

The name of the missing output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MissingStartingNodesFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about missing starting nodes (such as FlowInputNode) in the flow.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MongoDbAtlasConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in MongoDB Atlas.

Contents

collectionName

The collection name of the knowledge base in MongoDB Atlas.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `.*`

Required: Yes

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that contains user credentials for your MongoDB Atlas cluster.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

databaseName

The database name in your MongoDB Atlas cluster for your knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `.*`

Required: Yes

endpoint

The endpoint URL of your MongoDB Atlas cluster for your knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: MongoDbAtlasFieldMapping object

Required: Yes

vectorIndexName

The name of the MongoDB Atlas vector search index.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

endpointServiceName

The name of the VPC endpoint service in your account that is connected to your MongoDB Atlas cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `(?:arn:aws(?:-us-gov|-cn|-iso|-iso-[a-z])*:.+:.*:\d+:.+/.+$|[a-zA-Z0-9*]+[a-zA-Z0-9._-]*)`

Required: No

textIndexName

The name of the text search index in the MongoDB collection. This is required for using the hybrid search feature.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MongoDbAtlasFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MonthlySchedule

Service: Agents for Amazon Bedrock

A monthly sync on a specified day of the month.

Contents

dayOfMonth

The day of the month on which the monthly sync runs.

Type: DayOfMonth object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MultipleLoopControllerNodesFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a flow that contains multiple `LoopController` nodes in a DoWhile loop.

Contents

loopNode

The DoWhile loop in a flow that contains multiple `LoopController` nodes.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MultipleLoopInputNodesFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about a flow that contains multiple `LoopInput` nodes in a DoWhile loop.

Contents

loopNode

The DoWhile loop in a flow that contains multiple `LoopInput` nodes.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MultipleNodeInputConnectionsFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about multiple connections to a single node input.

Contents

input

The name of the input with multiple connections to it.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the node containing the input with multiple connections.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NeptuneAnalyticsConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Amazon Neptune Analytics. For more information, see Create a vector index in Amazon Neptune Analytics.

Contents

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: NeptuneAnalyticsFieldMapping object

Required: Yes

graphArn

The Amazon Resource Name (ARN) of the Neptune Analytics vector store.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `arn:aws(|-cn|-us-gov):neptune-graph:[a-zA-Z0-9-]*:[0-9]{12}:graph/g-[a-zA-Z0-9]{10}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NeptuneAnalyticsFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OpenSearchManagedClusterConfiguration

Service: Agents for Amazon Bedrock

Contains details about the Managed Cluster configuration of the knowledge base in Amazon OpenSearch Service. For more information, see Create a vector index in OpenSearch Managed Cluster.

Contents

domainArn

The Amazon Resource Name (ARN) of the OpenSearch domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov|-iso):es:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:domain/[a-z][a-z0-9-]{3,28}`

Required: Yes

domainEndpoint

The endpoint URL the OpenSearch domain.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `https://.*`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: OpenSearchManagedClusterFieldMapping object

Required: Yes

vectorIndexName

The name of the vector store.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(?![\-_+.])[a-z0-9][a-z0-9\-_\.]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OpenSearchManagedClusterFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OpenSearchServerlessConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Amazon OpenSearch Service. For more information, see Create a vector index in Amazon OpenSearch Service.

Contents

collectionArn

The Amazon Resource Name (ARN) of the OpenSearch Service vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws:aoss:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:collection/[a-z0-9-]{3,32}`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: OpenSearchServerlessFieldMapping object

Required: Yes

vectorIndexName

The name of the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OpenSearchServerlessFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OrchestrationExecutor

Service: Agents for Amazon Bedrock

Contains details about the Lambda function containing the orchestration logic carried out upon invoking the custom orchestration.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

lambda

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for an output flow node in the flow. You specify the data type expected for the input into the node in the `type` field and how to return the final output in the `expression` field.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParameterDetail

Service: Agents for Amazon Bedrock

Contains details about a parameter in a function for an action group.

This data type is used in the following API operations:

- CreateAgentActionGroup request
- CreateAgentActionGroup response
- UpdateAgentActionGroup request
- UpdateAgentActionGroup response
- GetAgentActionGroup response

Contents

type

The data type of the parameter.

Type: String

Valid Values: `string | number | integer | boolean | array`

Required: Yes

description

A description of the parameter. Helps the foundation model determine how to elicit the parameters from the user.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: No

required

Whether the parameter is required for the agent to complete the function for action group invocation.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParsingConfiguration

Service: Agents for Amazon Bedrock

Settings for parsing document contents. If you exclude this field, the default parser converts the contents of each document into text before splitting it into chunks. Specify the parsing strategy to use in the `parsingStrategy` field and include the relevant configuration, or omit it to use the Amazon Bedrock default parser. For more information, see Parsing options for your data source.

> **Note**
>
> If you specify `BEDROCK_DATA_AUTOMATION` or `BEDROCK_FOUNDATION_MODEL` and it fails to parse a file, the Amazon Bedrock default parser will be used instead.

Contents

parsingStrategy

The parsing strategy for the data source.

For managed knowledge bases, the strategy that you can select depends on the embedding model that your knowledge base uses:

- If your knowledge base uses a native multimodal embedding model, specify `MULTI_MODAL_EMBEDDINGS`. With this strategy, files are sent directly to the embedding model instead of being parsed into text. This is the only strategy that is supported for these knowledge bases.
- Otherwise, specify `SMART_PARSING`.

For more information, see Customize ingestion for managed knowledge bases.

Type: String

Valid Values: `BEDROCK_FOUNDATION_MODEL | BEDROCK_DATA_AUTOMATION | SMART_PARSING | MULTI_MODAL_EMBEDDINGS`

Required: Yes

bedrockDataAutomationConfiguration

If you specify `BEDROCK_DATA_AUTOMATION` as the parsing strategy for ingesting your data source, use this object to modify configurations for using the Amazon Bedrock Data Automation parser.

Type: BedrockDataAutomationConfiguration object

Required: No

bedrockFoundationModelConfiguration

If you specify `BEDROCK_FOUNDATION_MODEL` as the parsing strategy for ingesting your data source, use this object to modify configurations for using a foundation model to parse documents.

Type: BedrockFoundationModelConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParsingPrompt

Service: Agents for Amazon Bedrock

Instructions for interpreting the contents of a document.

Contents

parsingPromptText

Instructions for interpreting the contents of a document.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 10000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PatternObjectFilter

Service: Agents for Amazon Bedrock

The specific filters applied to your data source content. You can filter out or include certain content.

Contents

objectType

The supported object type or content type of the data source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Required: Yes

exclusionFilters

A list of one or more exclusion regular expression patterns to exclude certain object types that adhere to the pattern. If you specify an inclusion and exclusion filter/pattern and both match a document, the exclusion filter takes precedence and the document isn't crawled.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

inclusionFilters

A list of one or more inclusion regular expression patterns to include certain object types that adhere to the pattern. If you specify an inclusion and exclusion filter/pattern and both match a document, the exclusion filter takes precedence and the document isn't crawled.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PatternObjectFilterConfiguration

Service: Agents for Amazon Bedrock

The configuration of filtering certain objects or content types of the data source.

Contents

filters

The configuration of specific filters applied to your data source content. You can filter out or include certain content.

Type: Array of PatternObjectFilter objects

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PerformanceConfiguration

Service: Agents for Amazon Bedrock

The performance-related configuration options for the knowledge base retrieval and generation process.

Contents

latency

The latency optimization setting.

Type: String

Valid Values: `standard | optimized`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PineconeConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Pinecone. For more information, see Create a vector index in Pinecone.

Contents

connectionString

The endpoint URL for your index management page.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that is linked to your Pinecone API key.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: PineconeFieldMapping object

Required: Yes

namespace

The namespace to be used to write new data to your database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PineconeFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptAgentResource

Service: Agents for Amazon Bedrock

Contains specifications for an Amazon Bedrock agent with which to use the prompt. For more information, see Create a prompt using Prompt management and Automate tasks in your application using conversational agents.

Contents

agentIdentifier

The ARN of the agent with which to use the prompt.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptConfiguration

Service: Agents for Amazon Bedrock

Contains configurations to override a prompt template in one part of an agent sequence. For more information, see Advanced prompts.

Contents

additionalModelRequestFields

If the Converse or ConverseStream operations support the model, `additionalModelRequestFields` contains additional inference parameters, beyond the base set of inference parameters in the `inferenceConfiguration` field.

For more information, see Inference request parameters and response fields for foundation models.

Type: JSON value

Required: No

basePromptTemplate

Defines the prompt template with which to replace the default prompt template. You can use placeholder variables in the base prompt template to customize the prompt. For more information, see Prompt template placeholder variables. For more information, see Configure the prompt templates.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: No

foundationModel

The agent's foundation model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

inferenceConfiguration

Contains inference parameters to use when the agent invokes a foundation model in the part of the agent sequence defined by the `promptType`. For more information, see Inference parameters for foundation models.

Type: InferenceConfiguration object

Required: No

parserMode

Specifies whether to override the default parser Lambda function when parsing the raw foundation model output in the part of the agent sequence defined by the `promptType`. If you set the field as `OVERRIDDEN`, the `overrideLambda` field in the PromptOverrideConfiguration must be specified with the ARN of a Lambda function.

Type: String

Valid Values: `DEFAULT | OVERRIDDEN`

Required: No

promptCreationMode

Specifies whether to override the default prompt template for this `promptType`. Set this value to `OVERRIDDEN` to use the prompt that you provide in the `basePromptTemplate`. If you leave it as `DEFAULT`, the agent uses a default prompt template.

Type: String

Valid Values: `DEFAULT | OVERRIDDEN`

Required: No

promptState

Specifies whether to allow the agent to carry out the step specified in the `promptType`. If you set this value to `DISABLED`, the agent skips that step. The default state for each `promptType` is as follows.

- PRE_PROCESSING – DISABLED
- ORCHESTRATION – ENABLED
- KNOWLEDGE_BASE_RESPONSE_GENERATION – ENABLED
- POST_PROCESSING – DISABLED

Type: String

Valid Values: ENABLED | DISABLED

Required: No

promptType

The step in the agent sequence that this prompt configuration applies to.

Type: String

Valid Values: PRE_PROCESSING | ORCHESTRATION | POST_PROCESSING | KNOWLEDGE_BASE_RESPONSE_GENERATION | MEMORY_SUMMARIZATION

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a prompt node in the flow. You can use a prompt from Prompt management or you can define one in this node. If the prompt contains variables, the inputs into this node will fill in the variables. The output from this node is the response generated by the model. For more information, see Node types in a flow in the Amazon Bedrock User Guide.

Contents

sourceConfiguration

Specifies whether the prompt is from Prompt management or defined inline.

Type: PromptFlowNodeSourceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

guardrailConfiguration

Contains configurations for a guardrail to apply to the prompt in this node and the response generated from it.

Type: GuardrailConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptFlowNodeInlineConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a prompt defined inline in the node.

Contents

modelId

The unique identifier of the model or inference profile to run inference with.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `$|^(arn:aws(-cn|-us-gov|-eusc|-iso(-[b-f])?)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

templateConfiguration

Contains a prompt and variables in the prompt that can be replaced with values at runtime.

Type: PromptTemplateConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

templateType

The type of prompt template.

Type: String

Valid Values: `TEXT | CHAT`

Required: Yes

additionalModelRequestFields

Additional fields to be included in the model request for the Prompt node.

Type: JSON value

Required: No

inferenceConfiguration

Contains inference configurations for the prompt.

Type: PromptInferenceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptFlowNodeResourceConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a prompt from Prompt management to use in a node.

Contents

promptArn

The Amazon Resource Name (ARN) of the prompt from Prompt management.

Type: String

Pattern: `$|^(arn:aws(-cn|-us-gov|-eusc|-iso(-[b-f])?)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptFlowNodeSourceConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a prompt and whether it is from Prompt management or defined inline.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

inline

Contains configurations for a prompt that is defined inline

Type: PromptFlowNodeInlineConfiguration object

Required: No

resource

Contains configurations for a prompt from Prompt management.

Type: PromptFlowNodeResourceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptGenAiResource

Service: Agents for Amazon Bedrock

Contains specifications for a generative AI resource with which to use the prompt. For more information, see Create a prompt using Prompt management.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

agent

Specifies an Amazon Bedrock agent with which to use the prompt.

Type: PromptAgentResource object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptInferenceConfiguration

Service: Agents for Amazon Bedrock

Contains inference configurations for the prompt.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

Contains inference configurations for a text prompt.

Type: PromptModelInferenceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptInputVariable

Service: Agents for Amazon Bedrock

Contains information about a variable in the prompt.

Contents

name

The name of the variable.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptMetadataEntry

Service: Agents for Amazon Bedrock

Contains a key-value pair that defines a metadata tag and value to attach to a prompt variant. For more information, see Create a prompt using Prompt management.

Contents

key

The key of a metadata tag for a prompt variant.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

value

The value of a metadata tag for a prompt variant.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptModelInferenceConfiguration

Service: Agents for Amazon Bedrock

Contains inference configurations related to model inference for a prompt. For more information, see Inference parameters.

Contents

maxTokens

The maximum number of tokens to return in the response.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

stopSequences

A list of strings that define sequences after which the model will stop generating.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

temperature

Controls the randomness of the response. Choose a lower value for more predictable outputs and a higher value for more surprising outputs.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topP

The percentage of most-likely candidates that the model considers for the next token.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptOverrideConfiguration

Service: Agents for Amazon Bedrock

Contains configurations to override prompts in different parts of an agent sequence. For more information, see Advanced prompts.

Contents

promptConfigurations

Contains configurations to override a prompt template in one part of an agent sequence. For more information, see Advanced prompts.

Type: Array of PromptConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

overrideLambda

The ARN of the Lambda function to use when parsing the raw foundation model output in parts of the agent sequence. If you specify this field, at least one of the `promptConfigurations` must contain a `parserMode` value that is set to `OVERRIDDEN`. For more information, see Parser Lambda function in Amazon Bedrock Agents.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptSummary

Service: Agents for Amazon Bedrock

Contains information about a prompt in your Prompt management tool.

This data type is used in the following API operations:

- ListPrompts response

Contents

arn

The Amazon Resource Name (ARN) of the prompt or the prompt version (if you specified a version in the request).

Type: String

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:prompt/[0-9a-zA-Z]{10}(?::[0-9]{1,5})?)`

Required: Yes

createdAt

The time at which the prompt was created.

Type: Timestamp

Required: Yes

id

The unique identifier of the prompt.

Type: String

Pattern: `[0-9a-zA-Z]{10}`

Required: Yes

name

The name of the prompt.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

updatedAt

The time at which the prompt was last updated.

Type: Timestamp

Required: Yes

version

The version of the prompt that this summary applies to.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

description

The description of the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptTemplateConfiguration

Service: Agents for Amazon Bedrock

Contains the message for a prompt. For more information, see Construct and store reusable prompts with Prompt management in Amazon Bedrock.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

chat

Contains configurations to use the prompt in a conversational format.

Type: ChatPromptTemplateConfiguration object

Required: No

text

Contains configurations for the text in a message for a prompt.

Type: TextPromptTemplateConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptVariant

Service: Agents for Amazon Bedrock

Contains details about a variant of the prompt.

Contents

name

The name of the prompt variant.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

templateConfiguration

Contains configurations for the prompt template.

Type: PromptTemplateConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

templateType

The type of prompt template to use.

Type: String

Valid Values: `TEXT | CHAT`

Required: Yes

additionalModelRequestFields

Contains model-specific inference configurations that aren't in the `inferenceConfiguration` field. To see model-specific inference parameters, see Inference request parameters and response fields for foundation models.

Type: JSON value

Required: No

genAiResource

Specifies a generative AI resource with which to use the prompt.

Type: PromptGenAiResource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

inferenceConfiguration

Contains inference configurations for the prompt variant.

Type: PromptInferenceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

metadata

An array of objects, each containing a key-value pair that defines a metadata tag and value to attach to a prompt variant.

Type: Array of PromptMetadataEntry objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

modelId

The unique identifier of the model or inference profile with which to run inference on the prompt.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryGenerationColumn

Service: Agents for Amazon Bedrock

Contains information about a column in the current table for the query engine to consider.

Contents

description

A description of the column that helps the query engine understand the contents of the column.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

inclusion

Specifies whether to include or exclude the column during query generation. If you specify `EXCLUDE`, the column will be ignored. If you specify `INCLUDE`, all other columns in the table will be ignored.

Type: String

Valid Values: `INCLUDE | EXCLUDE`

Required: No

name

The name of the column for which the other fields in this object apply.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 127.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryGenerationConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for query generation. For more information, see Build a knowledge base by connecting to a structured data source in the Amazon Bedrock User Guide..

Contents

executionTimeoutSeconds

The time after which query generation will time out.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 200.

Required: No

generationContext

Specifies configurations for context to use during query generation.

Type: QueryGenerationContext object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryGenerationContext

Service: Agents for Amazon Bedrock

>Contains configurations for context to use during query generation.

Contents

curatedQueries

An array of objects, each of which defines information about example queries to help the query engine generate appropriate SQL queries.

Type: Array of CuratedQuery objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: No

tables

An array of objects, each of which defines information about a table in the database.

Type: Array of QueryGenerationTable objects

Array Members: Minimum number of 0 items. Maximum number of 50 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryGenerationTable

Service: Agents for Amazon Bedrock

Contains information about a table for the query engine to consider.

Contents

name

The name of the table for which the other fields in this object apply.

Type: String

Pattern: `.*\..*\..*`

Required: Yes

columns

An array of objects, each of which defines information about a column in the table.

Type: Array of QueryGenerationColumn objects

Required: No

description

A description of the table that helps the query engine understand the contents of the table.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

inclusion

Specifies whether to include or exclude the table during query generation. If you specify `EXCLUDE`, the table will be ignored. If you specify `INCLUDE`, all other tables will be ignored.

Type: String

Valid Values: `INCLUDE | EXCLUDE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RdsConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Amazon RDS. For more information, see Create a vector index in Amazon RDS.

Contents

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that is linked to your Amazon RDS database.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

databaseName

The name of your Amazon RDS database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\-]+`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: RdsFieldMapping object

Required: Yes

resourceArn

The Amazon Resource Name (ARN) of the vector store.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):rds:[a-zA-Z0-9-]*:[0-9]{12}:cluster:[a-zA-Z0-9-]{1,63}`

Required: Yes

tableName

The name of the table in the database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\.\-]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RdsFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\-]+`

Required: Yes

primaryKeyField

The name of the field in which Amazon Bedrock stores the ID for each entry.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\-]+`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\-]+`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\-]+`

Required: Yes

customMetadataField

Provide a name for the universal metadata field where Amazon Bedrock will store any custom metadata from your data source.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `[a-zA-Z0-9_\-]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedisEnterpriseCloudConfiguration

Service: Agents for Amazon Bedrock

Contains details about the storage configuration of the knowledge base in Redis Enterprise Cloud. For more information, see Create a vector index in Redis Enterprise Cloud.

Contents

credentialsSecretArn

The Amazon Resource Name (ARN) of the secret that you created in AWS Secrets Manager that is linked to your Redis Enterprise Cloud database.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

endpoint

The endpoint URL of the Redis Enterprise Cloud database.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

fieldMapping

Contains the names of the fields to which to map information about the vector store.

Type: RedisEnterpriseCloudFieldMapping object

Required: Yes

vectorIndexName

The name of the vector index.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedisEnterpriseCloudFieldMapping

Service: Agents for Amazon Bedrock

Contains the names of the fields to which to map information about the vector store.

Contents

metadataField

The name of the field in which Amazon Bedrock stores metadata about the vector store.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

textField

The name of the field in which Amazon Bedrock stores the raw text from your data. The text is split according to the chunking strategy you choose.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

vectorField

The name of the field in which Amazon Bedrock stores the vector embeddings for your data sources.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for an Amazon Redshift database. For more information, see Build a knowledge base by connecting to a structured data source in the Amazon Bedrock User Guide.

Contents

queryEngineConfiguration

Specifies configurations for an Amazon Redshift query engine.

Type: RedshiftQueryEngineConfiguration object

Required: Yes

storageConfigurations

Specifies configurations for Amazon Redshift database storage.

Type: Array of RedshiftQueryEngineStorageConfiguration objects

Array Members: Fixed number of 1 item.

Required: Yes

queryGenerationConfiguration

Specifies configurations for generating queries.

Type: QueryGenerationConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftProvisionedAuthConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for authentication to an Amazon Redshift provisioned data warehouse. Specify the type of authentication to use in the `type` field and include the corresponding field. If you specify IAM authentication, you don't need to include another field.

Contents

type

The type of authentication to use.

Type: String

Valid Values: `IAM | USERNAME_PASSWORD | USERNAME`

Required: Yes

databaseUser

The database username for authentication to an Amazon Redshift provisioned data warehouse.

Type: String

Required: No

usernamePasswordSecretArn

The ARN of an Secrets Manager secret for authentication.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftProvisionedConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a provisioned Amazon Redshift query engine.

Contents

authConfiguration

Specifies configurations for authentication to Amazon Redshift.

Type: RedshiftProvisionedAuthConfiguration object

Required: Yes

clusterIdentifier

The ID of the Amazon Redshift cluster.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftQueryEngineAwsDataCatalogStorageConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for storage in AWS Glue Data Catalog.

Contents

tableNames

A list of names of the tables to use.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 1000 items.

Length Constraints: Minimum length of 1. Maximum length of 200.

Pattern: `.*\.*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftQueryEngineConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for an Amazon Redshift query engine. Specify the type of query engine in `type` and include the corresponding field. For more information, see Build a knowledge base by connecting to a structured data source in the Amazon Bedrock User Guide.

Contents

type

The type of query engine.

Type: String

Valid Values: `SERVERLESS | PROVISIONED`

Required: Yes

provisionedConfiguration

Specifies configurations for a provisioned Amazon Redshift query engine.

Type: RedshiftProvisionedConfiguration object

Required: No

serverlessConfiguration

Specifies configurations for a serverless Amazon Redshift query engine.

Type: RedshiftServerlessConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

RedshiftQueryEngineRedshiftStorageConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for storage in Amazon Redshift.

Contents

databaseName

The name of the Amazon Redshift database.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftQueryEngineStorageConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for Amazon Redshift data storage. Specify the data storage service to use in the `type` field and include the corresponding field. For more information, see Build a knowledge base by connecting to a structured data source in the Amazon Bedrock User Guide.

Contents

type

The data storage service to use.

Type: String

Valid Values: `REDSHIFT | AWS_DATA_CATALOG`

Required: Yes

awsDataCatalogConfiguration

Specifies configurations for storage in AWS Glue Data Catalog.

Type: RedshiftQueryEngineAwsDataCatalogStorageConfiguration object

Required: No

redshiftConfiguration

Specifies configurations for storage in Amazon Redshift.

Type: RedshiftQueryEngineRedshiftStorageConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

RedshiftServerlessAuthConfiguration

Service: Agents for Amazon Bedrock

Specifies configurations for authentication to a Redshift Serverless. Specify the type of authentication to use in the `type` field and include the corresponding field. If you specify IAM authentication, you don't need to include another field.

Contents

type

The type of authentication to use.

Type: String

Valid Values: `IAM | USERNAME_PASSWORD`

Required: Yes

usernamePasswordSecretArn

The ARN of an Secrets Manager secret for authentication.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RedshiftServerlessConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for authentication to Amazon Redshift Serverless.

Contents

authConfiguration

Specifies configurations for authentication to an Amazon Redshift provisioned data warehouse.

Type: RedshiftServerlessAuthConfiguration object

Required: Yes

workgroupArn

The ARN of the Amazon Redshift workgroup.

Type: String

Pattern: `(arn:(aws(-[a-z]+)*):redshift-serverless:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:workgroup/[0-9a-f]{8}-[0-9a-f]{4}-[0-9a-f]{4}-[0-9a-f]{4}-[0-9a-f]{12})`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankingMetadataSelectiveModeConfiguration

Service: Agents for Amazon Bedrock

Configures the metadata fields to include or exclude during the reranking process when using selective mode.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

fieldsToExclude

Specifies the metadata fields to exclude from the reranking process.

Type: Array of FieldForReranking objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

fieldsToInclude

Specifies the metadata fields to include in the reranking process.

Type: Array of FieldForReranking objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

RetrievalFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a Retrieval node in a flow. This node retrieves data from the Amazon S3 location that you specify and returns it as the output.

Contents

serviceConfiguration

Contains configurations for the service to use for retrieving data to return as the output from the node.

Type: RetrievalFlowNodeServiceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalFlowNodeS3Configuration

Service: Agents for Amazon Bedrock

Contains configurations for the Amazon S3 location from which to retrieve data to return as the output from the node.

Contents

bucketName

The name of the Amazon S3 bucket from which to retrieve data.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `$|^[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9]`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalFlowNodeServiceConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for the service to use for retrieving data to return as the output from the node.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3

Contains configurations for the Amazon S3 location from which to retrieve data to return as the output from the node.

Type: RetrievalFlowNodeS3Configuration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Content

Service: Agents for Amazon Bedrock

Contains information about the content to ingest into a knowledge base connected to an Amazon S3 data source.

Contents

s3Location

The S3 location of the file containing the content to ingest.

Type: S3Location object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3DataSourceConfiguration

Service: Agents for Amazon Bedrock

The configuration information to connect to Amazon S3 as your data source for self-managed knowledge bases. To configure this data source for managed knowledge bases, use managedKnowledgeBaseConnectorConfiguration.

Contents

bucketArn

The Amazon Resource Name (ARN) of the S3 bucket that contains your data.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):s3:::[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]`

Required: Yes

bucketOwnerAccountId

The account ID for the owner of the S3 bucket.

Type: String

Length Constraints: Fixed length of 12.

Pattern: `[0-9]{12}`

Required: No

inclusionPrefixes

A list of S3 prefixes to include certain files or content. For more information, see Organizing objects using prefixes.

Type: Array of strings

Array Members: Fixed number of 1 item.

Length Constraints: Minimum length of 1. Maximum length of 300.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Identifier

Service: Agents for Amazon Bedrock

The identifier information for an Amazon S3 bucket.

Contents

s3BucketName

The name of the S3 bucket.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Pattern: `[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9]`

Required: No

s3ObjectKey

The S3 object key for the S3 resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\.\-\!\*\_\'\(\)a-zA-Z0-9][\.\-\!\*\_\'\(\)\/a-zA-Z0-9]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Location

Service: Agents for Amazon Bedrock

An Amazon S3 location.

Contents

uri

The location's URI. For example, `s3://my-bucket/chunk-processor/`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `s3://.{1,128}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3VectorsConfiguration

Service: Agents for Amazon Bedrock

Contains the storage configuration of the knowledge base for S3 vectors.

Contents

indexArn

The Amazon Resource Name (ARN) of the vector index used for the knowledge base. This ARN identifies the specific vector index resource within Amazon Bedrock.

Type: String

Required: No

indexName

The name of the vector index used for the knowledge base. This name identifies the vector index within the Amazon Bedrock service.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Required: No

vectorBucketArn

The Amazon Resource Name (ARN) of the S3 bucket where vector embeddings are stored. This bucket contains the vector data used by the knowledge base.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

SalesforceCrawlerConfiguration

Service: Agents for Amazon Bedrock

The configuration of the Salesforce content. For example, configuring specific types of Salesforce content.

Contents

filterConfiguration

The configuration of filtering the Salesforce content. For example, configuring regular expression patterns to include or exclude certain content.

Type: CrawlFilterConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SalesforceDataSourceConfiguration

Service: Agents for Amazon Bedrock

The configuration information to connect to Salesforce as your data source.

Contents

sourceConfiguration

The endpoint information to connect to your Salesforce data source.

Type: SalesforceSourceConfiguration object

Required: Yes

crawlerConfiguration

The configuration of the Salesforce content. For example, configuring specific types of Salesforce content.

Type: SalesforceCrawlerConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SalesforceSourceConfiguration

Service: Agents for Amazon Bedrock

The endpoint information to connect to your Salesforce data source.

Contents

authType

The supported authentication type to authenticate and connect to your Salesforce instance.

Type: String

Valid Values: `OAUTH2_CLIENT_CREDENTIALS`

Required: Yes

credentialsSecretArn

The Amazon Resource Name of an AWS Secrets Manager secret that stores your authentication credentials for your Salesforce instance URL. For more information on the key-value pairs that must be included in your secret, depending on your authentication type, see Salesforce connection configuration.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

hostUrl

The Salesforce host URL or instance URL.

Type: String

Pattern: `https://[A-Za-z0-9][^\s]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SeedUrl

Service: Agents for Amazon Bedrock

The seed or starting point URL. You should be authorized to crawl the URL.

Contents

url

A seed or starting point URL.

Type: String

Pattern: `https?://[A-Za-z0-9][^\s]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SemanticChunkingConfiguration

Service: Agents for Amazon Bedrock

Settings for semantic document chunking for a data source. Semantic chunking splits a document into into smaller documents based on groups of similar content derived from the text with natural language processing.

With semantic chunking, each sentence is compared to the next to determine how similar they are. You specify a threshold in the form of a percentile, where adjacent sentences that are less similar than that percentage of sentence pairs are divided into separate chunks. For example, if you set the threshold to 90, then the 10 percent of sentence pairs that are least similar are split. So if you have 101 sentences, 100 sentence pairs are compared, and the 10 with the least similarity are split, creating 11 chunks. These chunks are further split if they exceed the max token size.

You must also specify a buffer size, which determines whether sentences are compared in isolation, or within a moving context window that includes the previous and following sentence. For example, if you set the buffer size to 1, the embedding for sentence 10 is derived from sentences 9, 10, and 11 combined.

Contents

breakpointPercentileThreshold

The dissimilarity threshold for splitting chunks.

Type: Integer

Valid Range: Minimum value of 50. Maximum value of 99.

Required: Yes

bufferSize

The buffer size.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 1.

Required: Yes

maxTokens

The maximum number of tokens that a chunk can contain.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 8192.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ServerSideEncryptionConfiguration

Service: Agents for Amazon Bedrock

Contains the configuration for server-side encryption for your managed knowledge base.

Contents

kmsKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key used to encrypt the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SessionSummaryConfiguration

Service: Agents for Amazon Bedrock

Configuration for SESSION_SUMMARY memory type enabled for the agent.

Contents

maxRecentSessions

Maximum number of recent session summaries to include in the agent's prompt context.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SharePointCrawlerConfiguration

Service: Agents for Amazon Bedrock

The configuration of the SharePoint content. For example, configuring specific types of SharePoint content.

Contents

filterConfiguration

The configuration of filtering the SharePoint content. For example, configuring regular expression patterns to include or exclude certain content.

Type: CrawlFilterConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SharePointDataSourceConfiguration

Service: Agents for Amazon Bedrock

The configuration information to connect to SharePoint as your data source for self-managed knowledge bases.

Contents

sourceConfiguration

The endpoint information to connect to your SharePoint data source.

Type: SharePointSourceConfiguration object

Required: Yes

crawlerConfiguration

The configuration of the SharePoint content. For example, configuring specific types of SharePoint content.

Type: SharePointCrawlerConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SharePointSourceConfiguration

Service: Agents for Amazon Bedrock

The endpoint information to connect to your SharePoint data source.

Contents

authType

The supported authentication type to authenticate and connect to your SharePoint site/sites.

Type: String

Valid Values: `OAUTH2_CLIENT_CREDENTIALS | OAUTH2_SHAREPOINT_APP_ONLY_CLIENT_CREDENTIALS`

Required: Yes

credentialsSecretArn

The Amazon Resource Name of an AWS Secrets Manager secret that stores your authentication credentials for your SharePoint site/sites. For more information on the key-value pairs that must be included in your secret, depending on your authentication type, see SharePoint connection configuration.

Type: String

Pattern: `arn:aws(|-cn|-us-gov):secretsmanager:[a-z0-9-]{1,20}:([0-9]{12}|):secret:[a-zA-Z0-9!/_+=.@-]{1,512}`

Required: Yes

domain

The domain of your SharePoint instance or site URL/URLs.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 50.

Required: Yes

hostType

The supported host type, whether online/cloud or server/on-premises.

Type: String

Valid Values: `ONLINE`

Required: Yes

siteUrls

A list of one or more SharePoint site URLs.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Pattern: `https://[A-Za-z0-9][^\s]*`

Required: Yes

tenantId

The identifier of your Microsoft 365 tenant.

Type: String

Length Constraints: Fixed length of 36.

Pattern: `[0-9a-f]{8}-[0-9a-f]{4}-[0-9a-f]{4}-[0-9a-f]{4}-[0-9a-f]{12}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpecificToolChoice

Service: Agents for Amazon Bedrock

Defines a specific tool that the model must request. No text is generated but the results of tool use are sent back to the model to help generate a response. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

name

The name of the tool.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z][a-zA-Z0-9_]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SqlKnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a knowledge base connected to an SQL database. Specify the SQL database type in the `type` field and include the corresponding field. For more information, see Build a knowledge base by connecting to a structured data source in the Amazon Bedrock User Guide.

Contents

type

The type of SQL database to connect to the knowledge base.

Type: String

Valid Values: `REDSHIFT`

Required: Yes

redshiftConfiguration

Specifies configurations for a knowledge base connected to an Amazon Redshift database.

Type: RedshiftConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StorageConfiguration

Service: Agents for Amazon Bedrock

Contains the storage configuration of the knowledge base.

Contents

type

The vector store service in which the knowledge base is stored.

Type: String

Valid Values: `OPENSEARCH_SERVERLESS | PINECONE | REDIS_ENTERPRISE_CLOUD | RDS | MONGO_DB_ATLAS | NEPTUNE_ANALYTICS | OPENSEARCH_MANAGED_CLUSTER | S3_VECTORS`

Required: Yes

mongoDbAtlasConfiguration

Contains the storage configuration of the knowledge base in MongoDB Atlas.

Type: MongoDbAtlasConfiguration object

Required: No

neptuneAnalyticsConfiguration

Contains details about the Neptune Analytics configuration of the knowledge base in Amazon Neptune. For more information, see Create a vector index in Amazon Neptune Analytics..

Type: NeptuneAnalyticsConfiguration object

Required: No

opensearchManagedClusterConfiguration

Contains details about the storage configuration of the knowledge base in OpenSearch Managed Cluster. For more information, see Create a vector index in Amazon OpenSearch Service.

Type: OpenSearchManagedClusterConfiguration object

Required: No

opensearchServerlessConfiguration

Contains the storage configuration of the knowledge base in Amazon OpenSearch Service.

Type: OpenSearchServerlessConfiguration object

Required: No

pineconeConfiguration

Contains the storage configuration of the knowledge base in Pinecone.

Type: PineconeConfiguration object

Required: No

rdsConfiguration

Contains details about the storage configuration of the knowledge base in Amazon RDS. For more information, see Create a vector index in Amazon RDS.

Type: RdsConfiguration object

Required: No

redisEnterpriseCloudConfiguration

Contains the storage configuration of the knowledge base in Redis Enterprise Cloud.

Type: RedisEnterpriseCloudConfiguration object

Required: No

s3VectorsConfiguration

The configuration settings for storing knowledge base data using S3 vectors. This includes vector index information and S3 bucket details for vector storage.

Type: S3VectorsConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StorageFlowNodeConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a Storage node in a flow. This node stores the input in an Amazon S3 location that you specify.

Contents

serviceConfiguration

Contains configurations for the service to use for storing the input into the node.

Type: StorageFlowNodeServiceConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StorageFlowNodeS3Configuration

Service: Agents for Amazon Bedrock

Contains configurations for the Amazon S3 location in which to store the input into the node.

Contents

bucketName

The name of the Amazon S3 bucket in which to store the input into the node.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 63.

Pattern: `$|^[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9]`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StorageFlowNodeServiceConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for the service to use for storing the input into the node.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3

Contains configurations for the Amazon S3 location in which to store the input into the node.

Type: StorageFlowNodeS3Configuration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SupplementalDataStorageConfiguration

Service: Agents for Amazon Bedrock

Specifies configurations for the storage location of multimedia content (images, audio, and video) extracted from multimodal documents in your data source. This content can be retrieved and returned to the end user with timestamp references for audio and video segments.

Contents

storageLocations

A list of objects specifying storage locations for multimedia content (images, audio, and video) extracted from multimodal documents in your data source.

Type: Array of SupplementalDataStorageLocation objects

Array Members: Fixed number of 1 item.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SupplementalDataStorageLocation

Service: Agents for Amazon Bedrock

Contains information about a storage location for multimedia content (images, audio, and video) extracted from multimodal documents in your data source.

Contents

type

Specifies the storage service used for this location.

Type: String

Valid Values: S3

Required: Yes

s3Location

Contains information about the Amazon S3 location for the extracted multimedia content.

Type: S3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SyncSchedule

Service: Agents for Amazon Bedrock

The recurring schedule on which a managed knowledge base connector automatically syncs its data source. Specify exactly one of `daily`, `weekly`, or `monthly`.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

daily

A daily sync that runs once a day at a system-chosen off-peak time. The run time is not configurable.

Type: DailySchedule object

Required: No

monthly

A monthly sync that runs once a month on the specified day of the month.

Type: MonthlySchedule object

Required: No

weekly

A weekly sync that runs once a week on the specified day of the week.

Type: WeeklySchedule object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SystemContentBlock

Service: Agents for Amazon Bedrock

Contains a system prompt to provide context to the model or to describe how it should behave. For more information, see Create a prompt using Prompt management.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

cachePoint

Creates a cache checkpoint within a tool designation

Type: CachePointBlock object

Required: No

text

The text in the system prompt.

Type: String

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextContentDoc

Service: Agents for Amazon Bedrock

Contains information about content defined inline in text.

Contents

data

The text of the content.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5242880.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextPromptTemplateConfiguration

Service: Agents for Amazon Bedrock

Contains configurations for a text prompt template. To include a variable, enclose a word in double curly braces as in `{{variable}}`.

Contents

text

The message for the prompt.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

cachePoint

A cache checkpoint within a template configuration.

Type: CachePointBlock object

Required: No

inputVariables

An array of the variables in the prompt template.

Type: Array of PromptInputVariable objects

Array Members: Minimum number of 0 items. Maximum number of 20 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

Tool

Service: Agents for Amazon Bedrock

Contains configurations for a tool that a model can use when generating a response. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

cachePoint

Creates a cache checkpoint within a tool designation

Type: CachePointBlock object

Required: No

toolSpec

The specification for the tool.

Type: ToolSpecification object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolChoice

Service: Agents for Amazon Bedrock

Defines which tools the model should request when invoked. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

any

Defines tools, at least one of which must be requested by the model. No text is generated but the results of tool use are sent back to the model to help generate a response.

Type: AnyToolChoice object

Required: No

auto

Defines tools. The model automatically decides whether to call a tool or to generate text instead.

Type: AutoToolChoice object

Required: No

tool

Defines a specific tool that the model must request. No text is generated but the results of tool use are sent back to the model to help generate a response.

Type: SpecificToolChoice object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolConfiguration

Service: Agents for Amazon Bedrock

Configuration information for the tools that the model can use when generating a response. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

tools

An array of tools to pass to a model.

Type: Array of Tool objects

Array Members: Minimum number of 1 item.

Required: Yes

toolChoice

Defines which tools the model should request when invoked.

Type: ToolChoice object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolInputSchema

Service: Agents for Amazon Bedrock

The input schema for the tool. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

json

A JSON object defining the input schema for the tool.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolSpecification

Service: Agents for Amazon Bedrock

Contains a specification for a tool. For more information, see Use a tool to complete an Amazon Bedrock model response.

Contents

inputSchema

The input schema for the tool.

Type: ToolInputSchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

name

The name of the tool.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z][a-zA-Z0-9_]*`

Required: Yes

description

The description of the tool.

Type: String

Length Constraints: Minimum length of 1.

Required: No

strict

Whether the tool schema is strictly enforced.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Transformation

Service: Agents for Amazon Bedrock

A custom processing step for documents moving through a data source ingestion pipeline. To process documents after they have been converted into chunks, set the step to apply to `POST_CHUNKING`.

Contents

stepToApply

When the service applies the transformation.

Type: String

Valid Values: `POST_CHUNKING`

Required: Yes

transformationFunction

A Lambda function that processes documents.

Type: TransformationFunction object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformationFunction

Service: Agents for Amazon Bedrock

A Lambda function that processes documents.

Contents

transformationLambdaConfiguration

The Lambda function.

Type: TransformationLambdaConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformationLambdaConfiguration

Service: Agents for Amazon Bedrock

A Lambda function that processes documents.

Contents

lambdaArn

The function's ARN identifier.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnfulfilledNodeInputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unfulfilled node input with no valid connections.

Contents

input

The name of the unfulfilled input. An input is unfulfilled if there are no data connections to it.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the node containing the unfulfilled input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownConnectionConditionFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown condition for a connection.

Contents

connection

The name of the connection with the unknown condition.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownConnectionSourceFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown source node for a connection.

Contents

connection

The name of the connection with the unknown source.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownConnectionSourceOutputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown source output for a connection.

Contents

connection

The name of the connection with the unknown source output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownConnectionTargetFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown target node for a connection.

Contents

connection

The name of the connection with the unknown target.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownConnectionTargetInputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown target input for a connection.

Contents

connection

The name of the connection with the unknown target input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownNodeInputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown input for a node.

Contents

input

The name of the node with the unknown input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

node

The name of the unknown input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnknownNodeOutputFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unknown output for a node.

Contents

node

The name of the node with the unknown output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

output

The name of the unknown output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnreachableNodeFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unreachable node in the flow. A node is unreachable when there are no paths to it from any starting node.

Contents

node

The name of the unreachable node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnsatisfiedConnectionConditionsFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about unsatisfied conditions for a connection. A condition is unsatisfied if it can never be true, for example two branches of condition node cannot be simultaneously true.

Contents

connection

The name of the connection with unsatisfied conditions.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,100}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UnspecifiedFlowValidationDetails

Service: Agents for Amazon Bedrock

Details about an unspecified validation that doesn't fit other categories.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UrlConfiguration

Service: Agents for Amazon Bedrock

The configuration of web URLs that you want to crawl. You should be authorized to crawl the URLs.

Contents

seedUrls

One or more seed or starting point URLs.

Type: Array of SeedUrl objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidationExceptionField

Service: Agents for Amazon Bedrock

Stores information about a field passed inside a request that resulted in an validation error.

Contents

message

A message describing why this field failed validation.

Type: String

Pattern: `[\s\S]+`

Required: Yes

name

The name of the field.

Type: String

Pattern: `[\s\S]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorIngestionConfiguration

Service: Agents for Amazon Bedrock

Contains details about how to ingest the documents in a data source.

Contents

chunkingConfiguration

Details about how to chunk the documents in the data source. A *chunk* refers to an excerpt from a data source that is returned when the knowledge base that it belongs to is queried.

Type: ChunkingConfiguration object

Required: No

contextEnrichmentConfiguration

The context enrichment configuration used for ingestion of the data into the vector store.

Type: ContextEnrichmentConfiguration object

Required: No

customTransformationConfiguration

A custom document transformer for parsed data source documents.

Type: CustomTransformationConfiguration object

Required: No

parsingConfiguration

Configurations for a parser to use for parsing documents in your data source. If you exclude this field, the default parser will be used.

Type: ParsingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorKnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock

Contains details about the model used to create vector embeddings for the knowledge base.

Contents

embeddingModelArn

The Amazon Resource Name (ARN) of the model used to create vector embeddings for the knowledge base.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]{1,12})?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-zA-Z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

embeddingModelConfiguration

The embeddings model configuration details for the vector model used in Knowledge Base.

Type: EmbeddingModelConfiguration object

Required: No

supplementalDataStorageConfiguration

If you include multimodal data from your data source, use this object to specify configurations for the storage location of the images extracted from your documents. These images can be retrieved and returned to the end user. They can also be used in generation when using RetrieveAndGenerate.

Type: SupplementalDataStorageConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchBedrockRerankingConfiguration

Service: Agents for Amazon Bedrock

Configures the Amazon Bedrock reranker model to improve the relevance of retrieved results.

Contents

modelConfiguration

Specifies the configuration for the Amazon Bedrock reranker model.

Type: VectorSearchBedrockRerankingModelConfiguration object

Required: Yes

metadataConfiguration

Specifies how metadata fields should be handled during the reranking process.

Type: MetadataConfigurationForReranking object

Required: No

numberOfRerankedResults

Specifies the number of results to return after reranking.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchBedrockRerankingModelConfiguration

Service: Agents for Amazon Bedrock

Configures the Amazon Bedrock model used for reranking retrieved results.

Contents

modelArn

The Amazon Resource Name (ARN) of the Amazon Bedrock reranker model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/(.*))?`

Required: Yes

additionalModelRequestFields

Specifies additional model-specific request parameters as key-value pairs that are included in the request to the Amazon Bedrock reranker model.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchRerankingConfiguration

Service: Agents for Amazon Bedrock

Specifies how retrieved results from a knowledge base are reranked to improve relevance.

Contents

type

Specifies the type of reranking model to use. Currently, the only supported value is `BEDROCK_RERANKING_MODEL`.

Type: String

Valid Values: `BEDROCK_RERANKING_MODEL`

Required: Yes

bedrockRerankingConfiguration

Specifies the configuration for using an Amazon Bedrock reranker model to rerank retrieved results.

Type: VectorSearchBedrockRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoConfiguration

Service: Agents for Amazon Bedrock

Configuration settings for processing video content in multimodal knowledge bases.

Contents

segmentationConfiguration

Configuration for segmenting video content during processing.

Type: VideoSegmentationConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoExtractionConfiguration

Service: Agents for Amazon Bedrock

Configuration for video extraction.

Contents

videoExtractionStatus

Whether video extraction is enabled or disabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoSegmentationConfiguration

Service: Agents for Amazon Bedrock

Configuration for segmenting video content during multimodal knowledge base ingestion. Determines how video files are divided into chunks for processing.

Contents

fixedLengthDuration

The duration in seconds for each video segment. Video files will be divided into chunks of this length for processing.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 30.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VpcConfiguration

Service: Agents for Amazon Bedrock

Contains the details of a VPC configuration, including its connection settings, resolution mode, and current lifecycle status.

Contents

createdAt

The time at which the VPC configuration was created.

Type: Timestamp

Required: Yes

port

The port on which the resource is reached.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 65535.

Required: Yes

protocol

The protocol used to connect to the resource.

Type: String

Valid Values: `HTTP | HTTPS`

Required: Yes

resolutionMode

Specifies how the resource target is resolved.

Type: String

Valid Values: `PUBLIC | IN_VPC`

Required: Yes

resourceTarget

The private IPv4 address or DNS name of the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: Yes

status

The current lifecycle status of the VPC configuration.

Type: String

Valid Values: `CREATING | CREATED | DELETING | CREATE_FAILED | DELETE_FAILED`

Required: Yes

subnetIds

The subnets that the knowledge base uses to connect to the resource.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 6 items.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `subnet-[a-zA-Z0-9]+`

Required: Yes

updatedAt

The time at which the VPC configuration was last updated.

Type: Timestamp

Required: Yes

vpcConfigurationId

The unique identifier of the VPC configuration.

Type: String

Length Constraints: Fixed length of 32.

Pattern: `[a-z0-9](?:[a-z0-9-]{30}[a-z0-9])`

Required: Yes

vpcId

The identifier of the VPC that the knowledge base connects through to reach the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `vpc-[a-zA-Z0-9]+`

Required: Yes

description

The description of the VPC configuration, if provided.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `[^\p{C}]*`

Required: No

hostHeader

The HTTP `Host` header value sent when invoking the resource, if configured.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[A-Za-z0-9._:\[\]-]{1,255}`

Required: No

name

The human-readable name of the VPC configuration, if provided.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9 _-]*[a-zA-Z0-9])?`

Required: No

statusMessage

Additional detail about the current status, such as the cause of a `CREATE_FAILED` or `DELETE_FAILED` status.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: No

tlsServerName

The expected TLS server name that the service matches against the Subject Alternative Names on the resource's TLS certificate. Present when `protocol` is `HTTPS`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 253.

Pattern: `(\*\.)?([A-Za-z0-9]([A-Za-z0-9-]{0,61}[A-Za-z0-9])?\.)*[A-Za-z0-9]([A-Za-z0-9-]{0,61}[A-Za-z0-9])?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VpcConfigurationSummary

Service: Agents for Amazon Bedrock

A summary of a VPC configuration returned by `ListVpcConfigurations`.

Contents

createdAt

The time at which the VPC configuration was created.

Type: Timestamp

Required: Yes

port

The port on which the resource is reached.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 65535.

Required: Yes

protocol

The protocol used to connect to the resource.

Type: String

Valid Values: `HTTP | HTTPS`

Required: Yes

resolutionMode

Specifies how the resource target is resolved.

Type: String

Valid Values: `PUBLIC | IN_VPC`

Required: Yes

resourceTarget

The private IPv4 address or DNS name of the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Required: Yes

status

The current lifecycle status of the VPC configuration.

Type: String

Valid Values: `CREATING | CREATED | DELETING | CREATE_FAILED | DELETE_FAILED`

Required: Yes

vpcConfigurationId

The unique identifier of the VPC configuration.

Type: String

Length Constraints: Fixed length of 32.

Pattern: `[a-z0-9](?:[a-z0-9-]{30}[a-z0-9])`

Required: Yes

vpcId

The identifier of the VPC that the knowledge base connects through to reach the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `vpc-[a-zA-Z0-9]+`

Required: Yes

description

The description of the VPC configuration, if provided.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 512.

Pattern: `[^\p{C}]*`

Required: No

hostHeader

The HTTP `Host` header value sent when invoking the resource, if configured.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[A-Za-z0-9._:\[\]-]{1,255}`

Required: No

name

The human-readable name of the VPC configuration, if provided.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9]([a-zA-Z0-9 _-]*[a-zA-Z0-9])?`

Required: No

statusMessage

Additional detail about the current status, such as the cause of a failure.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Required: No

tlsServerName

The expected TLS server name that the service matches against the Subject Alternative Names on the resource's TLS certificate. Present when `protocol` is `HTTPS`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 253.

Pattern: `(\*\.)?([A-Za-z0-9]([A-Za-z0-9-]{0,61}[A-Za-z0-9])?\.)*[A-Za-z0-9]([A-Za-z0-9-]{0,61}[A-Za-z0-9])?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WebCrawlerConfiguration

Service: Agents for Amazon Bedrock

The configuration of web URLs that you want to crawl. You should be authorized to crawl the URLs.

Contents

crawlerLimits

The configuration of crawl limits for the web URLs.

Type: WebCrawlerLimits object

Required: No

exclusionFilters

A list of one or more exclusion regular expression patterns to exclude certain URLs. If you specify an inclusion and exclusion filter/pattern and both match a URL, the exclusion filter takes precedence and the web content of the URL isn't crawled.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

inclusionFilters

A list of one or more inclusion regular expression patterns to include certain URLs. If you specify an inclusion and exclusion filter/pattern and both match a URL, the exclusion filter takes precedence and the web content of the URL isn't crawled.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

scope

The scope of what is crawled for your URLs.

You can choose to crawl only web pages that belong to the same host or primary domain. For example, only web pages that contain the seed URL "https://docs.aws.amazon.com/bedrock/latest/userguide/" and no other domains. You can choose to include sub domains in addition to the host or primary domain. For example, web pages that contain "aws.amazon.com" can also include sub domain "docs.aws.amazon.com".

Type: String

Valid Values: `HOST_ONLY | SUBDOMAINS`

Required: No

userAgent

Returns the user agent suffix for your web crawler.

Type: String

Length Constraints: Minimum length of 15. Maximum length of 40.

Required: No

userAgentHeader

A string used for identifying the crawler or bot when it accesses a web server. The user agent header value consists of the `bedrockbot`, UUID, and a user agent suffix for your crawler (if one is provided). By default, it is set to `bedrockbot_UUID`. You can optionally append a custom suffix to `bedrockbot_UUID` to allowlist a specific user agent permitted to access your source URLs.

Type: String

Length Constraints: Minimum length of 61. Maximum length of 86.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

WebCrawlerLimits

Service: Agents for Amazon Bedrock

The rate limits for the URLs that you want to crawl. You should be authorized to crawl the URLs.

Contents

maxPages

The max number of web pages crawled from your source URLs, up to 25,000 pages. If the web pages exceed this limit, the data source sync will fail and no web pages will be ingested.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

rateLimit

The max rate at which pages are crawled, up to 300 per minute per host.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 300.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WebDataSourceConfiguration

Service: Agents for Amazon Bedrock

The configuration details for the web data source.

Contents

sourceConfiguration

The source configuration details for the web data source.

Type: WebSourceConfiguration object

Required: Yes

crawlerConfiguration

The Web Crawler configuration details for the web data source.

Type: WebCrawlerConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WebSourceConfiguration

Service: Agents for Amazon Bedrock

The configuration of the URL/URLs for the web content that you want to crawl. You should be authorized to crawl the URLs.

Contents

urlConfiguration

The configuration of the URL/URLs.

Type: UrlConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

WeeklySchedule

Service: Agents for Amazon Bedrock

A weekly sync on a specified day of the week.

Contents

dayOfWeek

The day of the week on which the weekly sync runs.

Type: String

Valid Values: `SUNDAY | MONDAY | TUESDAY | WEDNESDAY | THURSDAY | FRIDAY | SATURDAY`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Agents for Amazon Bedrock Runtime

The following data types are supported by Agents for Amazon Bedrock Runtime:

- ActionGroupExecutor
- ActionGroupInvocationInput
- ActionGroupInvocationOutput
- AgentActionGroup
- AgentCollaboratorInputPayload
- AgentCollaboratorInvocationInput

- AgentCollaboratorInvocationOutput
- AgentCollaboratorOutputPayload
- AgenticRetrieveAction
- AgenticRetrieveActionDetails
- AgenticRetrieveBedrockGuardrailConfiguration
- AgenticRetrieveBedrockRerankingConfiguration
- AgenticRetrieveBedrockRerankingModelConfiguration
- AgenticRetrieveCitation
- AgenticRetrieveCitationReference
- AgenticRetrieveConfiguration
- AgenticRetrieveFailure
- AgenticRetrieveFullDocExpansionDetails
- AgenticRetrieveGeneratedResponse
- AgenticRetrieveGuardrailWarning
- AgenticRetrieveMemoryConfiguration
- AgenticRetrieveMemoryMetadataFilter
- AgenticRetrieveMemoryMetadataFilterLeft
- AgenticRetrieveMemoryMetadataFilterRight
- AgenticRetrieveMemoryMetadataValue
- AgenticRetrieveMemoryRetrievalConfig
- AgenticRetrieveMemoryRetrieveDetails
- AgenticRetrieveMemorySessionBinding
- AgenticRetrieveMessage
- AgenticRetrieveMessageContent
- AgenticRetrievePolicyConfiguration
- AgenticRetriever
- AgenticRetrieveRerankingConfiguration
- AgenticRetrieveResponseEvent
- AgenticRetrieveResultEvent
- AgenticRetrieveResultItem

- AgenticRetrieveSourceMetadata
- AgenticRetrieveSourceRetriever
- AgenticRetrieveStreamResponseOutput
- AgenticRetrieveTraceEvent
- AgenticRetrieveTraceEventAttributes
- AgenticRetrieveTraceResultItem
- AgenticRetrieveWarning
- AgenticRetrieveWarningMessage
- AnalyzePromptEvent
- ApiInvocationInput
- ApiParameter
- ApiRequestBody
- ApiResult
- APISchema
- Attribution
- AudioSegment
- BedrockFoundationModelConfiguration
- BedrockFoundationModelModelConfiguration
- BedrockModelConfigurations
- BedrockRerankingConfiguration
- BedrockRerankingModelConfiguration
- BedrockSessionContentBlock
- ByteContentDoc
- ByteContentFile
- Caller
- Citation
- CitationEvent
- CodeInterpreterInvocationInput
- CodeInterpreterInvocationOutput
- Collaborator

- CollaboratorConfiguration
- ConditionResultEvent
- ContentBlock
- ContentBody
- ConversationHistory
- CustomOrchestration
- CustomOrchestrationTrace
- CustomOrchestrationTraceEvent
- DocumentAcl
- DocumentAclCondition
- DocumentAclGroup
- DocumentAclMembership
- DocumentAclUser
- ExternalSource
- ExternalSourcesGenerationConfiguration
- ExternalSourcesRetrieveAndGenerateConfiguration
- FailureTrace
- FieldForReranking
- FilePart
- FileSource
- FilterAttribute
- FinalResponse
- FlowCompletionEvent
- FlowExecutionContent
- FlowExecutionError
- FlowExecutionEvent
- FlowExecutionInputEvent
- FlowExecutionOutputEvent
- FlowExecutionSummary
- FlowFailureEvent

- FlowInput
- FlowInputContent
- FlowInputField
- FlowMultiTurnInputContent
- FlowMultiTurnInputRequestEvent
- FlowOutputContent
- FlowOutputEvent
- FlowOutputField
- FlowResponseStream
- FlowTrace
- FlowTraceCondition
- FlowTraceConditionNodeResultEvent
- FlowTraceDependencyEvent
- FlowTraceEvent
- FlowTraceNodeActionEvent
- FlowTraceNodeInputContent
- FlowTraceNodeInputEvent
- FlowTraceNodeInputExecutionChainItem
- FlowTraceNodeInputField
- FlowTraceNodeInputSource
- FlowTraceNodeOutputContent
- FlowTraceNodeOutputEvent
- FlowTraceNodeOutputField
- FlowTraceNodeOutputNext
- FoundationModelConfiguration
- FunctionDefinition
- FunctionInvocationInput
- FunctionParameter
- FunctionResult
- FunctionSchema

- GeneratedQuery
- GeneratedResponsePart
- GenerationConfiguration
- GuardrailAssessment
- GuardrailConfiguration
- GuardrailConfigurationWithArn
- GuardrailContentFilter
- GuardrailContentPolicyAssessment
- GuardrailCustomWord
- GuardrailEvent
- GuardrailManagedWord
- GuardrailPiiEntityFilter
- GuardrailRegexFilter
- GuardrailSensitiveInformationPolicyAssessment
- GuardrailTopic
- GuardrailTopicPolicyAssessment
- GuardrailTrace
- GuardrailWordPolicyAssessment
- ImageBlock
- ImageInput
- ImageInputSource
- ImageSource
- ImplicitFilterConfiguration
- InferenceConfig
- InferenceConfiguration
- InlineAgentFilePart
- InlineAgentPayloadPart
- InlineAgentResponseStream
- InlineAgentReturnControlPayload
- InlineAgentTracePart

- InlineBedrockModelConfigurations
- InlineSessionState
- InputFile
- InputImage
- InputPrompt
- InvocationInput
- InvocationInputMember
- InvocationResultMember
- InvocationStep
- InvocationStepPayload
- InvocationStepSummary
- InvocationSummary
- KnowledgeBase
- KnowledgeBaseConfiguration
- KnowledgeBaseLookupInput
- KnowledgeBaseLookupOutput
- KnowledgeBaseQuery
- KnowledgeBaseRetrievalConfiguration
- KnowledgeBaseRetrievalResult
- KnowledgeBaseRetrieveAndGenerateConfiguration
- KnowledgeBaseRetrieverConfiguration
- KnowledgeBaseVectorSearchConfiguration
- ManagedSearchBedrockRerankingConfiguration
- ManagedSearchBedrockRerankingModelConfiguration
- ManagedSearchConfiguration
- ManagedSearchRerankingConfiguration
- Memory
- MemorySessionSummary
- Message
- Metadata

- MetadataAttributeSchema
- MetadataConfigurationForReranking
- ModelInvocationInput
- ModelPerformanceConfiguration
- NodeActionEvent
- NodeDependencyEvent
- NodeExecutionContent
- NodeFailureEvent
- NodeInputEvent
- NodeInputExecutionChainItem
- NodeInputField
- NodeInputSource
- NodeOutputEvent
- NodeOutputField
- NodeOutputNext
- NodeTraceElements
- Observation
- OptimizedPrompt
- OptimizedPromptEvent
- OptimizedPromptStream
- OrchestrationConfiguration
- OrchestrationExecutor
- OrchestrationModelInvocationOutput
- OrchestrationTrace
- OutputFile
- Parameter
- ParameterDetail
- PayloadPart
- PerformanceConfiguration
- PostProcessingModelInvocationOutput

- PostProcessingParsedResponse
- PostProcessingTrace
- PreProcessingModelInvocationOutput
- PreProcessingParsedResponse
- PreProcessingTrace
- PromptConfiguration
- PromptCreationConfigurations
- PromptOverrideConfiguration
- PromptTemplate
- PropertyParameters
- QueryGenerationInput
- QueryTransformationConfiguration
- Rationale
- RawResponse
- ReasoningContentBlock
- ReasoningTextBlock
- RepromptResponse
- RequestBody
- RerankDocument
- RerankingConfiguration
- RerankingMetadataSelectiveModeConfiguration
- RerankQuery
- RerankResult
- RerankSource
- RerankTextDocument
- ResponseStream
- RetrievalContent
- RetrievalFilter
- RetrievalOverrides
- RetrievalResultConfluenceLocation

- RetrievalResultContent
- RetrievalResultContentColumn
- RetrievalResultCustomDocumentLocation
- RetrievalResultGoogleDriveLocation
- RetrievalResultKendraDocumentLocation
- RetrievalResultLocation
- RetrievalResultOneDriveLocation
- RetrievalResultS3Location
- RetrievalResultSalesforceLocation
- RetrievalResultSharePointLocation
- RetrievalResultSqlLocation
- RetrievalResultWebLocation
- RetrieveAndGenerateConfiguration
- RetrieveAndGenerateInput
- RetrieveAndGenerateOutput
- RetrieveAndGenerateOutputEvent
- RetrieveAndGenerateSessionConfiguration
- RetrieveAndGenerateStreamResponseOutput
- RetrievedReference
- RetrieverConfiguration
- ReturnControlPayload
- ReturnControlResults
- RoutingClassifierModelInvocationOutput
- RoutingClassifierTrace
- S3Identifier
- S3Location
- S3ObjectDoc
- S3ObjectFile
- SatisfiedCondition
- SessionState

- SessionSummary
- Span
- StreamingConfigurations
- TextInferenceConfig
- TextPrompt
- TextResponsePart
- TextToSqlConfiguration
- TextToSqlKnowledgeBaseConfiguration
- Trace
- TraceElements
- TracePart
- TransformationConfiguration
- Usage
- UserContext
- VectorSearchBedrockRerankingConfiguration
- VectorSearchBedrockRerankingModelConfiguration
- VectorSearchRerankingConfiguration
- VideoSegment

ActionGroupExecutor

Service: Agents for Amazon Bedrock Runtime

Contains details about the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

customControl

To return the action group invocation results directly in the `InvokeInlineAgent` response, specify `RETURN_CONTROL`.

Type: String

Valid Values: `RETURN_CONTROL`

Required: No

lambda

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ActionGroupInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the action group being invoked. For more information about the possible structures, see the InvocationInput tab in OrchestrationTrace in the Amazon Bedrock User Guide.

Contents

actionGroupName

The name of the action group.

Type: String

Required: No

apiPath

The path to the API to call, based off the action group.

Type: String

Required: No

executionType

How fulfillment of the action is handled. For more information, see Handling fulfillment of the action.

Type: String

Valid Values: `LAMBDA | RETURN_CONTROL`

Required: No

function

The function in the action group to call.

Type: String

Required: No

invocationId

The unique identifier of the invocation. Only returned if the `executionType` is `RETURN_CONTROL`.

Type: String

Required: No

parameters

The parameters in the Lambda input event.

Type: Array of Parameter objects

Required: No

requestBody

The parameters in the request body for the Lambda input event.

Type: RequestBody object

Required: No

verb

The API method being used, based off the action group.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ActionGroupInvocationOutput

Service: Agents for Amazon Bedrock Runtime

Contains the JSON-formatted string returned by the API invoked by the action group.

Contents

metadata

Contains information about the action group output.

Type: Metadata object

Required: No

text

The JSON-formatted string returned by the API invoked by the action group.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentActionGroup

Service: Agents for Amazon Bedrock Runtime

Contains details of the inline agent's action group.

Contents

actionGroupName

The name of the action group.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

actionGroupExecutor

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action or the custom control method for handling the information elicited from the user.

Type: ActionGroupExecutor object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

apiSchema

Contains either details about the S3 object containing the OpenAPI schema for the action group or the JSON or YAML-formatted payload defining the schema. For more information, see Action group OpenAPI schemas.

Type: APISchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

description

A description of the action group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: No

functionSchema

Contains details about the function schema for the action group or the JSON or YAML-formatted payload defining the schema.

Type: FunctionSchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

parentActionGroupSignature

Specify a built-in or computer use action for this action group. If you specify a value, you must leave the `description`, `apiSchema`, and `actionGroupExecutor` fields empty for this action group.

- To allow your agent to request the user for additional information when trying to complete a task, set this field to `AMAZON.UserInput`.
- To allow your agent to generate, run, and troubleshoot code when trying to complete a task, set this field to `AMAZON.CodeInterpreter`.
- To allow your agent to use an Anthropic computer use tool, specify one of the following values.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Anthropic Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. When operating computer use functionality, we recommend taking additional security precautions, such as executing computer actions in virtual environments with restricted data access and limited internet connectivity. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

  - `ANTHROPIC.Computer` - Gives the agent permission to use the mouse and keyboard and take screenshots.

- `ANTHROPIC.TextEditor` - Gives the agent permission to view, create and edit files.
- `ANTHROPIC.Bash` - Gives the agent permission to run commands in a bash shell.

Type: String

Valid Values: `AMAZON.UserInput | AMAZON.CodeInterpreter | ANTHROPIC.Computer | ANTHROPIC.Bash | ANTHROPIC.TextEditor`

Required: No

parentActionGroupSignatureParams

The configuration settings for a computer use action.

> **Important**
>
> Computer use is a new Anthropic Claude model capability (in beta) available with Claude 3.7 Sonnet and Claude 3.5 Sonnet v2 only. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Type: String to string map

Key Length Constraints: Minimum length of 0. Maximum length of 100.

Value Length Constraints: Minimum length of 0. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentCollaboratorInputPayload

Service: Agents for Amazon Bedrock Runtime

Input for an agent collaborator. The input can be text or an action invocation result.

Contents

returnControlResults

An action invocation result.

Type: ReturnControlResults object

Required: No

text

Input text.

Type: String

Required: No

type

The input type.

Type: String

Valid Values: `TEXT | RETURN_CONTROL`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentCollaboratorInvocationInput

Service: Agents for Amazon Bedrock Runtime

An agent collaborator invocation input.

Contents

agentCollaboratorAliasArn

The collaborator's alias ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: No

agentCollaboratorName

The collaborator's name.

Type: String

Required: No

input

Text or action invocation result input for the collaborator.

Type: AgentCollaboratorInputPayload object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AgentCollaboratorInvocationOutput

Service: Agents for Amazon Bedrock Runtime

Output from an agent collaborator.

Contents

agentCollaboratorAliasArn

The output's agent collaborator alias ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: No

agentCollaboratorName

The output's agent collaborator name.

Type: String

Required: No

metadata

Contains information about the output from the agent collaborator.

Type: Metadata object

Required: No

output

The output's output.

Type: AgentCollaboratorOutputPayload object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgentCollaboratorOutputPayload

Service: Agents for Amazon Bedrock Runtime

Output from an agent collaborator. The output can be text or an action invocation result.

Contents

returnControlPayload

An action invocation result.

Type: ReturnControlPayload object

Required: No

text

Text output.

Type: String

Required: No

type

The type of output.

Type: String

Valid Values: `TEXT | RETURN_CONTROL`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveAction

Service: Agents for Amazon Bedrock Runtime

An action taken during agentic retrieval.

Contents

fullDocumentExpansion

Details of a full document expansion action.

Type: AgenticRetrieveFullDocExpansionDetails object

Required: No

memoryRetrieve

The details of a long-term memory retrieval that the agent chose to perform.

Type: AgenticRetrieveMemoryRetrieveDetails object

Required: No

retrieve

Details of the retrieve action.

Type: AgenticRetrieveActionDetails object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveActionDetails

Service: Agents for Amazon Bedrock Runtime

Details of a retrieve action including the query and target retrievers.

Contents

inputQuery

The input query used for retrieval.

Type: AgenticRetrieveMessageContent object

Required: Yes

sourceRetrievers

The list of source retrievers targeted by this action.

Type: Array of AgenticRetrieveSourceRetriever objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveBedrockGuardrailConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for a Bedrock guardrail applied during agentic retrieval.

Contents

guardrailId

The unique identifier of the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

Required: Yes

guardrailVersion

The version of the guardrail to use.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(([1-9][0-9]{0,7})|(DRAFT))`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveBedrockRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for a Bedrock reranking model.

Contents

modelConfiguration

The model configuration containing the model ARN.

Type: AgenticRetrieveBedrockRerankingModelConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveBedrockRerankingModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Model configuration for a Bedrock reranking model.

Contents

modelArn

The ARN of the Bedrock reranking model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveCitation

Service: Agents for Amazon Bedrock Runtime

A citation mapping a span of the generated answer to supporting results.

Contents

endIndex

Character offset end (exclusive) in the answer text.

Type: Integer

Required: Yes

references

References to results that support this span.

Type: Array of AgenticRetrieveCitationReference objects

Required: Yes

startIndex

Character offset start in the answer text.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveCitationReference

Service: Agents for Amazon Bedrock Runtime

A reference to a specific result item.

Contents

resultIndex

Index into the results array on the same event.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration settings for the agentic retrieval operation.

Contents

foundationModelConfiguration

The foundation model configuration. Required when foundationModelType is CUSTOM.

Type: FoundationModelConfiguration object

Required: No

foundationModelType

The type of foundation model to use. CUSTOM uses a specified model, MANAGED uses the service default.

Type: String

Valid Values: `CUSTOM | MANAGED`

Required: No

maxAgentIteration

The maximum number of agent iterations for retrieval.

Type: Integer

Valid Range: Minimum value of 2.

Required: No

rerankingConfiguration

The reranking model configuration. Required when rerankingModelType is CUSTOM.

Type: AgenticRetrieveRerankingConfiguration object

Required: No

rerankingModelType

The type of reranking model to use. CUSTOM uses a specified model, MANAGED uses the service default. If not specified, defaults to MANAGED for managed embedding knowledge bases and NONE for custom embedding knowledge bases.

Type: String

Valid Values: `CUSTOM | MANAGED | NONE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveFailure

Service: Agents for Amazon Bedrock Runtime

A failure that occurred during agentic retrieval.

Contents

message

A message describing the failure.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveFullDocExpansionDetails

Service: Agents for Amazon Bedrock Runtime

Details of a full document expansion action.

Contents

documentId

The identifier of the document to expand.

Type: String

Required: No

sourceRetriever

The source retriever associated with the document.

Type: AgenticRetrieveSourceRetriever object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveGeneratedResponse

Service: Agents for Amazon Bedrock Runtime

The generated response synthesized from retrieved results.

Contents

answer

The generated answer text.

Type: String

Required: Yes

citations

Citations mapping spans of the answer to supporting results.

Type: Array of AgenticRetrieveCitation objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveGuardrailWarning

Service: Agents for Amazon Bedrock Runtime

A warning generated by a guardrail during agentic retrieval.

Contents

action

The action taken by the guardrail.

Type: String

Valid Values: `INTERVENED | NONE`

Required: Yes

id

The unique identifier of the guardrail.

Type: String

Required: Yes

version

The version of the guardrail.

Type: String

Required: Yes

message

A message describing the guardrail evaluation result.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryConfiguration

Service: Agents for Amazon Bedrock Runtime

Specifies an AgentCore Memory resource and how this retrieval uses it. Set sessionBinding to restore and continue a session. Set retrievalConfigs to let the agent retrieve from long-term memory. You must specify at least one of the two.

Contents

memoryId

The identifier of the AgentCore Memory resource to use. The resource must exist in your account and be in the ACTIVE state.

Type: String

Length Constraints: Minimum length of 12. Maximum length of 111.

Pattern: `[a-zA-Z][a-zA-Z0-9\-_]{0,99}-[a-zA-Z0-9]{10}`

Required: Yes

persistenceMode

Specifies whether the agent-generated answer is written back to the given short-term memory session, and applies only when sessionBinding is set. Valid values:

- `DEFAULT` (default) – Specifies that the question and the agent-generated answer are persisted to the session as a single event. This value requires generateResponse to be true.
- `NONE` – Specifies that the session is left unchanged.

Type: String

Valid Values: `DEFAULT | NONE`

Required: No

retrievalConfigs

Specifies the long-term memory configuration the agent can retrieve from. The agent decides whether to retrieve and composes its own query. This field currently accepts at most one entry.

Type: Array of AgenticRetrieveMemoryRetrievalConfig objects

Required: No

sessionBinding

The short-term memory session whose history is restored for this retrieval. To persist the agent-generated answer to the session, omit persistenceMode or set it to DEFAULT. To leave the session unchanged, set persistenceMode to NONE. Supply session history through the existing messages parameter or through short-term memory, but not both.

Type: AgenticRetrieveMemorySessionBinding object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryMetadataFilter

Service: Agents for Amazon Bedrock Runtime

A metadata filter expression, in the form accepted by the AgentCore Memory RetrieveMemoryRecords operation. The expression has a left operand that names the metadata key, an operator, and a right operand. For the EXISTS and NOT_EXISTS operators, omit the right operand.

Contents

left

The metadata key that the expression evaluates.

Type: AgenticRetrieveMemoryMetadataFilterLeft object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

operator

The relationship that the metadata key and value must have for a memory record to match.

Type: String

Valid Values: `EQUALS_TO | EXISTS | NOT_EXISTS | BEFORE | AFTER | CONTAINS | GREATER_THAN | GREATER_THAN_OR_EQUALS | LESS_THAN | LESS_THAN_OR_EQUALS`

Required: Yes

right

The value that the expression compares the metadata key against. Supply this value for every operator except EXISTS and NOT_EXISTS.

Type: AgenticRetrieveMemoryMetadataFilterRight object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryMetadataFilterLeft

Service: Agents for Amazon Bedrock Runtime

The left operand of a metadata filter expression. Set exactly one member.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

metadataKey

The metadata key to filter on.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryMetadataFilterRight

Service: Agents for Amazon Bedrock Runtime

The right operand of a metadata filter expression. Set exactly one member.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

metadataValue

The value to compare the metadata key against.

Type: AgenticRetrieveMemoryMetadataValue object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryMetadataValue

Service: Agents for Amazon Bedrock Runtime

A metadata value that a filter expression compares against. Set exactly one member.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

dateTimeValue

A timestamp value in ISO 8601 UTC format.

Type: Timestamp

Required: No

numberValue

A numeric value.

Type: Double

Required: No

stringListValue

A list of string values.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

stringValue

A string value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryRetrievalConfig

Service: Agents for Amazon Bedrock Runtime

The long-term memory namespace that the agent might retrieve memory records from, and the filters applied to that retrieval. You must specify either namespace or namespacePath.

Contents

metadataFilters

The metadata filter expressions that restrict retrieval to matching memory records. You can specify a maximum of 5 expressions.

Type: Array of AgenticRetrieveMemoryMetadataFilter objects

Required: No

namespace

The namespace prefix to filter memory records by. The agent retrieves memory records in namespaces that start with the provided prefix. You must specify either namespace or namespacePath.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[a-zA-Z0-9/*][a-zA-Z0-9\-_/*]*(?::[a-zA-Z0-9\-_/*]+)*[a-zA-Z0-9\-_/*]*`

Required: No

namespacePath

The parent namespace to use for hierarchical retrievals. The agent retrieves all memory records whose namespace falls under the same parent hierarchy. You must specify either namespace or namespacePath.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[a-zA-Z0-9/*][a-zA-Z0-9\-_/*]*(?::[a-zA-Z0-9\-_/*]+)*[a-zA-Z0-9\-_/*]*`

Required: No

strategyId

The extraction strategy ID that restricts retrieval to memory records produced by a single strategy. Omit this parameter to retrieve records from every strategy on the memory resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-_]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemoryRetrieveDetails

Service: Agents for Amazon Bedrock Runtime

A long-term memory retrieval that the agent chose to perform. The record reports the query and the namespace. The corresponding Retrieval step reports the results.

Contents

inputQuery

The query that the agent composed.

Type: AgenticRetrieveMessageContent object

Required: Yes

memoryId

The identifier of the AgentCore Memory resource retrieved from.

Type: String

Required: Yes

namespace

The namespace prefix retrieved from, as supplied in the request. This field is present when the request specified namespace.

Type: String

Required: No

namespacePath

The parent namespace retrieved from hierarchically, as supplied in the request. This field is present when the request specified namespacePath.

Type: String

Required: No

strategyId

The extraction strategy that restricted retrieval, if the request specified one.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMemorySessionBinding

Service: Agents for Amazon Bedrock Runtime

The short-term memory session that this retrieval reads from and writes to.

Contents

actorId

The identifier of the end user or agent that the session belongs to. This identifier scopes session history so that one actor's history is never returned for another. You are responsible for sending the correct actor value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 255.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-_/]*(?::[a-zA-Z0-9\-_/]+)*[a-zA-Z0-9\-_/]*`

Required: Yes

sessionId

The identifier of the session to restore and continue. You are responsible for sending the correct session value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Pattern: `[a-zA-Z0-9][a-zA-Z0-9\-_]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AgenticRetrieveMessage

Service: Agents for Amazon Bedrock Runtime

A message in the agentic retrieval conversation.

Contents

content

The content of the message.

Type: AgenticRetrieveMessageContent object

Required: Yes

role

The role of the message sender (e.g., user or assistant).

Type: String

Valid Values: `user | assistant`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveMessageContent

Service: Agents for Amazon Bedrock Runtime

The content of an agentic retrieval message.

Contents

text

The text content of the message.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrievePolicyConfiguration

Service: Agents for Amazon Bedrock Runtime

Policy configuration for agentic retrieval.

Contents

bedrockGuardrailConfiguration

Configuration for Bedrock guardrails to apply during retrieval.

Type: AgenticRetrieveBedrockGuardrailConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetriever

Service: Agents for Amazon Bedrock Runtime

A retriever used in agentic retrieval.

Contents

configuration

The configuration for this retriever.

Type: RetrieverConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

description

A description of the retriever's purpose.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for the reranking model.

Contents

type

The type of reranking configuration.

Type: String

Valid Values: `BEDROCK_RERANKING_MODEL`

Required: Yes

bedrockRerankingConfiguration

The Bedrock reranking model configuration.

Type: AgenticRetrieveBedrockRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveResponseEvent

Service: Agents for Amazon Bedrock Runtime

A chunk of the generated answer text.

Contents

text

The generated text chunk.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveResultEvent

Service: Agents for Amazon Bedrock Runtime

An event containing agentic retrieval results.

Contents

results

The list of retrieved result items.

Type: Array of AgenticRetrieveResultItem objects

Required: Yes

generatedResponse

The generated response. Present only when generateResponse is true.

Type: AgenticRetrieveGeneratedResponse object

Required: No

nextToken

Opaque continuation token for paginated results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: \S*

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveResultItem

Service: Agents for Amazon Bedrock Runtime

A single item from the agentic retrieval results.

Contents

content

The retrieved content.

Type: RetrievalContent object

Required: Yes

sourceRetriever

The source retriever that produced this result.

Type: AgenticRetrieveSourceRetriever object

Required: Yes

metadata

Metadata associated with the retrieved item.

Type: String to JSON value map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveSourceMetadata

Service: Agents for Amazon Bedrock Runtime

Metadata about a retrieval source.

Contents

identifier

The identifier of the retrieval source.

Type: String

Required: No

retrievalType

The type of retrieval source.

Type: String

Valid Values: `BedrockKnowledgeBase | BedrockAgentCoreMemory`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveSourceRetriever

Service: Agents for Amazon Bedrock Runtime

A source retriever that produced retrieval results.

Contents

identifier

The unique identifier of the source retriever.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveStreamResponseOutput

Service: Agents for Amazon Bedrock Runtime

The streaming output for agentic retrieval, containing results, traces, and errors.

Contents

accessDeniedException

Access to the resource was denied.

Type: Exception

HTTP Status Code: 403

Required: No

badGatewayException

A bad gateway error occurred.

Type: Exception

HTTP Status Code: 502

Required: No

conflictException

A conflict occurred with the current state of the resource.

Type: Exception

HTTP Status Code: 409

Required: No

dependencyFailedException

A dependency failed during the operation.

Type: Exception

HTTP Status Code: 424

Required: No

internalServerException

An internal server error occurred.

Type: Exception

HTTP Status Code: 500

Required: No

resourceNotFoundException

The specified resource was not found.

Type: Exception

HTTP Status Code: 404

Required: No

responseEvent

A chunk of the generated answer. Emitted only when generateResponse is true.

Type: AgenticRetrieveResponseEvent object

Required: No

result

A retrieval result event containing the retrieved items.

Type: AgenticRetrieveResultEvent object

Required: No

serviceQuotaExceededException

The service quota has been exceeded.

Type: Exception

HTTP Status Code: 400

Required: No

throttlingException

The request was throttled.

Type: Exception

HTTP Status Code: 429

Required: No

traceEvent

A trace event providing visibility into the retrieval process.

Type: AgenticRetrieveTraceEvent object

Required: No

validationException

The request validation failed.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveTraceEvent

Service: Agents for Amazon Bedrock Runtime

A trace event providing visibility into the agentic retrieval process.

Contents

attributes

The attributes describing the trace event details.

Type: AgenticRetrieveTraceEventAttributes object

Required: Yes

id

The unique identifier of the trace event.

Type: String

Required: Yes

timestamp

The timestamp when the trace event occurred.

Type: Long

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveTraceEventAttributes

Service: Agents for Amazon Bedrock Runtime

Attributes describing the details of an agentic retrieval trace event.

Contents

message

A human-readable message describing the trace event.

Type: String

Required: Yes

status

The status of the current step.

Type: String

Valid Values: `IN_PROGRESS | SUCCEEDED | FAILED`

Required: Yes

step

The current step in the retrieval process.

Type: String

Valid Values: `Planning | Retrieval | SpeculativeRetrieval | FullDocumentExpansion | SessionHistoryLoad`

Required: Yes

actions

The list of actions taken during this step.

Type: Array of AgenticRetrieveAction objects

Required: No

failures

Failures that occurred during this step.

Type: Array of AgenticRetrieveFailure objects

Array Members: Minimum number of 1 item.

Required: No

retrievalMetadata

Metadata about the retrieval sources used.

Type: Array of AgenticRetrieveSourceMetadata objects

Required: No

retrievalResponse

The retrieval results from this step.

Type: Array of AgenticRetrieveTraceResultItem objects

Required: No

warnings

Warnings generated during this step.

Type: Array of AgenticRetrieveWarning objects

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveTraceResultItem

Service: Agents for Amazon Bedrock Runtime

A result item from an agentic retrieval trace.

Contents

content

The retrieved content.

Type: RetrievalContent object

Required: No

metadata

Metadata associated with the retrieved item.

Type: String to JSON value map

Required: No

sourceRetriever

The source retriever that produced this result.

Type: AgenticRetrieveSourceRetriever object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveWarning

Service: Agents for Amazon Bedrock Runtime

A warning generated during agentic retrieval.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

guardrail

A warning from a guardrail evaluation.

Type: AgenticRetrieveGuardrailWarning object

Required: No

message

A general warning message.

Type: AgenticRetrieveWarningMessage object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AgenticRetrieveWarningMessage

Service: Agents for Amazon Bedrock Runtime

A general warning message from agentic retrieval.

Contents

message

The warning message text.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AnalyzePromptEvent

Service: Agents for Amazon Bedrock Runtime

An event in which the prompt was analyzed in preparation for optimization.

Contents

message

A message describing the analysis of the prompt.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ApiInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the API operation that the agent predicts should be called.

This data type is used in the following API operations:

- In the `returnControl` field of the InvokeAgent response

Contents

actionGroup

The action group that the API operation belongs to.

Type: String

Required: Yes

actionInvocationType

Contains information about the API operation to invoke.

Type: String

Valid Values: `RESULT | USER_CONFIRMATION | USER_CONFIRMATION_AND_RESULT`

Required: No

agentId

The agent's ID.

Type: String

Required: No

apiPath

The path to the API operation.

Type: String

Required: No

collaboratorName

The agent collaborator's name.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

httpMethod

The HTTP method of the API operation.

Type: String

Required: No

parameters

The parameters to provide for the API request, as the agent elicited from the user.

Type: Array of ApiParameter objects

Required: No

requestBody

The request body to provide for the API request, as the agent elicited from the user.

Type: ApiRequestBody object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ApiParameter

Service: Agents for Amazon Bedrock Runtime

Information about a parameter to provide to the API request.

This data type is used in the following API operations:

- InvokeAgent response

Contents

name

The name of the parameter.

Type: String

Required: No

type

The data type for the parameter.

Type: String

Required: No

value

The value of the parameter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ApiRequestBody

Service: Agents for Amazon Bedrock Runtime

The request body to provide for the API request, as the agent elicited from the user.

This data type is used in the following API operations:

- InvokeAgent response

Contents

content

The content of the request body. The key of the object in this field is a media type defining the format of the request body.

Type: String to PropertyParameters object map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ApiResult

Service: Agents for Amazon Bedrock Runtime

Contains information about the API operation that was called from the action group and the response body that was returned.

This data type is used in the following API operations:

- In the `returnControlInvocationResults` of the InvokeAgent request

Contents

actionGroup

The action group that the API operation belongs to.

Type: String

Required: Yes

agentId

The agent's ID.

Type: String

Required: No

apiPath

The path to the API operation.

Type: String

Required: No

confirmationState

Controls the API operations or functions to invoke based on the user confirmation.

Type: String

Valid Values: `CONFIRM | DENY`

Required: No

httpMethod

The HTTP method for the API operation.

Type: String

Required: No

httpStatusCode

http status code from API execution response (for example: 200, 400, 500).

Type: Integer

Required: No

responseBody

The response body from the API operation. The key of the object is the content type (currently, only TEXT is supported). The response may be returned directly or from the Lambda function.

Type: String to ContentBody object map

Required: No

responseState

Controls the final response state returned to end user when API/Function execution failed. When this state is FAILURE, the request would fail with dependency failure exception. When this state is REPROMPT, the API/function response will be sent to model for re-prompt

Type: String

Valid Values: `FAILURE | REPROMPT`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

APISchema

Service: Agents for Amazon Bedrock Runtime

Contains details about the OpenAPI schema for the action group. For more information, see Action group OpenAPI schemas. You can either include the schema directly in the payload field or you can upload it to an S3 bucket and specify the S3 bucket location in the s3 field.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

payload

The JSON or YAML-formatted payload defining the OpenAPI schema for the action group.

Type: String

Required: No

s3

Contains details about the S3 object containing the OpenAPI schema for the action group.

Type: S3Identifier object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Attribution

Service: Agents for Amazon Bedrock Runtime

Contains citations for a part of an agent response.

Contents

citations

A list of citations and related information for a part of an agent response.

Type: Array of Citation objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioSegment

Service: Agents for Amazon Bedrock Runtime

Contains information about an audio segment retrieved from a knowledge base, including its location and transcription.

This data type is used in the following API operations:

- Retrieve response – in the `audio` field

Contents

s3Uri

The S3 URI where this specific audio segment is stored in the multimodal storage destination.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

transcription

The text transcription of the audio segment content.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockFoundationModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for a Bedrock foundation model.

Contents

modelConfiguration

The model configuration containing the model ARN.

Type: BedrockFoundationModelModelConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockFoundationModelModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Model configuration for a Bedrock foundation model.

Contents

modelArn

The ARN of the Bedrock foundation model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockModelConfigurations

Service: Agents for Amazon Bedrock Runtime

Settings for a model called with InvokeAgent.

Contents

performanceConfig

The performance configuration for the model.

Type: PerformanceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for an Amazon Bedrock reranker model.

Contents

modelConfiguration

Contains configurations for a reranker model.

Type: BedrockRerankingModelConfiguration object

Required: Yes

numberOfResults

The number of results to return after reranking.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockRerankingModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for a reranker model.

Contents

modelArn

The ARN of the reranker model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

additionalModelRequestFields

A JSON object whose keys are request fields for the model and whose values are values for those fields.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BedrockSessionContentBlock

Service: Agents for Amazon Bedrock Runtime

A block of content that you pass to, or receive from, a Amazon Bedrock session in an invocation step. You pass the content to a session in the payLoad of the PutInvocationStep API operation. You retrieve the content with the GetInvocationStep API operation.

For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

image

The image in the invocation step.

Type: ImageBlock object

Required: No

text

The text in the invocation step.

Type: String

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ByteContentDoc

Service: Agents for Amazon Bedrock Runtime

This property contains the document to chat with, along with its attributes.

Contents

contentType

The MIME type of the document contained in the wrapper object.

Type: String

Pattern: `.*[a-z]{1,20}/.{1,20}.*`

Required: Yes

data

The byte value of the file to upload, encoded as a Base-64 string.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 10485760.

Required: Yes

identifier

The file name of the document contained in the wrapper object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ByteContentFile

Service: Agents for Amazon Bedrock Runtime

The property contains the file to chat with, along with its attributes.

Contents

data

The raw bytes of the file to attach. The maximum size of all files that is attached is 10MB. You can attach a maximum of 5 files.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 10485760.

Required: Yes

mediaType

The MIME type of data contained in the file used for chat.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Caller

Service: Agents for Amazon Bedrock Runtime

Details about a caller.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

agentAliasArn

The caller's agent alias ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Citation

Service: Agents for Amazon Bedrock Runtime

An object containing a segment of the generated response that is based on a source in the knowledge base, alongside information about the source.

This data type is used in the following API operations:

- InvokeAgent response – in the `citations` field
- RetrieveAndGenerate response – in the `citations` field

Contents

generatedResponsePart

Contains the generated response and metadata

Type: GeneratedResponsePart object

Required: No

retrievedReferences

Contains metadata about the sources cited for the generated response.

Type: Array of RetrievedReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationEvent

Service: Agents for Amazon Bedrock Runtime

A citation event.

Contents

citation

This member has been deprecated.

The citation.

Type: Citation object

Required: No

generatedResponsePart

The generated response to the citation event.

Type: GeneratedResponsePart object

Required: No

retrievedReferences

The retrieved references of the citation event.

Type: Array of RetrievedReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CodeInterpreterInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the code interpreter being invoked.

Contents

code

The code for the code interpreter to use.

Type: String

Required: No

files

Files that are uploaded for code interpreter to use.

Type: Array of strings

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CodeInterpreterInvocationOutput

Service: Agents for Amazon Bedrock Runtime

Contains the JSON-formatted string returned by the API invoked by the code interpreter.

Contents

executionError

Contains the error returned from code execution.

Type: String

Required: No

executionOutput

Contains the successful output returned from code execution

Type: String

Required: No

executionTimeout

Indicates if the execution of the code timed out.

Type: Boolean

Required: No

files

Contains output files, if generated by code execution.

Type: Array of strings

Required: No

metadata

Contains information about the output from the code interpreter.

Type: Metadata object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Collaborator

Service: Agents for Amazon Bedrock Runtime

List of inline collaborators.

Contents

foundationModel

The foundation model used by the inline collaborator agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+))$|(^arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{1,20}):(|[0-9]{12}):inference-profile/[a-zA-Z0-9-:.]+)`

Required: Yes

instruction

Instruction that tell the inline collaborator agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 40.

Required: Yes

actionGroups

List of action groups with each action group defining tasks the inline collaborator agent needs to carry out.

Type: Array of AgentActionGroup objects

Required: No

agentCollaboration

Defines how the inline supervisor agent handles information across multiple collaborator agents to coordinate a final response.

Type: String

Valid Values: `SUPERVISOR | SUPERVISOR_ROUTER | DISABLED`

Required: No

agentName

Name of the inline collaborator agent which must be the same name as specified for `collaboratorName`.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

collaboratorConfigurations

Settings of the collaborator agent.

Type: Array of CollaboratorConfiguration objects

Required: No

customerEncryptionKeyArn

The Amazon Resource Name (ARN) of the AWS KMS key that encrypts the inline collaborator.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: No

guardrailConfiguration

Details of the guardwrail associated with the inline collaborator.

Type: GuardrailConfigurationWithArn object

Required: No

idleSessionTTLInSeconds

The number of seconds for which the Amazon Bedrock keeps information about the user's conversation with the inline collaborator agent.

A user interaction remains active for the amount of time specified. If no conversation occurs during this time, the session expires and Amazon Bedrock deletes any data provided before the timeout.

Type: Integer

Valid Range: Minimum value of 60. Maximum value of 3600.

Required: No

knowledgeBases

Knowledge base associated with the inline collaborator agent.

Type: Array of KnowledgeBase objects

Required: No

promptOverrideConfiguration

Contains configurations to override prompt templates in different parts of an inline collaborator sequence. For more information, see Advanced prompts.

Type: PromptOverrideConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

CollaboratorConfiguration

Service: Agents for Amazon Bedrock Runtime

Settings of an inline collaborator agent.

Contents

collaboratorInstruction

Instructions that tell the inline collaborator agent what it should do and how it should interact with users.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: Yes

collaboratorName

Name of the inline collaborator agent which must be the same name as specified for `agentName`.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

agentAliasArn

The Amazon Resource Name (ARN) of the inline collaborator agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:agent-alias/[0-9a-zA-Z]{10}/[0-9a-zA-Z]{10}`

Required: No

relayConversationHistory

A relay conversation history for the inline collaborator agent.

Type: String

Valid Values: `TO_COLLABORATOR | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConditionResultEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about a condition evaluation result during a flow execution. This event is generated when a condition node in the flow evaluates its conditions.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

nodeName

The name of the condition node that evaluated the conditions.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

satisfiedConditions

A list of conditions that were satisfied during the evaluation.

Type: Array of SatisfiedCondition objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

timestamp

The timestamp when the condition evaluation occurred.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlock

Service: Agents for Amazon Bedrock Runtime

A content block.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

The block's text.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBody

Service: Agents for Amazon Bedrock Runtime

Contains the body of the API response.

This data type is used in the following API operations:

- In the `returnControlInvocationResults` field of the InvokeAgent request

Contents

body

The body of the API response.

Type: String

Required: No

images

Lists details, including format and source, for the image in the response from the function call. You can specify only one image and the function in the `returnControlInvocationResults` must be a computer use action. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Type: Array of ImageInput objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConversationHistory

Service: Agents for Amazon Bedrock Runtime

A conversation history.

Contents

messages

The conversation's messages.

Type: Array of Message objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomOrchestration

Service: Agents for Amazon Bedrock Runtime

Contains details of the custom orchestration configured for the agent.

Contents

executor

The structure of the executor invoking the actions in custom orchestration.

Type: OrchestrationExecutor object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomOrchestrationTrace

Service: Agents for Amazon Bedrock Runtime

The trace behavior for the custom orchestration.

Contents

event

The event details used with the custom orchestration.

Type: CustomOrchestrationTraceEvent object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomOrchestrationTraceEvent

Service: Agents for Amazon Bedrock Runtime

The event in the custom orchestration sequence. Events are the responses which the custom orchestration Lambda function sends as response to the agent.

Contents

text

The text that prompted the event at this step.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentAcl

Service: Agents for Amazon Bedrock Runtime

The access control list for a document, containing allow and deny membership lists. Each list specifies conditions that determine which users and groups are granted or denied access.

Contents

allowList

The list of principals allowed access to the document.

Type: DocumentAclMembership object

Required: No

denyList

The list of principals denied access to the document.

Type: DocumentAclMembership object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentAclCondition

Service: Agents for Amazon Bedrock Runtime

A condition within a document access control list (ACL) membership, specifying users and groups that are evaluated together.

Contents

conditionOperator

The logical operator for combining users and groups within this condition. Valid values: `AND` – Both a user match and a group match are required. `OR` – Either a user match or a group match is sufficient.

Type: String

Valid Values: `AND | OR`

Required: No

groups

The list of group entries in this condition.

Type: Array of DocumentAclGroup objects

Array Members: Minimum number of 0 items. Maximum number of 3000 items.

Required: No

users

The list of user entries in this condition.

Type: Array of DocumentAclUser objects

Array Members: Minimum number of 0 items. Maximum number of 3000 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentAclGroup

Service: Agents for Amazon Bedrock Runtime

A group entry within a document access control list (ACL) condition.

Contents

id

The identifier of the group.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `\P{C}*`

Required: Yes

type

The membership type indicating the scope of the group entry.

Type: String

Valid Values: `KNOWLEDGE_BASE | DATA_SOURCE`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentAclMembership

Service: Agents for Amazon Bedrock Runtime

The membership entry for a document access control list (ACL), containing conditions and their logical relation.

Contents

conditions

The list of conditions that determine membership.

Type: Array of DocumentAclCondition objects

Array Members: Minimum number of 0 items. Maximum number of 100 items.

Required: No

memberRelation

The logical relation between conditions. Valid values: `AND` – All conditions must match. `OR` – At least one condition must match.

Type: String

Valid Values: `AND | OR`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentAclUser

Service: Agents for Amazon Bedrock Runtime

A user entry within a document access control list (ACL) condition.

Contents

id

The identifier of the user.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `\P{C}*`

Required: Yes

type

The membership type indicating the scope of the user entry.

Type: String

Valid Values: `KNOWLEDGE_BASE | DATA_SOURCE`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExternalSource

Service: Agents for Amazon Bedrock Runtime

The unique external source of the content contained in the wrapper object.

Contents

sourceType

The source type of the external source wrapper object.

Type: String

Valid Values: `S3 | BYTE_CONTENT`

Required: Yes

byteContent

The identifier, contentType, and data of the external source wrapper object.

Type: ByteContentDoc object

Required: No

s3Location

The S3 location of the external source wrapper object.

Type: S3ObjectDoc object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExternalSourcesGenerationConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains the generation configuration of the external source wrapper object.

Contents

additionalModelRequestFields

Additional model parameters and their corresponding values not included in the textInferenceConfig structure for an external source. Takes in custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

guardrailConfiguration

The configuration details for the guardrail.

Type: GuardrailConfiguration object

Required: No

inferenceConfig

Configuration settings for inference when using RetrieveAndGenerate to generate responses while using an external source.

Type: InferenceConfig object

Required: No

performanceConfig

The latency configuration for the model.

Type: PerformanceConfiguration object

Required: No

promptTemplate

Contain the textPromptTemplate string for the external source wrapper object.

Type: PromptTemplate object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ExternalSourcesRetrieveAndGenerateConfiguration

Service: Agents for Amazon Bedrock Runtime

The configurations of the external source wrapper object in the `retrieveAndGenerate` function.

Contents

modelArn

The model Amazon Resource Name (ARN) for the external source wrapper object in the `retrieveAndGenerate` function.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

sources

The document for the external source wrapper object in the `retrieveAndGenerate` function.

Type: Array of ExternalSource objects

Array Members: Fixed number of 1 item.

Required: Yes

generationConfiguration

The prompt used with the external source wrapper object with the `retrieveAndGenerate` function.

Type: ExternalSourcesGenerationConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FailureTrace

Service: Agents for Amazon Bedrock Runtime

Contains information about the failure of the interaction.

Contents

failureCode

The failure code for the trace.

Type: Integer

Required: No

failureReason

The reason the interaction failed.

Type: String

Required: No

metadata

Information about the failure that occurred.

Type: Metadata object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FieldForReranking

Service: Agents for Amazon Bedrock Runtime

Contains information for a metadata field to include in or exclude from consideration when reranking.

Contents

fieldName

The name of a metadata field to include in or exclude from consideration when reranking.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FilePart

Service: Agents for Amazon Bedrock Runtime

Contains intermediate response for code interpreter if any files have been generated.

Contents

files

Files containing intermediate response for the user.

Type: Array of OutputFile objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FileSource

Service: Agents for Amazon Bedrock Runtime

The source file of the content contained in the wrapper object.

Contents

sourceType

The source type of the files to attach.

Type: String

Valid Values: `S3 | BYTE_CONTENT`

Required: Yes

byteContent

The data and the text of the attached files.

Type: ByteContentFile object

Required: No

s3Location

The s3 location of the files to attach.

Type: S3ObjectFile object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FilterAttribute

Service: Agents for Amazon Bedrock Runtime

Specifies the name that the metadata attribute must match and the value to which to compare the value of the metadata attribute. For more information, see Query configurations.

This data type is used in the following API operations:

- RetrieveAndGenerate request

Contents

key

The name that the metadata attribute must match.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100.

Required: Yes

value

The value to which to compare the value of the metadata attribute.

Type: JSON value

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FinalResponse

Service: Agents for Amazon Bedrock Runtime

Contains details about the response to the user.

Contents

metadata

Contains information about the invoke agent operation.

Type: Metadata object

Required: No

text

The text in the response to the user.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowCompletionEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about why a flow completed.

Contents

completionReason

The reason that the flow completed.

Type: String

Valid Values: `SUCCESS | INPUT_REQUIRED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowExecutionContent

Service: Agents for Amazon Bedrock Runtime

Contains the content of an flow execution input or output field.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The document content of the field, which can contain text or structured data.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowExecutionError

Service: Agents for Amazon Bedrock Runtime

Contains information about an error that occurred during an flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

error

The error code for the type of error that occurred.

Type: String

Valid Values: `ExecutionTimedOut`

Required: No

message

A descriptive message that provides details about the error.

Type: String

Required: No

nodeName

The name of the node in the flow where the error occurred (if applicable).

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowExecutionEvent

Service: Agents for Amazon Bedrock Runtime

Represents an event that occurred during an flow execution. This is a union type that can contain one of several event types, such as node input and output events; flow input and output events; condition node result events, or failure events.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

conditionResultEvent

Contains information about a condition evaluation result during the flow execution. This event is generated when a condition node in the flow evaluates its conditions.

Type: ConditionResultEvent object

Required: No

flowFailureEvent

Contains information about a failure that occurred at the flow level during execution.

Type: FlowFailureEvent object

Required: No

flowInputEvent

Contains information about the inputs provided to the flow at the start of execution.

Type: FlowExecutionInputEvent object

Required: No

flowOutputEvent

Contains information about the outputs produced by the flow at the end of execution.

Type: FlowExecutionOutputEvent object

Required: No

nodeActionEvent

Contains information about an action (operation) called by a node during execution.

Type: NodeActionEvent object

Required: No

nodeDependencyEvent

Contains information about an internal trace of a specific node during execution.

Type: NodeDependencyEvent object

Required: No

nodeFailureEvent

Contains information about a failure that occurred at a specific node during execution.

Type: NodeFailureEvent object

Required: No

nodeInputEvent

Contains information about the inputs provided to a specific node during execution.

Type: NodeInputEvent object

Required: No

nodeOutputEvent

Contains information about the outputs produced by a specific node during execution.

Type: NodeOutputEvent object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowExecutionInputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about the inputs provided to the flow at the start of a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

fields

A list of input fields provided to the flow.

Type: Array of FlowInputField objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

nodeName

The name of the node that receives the inputs.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The timestamp when the inputs are provided.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowExecutionOutputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about the outputs produced by the flow during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

fields

A list of output fields produced by the flow.

Type: Array of FlowOutputField objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

nodeName

The name of the node that produces the outputs.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The timestamp when the outputs are produced.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowExecutionSummary

Service: Agents for Amazon Bedrock Runtime

Contains summary information about a flow execution, including its status, timestamps, and identifiers.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

createdAt

The timestamp when the flow execution was created.

Type: Timestamp

Required: Yes

executionArn

The Amazon Resource Name (ARN) that uniquely identifies the flow execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `[a-zA-Z0-9-]{1,36}$|^(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10}/execution/[a-zA-Z0-9-]{1,36})`

Required: Yes

flowAliasIdentifier

The unique identifier of the flow alias used for the execution.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10}/alias/[0-9a-zA-Z]{10})|(\bTSTALIASID\b|[0-9a-zA-Z]+)`

Required: Yes

flowIdentifier

The unique identifier of the flow.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:flow/[0-9a-zA-Z]{10})|([0-9a-zA-Z]{10})`

Required: Yes

flowVersion

The version of the flow used for the execution.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: Yes

status

The current status of the flow execution.

Flow executions time out after 24 hours.

Type: String

Valid Values: `Running | Succeeded | Failed | TimedOut | Aborted`

Required: Yes

endedAt

The timestamp when the flow execution ended. This field is only populated when the execution has completed, failed, timed out, or been aborted.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowFailureEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about a failure that occurred at the flow level during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

errorCode

The error code that identifies the type of failure that occurred.

Type: String

Valid Values: `VALIDATION | INTERNAL_SERVER | NODE_EXECUTION_FAILED`

Required: Yes

errorMessage

A descriptive message that provides details about the failure.

Type: String

Required: Yes

timestamp

The timestamp when the failure occurred.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowInput

Service: Agents for Amazon Bedrock Runtime

Contains information about an input into the prompt flow and where to send it.

Contents

content

Contains information about an input into the prompt flow.

Type: FlowInputContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

nodeName

The name of the flow input node that begins the prompt flow.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

nodeInputName

The name of the input from the flow input node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: No

nodeOutputName

The name of the output from the flow input node that begins the prompt flow.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowInputContent

Service: Agents for Amazon Bedrock Runtime

Contains information about an input into the flow.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The input to send to the prompt flow input node.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowInputField

Service: Agents for Amazon Bedrock Runtime

Represents an input field provided to a flow during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

content

The content of the input field, which can contain text or structured data.

Type: FlowExecutionContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

name

The name of the input field as defined in the flow's input schema.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowMultiTurnInputContent

Service: Agents for Amazon Bedrock Runtime

The content structure containing input information for multi-turn flow interactions.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The requested additional input to send back to the multi-turn flow node.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowMultiTurnInputRequestEvent

Service: Agents for Amazon Bedrock Runtime

Response object from the flow multi-turn node requesting additional information.

Contents

content

The content payload containing the input request details for the multi-turn interaction.

Type: FlowMultiTurnInputContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

nodeName

The name of the node in the flow that is requesting the input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

nodeType

The type of the node in the flow that is requesting the input.

Type: String

Valid Values: `FlowInputNode | FlowOutputNode | LambdaFunctionNode | KnowledgeBaseNode | PromptNode | ConditionNode | LexNode`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowOutputContent

Service: Agents for Amazon Bedrock Runtime

Contains information about the content in an output from prompt flow invocation.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The content in the output.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowOutputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about an output from prompt flow invoction.

Contents

content

The content in the output.

Type: FlowOutputContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

nodeName

The name of the flow output node that the output is from.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

nodeType

The type of the node that the output is from.

Type: String

Valid Values: `FlowInputNode | FlowOutputNode | LambdaFunctionNode | KnowledgeBaseNode | PromptNode | ConditionNode | LexNode`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowOutputField

Service: Agents for Amazon Bedrock Runtime

Represents an output field produced by a flow during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

content

The content of the output field, which can contain text or structured data.

Type: FlowExecutionContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

name

The name of the output field as defined in the flow's output schema.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowResponseStream

Service: Agents for Amazon Bedrock Runtime

The output of the flow.

Contents

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

Required: No

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

Required: No

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

Required: No

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

Required: No

flowCompletionEvent

Contains information about why the flow completed.

Type: FlowCompletionEvent object

Required: No

flowMultiTurnInputRequestEvent

The event stream containing the multi-turn input request information from the flow.

Type: FlowMultiTurnInputRequestEvent object

Required: No

flowOutputEvent

Contains information about an output from flow invocation.

Type: FlowOutputEvent object

Required: No

flowTraceEvent

Contains information about a trace, which tracks an input or output for a node in the flow.

Type: FlowTraceEvent object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception

HTTP Status Code: 404

Required: No

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception

HTTP Status Code: 400

Required: No

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTrace

Service: Agents for Amazon Bedrock Runtime

Contains information about an input or output for a node in the flow. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

conditionNodeResultTrace

Contains information about an output from a condition node.

Type: FlowTraceConditionNodeResultEvent object

Required: No

nodeActionTrace

Contains information about an action (operation) called by a node.

Type: FlowTraceNodeActionEvent object

Required: No

nodeDependencyTrace

Contains information about an internal trace of a node.

Type: FlowTraceDependencyEvent object

Required: No

nodeInputTrace

Contains information about the input into a node.

Type: FlowTraceNodeInputEvent object

Required: No

nodeOutputTrace

Contains information about the output from a node.

Type: FlowTraceNodeOutputEvent object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceCondition

Service: Agents for Amazon Bedrock Runtime

Contains information about a condition that was satisfied. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

conditionName

The name of the condition.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceConditionNodeResultEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about an output from a condition node. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

nodeName

The name of the condition node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

satisfiedConditions

An array of objects containing information about the conditions that were satisfied.

Type: Array of FlowTraceCondition objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

timestamp

The date and time that the trace was returned.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

FlowTraceDependencyEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about a dependency trace event in the flow.

Contents

nodeName

The name of the node that generated the dependency trace.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The date and time that the dependency trace was generated.

Type: Timestamp

Required: Yes

traceElements

The trace elements containing detailed information about the dependency.

Type: TraceElements object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about a trace, which tracks an input or output for a node in the flow. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

trace

The trace object containing information about an input or output for a node in the flow.

Type: FlowTrace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeActionEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about an action (operation) called by a node in an Amazon Bedrock flow. The service generates action events for calls made by prompt nodes, agent nodes, and AWS Lambda nodes.

Contents

nodeName

The name of the node that called the operation.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

operationName

The name of the operation that the node called.

Type: String

Required: Yes

requestId

The ID of the request that the node made to the operation.

Type: String

Required: Yes

serviceName

The name of the service that the node called.

Type: String

Required: Yes

timestamp

The date and time that the operation was called.

Type: Timestamp

Required: Yes

operationRequest

The request payload sent to the downstream service.

Type: JSON value

Required: No

operationResponse

The response payload received from the downstream service.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeInputContent

Service: Agents for Amazon Bedrock Runtime

Contains the content of the node input. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The content of the node input.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeInputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about the input into a node. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

fields

An array of objects containing information about each field in the input.

Type: Array of FlowTraceNodeInputField objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

nodeName

The name of the node that received the input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The date and time that the trace was returned.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

FlowTraceNodeInputExecutionChainItem

Service: Agents for Amazon Bedrock Runtime

Represents an item in the execution chain for flow trace node input tracking.

Contents

nodeName

The name of the node in the execution chain.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

type

The type of execution chain item. Supported values are Iterator and Loop.

Type: String

Valid Values: `Iterator | Loop`

Required: Yes

index

The index position of this item in the execution chain.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeInputField

Service: Agents for Amazon Bedrock Runtime

Contains information about a field in the input into a node. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

content

The content of the node input.

Type: FlowTraceNodeInputContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

nodeInputName

The name of the node input.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

category

The category of the input field.

Type: String

Valid Values: `LoopCondition | ReturnValueToLoopStart | ExitLoop`

Required: No

executionChain

The execution path through nested nodes like iterators and loops.

Type: Array of FlowTraceNodeInputExecutionChainItem objects

Required: No

source

The source node that provides input data to this field.

Type: FlowTraceNodeInputSource object

Required: No

type

The data type of the input field for compatibility validation.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeInputSource

Service: Agents for Amazon Bedrock Runtime

Represents the source of input data for a flow trace node field.

Contents

expression

The expression used to extract data from the source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

nodeName

The name of the source node that provides the input data.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

outputFieldName

The name of the output field from the source node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeOutputContent

Service: Agents for Amazon Bedrock Runtime

Contains the content of the node output. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The content of the node output.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeOutputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about the output from a node. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

fields

An array of objects containing information about each field in the output.

Type: Array of FlowTraceNodeOutputField objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: Yes

nodeName

The name of the node that yielded the output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The date and time that the trace was returned.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

FlowTraceNodeOutputField

Service: Agents for Amazon Bedrock Runtime

Contains information about a field in the output from a node. For more information, see Track each step in your prompt flow by viewing its trace in Amazon Bedrock.

Contents

content

The content of the node output.

Type: FlowTraceNodeOutputContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

nodeOutputName

The name of the node output.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

next

The next node that receives output data from this field.

Type: Array of FlowTraceNodeOutputNext objects

Required: No

type

The data type of the output field for compatibility validation.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FlowTraceNodeOutputNext

Service: Agents for Amazon Bedrock Runtime

Represents the next node that receives output data from a flow trace.

Contents

inputFieldName

The name of the input field in the next node that receives the data.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

nodeName

The name of the next node that receives the output data.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FoundationModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for the foundation model.

Contents

type

The type of foundation model configuration.

Type: String

Valid Values: `BEDROCK_FOUNDATION_MODEL`

Required: Yes

bedrockFoundationModelConfiguration

The Bedrock foundation model configuration.

Type: BedrockFoundationModelConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FunctionDefinition

Service: Agents for Amazon Bedrock Runtime

Defines parameters that the agent needs to invoke from the user to complete the function. Corresponds to an action in an action group.

Contents

name

A name for the function.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: Yes

description

A description of the function and its purpose.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1200.

Required: No

parameters

The parameters that the agent elicits from the user to fulfill the function.

Type: String to ParameterDetail object map

Key Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

requireConfirmation

Contains information if user confirmation is required to invoke the function.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FunctionInvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about the function that the agent predicts should be called.

This data type is used in the following API operations:

- In the `returnControl` field of the InvokeAgent response

Contents

actionGroup

The action group that the function belongs to.

Type: String

Required: Yes

actionInvocationType

Contains information about the function to invoke,

Type: String

Valid Values: `RESULT | USER_CONFIRMATION | USER_CONFIRMATION_AND_RESULT`

Required: No

agentId

The agent's ID.

Type: String

Required: No

collaboratorName

The collaborator's name.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

function

The name of the function.

Type: String

Required: No

parameters

A list of parameters of the function.

Type: Array of FunctionParameter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FunctionParameter

Service: Agents for Amazon Bedrock Runtime

Contains information about a parameter of the function.

This data type is used in the following API operations:

- In the `returnControl` field of the InvokeAgent response

Contents

name

The name of the parameter.

Type: String

Required: No

type

The data type of the parameter.

Type: String

Required: No

value

The value of the parameter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

FunctionResult

Service: Agents for Amazon Bedrock Runtime

Contains information about the function that was called from the action group and the response that was returned.

This data type is used in the following API operations:

- In the `returnControlInvocationResults` of the InvokeAgent request

Contents

actionGroup

The action group that the function belongs to.

Type: String

Required: Yes

agentId

The agent's ID.

Type: String

Required: No

confirmationState

Contains the user confirmation information about the function that was called.

Type: String

Valid Values: `CONFIRM | DENY`

Required: No

function

The name of the function that was called.

Type: String

Required: No

responseBody

The response from the function call using the parameters. The response might be returned directly or from the Lambda function. Specify `TEXT` or `IMAGES`. The key of the object is the content type. You can only specify one type. If you specify `IMAGES`, you can specify only one image. You can specify images only when the function in the `returnControlInvocationResults` is a computer use action. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Type: String to ContentBody object map

Required: No

responseState

Controls the final response state returned to end user when API/Function execution failed. When this state is FAILURE, the request would fail with dependency failure exception. When this state is REPROMPT, the API/function response will be sent to model for re-prompt

Type: String

Valid Values: `FAILURE | REPROMPT`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FunctionSchema

Service: Agents for Amazon Bedrock Runtime

Contains details about the function schema for the action group or the JSON or YAML-formatted payload defining the schema.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

functions

A list of functions that each define an action in the action group.

Type: Array of FunctionDefinition objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GeneratedQuery

Service: Agents for Amazon Bedrock Runtime

Contains information about a query generated for a natural language query.

Contents

sql

An SQL query that corresponds to the natural language query.

Type: String

Required: No

type

The type of transformed query.

Type: String

Valid Values: `REDSHIFT_SQL`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GeneratedResponsePart

Service: Agents for Amazon Bedrock Runtime

Contains metadata about a part of the generated response that is accompanied by a citation.

This data type is used in the following API operations:

- InvokeAgent response – in the `generatedResponsePart` field
- RetrieveAndGenerate response – in the `generatedResponsePart` field

Contents

textResponsePart

Contains metadata about a textual part of the generated response that is accompanied by a citation.

Type: TextResponsePart object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GenerationConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for response generation based on the knowledge base query results.

This data type is used in the following API operations:

- RetrieveAndGenerate request

Contents

additionalModelRequestFields

Additional model parameters and corresponding values not included in the textInferenceConfig structure for a knowledge base. This allows users to provide custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

guardrailConfiguration

The configuration details for the guardrail.

Type: GuardrailConfiguration object

Required: No

inferenceConfig

Configuration settings for inference when using RetrieveAndGenerate to generate responses while using a knowledge base as a source.

Type: InferenceConfig object

Required: No

performanceConfig

The latency configuration for the model.

Type: PerformanceConfiguration object

Required: No

promptTemplate

Contains the template for the prompt that's sent to the model for response generation. Generation prompts must include the `$search_results$` variable. For more information, see Use placeholder variables in the user guide.

Type: PromptTemplate object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAssessment

Service: Agents for Amazon Bedrock Runtime

Assessment details of the content analyzed by Guardrails.

Contents

contentPolicy

Content policy details of the Guardrail.

Type: GuardrailContentPolicyAssessment object

Required: No

sensitiveInformationPolicy

Sensitive Information policy details of Guardrail.

Type: GuardrailSensitiveInformationPolicyAssessment object

Required: No

topicPolicy

Topic policy details of the Guardrail.

Type: GuardrailTopicPolicyAssessment object

Required: No

wordPolicy

Word policy details of the Guardrail.

Type: GuardrailWordPolicyAssessment object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConfiguration

Service: Agents for Amazon Bedrock Runtime

The configuration details for the guardrail.

Contents

guardrailId

The unique identifier for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 64.

Pattern: `[a-z0-9]+`

Required: Yes

guardrailVersion

The version of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(([1-9][0-9]{0,7})|(DRAFT))`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConfigurationWithArn

Service: Agents for Amazon Bedrock Runtime

The configuration details for the guardrail.

Contents

guardrailIdentifier

The unique identifier for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: Yes

guardrailVersion

The version of the guardrail.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(([1-9][0-9]{0,7})|(DRAFT))`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentFilter

Service: Agents for Amazon Bedrock Runtime

Details of the content filter used in the Guardrail.

Contents

action

The action placed on the content by the Guardrail filter.

Type: String

Valid Values: `BLOCKED`

Required: No

confidence

The confidence level regarding the content detected in the filter by the Guardrail.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: No

type

The type of content detected in the filter by the Guardrail.

Type: String

Valid Values: `INSULTS | HATE | SEXUAL | VIOLENCE | MISCONDUCT | PROMPT_ATTACK`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentPolicyAssessment

Service: Agents for Amazon Bedrock Runtime

The details of the policy assessment in the Guardrails filter.

Contents

filters

The filter details of the policy assessment used in the Guardrails filter.

Type: Array of GuardrailContentFilter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailCustomWord

Service: Agents for Amazon Bedrock Runtime

The custom word details for the filter in the Guardrail.

Contents

action

The action details for the custom word filter in the Guardrail.

Type: String

Valid Values: BLOCKED

Required: No

match

The match details for the custom word filter in the Guardrail.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailEvent

Service: Agents for Amazon Bedrock Runtime

A guardrail event.

Contents

action

The guardrail action.

Type: String

Valid Values: `INTERVENED | NONE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailManagedWord

Service: Agents for Amazon Bedrock Runtime

The managed word details for the filter in the Guardrail.

Contents

action

The action details for the managed word filter in the Guardrail.

Type: String

Valid Values: `BLOCKED`

Required: No

match

The match details for the managed word filter in the Guardrail.

Type: String

Required: No

type

The type details for the managed word filter in the Guardrail.

Type: String

Valid Values: `PROFANITY`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailPiiEntityFilter

Service: Agents for Amazon Bedrock Runtime

The Guardrail filter to identify and remove personally identifiable information (PII).

Contents

action

The action of the Guardrail filter to identify and remove PII.

Type: String

Valid Values: `BLOCKED | ANONYMIZED`

Required: No

match

The match to settings in the Guardrail filter to identify and remove PII.

Type: String

Required: No

type

The type of PII the Guardrail filter has identified and removed.

Type: String

Valid Values: `ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailRegexFilter

Service: Agents for Amazon Bedrock Runtime

The details for the regex filter used in the Guardrail.

Contents

action

The action details for the regex filter used in the Guardrail.

Type: String

Valid Values: `BLOCKED | ANONYMIZED`

Required: No

match

The match details for the regex filter used in the Guardrail.

Type: String

Required: No

name

The name details for the regex filter used in the Guardrail.

Type: String

Required: No

regex

The regex details for the regex filter used in the Guardrail.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailSensitiveInformationPolicyAssessment

Service: Agents for Amazon Bedrock Runtime

The details of the sensitive policy assessment used in the Guardrail.

Contents

piiEntities

The details of the PII entities used in the sensitive policy assessment for the Guardrail.

Type: Array of GuardrailPiiEntityFilter objects

Required: No

regexes

The details of the regexes used in the sensitive policy assessment for the Guardrail.

Type: Array of GuardrailRegexFilter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopic

Service: Agents for Amazon Bedrock Runtime

The details for a specific topic defined in the Guardrail.

Contents

action

The action details on a specific topic in the Guardrail.

Type: String

Valid Values: BLOCKED

Required: No

name

The name details on a specific topic in the Guardrail.

Type: String

Required: No

type

The type details on a specific topic in the Guardrail.

Type: String

Valid Values: DENY

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicPolicyAssessment

Service: Agents for Amazon Bedrock Runtime

The details of the policy assessment used in the Guardrail.

Contents

topics

The topic details of the policy assessment used in the Guardrail.

Type: Array of GuardrailTopic objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTrace

Service: Agents for Amazon Bedrock Runtime

The trace details used in the Guardrail.

Contents

action

The trace action details used with the Guardrail.

Type: String

Valid Values: `INTERVENED | NONE`

Required: No

inputAssessments

The details of the input assessments used in the Guardrail Trace.

Type: Array of GuardrailAssessment objects

Required: No

metadata

Contains information about the Guardrail output.

Type: Metadata object

Required: No

outputAssessments

The details of the output assessments used in the Guardrail Trace.

Type: Array of GuardrailAssessment objects

Required: No

traceId

The details of the trace Id used in the Guardrail Trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailWordPolicyAssessment

Service: Agents for Amazon Bedrock Runtime

The assessment details for words defined in the Guardrail filter.

Contents

customWords

The custom word details for words defined in the Guardrail filter.

Type: Array of GuardrailCustomWord objects

Required: No

managedWordLists

The managed word lists for words defined in the Guardrail filter.

Type: Array of GuardrailManagedWord objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageBlock

Service: Agents for Amazon Bedrock Runtime

Image content for an invocation step.

Contents

format

The format of the image.

Type: String

Valid Values: `png | jpeg | gif | webp`

Required: Yes

source

The source for the image.

Type: ImageSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageInput

Service: Agents for Amazon Bedrock Runtime

Details about an image in the result from a function in the action group invocation. You can specify images only when the function is a computer use action. For more information, see Configure an Amazon Bedrock Agent to complete tasks with computer use tools.

Contents

format

The type of image in the result.

Type: String

Valid Values: `png | jpeg | gif | webp`

Required: Yes

source

The source of the image in the result.

Type: ImageInputSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageInputSource

Service: Agents for Amazon Bedrock Runtime

Details about the source of an input image in the result from a function in the action group invocation.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

The raw image bytes for the image. If you use an AWS SDK, you don't need to encode the image bytes in base64.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageSource

Service: Agents for Amazon Bedrock Runtime

The source for an image.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

The raw image bytes for the image. If you use an AWS SDK, you don't need to encode the image bytes in base64.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

s3Location

The path to the Amazon S3 bucket where the image is stored.

Type: S3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImplicitFilterConfiguration

Service: Agents for Amazon Bedrock Runtime

Settings for implicit filtering, where a model generates a metadata filter based on the prompt.

Contents

metadataAttributes

Metadata that can be used in a filter.

Type: Array of MetadataAttributeSchema objects

Array Members: Minimum number of 1 item. Maximum number of 25 items.

Required: Yes

modelArn

The model that generates the filter.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceConfig

Service: Agents for Amazon Bedrock Runtime

The configuration for inference settings when generating responses using RetrieveAndGenerate.

Contents

textInferenceConfig

Configuration settings specific to text generation while generating responses using RetrieveAndGenerate.

Type: TextInferenceConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InferenceConfiguration

Service: Agents for Amazon Bedrock Runtime

Specifications about the inference parameters that were provided alongside the prompt. These are specified in the PromptOverrideConfiguration object that was set when the agent was created or updated. For more information, see Inference parameters for foundation models.

Contents

maximumLength

The maximum number of tokens allowed in the generated response.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 8192.

Required: No

stopSequences

A list of stop sequences. A stop sequence is a sequence of characters that causes the model to stop generating the response.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Required: No

temperature

The likelihood of the model selecting higher-probability options while generating a response. A lower value makes the model more likely to choose higher-probability options, while a higher value makes the model more likely to choose lower-probability options.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topK

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for `topK` is the number of most-likely

candidates from which the model chooses the next token in the sequence. For example, if you set `topK` to 50, the model selects the next token from among the top 50 most likely choices.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 500.

Required: No

topP

While generating a response, the model determines the probability of the following token at each point of generation. The value that you set for `Top P` determines the number of most-likely candidates from which the model chooses the next token in the sequence. For example, if you set `topP` to 0.8, the model only selects the next token from the top 80% of the probability distribution of next tokens.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineAgentFilePart

Service: Agents for Amazon Bedrock Runtime

Contains intermediate response for code interpreter if any files have been generated.

Contents

files

Files containing intermediate response for the user.

Type: Array of OutputFile objects

Array Members: Minimum number of 0 items. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineAgentPayloadPart

Service: Agents for Amazon Bedrock Runtime

Contains a part of an agent response and citations for it.

Contents

attribution

Contains citations for a part of an agent response.

Type: Attribution object

Required: No

bytes

A part of the agent response in bytes.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineAgentResponseStream

Service: Agents for Amazon Bedrock Runtime

The response from invoking the agent and associated citations and trace information.

Contents

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

Required: No

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

Required: No

chunk

Contains a part of an agent response and citations for it.

Type: InlineAgentPayloadPart object

Required: No

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

Required: No

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

Required: No

files

Contains intermediate response for code interpreter if any files have been generated.

Type: InlineAgentFilePart object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception

HTTP Status Code: 404

Required: No

returnControl

Contains the parameters and information that the agent elicited from the customer to carry out an action. This information is returned to the system and can be used in your own setup for fulfilling the action.

Type: InlineAgentReturnControlPayload object

Required: No

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception

HTTP Status Code: 400

Required: No

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

Required: No

trace

Contains information about the agent and session, alongside the agent's reasoning process and results from calling actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace events.

Type: InlineAgentTracePart object

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineAgentReturnControlPayload

Service: Agents for Amazon Bedrock Runtime

Contains information to return from the action group that the agent has predicted to invoke.

This data type is used in the InvokeAgent response API operation.

Contents

invocationId

The identifier of the action group invocation.

Type: String

Required: No

invocationInputs

A list of objects that contain information about the parameters and inputs that need to be sent into the API operation or function, based on what the agent determines from its session with the user.

Type: Array of InvocationInputMember objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineAgentTracePart

Service: Agents for Amazon Bedrock Runtime

Contains information about the agent and session, alongside the agent's reasoning process and results from calling API actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace enablement.

Contents

callerChain

The caller chain for the trace part.

Type: Array of Caller objects

Required: No

collaboratorName

The collaborator name for the trace part.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

eventTime

The time that trace occurred.

Type: Timestamp

Required: No

sessionId

The unique identifier of the session with the agent.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

trace

Contains one part of the agent's reasoning process and results from calling API actions and querying knowledge bases. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace enablement.

Type: Trace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineBedrockModelConfigurations

Service: Agents for Amazon Bedrock Runtime

Settings for a model called with InvokeInlineAgent.

Contents

performanceConfig

The latency configuration for the model.

Type: PerformanceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlineSessionState

Service: Agents for Amazon Bedrock Runtime

Contains parameters that specify various attributes that persist across a session or prompt. You can define session state attributes as key-value pairs when writing a Lambda function for an action group or pass them when making an `InvokeInlineAgent` request. Use session state attributes to control and provide conversational context for your inline agent and to help customize your agent's behavior. For more information, see Control session context

Contents

conversationHistory

Contains the conversation history that persist across sessions.

Type: ConversationHistory object

Required: No

files

Contains information about the files used by code interpreter.

Type: Array of InputFile objects

Required: No

invocationId

The identifier of the invocation of an action. This value must match the `invocationId` returned in the `InvokeInlineAgent` response for the action whose results are provided in the `returnControlInvocationResults` field. For more information, see Return control to the agent developer.

Type: String

Required: No

promptSessionAttributes

Contains attributes that persist across a session and the values of those attributes.

Type: String to string map

Required: No

returnControlInvocationResults

Contains information about the results from the action group invocation. For more information, see Return control to the agent developer.

> **Note**
>
> If you include this field in the `sessionState` field, the `inputText` field will be ignored.

Type: Array of InvocationResultMember objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

sessionAttributes

Contains attributes that persist across a session and the values of those attributes.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputFile

Service: Agents for Amazon Bedrock Runtime

Contains details of the source files.

Contents

name

The name of the source file.

Type: String

Required: Yes

source

Specifies where the files are located.

Type: FileSource object

Required: Yes

useCase

Specifies how the source files will be used by the code interpreter.

Type: String

Valid Values: `CODE_INTERPRETER | CHAT`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputImage

Service: Agents for Amazon Bedrock Runtime

Contains the image data for multimodal knowledge base queries, including format and content.

This data type is used in the following API operations:

- Retrieve request – in the `image` field

Contents

format

The format of the input image. Supported formats include png, gif, jpeg, and webp.

Type: String

Valid Values: `png | jpeg | gif | webp`

Required: Yes

inlineContent

The base64-encoded image data for inline image content. Maximum size is 5MB.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1. Maximum length of 5242880.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputPrompt

Service: Agents for Amazon Bedrock Runtime

Contains information about the prompt to optimize.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

textPrompt

Contains information about the text prompt to optimize.

Type: TextPrompt object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationInput

Service: Agents for Amazon Bedrock Runtime

Contains information pertaining to the action group or knowledge base that is being invoked.

Contents

actionGroupInvocationInput

Contains information about the action group to be invoked.

Type: ActionGroupInvocationInput object

Required: No

agentCollaboratorInvocationInput

The collaborator's invocation input.

Type: AgentCollaboratorInvocationInput object

Required: No

codeInterpreterInvocationInput

Contains information about the code interpreter to be invoked.

Type: CodeInterpreterInvocationInput object

Required: No

invocationType

Specifies whether the agent is invoking an action group or a knowledge base.

Type: String

Valid Values: `ACTION_GROUP | KNOWLEDGE_BASE | FINISH | ACTION_GROUP_CODE_INTERPRETER | AGENT_COLLABORATOR`

Required: No

knowledgeBaseLookupInput

Contains details about the knowledge base to look up and the query to be made.

Type: KnowledgeBaseLookupInput object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationInputMember

Service: Agents for Amazon Bedrock Runtime

Contains details about the API operation or function that the agent predicts should be called.

This data type is used in the following API operations:

- In the `returnControl` field of the InvokeAgent response

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

apiInvocationInput

Contains information about the API operation that the agent predicts should be called.

Type: ApiInvocationInput object

Required: No

functionInvocationInput

Contains information about the function that the agent predicts should be called.

Type: FunctionInvocationInput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

InvocationResultMember

Service: Agents for Amazon Bedrock Runtime

A result from the invocation of an action. For more information, see Return control to the agent developer and Control session context.

This data type is used in the following API operations:

- InvokeAgent request

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

apiResult

The result from the API response from the action group invocation.

Type: ApiResult object

Required: No

functionResult

The result from the function from the action group invocation.

Type: FunctionResult object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationStep

Service: Agents for Amazon Bedrock Runtime

Stores fine-grained state checkpoints, including text and images, for each interaction in an invocation in a session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Contents

invocationId

The unique identifier (in UUID format) for the invocation that includes the invocation step.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

invocationStepId

The unique identifier (in UUID format) for the invocation step.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

invocationStepTime

The timestamp for when the invocation step was created.

Type: Timestamp

Required: Yes

payload

Payload content, such as text and images, for the invocation step.

Type: InvocationStepPayload object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

sessionId

The unique identifier of the session containing the invocation step.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationStepPayload

Service: Agents for Amazon Bedrock Runtime

Payload content, such as text and images, for the invocation step.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

contentBlocks

The content for the invocation step.

Type: Array of BedrockSessionContentBlock objects

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationStepSummary

Service: Agents for Amazon Bedrock Runtime

Contains details about an invocation step within an invocation in a session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Contents

invocationId

A unique identifier for the invocation in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

invocationStepId

The unique identifier (in UUID format) for the invocation step.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

invocationStepTime

The timestamp for when the invocation step was created.

Type: Timestamp

Required: Yes

sessionId

The unique identifier for the session associated with the invocation step.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvocationSummary

Service: Agents for Amazon Bedrock Runtime

Contains details about an invocation in a session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Contents

createdAt

The timestamp for when the invocation was created.

Type: Timestamp

Required: Yes

invocationId

A unique identifier for the invocation in UUID format.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

sessionId

The unique identifier for the session associated with the invocation.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

KnowledgeBase

Service: Agents for Amazon Bedrock Runtime

Details of the knowledge base associated withe inline agent.

Contents

description

The description of the knowledge base associated with the inline agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

knowledgeBaseId

The unique identifier for a knowledge base associated with the inline agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

retrievalConfiguration

The configurations to apply to the knowledge base during query. For more information, see Query configurations.

Type: KnowledgeBaseRetrievalConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock Runtime

Configurations to apply to a knowledge base attached to the agent during query. For more information, see Knowledge base retrieval configurations.

Contents

knowledgeBaseId

The unique identifier for a knowledge base attached to the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

retrievalConfiguration

The configurations to apply to the knowledge base during query. For more information, see Query configurations.

Type: KnowledgeBaseRetrievalConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseLookupInput

Service: Agents for Amazon Bedrock Runtime

Contains details about the knowledge base to look up and the query to be made.

Contents

knowledgeBaseId

The unique identifier of the knowledge base to look up.

Type: String

Required: No

text

The query made to the knowledge base.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseLookupOutput

Service: Agents for Amazon Bedrock Runtime

Contains details about the results from looking up the knowledge base.

Contents

metadata

Contains information about the knowledge base output.

Type: Metadata object

Required: No

retrievedReferences

Contains metadata about the sources cited for the generated response.

Type: Array of RetrievedReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseQuery

Service: Agents for Amazon Bedrock Runtime

Contains the query made to the knowledge base.

This data type is used in the following API operations:

- Retrieve request – in the `retrievalQuery` field

Contents

image

An image to include in the knowledge base query for multimodal retrieval.

Type: InputImage object

Required: No

text

The text of the query made to the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 20000.

Required: No

type

The type of query being performed.

Type: String

Valid Values: `TEXT | IMAGE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseRetrievalConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for knowledge base query. For more information, see Query configurations.

This data type is used in the following API operations:

- Retrieve request – in the `retrievalConfiguration` field
- RetrieveAndGenerate request – in the `retrievalConfiguration` field

Contents

managedSearchConfiguration

Contains configurations for managed search. For more information, see Query configurations.

Type: ManagedSearchConfiguration object

Required: No

vectorSearchConfiguration

Contains details about how the results from the vector search should be returned. For more information, see Query configurations.

Type: KnowledgeBaseVectorSearchConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseRetrievalResult

Service: Agents for Amazon Bedrock Runtime

Details about a result from querying the knowledge base.

This data type is used in the following API operations:

- Retrieve response – in the `retrievalResults` field

Contents

content

Contains information about the content of the chunk.

Type: RetrievalResultContent object

Required: Yes

documentId

The unique identifier of the document. Use with `GetDocumentContent` to retrieve the full document.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1825.

Pattern: `\P{C}*`

Required: No

location

Contains information about the location of the data source.

Type: RetrievalResultLocation object

Required: No

metadata

Contains metadata attributes and their values for the file in the data source. For more information, see Metadata and filtering.

Type: String to JSON value map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

score

The level of relevance of the result to the query.

Type: Double

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseRetrieveAndGenerateConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains details about the resource being queried.

This data type is used in the following API operations:

- Retrieve request – in the `knowledgeBaseConfiguration` field
- RetrieveAndGenerate request – in the `knowledgeBaseConfiguration` field

Contents

knowledgeBaseId

The unique identifier of the knowledge base that is queried.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

modelArn

The ARN of the foundation model or inference profile used to generate a response.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:(bedrock|sagemaker):[a-z0-9-]{1,20}:([0-9]{12})?:([a-z-]+/)?)?([a-z0-9.-]{1,63}){0,2}(([:][a-z0-9-]{1,63}){0,2})?(/[a-z0-9]{1,12})?`

Required: Yes

generationConfiguration

Contains configurations for response generation based on the knowledge base query results.

Type: GenerationConfiguration object

Required: No

orchestrationConfiguration

Settings for how the model processes the prompt prior to retrieval and generation.

Type: OrchestrationConfiguration object

Required: No

retrievalConfiguration

Contains configurations for how to retrieve and return the knowledge base query.

Type: KnowledgeBaseRetrievalConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseRetrieverConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for retrieving from a Bedrock knowledge base.

Contents

knowledgeBaseId

The unique identifier of the knowledge base.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: Yes

retrievalOverrides

Overrides for retrieval behavior such as filters and result limits.

Type: RetrievalOverrides object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

KnowledgeBaseVectorSearchConfiguration

Service: Agents for Amazon Bedrock Runtime

Configurations for how to perform the search query and return results. For more information, see Query configurations.

This data type is used in the following API operations:

- Retrieve request – in the `vectorSearchConfiguration` field
- RetrieveAndGenerate request – in the `vectorSearchConfiguration` field

Contents

filter

Specifies the filters to use on the metadata in the knowledge base data sources before returning results. For more information, see Query configurations.

Type: RetrievalFilter object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

implicitFilterConfiguration

Settings for implicit filtering.

Type: ImplicitFilterConfiguration object

Required: No

numberOfResults

The number of source chunks to retrieve.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

overrideSearchType

By default, Amazon Bedrock decides a search strategy for you. If you're using an Amazon OpenSearch Serverless vector store that contains a filterable text field, you can specify whether to query the knowledge base with a `HYBRID` search using both vector embeddings and raw text, or `SEMANTIC` search using only vector embeddings. For other vector store configurations, only `SEMANTIC` search is available. For more information, see Test a knowledge base.

Type: String

Valid Values: `HYBRID` | `SEMANTIC`

Required: No

rerankingConfiguration

Contains configurations for reranking the retrieved results. For more information, see Improve the relevance of query responses with a reranker model.

Type: VectorSearchRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedSearchBedrockRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for a Bedrock reranking model used in managed search.

Contents

modelConfiguration

The model configuration containing the model ARN for reranking.

Type: ManagedSearchBedrockRerankingModelConfiguration object

Required: Yes

metadataConfiguration

The metadata configuration for reranking.

Type: MetadataConfigurationForReranking object

Required: No

numberOfRerankedResults

The number of results to return after reranking.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedSearchBedrockRerankingModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Model configuration for a Bedrock reranking model used in managed search.

Contents

modelArn

The ARN of the Bedrock reranking model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/(.*))?`

Required: Yes

additionalModelRequestFields

Additional request fields to pass to the reranking model.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedSearchConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for managed search in a knowledge base. Managed search automatically determines the best search strategy based on your data store configuration.

Contents

filter

Filters the metadata of the retrieved results so that Amazon Bedrock returns only results that match the filter.

Type: RetrievalFilter object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

numberOfResults

The number of results to retrieve.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

rerankingConfiguration

Contains configurations for reranking the results retrieved from the managed search.

Type: ManagedSearchRerankingConfiguration object

Required: No

rerankingModelType

The type of reranking model to use when reranking results retrieved from the managed search. Use `CUSTOM` to specify a model, `MANAGED` to use the service default, or `NONE` to disable reranking.

Type: String

Valid Values: `CUSTOM | MANAGED | NONE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ManagedSearchRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for the reranking model used in managed search.

Contents

type

The type of reranking configuration.

Type: String

Valid Values: `BEDROCK_RERANKING_MODEL`

Required: Yes

bedrockRerankingConfiguration

The Bedrock reranking model configuration for managed search.

Type: ManagedSearchBedrockRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Memory

Service: Agents for Amazon Bedrock Runtime

Contains sessions summaries.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

sessionSummary

Contains summary of a session.

Type: MemorySessionSummary object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MemorySessionSummary

Service: Agents for Amazon Bedrock Runtime

Contains details of a session summary.

Contents

memoryId

The unique identifier of the memory where the session summary is stored.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

sessionExpiryTime

The time when the memory duration for the session is set to end.

Type: Timestamp

Required: No

sessionId

The identifier for this session.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

sessionStartTime

The start time for this session.

Type: Timestamp

Required: No

summaryText

The summarized text for this session.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Message

Service: Agents for Amazon Bedrock Runtime

Details about a message.

Contents

content

The message's content.

Type: Array of ContentBlock objects

Required: Yes

role

The message's role.

Type: String

Valid Values: `user | assistant`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Metadata

Service: Agents for Amazon Bedrock Runtime

Provides information about the execution process for different types of invocations, such as model invocation, knowledge base invocation, agent collaborator invocation, guardrail invocation, and code interpreter Invocation.

Contents

clientRequestId

A unique identifier associated with the downstream invocation. This ID can be used for tracing, debugging, and identifying specific invocations in customer logs or systems.

Type: String

Required: No

endTime

In the final response, `endTime` is the end time of the agent invocation operation.

Type: Timestamp

Required: No

operationTotalTimeMs

The total time it took for the agent to complete execution. This field is only set for the final response.

Type: Long

Required: No

startTime

In the final response, `startTime` is the start time of the agent invocation operation.

Type: Timestamp

Required: No

totalTimeMs

The total execution time for the specific invocation being processed (model, knowledge base, guardrail, agent collaborator, or code interpreter). It represents how long the individual invocation took.

Type: Long

Required: No

usage

Specific to model invocation and contains details about the usage of a foundation model.

Type: Usage object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataAttributeSchema

Service: Agents for Amazon Bedrock Runtime

Details about a metadata attribute.

Contents

description

The attribute's description.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\s\S]+`

Required: Yes

key

The attribute's key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[\s\S]+`

Required: Yes

type

The attribute's type.

Type: String

Valid Values: `STRING | NUMBER | BOOLEAN | STRING_LIST`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetadataConfigurationForReranking

Service: Agents for Amazon Bedrock Runtime

Contains configurations for the metadata to use in reranking.

Contents

selectionMode

Specifies whether to consider all metadata when reranking, or only the metadata that you select. If you specify `SELECTIVE`, include the `selectiveModeConfiguration` field.

Type: String

Valid Values: `SELECTIVE | ALL`

Required: Yes

selectiveModeConfiguration

Contains configurations for the metadata fields to include or exclude when considering reranking.

Type: RerankingMetadataSelectiveModeConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelInvocationInput

Service: Agents for Amazon Bedrock Runtime

The input for the pre-processing step.

- The `type` matches the agent step.
- The `text` contains the prompt.
- The `inferenceConfiguration`, `parserMode`, and `overrideLambda` values are set in the PromptOverrideConfiguration object that was set when the agent was created or updated.

Contents

foundationModel

The identifier of a foundation model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+))$|(^arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{1,20}):(|[0-9]{12}):inference-profile/[a-zA-Z0-9-:.]+)`

Required: No

inferenceConfiguration

Specifications about the inference parameters that were provided alongside the prompt. These are specified in the PromptOverrideConfiguration object that was set when the agent was created or updated. For more information, see Inference parameters for foundation models.

Type: InferenceConfiguration object

Required: No

overrideLambda

The ARN of the Lambda function to use when parsing the raw foundation model output in parts of the agent sequence.

Type: String

Required: No

parserMode

Specifies whether to override the default parser Lambda function when parsing the raw foundation model output in the part of the agent sequence defined by the `promptType`.

Type: String

Valid Values: `DEFAULT | OVERRIDDEN`

Required: No

promptCreationMode

Specifies whether the default prompt template was `OVERRIDDEN`. If it was, the `basePromptTemplate` that was set in the PromptOverrideConfiguration object when the agent was created or updated is used instead.

Type: String

Valid Values: `DEFAULT | OVERRIDDEN`

Required: No

text

The text that prompted the agent at this step.

Type: String

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

type

The step in the agent sequence.

Type: String

Valid Values: `PRE_PROCESSING | ORCHESTRATION | KNOWLEDGE_BASE_RESPONSE_GENERATION | POST_PROCESSING | ROUTING_CLASSIFIER`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelPerformanceConfiguration

Service: Agents for Amazon Bedrock Runtime

The performance configuration for a model called with InvokeFlow.

Contents

performanceConfig

The latency configuration for the model.

Type: PerformanceConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeActionEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about an action (operation) called by a node during execution.

Contents

nodeName

The name of the node that called the operation.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

operationName

The name of the operation that the node called.

Type: String

Required: Yes

requestId

The ID of the request that the node made to the operation.

Type: String

Required: Yes

serviceName

The name of the service that the node called.

Type: String

Required: Yes

timestamp

The date and time that the operation was called.

Type: Timestamp

Required: Yes

operationRequest

The request payload sent to the downstream service.

Type: JSON value

Required: No

operationResponse

The response payload received from the downstream service.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeDependencyEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about an internal trace of a specific node during execution.

Contents

nodeName

The name of the node that generated the dependency trace.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The date and time that the dependency trace was generated.

Type: Timestamp

Required: Yes

traceElements

The trace elements containing detailed information about the node execution.

Type: NodeTraceElements object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeExecutionContent

Service: Agents for Amazon Bedrock Runtime

Contains the content of a flow node's input or output field for a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

document

The document content of the field, which can contain text or structured data.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeFailureEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about a failure that occurred at a specific node during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

errorCode

The error code that identifies the type of failure that occurred at the node.

Type: String

Valid Values: `VALIDATION | DEPENDENCY_FAILED | BAD_GATEWAY | INTERNAL_SERVER`

Required: Yes

errorMessage

A descriptive message that provides details about the node failure.

Type: String

Required: Yes

nodeName

The name of the node where the failure occurred.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The timestamp when the node failure occurred.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeInputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about the inputs provided to a specific node during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

fields

A list of input fields provided to the node.

Type: Array of NodeInputField objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

nodeName

The name of the node that received the inputs.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The timestamp when the inputs were provided to the node.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeInputExecutionChainItem

Service: Agents for Amazon Bedrock Runtime

Represents an item in the execution chain for node input tracking.

Contents

nodeName

The name of the node in the execution chain.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

type

The type of execution chain item. Supported values are Iterator and Loop.

Type: String

Valid Values: `Iterator | Loop`

Required: Yes

index

The index position of this item in the execution chain.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeInputField

Service: Agents for Amazon Bedrock Runtime

Represents an input field provided to a node during a flow execution.

Contents

content

The content of the input field, which can contain text or structured data.

Type: NodeExecutionContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

name

The name of the input field as defined in the node's input schema.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

category

The category of the input field.

Type: String

Valid Values: `LoopCondition | ReturnValueToLoopStart | ExitLoop`

Required: No

executionChain

The execution path through nested nodes like iterators and loops.

Type: Array of NodeInputExecutionChainItem objects

Required: No

source

The source node that provides input data to this field.

Type: NodeInputSource object

Required: No

type

The data type of the input field for compatibility validation.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeInputSource

Service: Agents for Amazon Bedrock Runtime

Represents the source of input data for a node field.

Contents

expression

The expression used to extract data from the source.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Required: Yes

nodeName

The name of the source node that provides the input data.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

outputFieldName

The name of the output field from the source node.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeOutputEvent

Service: Agents for Amazon Bedrock Runtime

Contains information about the outputs produced by a specific node during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

fields

A list of output fields produced by the node.

Type: Array of NodeOutputField objects

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Required: Yes

nodeName

The name of the node that produced the outputs.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

timestamp

The timestamp when the outputs were produced by the node.

Type: Timestamp

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeOutputField

Service: Agents for Amazon Bedrock Runtime

Represents an output field produced by a node during a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

content

The content of the output field, which can contain text or structured data.

Type: NodeExecutionContent object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

name

The name of the output field as defined in the node's output schema.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

next

The next node that receives output data from this field.

Type: Array of NodeOutputNext objects

Required: No

type

The data type of the output field for compatibility validation.

Type: String

Valid Values: `String | Number | Boolean | Object | Array`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeOutputNext

Service: Agents for Amazon Bedrock Runtime

Represents the next node that receives output data.

Contents

inputFieldName

The name of the input field in the next node that receives the data.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){1,50}`

Required: Yes

nodeName

The name of the next node that receives the output data.

Type: String

Pattern: `[a-zA-Z]([_]?[0-9a-zA-Z]){0,99}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NodeTraceElements

Service: Agents for Amazon Bedrock Runtime

Contains trace elements for node execution tracking.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

agentTraces

Agent trace information for the node execution.

Type: Array of TracePart objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Observation

Service: Agents for Amazon Bedrock Runtime

Contains the result or output of an action group or knowledge base, or the response to the user.

Contents

actionGroupInvocationOutput

Contains the JSON-formatted string returned by the API invoked by the action group.

Type: ActionGroupInvocationOutput object

Required: No

agentCollaboratorInvocationOutput

A collaborator's invocation output.

Type: AgentCollaboratorInvocationOutput object

Required: No

codeInterpreterInvocationOutput

Contains the JSON-formatted string returned by the API invoked by the code interpreter.

Type: CodeInterpreterInvocationOutput object

Required: No

finalResponse

Contains details about the response to the user.

Type: FinalResponse object

Required: No

knowledgeBaseLookupOutput

Contains details about the results from looking up the knowledge base.

Type: KnowledgeBaseLookupOutput object

Required: No

repromptResponse

Contains details about the response to reprompt the input.

Type: RepromptResponse object

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

type

Specifies what kind of information the agent returns in the observation. The following values are possible.

- `ACTION_GROUP` – The agent returns the result of an action group.
- `KNOWLEDGE_BASE` – The agent returns information from a knowledge base.
- `FINISH` – The agent returns a final response to the user with no follow-up.
- `ASK_USER` – The agent asks the user a question.
- `REPROMPT` – The agent prompts the user again for the same information.

Type: String

Valid Values: `ACTION_GROUP | AGENT_COLLABORATOR | KNOWLEDGE_BASE | FINISH | ASK_USER | REPROMPT`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizedPrompt

Service: Agents for Amazon Bedrock Runtime

Contains information about the optimized prompt.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

textPrompt

Contains information about the text in the prompt that was optimized.

Type: TextPrompt object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizedPromptEvent

Service: Agents for Amazon Bedrock Runtime

An event in which the prompt was optimized.

Contents

optimizedPrompt

Contains information about the optimized prompt.

Type: OptimizedPrompt object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OptimizedPromptStream

Service: Agents for Amazon Bedrock Runtime

The stream containing events in the prompt optimization process.

Contents

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

Required: No

analyzePromptEvent

An event in which the prompt was analyzed in preparation for optimization.

Type: AnalyzePromptEvent object

Required: No

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

Required: No

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

optimizedPromptEvent

An event in which the prompt was optimized.

Type: OptimizedPromptEvent object

Required: No

throttlingException

Your request was throttled because of service-wide limitations. Resubmit your request later or in a different region. You can also purchase Provisioned Throughput to increase the rate or number of tokens you can process.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OrchestrationConfiguration

Service: Agents for Amazon Bedrock Runtime

Settings for how the model processes the prompt prior to retrieval and generation.

Contents

additionalModelRequestFields

Additional model parameters and corresponding values not included in the textInferenceConfig structure for a knowledge base. This allows users to provide custom model parameters specific to the language model being used.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

inferenceConfig

Configuration settings for inference when using RetrieveAndGenerate to generate responses while using a knowledge base as a source.

Type: InferenceConfig object

Required: No

performanceConfig

The latency configuration for the model.

Type: PerformanceConfiguration object

Required: No

promptTemplate

Contains the template for the prompt that's sent to the model. Orchestration prompts must include the `$conversation_history$` and `$output_format_instructions$` variables. For more information, see Use placeholder variables in the user guide.

Type: PromptTemplate object

Required: No

queryTransformationConfiguration

To split up the prompt and retrieve multiple sources, set the transformation type to `QUERY_DECOMPOSITION`.

Type: QueryTransformationConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OrchestrationExecutor

Service: Agents for Amazon Bedrock Runtime

The structure of the executor invoking the actions in custom orchestration.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

lambda

The Amazon Resource Name (ARN) of the Lambda function containing the business logic that is carried out upon invoking the action.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OrchestrationModelInvocationOutput

Service: Agents for Amazon Bedrock Runtime

The foundation model output from the orchestration step.

Contents

metadata

Contains information about the foundation model output from the orchestration step.

Type: Metadata object

Required: No

rawResponse

Contains details of the raw response from the foundation model output.

Type: RawResponse object

Required: No

reasoningContent

Contains content about the reasoning that the model made during the orchestration step.

Type: ReasoningContentBlock object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OrchestrationTrace

Service: Agents for Amazon Bedrock Runtime

Details about the orchestration step, in which the agent determines the order in which actions are executed and which knowledge bases are retrieved.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

invocationInput

Contains information pertaining to the action group or knowledge base that is being invoked.

Type: InvocationInput object

Required: No

modelInvocationInput

The input for the orchestration step.

- The `type` is `ORCHESTRATION`.
- The `text` contains the prompt.
- The `inferenceConfiguration`, `parserMode`, and `overrideLambda` values are set in the PromptOverrideConfiguration object that was set when the agent was created or updated.

Type: ModelInvocationInput object

Required: No

modelInvocationOutput

Contains information pertaining to the output from the foundation model that is being invoked.

Type: OrchestrationModelInvocationOutput object

Required: No

observation

Details about the observation (the output of the action group Lambda or knowledge base) made by the agent.

Type: Observation object

Required: No

rationale

Details about the reasoning, based on the input, that the agent uses to justify carrying out an action group or getting information from a knowledge base.

Type: Rationale object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputFile

Service: Agents for Amazon Bedrock Runtime

Contains details of the response from code interpreter.

Contents

bytes

The byte count of files that contains response from code interpreter.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

name

The name of the file containing response from code interpreter.

Type: String

Required: No

type

The type of file that contains response from the code interpreter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Parameter

Service: Agents for Amazon Bedrock Runtime

A parameter for the API request or function.

Contents

name

The name of the parameter.

Type: String

Required: No

type

The type of the parameter.

Type: String

Required: No

value

The value of the parameter.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ParameterDetail

Service: Agents for Amazon Bedrock Runtime

Contains details about a parameter in a function for an action group.

Contents

type

The data type of the parameter.

Type: String

Valid Values: `string | number | integer | boolean | array`

Required: Yes

description

A description of the parameter. Helps the foundation model determine how to elicit the parameters from the user.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 500.

Required: No

required

Whether the parameter is required for the agent to complete the function for action group invocation.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

PayloadPart

Service: Agents for Amazon Bedrock Runtime

Contains a part of an agent response and citations for it.

Contents

attribution

Contains citations for a part of an agent response.

Type: Attribution object

Required: No

bytes

A part of the agent response in bytes.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PerformanceConfiguration

Service: Agents for Amazon Bedrock Runtime

Performance settings for a model.

Contents

latency

To use a latency-optimized version of the model, set to `optimized`.

Type: String

Valid Values: `standard | optimized`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PostProcessingModelInvocationOutput

Service: Agents for Amazon Bedrock Runtime

The foundation model output from the post-processing step.

Contents

metadata

Contains information about the foundation model output from the post-processing step.

Type: Metadata object

Required: No

parsedResponse

Details about the response from the Lambda parsing of the output of the post-processing step.

Type: PostProcessingParsedResponse object

Required: No

rawResponse

Details of the raw response from the foundation model output.

Type: RawResponse object

Required: No

reasoningContent

Contains content about the reasoning that the model made during the post-processing step.

Type: ReasoningContentBlock object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PostProcessingParsedResponse

Service: Agents for Amazon Bedrock Runtime

Details about the response from the Lambda parsing of the output from the post-processing step.

Contents

text

The text returned by the parser.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PostProcessingTrace

Service: Agents for Amazon Bedrock Runtime

Details about the post-processing step, in which the agent shapes the response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

modelInvocationInput

The input for the post-processing step.

- The `type` is `POST_PROCESSING`.
- The `text` contains the prompt.
- The `inferenceConfiguration`, `parserMode`, and `overrideLambda` values are set in the PromptOverrideConfiguration object that was set when the agent was created or updated.

Type: ModelInvocationInput object

Required: No

modelInvocationOutput

The foundation model output from the post-processing step.

Type: PostProcessingModelInvocationOutput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

PreProcessingModelInvocationOutput

Service: Agents for Amazon Bedrock Runtime

The foundation model output from the pre-processing step.

Contents

metadata

Contains information about the foundation model output from the pre-processing step.

Type: Metadata object

Required: No

parsedResponse

Details about the response from the Lambda parsing of the output of the pre-processing step.

Type: PreProcessingParsedResponse object

Required: No

rawResponse

Details of the raw response from the foundation model output.

Type: RawResponse object

Required: No

reasoningContent

Contains content about the reasoning that the model made during the pre-processing step.

Type: ReasoningContentBlock object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

traceId

The unique identifier of the trace.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PreProcessingParsedResponse

Service: Agents for Amazon Bedrock Runtime

Details about the response from the Lambda parsing of the output from the pre-processing step.

Contents

isValid

Whether the user input is valid or not. If `false`, the agent doesn't proceed to orchestration.

Type: Boolean

Required: No

rationale

The text returned by the parsing of the pre-processing step, explaining the steps that the agent plans to take in orchestration, if the user input is valid.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PreProcessingTrace

Service: Agents for Amazon Bedrock Runtime

Details about the pre-processing step, in which the agent contextualizes and categorizes user inputs.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

modelInvocationInput

The input for the pre-processing step.

- The `type` is `PRE_PROCESSING`.
- The `text` contains the prompt.
- The `inferenceConfiguration`, `parserMode`, and `overrideLambda` values are set in the PromptOverrideConfiguration object that was set when the agent was created or updated.

Type: ModelInvocationInput object

Required: No

modelInvocationOutput

The foundation model output from the pre-processing step.

Type: PreProcessingModelInvocationOutput object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations to override a prompt template in one part of an agent sequence. For more information, see Advanced prompts.

Contents

additionalModelRequestFields

If the Converse or ConverseStream operations support the model, `additionalModelRequestFields` contains additional inference parameters, beyond the base set of inference parameters in the `inferenceConfiguration` field.

For more information, see Inference request parameters and response fields for foundation models.

Type: JSON value

Required: No

basePromptTemplate

Defines the prompt template with which to replace the default prompt template. You can use placeholder variables in the base prompt template to customize the prompt. For more information, see Prompt template placeholder variables. For more information, see Configure the prompt templates.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: No

foundationModel

The foundation model to use.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `.*(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/`

```
[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}
([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}
[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|
(([0-9a-zA-Z][_-]?)+))$|(^arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|
[0-9a-z-]{1,20}):(|[0-9]{12}):inference-profile/[a-zA-Z0-9-:.]+)
```

Required: No

inferenceConfiguration

Contains inference parameters to use when the agent invokes a foundation model in the part of the agent sequence defined by the `promptType`. For more information, see Inference parameters for foundation models.

Type: InferenceConfiguration object

Required: No

parserMode

Specifies whether to override the default parser Lambda function when parsing the raw foundation model output in the part of the agent sequence defined by the `promptType`. If you set the field as `OVERRIDDEN`, the `overrideLambda` field in the PromptOverrideConfiguration must be specified with the ARN of a Lambda function.

Type: String

Valid Values: `DEFAULT | OVERRIDDEN`

Required: No

promptCreationMode

Specifies whether to override the default prompt template for this `promptType`. Set this value to `OVERRIDDEN` to use the prompt that you provide in the `basePromptTemplate`. If you leave it as `DEFAULT`, the agent uses a default prompt template.

Type: String

Valid Values: `DEFAULT | OVERRIDDEN`

Required: No

promptState

Specifies whether to allow the inline agent to carry out the step specified in the `promptType`. If you set this value to `DISABLED`, the agent skips that step. The default state for each `promptType` is as follows.

- `PRE_PROCESSING` – `ENABLED`
- `ORCHESTRATION` – `ENABLED`
- `KNOWLEDGE_BASE_RESPONSE_GENERATION` – `ENABLED`
- `POST_PROCESSING` – `DISABLED`

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

promptType

The step in the agent sequence that this prompt configuration applies to.

Type: String

Valid Values: `PRE_PROCESSING | ORCHESTRATION | KNOWLEDGE_BASE_RESPONSE_GENERATION | POST_PROCESSING | ROUTING_CLASSIFIER`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptCreationConfigurations

Service: Agents for Amazon Bedrock Runtime

Specifies parameters that control how the service populates the agent prompt for an `InvokeAgent` or `InvokeInlineAgent` request. You can control which aspects of previous invocations in the same agent session the service uses to populate the agent prompt. This gives you more granular control over the contextual history that is used to process the current request.

Contents

excludePreviousThinkingSteps

If `true`, the service removes any content between `<thinking>` tags from previous conversations in an agent session. The service will only remove content from already processed turns. This helps you remove content which might not be useful for current and subsequent invocations. This can reduce the input token count and potentially save costs. The default value is `false`.

Type: Boolean

Required: No

previousConversationTurnsToInclude

The number of previous conversations from the ongoing agent session to include in the conversation history of the agent prompt, during the current invocation. This gives you more granular control over the context that the model is made aware of, and helps the model remove older context which is no longer useful during the ongoing agent session.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptOverrideConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations to override prompts in different parts of an agent sequence. For more information, see Advanced prompts.

Contents

promptConfigurations

Contains configurations to override a prompt template in one part of an agent sequence. For more information, see Advanced prompts.

Type: Array of PromptConfiguration objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

overrideLambda

The ARN of the Lambda function to use when parsing the raw foundation model output in parts of the agent sequence. If you specify this field, at least one of the `promptConfigurations` must contain a `parserMode` value that is set to `OVERRIDDEN`. For more information, see Parser Lambda function in Amazon Bedrock Agents.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:(aws[a-zA-Z-]*)?:lambda:[a-z]{2}(-gov)?-[a-z]+-\d{1}:\d{12}:function:[a-zA-Z0-9-_\.]+(:(\$LATEST|[a-zA-Z0-9-_]+))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptTemplate

Service: Agents for Amazon Bedrock Runtime

Contains the template for the prompt that's sent to the model for response generation. For more information, see Knowledge base prompt templates.

This data type is used in the following API operations:

- RetrieveAndGenerate request – in the `filter` field

Contents

textPromptTemplate

The template for the prompt that's sent to the model for response generation. You can include prompt placeholders, which become replaced before the prompt is sent to the model to provide instructions and context to the model. In addition, you can include XML tags to delineate meaningful sections of the prompt template.

For more information, see the following resources:

- Knowledge base prompt templates
- Use XML tags with Anthropic Claude models

Type: String

Length Constraints: Minimum length of 1. Maximum length of 4000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PropertyParameters

Service: Agents for Amazon Bedrock Runtime

Contains the parameters in the request body.

Contents

properties

A list of parameters in the request body.

Type: Array of Parameter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryGenerationInput

Service: Agents for Amazon Bedrock Runtime

Contains information about a natural language query to transform into SQL.

Contents

text

The text of the query.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20000.

Pattern: `(?!\s*$).+.*`

Required: Yes

type

The type of the query.

Type: String

Valid Values: TEXT

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueryTransformationConfiguration

Service: Agents for Amazon Bedrock Runtime

To split up the prompt and retrieve multiple sources, set the transformation type to `QUERY_DECOMPOSITION`.

Contents

type

The type of transformation to apply to the prompt.

Type: String

Valid Values: `QUERY_DECOMPOSITION`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Rationale

Service: Agents for Amazon Bedrock Runtime

Contains the reasoning, based on the input, that the agent uses to justify carrying out an action group or getting information from a knowledge base.

Contents

text

The reasoning or thought process of the agent, based on the input.

Type: String

Required: No

traceId

The unique identifier of the trace step.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RawResponse

Service: Agents for Amazon Bedrock Runtime

Contains the raw output from the foundation model.

Contents

content

The foundation model's raw output content.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReasoningContentBlock

Service: Agents for Amazon Bedrock Runtime

Contains content regarding the reasoning that the foundation model made with respect to the content in the content block. Reasoning refers to a Chain of Thought (CoT) that the model generates to enhance the accuracy of its final response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

reasoningText

Contains information about the reasoning that the model used to return the content in the content block.

Type: ReasoningTextBlock object

Required: No

redactedContent

The content in the reasoning that was encrypted by the model provider for trust and safety reasons.

Type: Base64-encoded binary data object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ReasoningTextBlock

Service: Agents for Amazon Bedrock Runtime

Contains information about the reasoning that the model used to return the content in the content block.

Contents

text

Text describing the reasoning that the model used to return the content in the content block.

Type: String

Required: Yes

signature

A hash of all the messages in the conversation to ensure that the content in the reasoning text block isn't tampered with. You must submit the signature in subsequent `Converse` requests, in addition to the previous messages. If the previous messages are tampered with, the response throws an error.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RepromptResponse

Service: Agents for Amazon Bedrock Runtime

Contains details about the agent's response to reprompt the input.

Contents

source

Specifies what output is prompting the agent to reprompt the input.

Type: String

Valid Values: `ACTION_GROUP | KNOWLEDGE_BASE | PARSER`

Required: No

text

The text reprompting the input.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RequestBody

Service: Agents for Amazon Bedrock Runtime

The parameters in the API request body.

Contents

content

The content in the request body.

Type: String to array of Parameter objects map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankDocument

Service: Agents for Amazon Bedrock Runtime

Contains information about a document to rerank. Choose the `type` to define and include the field that corresponds to the type.

Contents

type

The type of document to rerank.

Type: String

Valid Values: `TEXT | JSON`

Required: Yes

jsonDocument

Contains a JSON document to rerank.

Type: JSON value

Required: No

textDocument

Contains information about a text document to rerank.

Type: RerankTextDocument object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for reranking.

Contents

bedrockRerankingConfiguration

Contains configurations for an Amazon Bedrock reranker.

Type: BedrockRerankingConfiguration object

Required: Yes

type

The type of reranker that the configurations apply to.

Type: String

Valid Values: `BEDROCK_RERANKING_MODEL`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankingMetadataSelectiveModeConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for the metadata fields to include or exclude when considering reranking. If you include the `fieldsToExclude` field, the reranker ignores all the metadata fields that you specify. If you include the `fieldsToInclude` field, the reranker uses only the metadata fields that you specify and ignores all others. You can include only one of these fields.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

fieldsToExclude

An array of objects, each of which specifies a metadata field to exclude from consideration when reranking.

Type: Array of FieldForReranking objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

fieldsToInclude

An array of objects, each of which specifies a metadata field to include in consideration when reranking. The remaining metadata fields are ignored.

Type: Array of FieldForReranking objects

Array Members: Minimum number of 1 item. Maximum number of 100 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankQuery

Service: Agents for Amazon Bedrock Runtime

Contains information about a query to submit to the reranker model.

Contents

textQuery

Contains information about a text query.

Type: RerankTextDocument object

Required: Yes

type

The type of the query.

Type: String

Valid Values: TEXT

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankResult

Service: Agents for Amazon Bedrock Runtime

Contains information about a document that was reranked.

Contents

index

The original index of the document from the input sources array.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 1000.

Required: Yes

relevanceScore

The relevance score of the document.

Type: Float

Required: Yes

document

Contains information about the document.

Type: RerankDocument object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankSource

Service: Agents for Amazon Bedrock Runtime

Contains information about a source for reranking.

Contents

inlineDocumentSource

Contains an inline definition of a source for reranking.

Type: RerankDocument object

Required: Yes

type

The type of the source.

Type: String

Valid Values: `INLINE`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RerankTextDocument

Service: Agents for Amazon Bedrock Runtime

Contains information about a text document to rerank.

Contents

text

The text of the document.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 32000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResponseStream

Service: Agents for Amazon Bedrock Runtime

The response from invoking the agent and associated citations and trace information.

Contents

accessDeniedException

The request is denied because of missing access permissions. Check your permissions and retry your request.

Type: Exception

HTTP Status Code: 403

Required: No

badGatewayException

There was an issue with a dependency due to a server issue. Retry your request.

Type: Exception

HTTP Status Code: 502

Required: No

chunk

Contains a part of an agent response and citations for it.

Type: PayloadPart object

Required: No

conflictException

There was a conflict performing an operation. Resolve the conflict and retry your request.

Type: Exception

HTTP Status Code: 409

Required: No

dependencyFailedException

There was an issue with a dependency. Check the resource configurations and retry the request.

Type: Exception

HTTP Status Code: 424

Required: No

files

Contains intermediate response for code interpreter if any files have been generated.

Type: FilePart object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

modelNotReadyException

The model specified in the request is not ready to serve Inference requests. The AWS SDK will automatically retry the operation up to 5 times. For information about configuring automatic retries, see Retry behavior in the *AWS SDKs and Tools* reference guide.

Type: Exception

HTTP Status Code: 424

Required: No

resourceNotFoundException

The specified resource Amazon Resource Name (ARN) was not found. Check the Amazon Resource Name (ARN) and try your request again.

Type: Exception

HTTP Status Code: 404

Required: No

returnControl

Contains the parameters and information that the agent elicited from the customer to carry out an action. This information is returned to the system and can be used in your own setup for fulfilling the action.

Type: ReturnControlPayload object

Required: No

serviceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

Type: Exception

HTTP Status Code: 400

Required: No

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception

HTTP Status Code: 429

Required: No

trace

Contains information about the agent and session, alongside the agent's reasoning process and results from calling actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace events.

Type: TracePart object

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalContent

Service: Agents for Amazon Bedrock Runtime

The content retrieved from a knowledge source.

Contents

mimeType

The MIME type of the retrieved content.

Type: String

Required: Yes

byteContent

The binary content of the retrieved item.

Type: Base64-encoded binary data object

Required: No

text

The text content of the retrieved item.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalFilter

Service: Agents for Amazon Bedrock Runtime

Specifies the filters to use on the metadata attributes in the knowledge base data sources before returning results. For more information, see Query configurations. See the examples below to see how to use these filters.

This data type is used in the following API operations:

- Retrieve request – in the `filter` field
- RetrieveAndGenerate request – in the `filter` field

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

andAll

Knowledge base data sources are returned if their metadata attributes fulfill all the filter conditions inside this list.

Type: Array of RetrievalFilter objects

Array Members: Minimum number of 2 items.

Required: No

equals

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value matches the `value` in this object.

The following example would return data sources with an `animal` attribute whose value is `cat`:

```
"equals": { "key": "animal", "value": "cat" }
```

Type: FilterAttribute object

Required: No

greaterThan

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is greater than the `value` in this object.

The following example would return data sources with an `year` attribute whose value is greater than 1989:

```
"greaterThan": { "key": "year", "value": 1989 }
```

Type: FilterAttribute object

Required: No

greaterThanOrEquals

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is greater than or equal to the `value` in this object.

The following example would return data sources with an `year` attribute whose value is greater than or equal to 1989:

```
"greaterThanOrEquals": { "key": "year", "value": 1989 }
```

Type: FilterAttribute object

Required: No

in

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is in the list specified in the `value` in this object.

The following example would return data sources with an `animal` attribute that is either `cat` or dog:

```
"in": { "key": "animal", "value": ["cat", "dog"] }
```

Type: FilterAttribute object

Required: No

lessThan

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is less than the `value` in this object.

The following example would return data sources with an `year` attribute whose value is less than to 1989.

```
"lessThan": { "key": "year", "value": 1989 }
```

Type: FilterAttribute object

Required: No

lessThanOrEquals

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is less than or equal to the `value` in this object.

The following example would return data sources with an `year` attribute whose value is less than or equal to 1989.

```
"lessThanOrEquals": { "key": "year", "value": 1989 }
```

Type: FilterAttribute object

Required: No

listContains

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is a list that contains the `value` as one of its members.

The following example would return data sources with an `animals` attribute that is a list containing a `cat` member (for example `["dog", "cat"]`).

```
"listContains": { "key": "animals", "value": "cat" }
```

Type: FilterAttribute object

Required: No

notEquals

Knowledge base data sources are returned when:

- It contains a metadata attribute whose name matches the `key` and whose value doesn't match the `value` in this object.
- The key is not present in the document.

The following example would return data sources that don't contain an `animal` attribute whose value is `cat`.

```
"notEquals": { "key": "animal", "value": "cat" }
```

Type: FilterAttribute object

Required: No

notIn

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value isn't in the list specified in the `value` in this object.

The following example would return data sources whose `animal` attribute is neither `cat` nor dog.

```
"notIn": { "key": "animal", "value": ["cat", "dog"] }
```

Type: FilterAttribute object

Required: No

orAll

Knowledge base data sources are returned if their metadata attributes fulfill at least one of the filter conditions inside this list.

Type: Array of RetrievalFilter objects

Array Members: Minimum number of 2 items.

Required: No

startsWith

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value starts with the `value` in this object. This filter is currently only supported for Amazon OpenSearch Serverless vector stores.

The following example would return data sources with an `animal` attribute starts with `ca` (for example, `cat` or `camel`).

```
"startsWith": { "key": "animal", "value": "ca" }
```

Type: FilterAttribute object

Required: No

stringContains

Knowledge base data sources are returned if they contain a metadata attribute whose name matches the `key` and whose value is one of the following:

- A string that contains the `value` as a substring. The following example would return data sources with an `animal` attribute that contains the substring `at` (for example `cat`).

  ```
  "stringContains": { "key": "animal", "value": "at" }
  ```

- A list with a member that contains the `value` as a substring. The following example would return data sources with an `animals` attribute that is a list containing a member that contains the substring `at` (for example `["dog", "cat"]`).

  ```
  "stringContains": { "key": "animals", "value": "at" }
  ```

Type: FilterAttribute object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalOverrides

Service: Agents for Amazon Bedrock Runtime

Overrides for retrieval behavior.

Contents

filter

A filter to apply to the retrieval results.

Type: RetrievalFilter object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

maxNumberOfResults

The maximum number of results to return.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultConfluenceLocation

Service: Agents for Amazon Bedrock Runtime

The Confluence data source location.

Contents

url

The Confluence host URL for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultContent

Service: Agents for Amazon Bedrock Runtime

Contains information about a chunk of text from a data source in the knowledge base. If the result is from a structured data source, the cell in the database and the type of the value is also identified.

This data type is used in the following API operations:

- Retrieve response – in the `content` field
- RetrieveAndGenerate response – in the `content` field
- InvokeAgent response – in the `content` field

Contents

audio

Audio segment information when the retrieval result contains audio content.

Type: AudioSegment object

Required: No

byteContent

A data URI with base64-encoded content from the data source. The URI is in the following format: returned in the following format: `${base64-encoded string}`.

Type: String

Required: No

row

Specifies information about the rows with the cells to return in retrieval.

Type: Array of RetrievalResultContentColumn objects

Required: No

text

The cited text from the data source.

Type: String

Required: No

type

The type of content in the retrieval result.

Type: String

Valid Values: `TEXT | IMAGE | ROW | AUDIO | VIDEO`

Required: No

video

Video segment information when the retrieval result contains video content.

Type: VideoSegment object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultContentColumn

Service: Agents for Amazon Bedrock Runtime

Contains information about a column with a cell to return in retrieval.

Contents

columnName

The name of the column.

Type: String

Required: No

columnValue

The value in the column.

Type: String

Required: No

type

The data type of the value.

Type: String

Valid Values: `BLOB | BOOLEAN | DOUBLE | NULL | LONG | STRING`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultCustomDocumentLocation

Service: Agents for Amazon Bedrock Runtime

Contains information about the location of a document in a custom data source.

Contents

id

The ID of the document.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultGoogleDriveLocation

Service: Agents for Amazon Bedrock Runtime

The Google Drive location of a retrieval result.

Contents

url

The Google Drive URL for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultKendraDocumentLocation

Service: Agents for Amazon Bedrock Runtime

The location of a result in Amazon Kendra.

Contents

uri

The document's uri.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultLocation

Service: Agents for Amazon Bedrock Runtime

Contains information about the data source location.

This data type is used in the following API operations:

- Retrieve response – in the `location` field
- RetrieveAndGenerate response – in the `location` field
- InvokeAgent response – in the `location` field

Contents

type

The type of data source location.

Type: String

Valid Values: `S3 | WEB | CONFLUENCE | SALESFORCE | SHAREPOINT | CUSTOM | KENDRA | SQL | ONEDRIVE | GOOGLEDRIVE`

Required: Yes

confluenceLocation

The Confluence data source location.

Type: RetrievalResultConfluenceLocation object

Required: No

customDocumentLocation

Specifies the location of a document in a custom data source.

Type: RetrievalResultCustomDocumentLocation object

Required: No

googleDriveLocation

The Google Drive data source location.

Type: RetrievalResultGoogleDriveLocation object

Required: No

kendraDocumentLocation

The location of a document in Amazon Kendra.

Type: RetrievalResultKendraDocumentLocation object

Required: No

oneDriveLocation

The Microsoft OneDrive data source location.

Type: RetrievalResultOneDriveLocation object

Required: No

s3Location

The S3 data source location.

Type: RetrievalResultS3Location object

Required: No

salesforceLocation

The Salesforce data source location.

Type: RetrievalResultSalesforceLocation object

Required: No

sharePointLocation

The SharePoint data source location.

Type: RetrievalResultSharePointLocation object

Required: No

sqlLocation

Specifies information about the SQL query used to retrieve the result.

Type: RetrievalResultSqlLocation object

Required: No

webLocation

The web URL/URLs data source location.

Type: RetrievalResultWebLocation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultOneDriveLocation

Service: Agents for Amazon Bedrock Runtime

The Microsoft OneDrive location of a retrieval result.

Contents

url

The OneDrive URL for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultS3Location

Service: Agents for Amazon Bedrock Runtime

The S3 data source location.

This data type is used in the following API operations:

- Retrieve response – in the `s3Location` field
- RetrieveAndGenerate response – in the `s3Location` field
- InvokeAgent response – in the `s3Location` field

Contents

uri

The S3 URI for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultSalesforceLocation

Service: Agents for Amazon Bedrock Runtime

The Salesforce data source location.

Contents

url

The Salesforce host URL for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultSharePointLocation

Service: Agents for Amazon Bedrock Runtime

The SharePoint data source location.

Contents

url

The SharePoint site URL for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultSqlLocation

Service: Agents for Amazon Bedrock Runtime

Contains information about the SQL query used to retrieve the result.

Contents

query

The SQL query used to retrieve the result.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievalResultWebLocation

Service: Agents for Amazon Bedrock Runtime

The web URL/URLs data source location.

Contents

url

The web URL/URLs for the data source location.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains details about the resource being queried.

This data type is used in the following API operations:

- RetrieveAndGenerate request – in the `retrieveAndGenerateConfiguration` field

Contents

type

The type of resource that contains your data for retrieving information and generating responses.

> **Note**
>
> If you choose to use `EXTERNAL_SOURCES`, then currently only Anthropic Claude 3 Sonnet models for knowledge bases are supported.

Type: String

Valid Values: `KNOWLEDGE_BASE | EXTERNAL_SOURCES`

Required: Yes

externalSourcesConfiguration

The configuration for the external source wrapper object in the `retrieveAndGenerate` function.

Type: ExternalSourcesRetrieveAndGenerateConfiguration object

Required: No

knowledgeBaseConfiguration

Contains details about the knowledge base for retrieving information and generating responses.

Type: KnowledgeBaseRetrieveAndGenerateConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateInput

Service: Agents for Amazon Bedrock Runtime

Contains the query made to the knowledge base.

This data type is used in the following API operations:

- RetrieveAndGenerate request – in the `input` field

Contents

text

The query made to the knowledge base, in characters.

Type: String

Length Constraints: Maximum length of 20000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateOutput

Service: Agents for Amazon Bedrock Runtime

Contains the response generated from querying the knowledge base.

This data type is used in the following API operations:

- RetrieveAndGenerate response – in the `output` field

Contents

text

The response generated from querying the knowledge base.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateOutputEvent

Service: Agents for Amazon Bedrock Runtime

A retrieve and generate output event.

Contents

text

A text response.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateSessionConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configuration about the session with the knowledge base.

This data type is used in the following API operations:

- RetrieveAndGenerate request – in the `sessionConfiguration` field

Contents

kmsKeyArn

The ARN of the AWS KMS key encrypting the session.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(|-cn|-us-gov):kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieveAndGenerateStreamResponseOutput

Service: Agents for Amazon Bedrock Runtime

A retrieve and generate stream response output.

Contents

accessDeniedException

The request is denied because you do not have sufficient permissions to perform the requested action. For troubleshooting this error, see AccessDeniedException in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 403

Required: No

badGatewayException

The request failed due to a bad gateway error.

Type: Exception

HTTP Status Code: 502

Required: No

citation

A citation event.

Type: CitationEvent object

Required: No

conflictException

Error occurred because of a conflict while performing an operation.

Type: Exception

HTTP Status Code: 409

Required: No

dependencyFailedException

The request failed due to a dependency error.

Type: Exception

HTTP Status Code: 424

Required: No

guardrail

A guardrail event.

Type: GuardrailEvent object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

output

An output event.

Type: RetrieveAndGenerateOutputEvent object

Required: No

resourceNotFoundException

The specified resource ARN was not found. For troubleshooting this error, see ResourceNotFound in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 404

Required: No

serviceQuotaExceededException

Your request exceeds the service quota for your account. You can view your quotas at Viewing service quotas. You can resubmit your request later.

Type: Exception

HTTP Status Code: 400

Required: No

throttlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrievedReference

Service: Agents for Amazon Bedrock Runtime

Contains metadata about a source cited for the generated response.

This data type is used in the following API operations:

- RetrieveAndGenerate response – in the `retrievedReferences` field
- InvokeAgent response – in the `retrievedReferences` field

Contents

content

Contains the cited text from the data source.

Type: RetrievalResultContent object

Required: No

location

Contains information about the location of the data source.

Type: RetrievalResultLocation object

Required: No

metadata

Contains metadata attributes and their values for the file in the data source. For more information, see Metadata and filtering.

Type: String to JSON value map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RetrieverConfiguration

Service: Agents for Amazon Bedrock Runtime

Configuration for a retriever, specified as a union of retriever types.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

knowledgeBase

Configuration for a knowledge base retriever.

Type: KnowledgeBaseRetrieverConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReturnControlPayload

Service: Agents for Amazon Bedrock Runtime

Contains information to return from the action group that the agent has predicted to invoke.

This data type is used in the following API operations:

- InvokeAgent response

Contents

invocationId

The identifier of the action group invocation.

Type: String

Required: No

invocationInputs

A list of objects that contain information about the parameters and inputs that need to be sent into the API operation or function, based on what the agent determines from its session with the user.

Type: Array of InvocationInputMember objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReturnControlResults

Service: Agents for Amazon Bedrock Runtime

An action invocation result.

Contents

invocationId

The action's invocation ID.

Type: String

Required: No

returnControlInvocationResults

The action invocation result.

Type: Array of InvocationResultMember objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RoutingClassifierModelInvocationOutput

Service: Agents for Amazon Bedrock Runtime

Invocation output from a routing classifier model.

Contents

metadata

The invocation's metadata.

Type: Metadata object

Required: No

rawResponse

The invocation's raw response.

Type: RawResponse object

Required: No

traceId

The invocation's trace ID.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 16.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

RoutingClassifierTrace

Service: Agents for Amazon Bedrock Runtime

A trace for a routing classifier.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

invocationInput

The classifier's invocation input.

Type: InvocationInput object

Required: No

modelInvocationInput

The classifier's model invocation input.

Type: ModelInvocationInput object

Required: No

modelInvocationOutput

The classifier's model invocation output.

Type: RoutingClassifierModelInvocationOutput object

Required: No

observation

The classifier's observation.

Type: Observation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Identifier

Service: Agents for Amazon Bedrock Runtime

The identifier information for an Amazon S3 bucket.

Contents

s3BucketName

The name of the S3 bucket.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Pattern: `[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9]`

Required: No

s3ObjectKey

The S3 object key for the S3 resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `[\.\-\!\*\_\'\(\)a-zA-Z0-9][\.\-\!\*\_\'\(\)\/a-zA-Z0-9]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Location

Service: Agents for Amazon Bedrock Runtime

Information about the Amazon S3 bucket where the image is stored.

Contents

uri

The path to the Amazon S3 bucket where the image is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]/.{1,1024}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3ObjectDoc

Service: Agents for Amazon Bedrock Runtime

The unique wrapper object of the document from the S3 location.

Contents

uri

The file location of the S3 wrapper object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]/.{1,1024}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3ObjectFile

Service: Agents for Amazon Bedrock Runtime

Contains details of the s3 object where the source file is located.

Contents

uri

The uri of the s3 object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][a-z0-9.-]{1,61}[a-z0-9]/.{1,1024}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SatisfiedCondition

Service: Agents for Amazon Bedrock Runtime

Represents a condition that was satisfied during a condition node evaluation in a flow execution.

> **Note**
>
> Flow executions is in preview release for Amazon Bedrock and is subject to change.

Contents

conditionName

The name of the condition that was satisfied.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SessionState

Service: Agents for Amazon Bedrock Runtime

Contains parameters that specify various attributes that persist across a session or prompt. You can define session state attributes as key-value pairs when writing a Lambda function for an action group or pass them when making an InvokeAgent request. Use session state attributes to control and provide conversational context for your agent and to help customize your agent's behavior. For more information, see Control session context.

Contents

conversationHistory

The state's conversation history.

Type: ConversationHistory object

Required: No

files

Contains information about the files used by code interpreter.

Type: Array of InputFile objects

Required: No

invocationId

The identifier of the invocation of an action. This value must match the `invocationId` returned in the `InvokeAgent` response for the action whose results are provided in the `returnControlInvocationResults` field. For more information, see Return control to the agent developer and Control session context.

Type: String

Required: No

knowledgeBaseConfigurations

An array of configurations, each of which applies to a knowledge base attached to the agent.

Type: Array of KnowledgeBaseConfiguration objects

Array Members: Minimum number of 1 item.

Required: No

promptSessionAttributes

Contains attributes that persist across a prompt and the values of those attributes.

- In orchestration prompt template, these attributes replace the $prompt_session_attributes$ placeholder variable. For more information, see Prompt template placeholder variables.
- In multi-agent collaboration, the `promptSessionAttributes` will only be used by supervisor agent when $prompt_session_attributes$ is present in prompt template.

Type: String to string map

Required: No

returnControlInvocationResults

Contains information about the results from the action group invocation. For more information, see Return control to the agent developer and Control session context.

> **Note**
>
> If you include this field, the `inputText` field will be ignored.

Type: Array of InvocationResultMember objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: No

sessionAttributes

Contains attributes that persist across a session and the values of those attributes. If `sessionAttributes` are passed to a supervisor agent in multi-agent collaboration, it will be forwarded to all agent collaborators.

Type: String to string map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SessionSummary

Service: Agents for Amazon Bedrock Runtime

Contains details about a session. For more information about sessions, see Store and retrieve conversation history and context with Amazon Bedrock sessions.

Contents

createdAt

The timestamp for when the session was created.

Type: Timestamp

Required: Yes

lastUpdatedAt

The timestamp for when the session was last modified.

Type: Timestamp

Required: Yes

sessionArn

The Amazon Resource Name (ARN) of the session.

Type: String

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]+:[0-9]{12}:session/[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

sessionId

The unique identifier for the session.

Type: String

Pattern: `[a-f0-9]{8}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{4}-[a-f0-9]{12}`

Required: Yes

sessionStatus

The current status of the session.

Type: String

Valid Values: `ACTIVE | EXPIRED | ENDED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Span

Service: Agents for Amazon Bedrock Runtime

Contains information about where the text with a citation begins and ends in the generated output.

This data type is used in the following API operations:

- RetrieveAndGenerate response – in the `span` field
- InvokeAgent response – in the `span` field

Contents

end

Where the text with a citation ends in the generated output.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

start

Where the text with a citation starts in the generated output.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StreamingConfigurations

Service: Agents for Amazon Bedrock Runtime

Configurations for streaming.

Contents

applyGuardrailInterval

The guardrail interval to apply as response is generated. By default, the guardrail interval is set to 50 characters. If a larger interval is specified, the response will be generated in larger chunks with fewer `ApplyGuardrail` calls. The following examples show the response generated for *Hello, I am an agent* input string.

Example response in chunks: Interval set to 3 characters

```
'Hel', 'lo, ','I am', ' an', ' Age', 'nt'
```

Each chunk has at least 3 characters except for the last chunk

Example response in chunks: Interval set to 20 or more characters

```
Hello, I am an Agent
```

Type: Integer

Valid Range: Minimum value of 1.

Required: No

streamFinalResponse

Specifies whether to enable streaming for the final response. This is set to `false` by default.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextInferenceConfig

Service: Agents for Amazon Bedrock Runtime

Configuration settings for text generation using a language model via the RetrieveAndGenerate operation. Includes parameters like temperature, top-p, maximum token count, and stop sequences.

> **Note**
>
> The valid range of `maxTokens` depends on the accepted values for your chosen model's inference parameters. To see the inference parameters for your model, see Inference parameters for foundation models.

Contents

maxTokens

The maximum number of tokens to generate in the output text. Do not use the minimum of 0 or the maximum of 65536. The limit values described here are arbitary values, for actual values consult the limits defined by your specific model.

Type: Integer

Valid Range: Minimum value of 0. Maximum value of 65536.

Required: No

stopSequences

A list of sequences of characters that, if generated, will cause the model to stop generating further tokens. Do not use a minimum length of 1 or a maximum length of 1000. The limit values described here are arbitary values, for actual values consult the limits defined by your specific model.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 4 items.

Length Constraints: Minimum length of 1. Maximum length of 1000.

Required: No

temperature

Controls the random-ness of text generated by the language model, influencing how much the model sticks to the most predictable next words versus exploring more surprising options. A lower temperature value (e.g. 0.2 or 0.3) makes model outputs more deterministic or predictable, while a higher temperature (e.g. 0.8 or 0.9) makes the outputs more creative or unpredictable.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topP

A probability distribution threshold which controls what the model considers for the set of possible next tokens. The model will only consider the top p% of the probability distribution when generating the next token.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextPrompt

Service: Agents for Amazon Bedrock Runtime

Contains information about the text prompt to optimize.

Contents

text

The text in the text prompt to optimize.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextResponsePart

Service: Agents for Amazon Bedrock Runtime

Contains the part of the generated text that contains a citation, alongside where it begins and ends.

This data type is used in the following API operations:

- RetrieveAndGenerate response – in the `textResponsePart` field
- InvokeAgent response – in the `textResponsePart` field

Contents

span

Contains information about where the text with a citation begins and ends in the generated output.

Type: Span object

Required: No

text

The part of the generated text that contains a citation.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextToSqlConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for transforming text to SQL.

Contents

type

The type of resource to use in transformation.

Type: String

Valid Values: `KNOWLEDGE_BASE`

Required: Yes

knowledgeBaseConfiguration

Specifies configurations for a knowledge base to use in transformation.

Type: TextToSqlKnowledgeBaseConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TextToSqlKnowledgeBaseConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for a knowledge base to use in transformation.

Contents

knowledgeBaseArn

The ARN of the knowledge base

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:knowledge-base/[0-9a-zA-Z]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Trace

Service: Agents for Amazon Bedrock Runtime

Contains one part of the agent's reasoning process and results from calling API actions and querying knowledge bases. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace enablement.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

customOrchestrationTrace

Details about the custom orchestration step in which the agent determines the order in which actions are executed.

Type: CustomOrchestrationTrace object

Required: No

failureTrace

Contains information about the failure of the interaction.

Type: FailureTrace object

Required: No

guardrailTrace

The trace details for a trace defined in the Guardrail filter.

Type: GuardrailTrace object

Required: No

orchestrationTrace

Details about the orchestration step, in which the agent determines the order in which actions are executed and which knowledge bases are retrieved.

Type: OrchestrationTrace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

postProcessingTrace

Details about the post-processing step, in which the agent shapes the response..

Type: PostProcessingTrace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

preProcessingTrace

Details about the pre-processing step, in which the agent contextualizes and categorizes user inputs.

Type: PreProcessingTrace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

routingClassifierTrace

A routing classifier's trace.

Type: RoutingClassifierTrace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

TraceElements

Service: Agents for Amazon Bedrock Runtime

Contains trace elements for flow execution tracking.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

agentTraces

Agent trace information for the flow execution.

Type: Array of TracePart objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TracePart

Service: Agents for Amazon Bedrock Runtime

Contains information about the agent and session, alongside the agent's reasoning process and results from calling API actions and querying knowledge bases and metadata about the trace. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace enablement.

Contents

agentAliasId

The unique identifier of the alias of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: No

agentId

The unique identifier of the agent.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 10.

Pattern: `[0-9a-zA-Z]+`

Required: No

agentVersion

The version of the agent.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 5.

Pattern: `(DRAFT|[0-9]{0,4}[1-9][0-9]{0,4})`

Required: No

callerChain

The part's caller chain.

Type: Array of Caller objects

Required: No

collaboratorName

The part's collaborator name.

Type: String

Pattern: `([0-9a-zA-Z][_-]?){1,100}`

Required: No

eventTime

The time of the trace.

Type: Timestamp

Required: No

sessionId

The unique identifier of the session with the agent.

Type: String

Length Constraints: Minimum length of 2. Maximum length of 100.

Pattern: `[0-9a-zA-Z._:-]+`

Required: No

trace

Contains one part of the agent's reasoning process and results from calling API actions and querying knowledge bases. You can use the trace to understand how the agent arrived at the response it provided the customer. For more information, see Trace enablement.

Type: Trace object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TransformationConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for transforming the natural language query into SQL.

Contents

mode

The mode of the transformation.

Type: String

Valid Values: `TEXT_TO_SQL`

Required: Yes

textToSqlConfiguration

Specifies configurations for transforming text to SQL.

Type: TextToSqlConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Usage

Service: Agents for Amazon Bedrock Runtime

Contains information of the usage of the foundation model.

Contents

inputTokens

Contains information about the input tokens from the foundation model usage.

Type: Integer

Required: No

outputTokens

Contains information about the output tokens from the foundation model usage.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UserContext

Service: Agents for Amazon Bedrock Runtime

Contains information about the user making the request. Use this to pass user identity information for access control filtering, so that retrieval results only include documents the user is authorized to access.

Contents

userId

The identifier of the user making the retrieval request.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchBedrockRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for reranking with an Amazon Bedrock reranker model.

Contents

modelConfiguration

Contains configurations for the reranker model.

Type: VectorSearchBedrockRerankingModelConfiguration object

Required: Yes

metadataConfiguration

Contains configurations for the metadata to use in reranking.

Type: MetadataConfigurationForReranking object

Required: No

numberOfRerankedResults

The number of results to return after reranking.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchBedrockRerankingModelConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for an Amazon Bedrock reranker model.

Contents

modelArn

The ARN of the reranker model to use.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/(.*))?`

Required: Yes

additionalModelRequestFields

A JSON object whose keys are request fields for the model and whose values are values for those fields.

Type: String to JSON value map

Key Length Constraints: Minimum length of 1. Maximum length of 100.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VectorSearchRerankingConfiguration

Service: Agents for Amazon Bedrock Runtime

Contains configurations for reranking the retrieved results.

Contents

type

The type of reranker model.

Type: String

Valid Values: `BEDROCK_RERANKING_MODEL`

Required: Yes

bedrockRerankingConfiguration

Contains configurations for an Amazon Bedrock reranker model.

Type: VectorSearchBedrockRerankingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoSegment

Service: Agents for Amazon Bedrock Runtime

Contains information about a video segment retrieved from a knowledge base, including its location and summary.

This data type is used in the following API operations:

- Retrieve response – in the `video` field

Contents

s3Uri

The S3 URI where this specific video segment is stored in the multimodal storage destination.

Type: String

Length Constraints: Minimum length of 1.

Required: Yes

summary

A text summary describing the content of the video segment.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Data Automation for Amazon Bedrock

The following data types are supported by Data Automation for Amazon Bedrock:

- AudioExtractionCategory
- AudioExtractionCategoryTypeConfiguration
- AudioLanguageConfiguration
- AudioOverrideConfiguration
- AudioStandardExtraction
- AudioStandardGenerativeField
- AudioStandardOutputConfiguration
- Blueprint
- BlueprintFilter
- BlueprintItem
- BlueprintOptimizationObject
- BlueprintOptimizationOutputConfiguration
- BlueprintOptimizationSample
- BlueprintSummary
- ChannelLabelingConfiguration
- CustomOutputConfiguration
- DataAutomationLibrary
- DataAutomationLibraryConfiguration
- DataAutomationLibraryEntitySummary
- DataAutomationLibraryFilter
- DataAutomationLibraryIngestionJob
- DataAutomationLibraryIngestionJobSummary
- DataAutomationLibraryItem
- DataAutomationLibrarySummary
- DataAutomationProject
- DataAutomationProjectFilter
- DataAutomationProjectSummary

- DeleteEntitiesInfo
- DocumentBoundingBox
- DocumentCustomOutputConfiguration
- DocumentExtractionGranularity
- DocumentOutputAdditionalFileFormat
- DocumentOutputFormat
- DocumentOutputTextFormat
- DocumentOverrideConfiguration
- DocumentStandardExtraction
- DocumentStandardGenerativeField
- DocumentStandardOutputConfiguration
- EncryptionConfiguration
- EntityDetails
- EntityTypeInfo
- EventBridgeConfiguration
- ImageBoundingBox
- ImageExtractionCategory
- ImageOverrideConfiguration
- ImageStandardExtraction
- ImageStandardGenerativeField
- ImageStandardOutputConfiguration
- InlinePayload
- InputConfiguration
- ModalityProcessingConfiguration
- ModalityRoutingConfiguration
- NotificationConfiguration
- OutputConfiguration
- OverrideConfiguration
- Phrase
- PIIEntitiesConfiguration

- S3Object
- SensitiveDataConfiguration
- SpeakerLabelingConfiguration
- SplitterConfiguration
- StandardOutputConfiguration
- Tag
- TranscriptConfiguration
- UpsertEntityInfo
- ValidationExceptionField
- VideoBoundingBox
- VideoExtractionCategory
- VideoOverrideConfiguration
- VideoStandardExtraction
- VideoStandardGenerativeField
- VideoStandardOutputConfiguration
- VocabularyEntity
- VocabularyEntityInfo
- VocabularyEntitySummary

AudioExtractionCategory

Service: Data Automation for Amazon Bedrock

Settings for generating data from audio.

Contents

state

Whether generating categorical data from audio is enabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

typeConfiguration

This element contains information about extractions from different types. Used to enable speaker and channel labeling for transcripts.

Type: AudioExtractionCategoryTypeConfiguration object

Required: No

types

The types of data to generate.

Type: Array of strings

Valid Values: `AUDIO_CONTENT_MODERATION | TRANSCRIPT | TOPIC_CONTENT_MODERATION`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioExtractionCategoryTypeConfiguration

Service: Data Automation for Amazon Bedrock

Allows configuration of extractions for different types of data, such as transcript and content moderation.

Contents

transcript

This element allows you to configure different extractions for your transcript data, such as speaker and channel labeling.

Type: TranscriptConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioLanguageConfiguration

Service: Data Automation for Amazon Bedrock

This allows you to set the input and output language of your audio. The input language can be set to any of the languages supported by Bedrock Data Automation. The output can either be set to english or whatever the dominant language is of the audio, determined by the language spoken for the most seconds.

Contents

generativeOutputLanguage

The output language of your processing results. This can either be set to `EN` (English) or `DEFAULT` which will output the results in the dominant language of the audio. The dominant language is determined as the language in the audio, spoken the longest in the input audio.

Type: String

Valid Values: `DEFAULT | EN`

Required: No

identifyMultipleLanguages

The toggle determining if you want to detect multiple languages from your audio.

Type: Boolean

Required: No

inputLanguages

The input language of your audio. This can be set to any of the currently supported languages via the language codes.

Type: Array of strings

Valid Values: `EN | DE | ES | FR | IT | PT | JA | KO | CN | TW | HK`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioOverrideConfiguration

Service: Data Automation for Amazon Bedrock

Sets whether your project will process audio or not.

Contents

languageConfiguration

The output and input language configuration for your audio.

Type: AudioLanguageConfiguration object

Required: No

modalityProcessing

Sets modality processing for audio files. All modalities are enabled by default.

Type: ModalityProcessingConfiguration object

Required: No

sensitiveDataConfiguration

Configuration for sensitive data detection and redaction for audio files.

Type: SensitiveDataConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioStandardExtraction

Service: Data Automation for Amazon Bedrock

Settings for generating data from audio.

Contents

category

Settings for generating categorical data.

Type: AudioExtractionCategory object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioStandardGenerativeField

Service: Data Automation for Amazon Bedrock

Settings for generating descriptions of audio.

Contents

state

Whether generating descriptions is enabled for audio.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

types

The types of description to generate.

Type: Array of strings

Valid Values: `AUDIO_SUMMARY | TOPIC_SUMMARY`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioStandardOutputConfiguration

Service: Data Automation for Amazon Bedrock

Output settings for processing audio.

Contents

extraction

Settings for populating data fields that describe the audio.

Type: AudioStandardExtraction object

Required: No

generativeField

Whether to generate descriptions of the data.

Type: AudioStandardGenerativeField object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Blueprint

Service: Data Automation for Amazon Bedrock

A data processing blueprint.

Contents

blueprintArn

The blueprint's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

blueprintName

The blueprint's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

creationTime

When the blueprint was created.

Type: Timestamp

Required: Yes

lastModifiedTime

When the blueprint was last updated.

Type: Timestamp

Required: Yes

schema

The blueprint's schema.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 100000.

Required: Yes

type

The blueprint's type.

Type: String

Valid Values: `DOCUMENT | IMAGE | AUDIO`

Required: Yes

blueprintStage

The blueprint's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

blueprintVersion

The blueprint's version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Required: No

kmsEncryptionContext

The AWS KMS encryption context to use for encryption.

Type: String to string map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 2000.

Key Pattern: `.*\S.*`

Value Length Constraints: Minimum length of 1. Maximum length of 2000.

Value Pattern: `.*\S.*`

Required: No

kmsKeyId

The AWS KMS key to use for encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[A-Za-z0-9][A-Za-z0-9:_/+=,@.-]+`

Required: No

optimizationSamples

Sample data used to optimize this blueprint.

Type: Array of BlueprintOptimizationSample objects

Required: No

optimizationTime

When the blueprint was last optimized.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueprintFilter

Service: Data Automation for Amazon Bedrock

A blueprint filter.

Contents

blueprintArn

Filter by blueprint ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

blueprintStage

Filter by blueprint stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

blueprintVersion

Filter by blueprint version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueprintItem

Service: Data Automation for Amazon Bedrock

An abbreviated summary of a blueprint.

Contents

blueprintArn

The blueprint's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

blueprintStage

The blueprint's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

blueprintVersion

The blueprint's version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueprintOptimizationObject

Service: Data Automation for Amazon Bedrock

A blueprint to be optimized.

Contents

blueprintArn

The blueprint's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

stage

The blueprint's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueprintOptimizationOutputConfiguration

Service: Data Automation for Amazon Bedrock

Configuration for where to store blueprint optimization results.

Contents

s3Object

The S3 location where optimization results will be stored.

Type: S3Object object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueprintOptimizationSample

Service: Data Automation for Amazon Bedrock

A sample asset and its corresponding ground truth for blueprint optimization.

Contents

assetS3Object

The S3 location of the sample asset.

Type: S3Object object

Required: Yes

groundTruthS3Object

The S3 location of the ground truth data for this asset.

Type: S3Object object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BlueprintSummary

Service: Data Automation for Amazon Bedrock

A summary of a blueprint.

Contents

blueprintArn

The blueprint's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

creationTime

When the blueprint was created.

Type: Timestamp

Required: Yes

blueprintName

The blueprint's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: No

blueprintStage

The blueprint's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

blueprintVersion

The blueprint's version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Required: No

lastModifiedTime

When the blueprint was last updated.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ChannelLabelingConfiguration

Service: Data Automation for Amazon Bedrock

Enables or disables channel labeling. Channel labeling, when enabled will assign a number to each audio channel, and indicate which channel is being used in each portion of the transcript. This appears in the response as "ch_0" for the first channel, and "ch_1" for the second.

Contents

state

State of channel labeling, either enabled or disabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomOutputConfiguration

Service: Data Automation for Amazon Bedrock

Blueprints to apply to objects processed by the project.

Contents

blueprints

A list of blueprints.

Type: Array of BlueprintItem objects

Required: No

document

Custom output configuration for document processing.

Type: DocumentCustomOutputConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibrary

Service: Data Automation for Amazon Bedrock

A data automation library.

Contents

creationTime

When the library was created.

Type: Timestamp

Required: Yes

libraryArn

The library's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

libraryName

The library's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

status

The library's status.

Type: String

Valid Values: `ACTIVE | DELETING`

Required: Yes

entityTypes

The entity types supported by the library.

Type: Array of EntityTypeInfo objects

Required: No

kmsEncryptionContext

The AWS KMS encryption context to use for encryption.

Type: String to string map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 2000.

Key Pattern: `.*\S.*`

Value Length Constraints: Minimum length of 1. Maximum length of 2000.

Value Pattern: `.*\S.*`

Required: No

kmsKeyId

The AWS KMS key to use for encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[A-Za-z0-9][A-Za-z0-9:_/+=,@.-]+`

Required: No

libraryDescription

The library's description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Pattern: `[a-zA-Z0-9\s!"\#\$%'&\(\)\*\+\,\-\./:;=\?@\[\\\]\^_`\{\|\}~><À-ÖØ-Üßà-öø-üß¿¡Œ-œº£¥₹€§©ª®™¹±-µ##-##₽###₣###₺]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibraryConfiguration

Service: Data Automation for Amazon Bedrock

Configuration for data automation libraries.

Contents

libraries

The libraries to use.

Type: Array of DataAutomationLibraryItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibraryEntitySummary

Service: Data Automation for Amazon Bedrock

A summary of a data automation library entity.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

vocabulary

Vocabulary entity details.

Type: VocabularyEntitySummary object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibraryFilter

Service: Data Automation for Amazon Bedrock

A data automation library filter.

Contents

libraryArn

Filter by library ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibraryIngestionJob

Service: Data Automation for Amazon Bedrock

A data automation library ingestion job.

Contents

creationTime

When the job was created.

Type: Timestamp

Required: Yes

entityType

The entity type.

Type: String

Valid Values: VOCABULARY

Required: Yes

jobArn

The job's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library-ingestion-job/[a-zA-Z0-9-]{12,36}`

Required: Yes

jobStatus

The job's status.

Type: String

Valid Values: `IN_PROGRESS | COMPLETED | COMPLETED_WITH_ERRORS | FAILED`

Required: Yes

operationType

The operation type.

Type: String

Valid Values: `UPSERT | DELETE`

Required: Yes

outputConfiguration

Output configuration for the job.

Type: OutputConfiguration object

Required: Yes

completionTime

When the job completed.

Type: Timestamp

Required: No

errorMessage

The error message if the job failed.

Type: String

Required: No

errorType

The error type if the job failed.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibraryIngestionJobSummary

Service: Data Automation for Amazon Bedrock

A summary of a data automation library ingestion job.

Contents

creationTime

When the job was created.

Type: Timestamp

Required: Yes

entityType

The entity type.

Type: String

Valid Values: VOCABULARY

Required: Yes

jobArn

The job's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library-ingestion-job/[a-zA-Z0-9-]{12,36}`

Required: Yes

jobStatus

The job's status.

Type: String

Valid Values: `IN_PROGRESS | COMPLETED | COMPLETED_WITH_ERRORS | FAILED`

Required: Yes

operationType

The operation type.

Type: String

Valid Values: `UPSERT | DELETE`

Required: Yes

completionTime

When the job completed.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibraryItem

Service: Data Automation for Amazon Bedrock

A data automation library item.

Contents

libraryArn

The library's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationLibrarySummary

Service: Data Automation for Amazon Bedrock

A summary of a data automation library.

Contents

creationTime

When the library was created.

Type: Timestamp

Required: Yes

libraryArn

The library's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-iso|-iso-[a-z]|-us-gov):bedrock:[a-zA-Z0-9-]*:[0-9]{12}:data-automation-library/[a-zA-Z0-9-]{12,36}`

Required: Yes

libraryName

The library's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationProject

Service: Data Automation for Amazon Bedrock

A data automation project.

Contents

creationTime

When the project was created.

Type: Timestamp

Required: Yes

lastModifiedTime

When the project was last updated.

Type: Timestamp

Required: Yes

projectArn

The project's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

Required: Yes

projectName

The project's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

status

The project's status.

Type: String

Valid Values: `COMPLETED | IN_PROGRESS | FAILED`

Required: Yes

customOutputConfiguration

Blueprints that are enabled for the project.

Type: CustomOutputConfiguration object

Required: No

dataAutomationLibraryConfiguration

Library configuration for the project.

Type: DataAutomationLibraryConfiguration object

Required: No

kmsEncryptionContext

The AWS KMS encryption context to use for encryption.

Type: String to string map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 2000.

Key Pattern: `.*\S.*`

Value Length Constraints: Minimum length of 1. Maximum length of 2000.

Value Pattern: `.*\S.*`

Required: No

kmsKeyId

The AWS KMS key to use for encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[A-Za-z0-9][A-Za-z0-9:_/+=,@.-]+`

Required: No

overrideConfiguration

Additional settings for the project.

Type: OverrideConfiguration object

Required: No

projectDescription

The project's description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Required: No

projectStage

The project's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

projectType

The type of bedrock data automation API that is compatible with this project.

Type: String

Valid Values: `ASYNC | SYNC`

Required: No

standardOutputConfiguration

The project's standard output configuration.

Type: StandardOutputConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationProjectFilter

Service: Data Automation for Amazon Bedrock

A data automation project filter.

Contents

projectArn

Filter by project ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

Required: Yes

projectStage

Filter by stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationProjectSummary

Service: Data Automation for Amazon Bedrock

A summary of a data automation project.

Contents

creationTime

The project's creation time.

Type: Timestamp

Required: Yes

projectArn

The project's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-]{12,36}`

Required: Yes

projectName

The project's name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: No

projectStage

The project's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

projectType

The project's type.

Type: String

Valid Values: `ASYNC | SYNC`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DeleteEntitiesInfo

Service: Data Automation for Amazon Bedrock

Information about entities to delete.

Contents

entityIds

The entity IDs to delete.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 1000 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentBoundingBox

Service: Data Automation for Amazon Bedrock

Bounding box settings for documents.

Contents

state

Whether bounding boxes are enabled for documents.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentCustomOutputConfiguration

Service: Data Automation for Amazon Bedrock

Custom output configuration for document processing.

Contents

fallbackBlueprints

The fallback blueprints to use for document processing. This will be used in the case where the service cannot find a match between the document and your list of blueprints.

Type: Array of BlueprintItem objects

Array Members: Minimum number of 0 items. Maximum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentExtractionGranularity

Service: Data Automation for Amazon Bedrock

Granularity settings for documents.

Contents

types

The levels of detail to extract.

Type: Array of strings

Valid Values: `DOCUMENT | PAGE | ELEMENT | WORD | LINE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentOutputAdditionalFileFormat

Service: Data Automation for Amazon Bedrock

Output settings for additional file formats.

Contents

state

Whether additional file formats are enabled for a project.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentOutputFormat

Service: Data Automation for Amazon Bedrock

A document output format.

Contents

additionalFileFormat

Output settings for additional file formats.

Type: DocumentOutputAdditionalFileFormat object

Required: Yes

textFormat

The text format.

Type: DocumentOutputTextFormat object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentOutputTextFormat

Service: Data Automation for Amazon Bedrock

An output text format.

Contents

types

The types of output text to generate.

Type: Array of strings

Valid Values: `PLAIN_TEXT | MARKDOWN | HTML | CSV`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentOverrideConfiguration

Service: Data Automation for Amazon Bedrock

Additional settings for a project.

Contents

modalityProcessing

Sets modality processing for document files. All modalities are enabled by default.

Type: ModalityProcessingConfiguration object

Required: No

sensitiveDataConfiguration

Configuration for sensitive data detection and redaction for document files.

Type: SensitiveDataConfiguration object

Required: No

splitter

Whether document splitter is enabled for a project.

Type: SplitterConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentStandardExtraction

Service: Data Automation for Amazon Bedrock

Settings for generating data from documents.

Contents

boundingBox

Whether to generate bounding boxes.

Type: DocumentBoundingBox object

Required: Yes

granularity

Which granularities to generate data for.

Type: DocumentExtractionGranularity object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentStandardGenerativeField

Service: Data Automation for Amazon Bedrock

Settings for generating descriptions of documents.

Contents

state

Whether generating descriptions is enabled for documents.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentStandardOutputConfiguration

Service: Data Automation for Amazon Bedrock

Output settings for processing documents.

Contents

extraction

Settings for populating data fields that describe the document.

Type: DocumentStandardExtraction object

Required: No

generativeField

Whether to generate descriptions.

Type: DocumentStandardGenerativeField object

Required: No

outputFormat

The output format to generate.

Type: DocumentOutputFormat object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EncryptionConfiguration

Service: Data Automation for Amazon Bedrock

Encryption settings for a project.

Contents

kmsKeyId

The AWS KMS key to use for encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[A-Za-z0-9][A-Za-z0-9:_/+=,@.-]+`

Required: Yes

kmsEncryptionContext

The AWS KMS encryption context to use for encryption.

Type: String to string map

Map Entries: Maximum number of items.

Key Length Constraints: Minimum length of 1. Maximum length of 2000.

Key Pattern: `.*\S.*`

Value Length Constraints: Minimum length of 1. Maximum length of 2000.

Value Pattern: `.*\S.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

EntityDetails

Service: Data Automation for Amazon Bedrock

Details about an entity.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

vocabulary

Vocabulary entity details.

Type: VocabularyEntity object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EntityTypeInfo

Service: Data Automation for Amazon Bedrock

Information about an entity type.

Contents

entityType

The entity type.

Type: String

Valid Values: VOCABULARY

Required: Yes

entityMetadata

Metadata about the entity type.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EventBridgeConfiguration

Service: Data Automation for Amazon Bedrock

EventBridge configuration for notifications.

Contents

eventBridgeEnabled

Whether EventBridge notifications are enabled.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageBoundingBox

Service: Data Automation for Amazon Bedrock

Bounding box settings for a project.

Contents

state

Whether bounding boxes are enabled for images.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageExtractionCategory

Service: Data Automation for Amazon Bedrock

Settings for generating categorical data from images.

Contents

state

Whether generating categorical data from images is enabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

types

The types of data to generate.

Type: Array of strings

Valid Values: `CONTENT_MODERATION | TEXT_DETECTION | LOGOS`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageOverrideConfiguration

Service: Data Automation for Amazon Bedrock

Sets whether your project will process images or not.

Contents

modalityProcessing

Sets modality processing for image files. All modalities are enabled by default.

Type: ModalityProcessingConfiguration object

Required: No

sensitiveDataConfiguration

Configuration for sensitive data detection and redaction for image files.

Type: SensitiveDataConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageStandardExtraction

Service: Data Automation for Amazon Bedrock

Settings for generating data from images.

Contents

boundingBox

Settings for generating bounding boxes.

Type: ImageBoundingBox object

Required: Yes

category

Settings for generating categorical data.

Type: ImageExtractionCategory object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageStandardGenerativeField

Service: Data Automation for Amazon Bedrock

Settings for generating descriptions of images.

Contents

state

Whether generating descriptions is enabled for images.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

types

The types of description to generate.

Type: Array of strings

Valid Values: `IMAGE_SUMMARY | IAB`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageStandardOutputConfiguration

Service: Data Automation for Amazon Bedrock

Output settings for processing images.

Contents

extraction

Settings for populating data fields that describe the image.

Type: ImageStandardExtraction object

Required: No

generativeField

Whether to generate descriptions of the data.

Type: ImageStandardGenerativeField object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InlinePayload

Service: Data Automation for Amazon Bedrock

Inline payload for ingestion jobs.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

deleteEntitiesInfo

Information about entities to delete.

Type: DeleteEntitiesInfo object

Required: No

upsertEntitiesInfo

Information about entities to upsert.

Type: Array of UpsertEntityInfo objects

Array Members: Minimum number of 1 item. Maximum number of 10 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputConfiguration

Service: Data Automation for Amazon Bedrock

Input configuration for ingestion jobs.

Only one member of this object can be specified or returned. If both inlinePayload and s3Object are specified, the request will fail with an exception.

Contents

inlinePayload

Inline payload input.

Type: InlinePayload object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

s3Object

S3 object input.

Type: S3Object object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModalityProcessingConfiguration

Service: Data Automation for Amazon Bedrock

This element is used to determine if the modality it is associated with is enabled or disabled. All modalities are enabled by default.

Contents

state

Stores the state of the modality for your project, set to either enabled or disabled

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModalityRoutingConfiguration

Service: Data Automation for Amazon Bedrock

This element allows you to set up where JPEG, PNG, MOV, and MP4 files get routed to for processing. JPEG routing applies to both "JPEG" and "JPG" file extensions.

Contents

jpeg

Sets whether JPEG files are routed to document or image processing.

Type: String

Valid Values: `IMAGE | DOCUMENT | AUDIO | VIDEO`

Required: No

mov

Sets whether MOV files are routed to audio or video processing.

Type: String

Valid Values: `IMAGE | DOCUMENT | AUDIO | VIDEO`

Required: No

mp4

Sets whether MP4 files are routed to audio or video processing.

Type: String

Valid Values: `IMAGE | DOCUMENT | AUDIO | VIDEO`

Required: No

png

Sets whether PNG files are routed to document or image processing.

Type: String

Valid Values: `IMAGE | DOCUMENT | AUDIO | VIDEO`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NotificationConfiguration

Service: Data Automation for Amazon Bedrock

Notification configuration for ingestion jobs.

Contents

eventBridgeConfiguration

EventBridge configuration.

Type: EventBridgeConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputConfiguration

Service: Data Automation for Amazon Bedrock

Output configuration for ingestion jobs.

Contents

s3Uri

S3 URI for output.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OverrideConfiguration

Service: Data Automation for Amazon Bedrock

Additional settings for a project.

Contents

audio

This element declares whether your project will process audio files.

Type: AudioOverrideConfiguration object

Required: No

document

Settings for processing documents.

Type: DocumentOverrideConfiguration object

Required: No

image

This element declares whether your project will process image files.

Type: ImageOverrideConfiguration object

Required: No

modalityRouting

Lets you set which modalities certain file types are processed as.

Type: ModalityRoutingConfiguration object

Required: No

video

This element declares whether your project will process video files.

Type: VideoOverrideConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Phrase

Service: Data Automation for Amazon Bedrock

A phrase in a vocabulary entity.

Contents

text

The phrase text.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*.+.*`

Required: Yes

displayAsText

The display text for the phrase.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `.*.+.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PIIEntitiesConfiguration

Service: Data Automation for Amazon Bedrock

Configuration for detecting and redacting Personally Identifiable Information (PII) entities. Specify which PII entity types to detect and the redaction mask mode. If not provided, defaults to ALL entity types with ENTITY_TYPE redaction mask mode.

Contents

piiEntityTypes

List of PII entity types to detect/redact in the output. Choose from specific entity types (such as ADDRESS, NAME, EMAIL, PHONE, US_SOCIAL_SECURITY_NUMBER) or specify ALL to detect all supported PII types. If not specified, defaults to ALL.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 32 items.

Valid Values: `ALL | ADDRESS | AGE | NAME | EMAIL | PHONE | USERNAME | PASSWORD | DRIVER_ID | LICENSE_PLATE | VEHICLE_IDENTIFICATION_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | PIN | INTERNATIONAL_BANK_ACCOUNT_NUMBER | SWIFT_CODE | IP_ADDRESS | MAC_ADDRESS | URL | AWS_ACCESS_KEY | AWS_SECRET_KEY | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER`

Required: No

redactionMaskMode

Defines how detected PII entities are masked in redacted output files. Set to PII to replace all detected entities with a generic [PII] marker regardless of entity type. Set to ENTITY_TYPE to replace each detected entity with its specific type marker (for example, [NAME], [EMAIL], [ADDRESS]). This setting only applies when detectionMode is set to DETECTION_AND_REDACTION. If not specified, defaults to ENTITY_TYPE.

Type: String

Valid Values: `PII | ENTITY_TYPE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Object

Service: Data Automation for Amazon Bedrock

An object stored in Amazon S3.

Contents

s3Uri

The S3 URI of the object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?`

Required: Yes

version

The version of the S3 object.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SensitiveDataConfiguration

Service: Data Automation for Amazon Bedrock

Configuration for detecting and redacting sensitive data in content. Use this to control whether sensitive data is detected only or both detected and redacted, specify the scope of detection (standard output, custom output, or both), and configure specific PII entity types to detect along with how they should be masked when redacted.

Contents

detectionMode

Specifies the mode for handling sensitive data detection. Set to DETECTION to only identify sensitive data without modifying content - this produces one output file per detection scope containing detection information with original unredacted content. Set to DETECTION_AND_REDACTION to both identify and mask sensitive data - this produces two output files per detection scope: one unredacted file with detection information and one redacted file with masking applied to sensitive content. For example, if detectionScope includes both STANDARD and CUSTOM with DETECTION_AND_REDACTION mode, four output files will be generated (two for standard output and two for custom output).

Type: String

Valid Values: `DETECTION | DETECTION_AND_REDACTION`

Required: Yes

detectionScope

Defines which BDA output types to apply sensitive data detection to. Specify STANDARD to detect sensitive data in standard output, CUSTOM to detect in custom output (blueprint-based extraction), or both to apply detection to both output types. If not specified, defaults to both STANDARD and CUSTOM. The number of output files generated depends on both the detection mode and the scopes selected - each scope specified will produce its own set of output files according to the detection mode configured.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 2 items.

Valid Values: `STANDARD | CUSTOM`

Required: No

piiEntitiesConfiguration

Configuration for detecting and redacting Personally Identifiable Information (PII) entities.

Type: PIIEntitiesConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpeakerLabelingConfiguration

Service: Data Automation for Amazon Bedrock

Enables or disables speaker labeling. Speaker labeling, when enabled will assign a number to each speaker, and indicate which speaker is talking in each portion of the transcript. This appears in the response as "spk_0" for the first speaker, "spk_1" for the second, and so on for up to 30 speakers.

Contents

state

State of speaker labeling, either enabled or disabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SplitterConfiguration

Service: Data Automation for Amazon Bedrock

Document splitter settings. If a document is too large to be processed in one pass, the document splitter splits it into smaller documents.

Contents

state

Whether document splitter is enabled for a project.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

StandardOutputConfiguration

Service: Data Automation for Amazon Bedrock

Output settings for processing data.

Contents

audio

Settings for processing audio.

Type: AudioStandardOutputConfiguration object

Required: No

document

Settings for processing documents.

Type: DocumentStandardOutputConfiguration object

Required: No

image

Settings for processing images.

Type: ImageStandardOutputConfiguration object

Required: No

video

Settings for processing video.

Type: VideoStandardOutputConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tag

Service: Data Automation for Amazon Bedrock

A tag associated with a resource. Made up of a Key and Value

Contents

key

The tag's key

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Required: Yes

value

The tag's value

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TranscriptConfiguration

Service: Data Automation for Amazon Bedrock

Configuration for transcript options. This option allows you to enable speaker labeling and channel labeling.

Contents

channelLabeling

Enables channel labeling. Each audio channel will be labeled with a number, and the transcript will indicate which channel is being used.

Type: ChannelLabelingConfiguration object

Required: No

speakerLabeling

Enables speaker labeling. Each speaker within a transcript will recieve a number, and the transcript will note which speaker is talking.

Type: SpeakerLabelingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpsertEntityInfo

Service: Data Automation for Amazon Bedrock

Information about entities to upsert.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

vocabulary

Vocabulary entity information.

Type: VocabularyEntityInfo object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidationExceptionField

Service: Data Automation for Amazon Bedrock

Error information for a field that failed validation.

Contents

message

The error message.

Type: String

Pattern: `[\s\S]+`

Required: Yes

name

The field's name.

Type: String

Pattern: `[\s\S]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoBoundingBox

Service: Data Automation for Amazon Bedrock

Bounding box settings for video.

Contents

state

Whether bounding boxes are enabled for video.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoExtractionCategory

Service: Data Automation for Amazon Bedrock

Settings for generating categorical data from video.

Contents

state

Whether generating categorical data from video is enabled.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

types

The types of data to generate.

Type: Array of strings

Valid Values: `CONTENT_MODERATION | TEXT_DETECTION | TRANSCRIPT | LOGOS`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoOverrideConfiguration

Service: Data Automation for Amazon Bedrock

Sets whether your project will process videos or not.

Contents

modalityProcessing

Sets modality processing for video files. All modalities are enabled by default.

Type: ModalityProcessingConfiguration object

Required: No

sensitiveDataConfiguration

Configuration for sensitive data detection and redaction for video files.

Type: SensitiveDataConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoStandardExtraction

Service: Data Automation for Amazon Bedrock

Settings for generating data from video.

Contents

boundingBox

Settings for generating bounding boxes.

Type: VideoBoundingBox object

Required: Yes

category

Settings for generating categorical data.

Type: VideoExtractionCategory object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoStandardGenerativeField

Service: Data Automation for Amazon Bedrock

Settings for generating descriptions of video.

Contents

state

Whether generating descriptions is enabled for video.

Type: String

Valid Values: `ENABLED | DISABLED`

Required: Yes

types

The types of description to generate.

Type: Array of strings

Valid Values: `VIDEO_SUMMARY | IAB | CHAPTER_SUMMARY`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoStandardOutputConfiguration

Service: Data Automation for Amazon Bedrock

Output settings for processing video.

Contents

extraction

Settings for populating data fields that describe the video.

Type: VideoStandardExtraction object

Required: No

generativeField

Whether to generate descriptions of the video.

Type: VideoStandardGenerativeField object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VocabularyEntity

Service: Data Automation for Amazon Bedrock

A vocabulary entity.

Contents

description

The entity description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Pattern: `[a-zA-Z0-9\s!"\#\$%'&\(\)\*\+\,\-\./:;=\?@\[\\\]\^_`\{\|\}~><À-ÖØ-Üßà-öø-üß¿¡Œ-œ°£¥₹€§©ª®™¹±-µ##-##₽###₣###₺]*`

Required: No

entityId

The entity ID.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: No

language

The entity language.

Type: String

Valid Values: `EN | DE | ES | FR | IT | PT | JA | KO | CN | TW | HK`

Required: No

lastModifiedTime

When the entity was last modified.

Type: Timestamp

Required: No

phrases

The entity phrases.

Type: Array of Phrase objects

Array Members: Minimum number of 1 item.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VocabularyEntityInfo

Service: Data Automation for Amazon Bedrock

Information about a vocabulary entity.

Contents

language

The entity language.

Type: String

Valid Values: `EN | DE | ES | FR | IT | PT | JA | KO | CN | TW | HK`

Required: Yes

phrases

The entity phrases.

Type: Array of Phrase objects

Array Members: Minimum number of 1 item.

Required: Yes

description

The entity description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Pattern: `[a-zA-Z0-9\s!"\#\$%'&\(\)\*\+\,\-\./:;=\?@\[\\\]\^_`\{\|\}~><À-ÖØ-Üßà-öø-üß¿¡Œ-œº£¥₹€§©ª®™¹±-µ##-##₽###₣###₺]*`

Required: No

entityId

The entity ID.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VocabularyEntitySummary

Service: Data Automation for Amazon Bedrock

A summary of a vocabulary entity.

Contents

description

The entity description.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 300.

Pattern: `[a-zA-Z0-9\s!"\#\$%'&\(\)\*\+\,\-\./:;=\?@\[\\\]\^_`\{\|\}~><À-ÖØ-Üßà-öø-üß¿¡Œ-œ°£¥₹€§©ª®™¹±-µ##-##₽###₣###₺]*`

Required: No

entityId

The entity ID.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9-_]+`

Required: No

language

The entity language.

Type: String

Valid Values: `EN | DE | ES | FR | IT | PT | JA | KO | CN | TW | HK`

Required: No

lastModifiedTime

When the entity was last modified.

Type: Timestamp

Required: No

numOfPhrases

The number of phrases in the entity.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Runtime for Amazon Bedrock Data Automation

The following data types are supported by Runtime for Amazon Bedrock Data Automation:

- AssetProcessingConfiguration
- Blueprint
- DataAutomationConfiguration
- EncryptionConfiguration
- EventBridgeConfiguration
- InputConfiguration
- NotificationConfiguration
- OutputConfiguration
- OutputSegment
- SyncInputConfiguration

- Tag
- TimestampSegment
- VideoAssetProcessingConfiguration
- VideoSegmentConfiguration

AssetProcessingConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Allows you to configure your input asset to narrow where you're looking for content. Works with passing segements of a larger video.

Contents

video

Configures what portion of a video you want to run analysis on.

Type: VideoAssetProcessingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Blueprint

Service: Runtime for Amazon Bedrock Data Automation

Details about a data automation blueprint.

Contents

blueprintArn

The blueprint's ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):blueprint/(bedrock-data-insights-public-[a-zA-Z0-9-_]{1,30}|bedrock-data-automation-public-[a-zA-Z0-9-_]{1,30}|[a-zA-Z0-9-]{12,36})`

Required: Yes

stage

The blueprint's stage.

Type: String

Valid Values: `DEVELOPMENT | LIVE`

Required: No

version

The blueprint's version.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[0-9]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DataAutomationConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Details about a data automation project.

Contents

dataAutomationProjectArn

The ARN of the project you're using in your configuration.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `arn:aws(|-cn|-us-gov):bedrock:[a-zA-Z0-9-]*:(aws|[0-9]{12}):data-automation-project/[a-zA-Z0-9-_]+`

Required: Yes

stage

The project's stage.

Type: String

Valid Values: `LIVE | DEVELOPMENT`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

EncryptionConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Encryption settings for an invocation.

Contents

kmsKeyId

A KMS key ID to use for encryption.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `[A-Za-z0-9][A-Za-z0-9:_/+=,@.-]+`

Required: Yes

kmsEncryptionContext

Name-value pairs to include as an encryption context.

Type: String to string map

Map Entries: Maximum number of 10 items.

Key Length Constraints: Minimum length of 1. Maximum length of 2000.

Key Pattern: `.*\S.*`

Value Length Constraints: Minimum length of 1. Maximum length of 2000.

Value Pattern: `.*\S.*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

EventBridgeConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Event bus settings for an invocation.

Contents

eventBridgeEnabled

Whether to send events to an event bus.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InputConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Input object settings for an invocation.

Contents

s3Uri

The objects S3 URI.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/[^\x00-\x1F\x7F\{^}%`\]">\[~<#|]*)?`

Required: Yes

assetProcessingConfiguration

Lets you configure what portion of content you want to analyze. Works with passing segements of a larger video.

Type: AssetProcessingConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NotificationConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Notification settings for an invocation.

Contents

eventBridgeConfiguration

Event bus settings.

Type: EventBridgeConfiguration object

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Output settings for an invocation. The service adds a unique invocation ID to the path that you specify and stores output objects under that path.

Contents

s3Uri

The S3 URI for invocation output.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/[^\x00-\x1F\x7F\{^}%`\]">\[~<#|]*)?`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputSegment

Service: Runtime for Amazon Bedrock Data Automation

A segment of output from a data automation invocation.

Contents

customOutput

The custom output content for this segment.

Type: String

Required: No

customOutputStatus

The status of the custom output for this segment.

Type: String

Valid Values: `MATCH | NO_MATCH`

Required: No

standardOutput

The standard output content for this segment.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SyncInputConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Input configuration for a synchronous data automation invocation.

Contents

bytes

The raw bytes of the input asset.

Type: Base64-encoded binary data object

Required: No

s3Uri

The S3 URI of the input asset.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/[^\x00-\x1F\x7F\{^}%`\]">\[~<#|]*)?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tag

Service: Runtime for Amazon Bedrock Data Automation

A tag associated with a resource. A tag consists of a key and value.

Contents

key

The key associated with a tag.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `(?!aws:)[\p{L}\p{Z}\p{N}_.:/=+\-@]*`

Required: Yes

value

The value associated with a tag.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `([\p{L}\p{Z}\p{N}_.:/=+\-@]*)`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TimestampSegment

Service: Runtime for Amazon Bedrock Data Automation

Contians the start and end timestamp of the segement of the larger video you want to detect. The minimum video segment length is 5 minutes.

Contents

endTimeMillis

The end time of the segment you want to extract information from.

Type: Long

Valid Range: Minimum value of 300000.

Required: Yes

startTimeMillis

The start time of the segement you want to extract information from.

Type: Long

Valid Range: Minimum value of 0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoAssetProcessingConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Sets up elements of your video asset. Lets you set what segement of a larger video you want to detect information from.

Contents

segmentConfiguration

Allows you to set up a portion of a video you want to detect information from. You can select a segment 5 minutes or longer, and your analysis will treat that as a full video.

Type: VideoSegmentConfiguration object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoSegmentConfiguration

Service: Runtime for Amazon Bedrock Data Automation

Used to set your start and end timestamp for what portion of a video you'd like to detect. The minimum segement length is 5 minutes.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

timestampSegment

Holds the timestamps for the beginning and end of your video segment.

Type: TimestampSegment object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon Bedrock Runtime

The following data types are supported by Amazon Bedrock Runtime:

- AnyToolChoice
- AppliedGuardrailDetails
- AsyncInvokeOutputDataConfig

- AsyncInvokeS3OutputDataConfig
- AsyncInvokeSummary
- AudioBlock
- AudioSource
- AutoToolChoice
- BidirectionalInputPayloadPart
- BidirectionalOutputPayloadPart
- CacheDetail
- CachePointBlock
- Citation
- CitationGeneratedContent
- CitationLocation
- CitationsConfig
- CitationsContentBlock
- CitationsDelta
- CitationSourceContent
- CitationSourceContentDelta
- ContentBlock
- ContentBlockDelta
- ContentBlockDeltaEvent
- ContentBlockStart
- ContentBlockStartEvent
- ContentBlockStopEvent
- ConverseMetrics
- ConverseOutput
- ConverseStreamMetadataEvent
- ConverseStreamMetrics
- ConverseStreamOutput
- ConverseStreamTrace
- ConverseTokensRequest

- ConverseTrace
- CountTokensInput
- DocumentBlock
- DocumentCharLocation
- DocumentChunkLocation
- DocumentContentBlock
- DocumentPageLocation
- DocumentSource
- ErrorBlock
- GuardrailAssessment
- GuardrailAutomatedReasoningFinding
- GuardrailAutomatedReasoningImpossibleFinding
- GuardrailAutomatedReasoningInputTextReference
- GuardrailAutomatedReasoningInvalidFinding
- GuardrailAutomatedReasoningLogicWarning
- GuardrailAutomatedReasoningNoTranslationsFinding
- GuardrailAutomatedReasoningPolicyAssessment
- GuardrailAutomatedReasoningRule
- GuardrailAutomatedReasoningSatisfiableFinding
- GuardrailAutomatedReasoningScenario
- GuardrailAutomatedReasoningStatement
- GuardrailAutomatedReasoningTooComplexFinding
- GuardrailAutomatedReasoningTranslation
- GuardrailAutomatedReasoningTranslationAmbiguousFinding
- GuardrailAutomatedReasoningTranslationOption
- GuardrailAutomatedReasoningValidFinding
- GuardrailChecksConfig
- GuardrailChecksContentBlock
- GuardrailChecksContentFilterCategoryConfig
- GuardrailChecksContentFilterConfig

- GuardrailChecksContentFilterResult
- GuardrailChecksContentFilterResultEntry
- GuardrailChecksContentFilterUsage
- GuardrailChecksMessage
- GuardrailChecksPromptAttackCategoryConfig
- GuardrailChecksPromptAttackConfig
- GuardrailChecksPromptAttackResult
- GuardrailChecksPromptAttackResultEntry
- GuardrailChecksPromptAttackUsage
- GuardrailChecksResults
- GuardrailChecksSensitiveInformationConfig
- GuardrailChecksSensitiveInformationEntityConfig
- GuardrailChecksSensitiveInformationResult
- GuardrailChecksSensitiveInformationResultEntry
- GuardrailChecksSensitiveInformationUsage
- GuardrailChecksUsageResults
- GuardrailConfiguration
- GuardrailContentBlock
- GuardrailContentFilter
- GuardrailContentPolicyAssessment
- GuardrailContextualGroundingFilter
- GuardrailContextualGroundingPolicyAssessment
- GuardrailConverseContentBlock
- GuardrailConverseImageBlock
- GuardrailConverseImageSource
- GuardrailConverseTextBlock
- GuardrailCoverage
- GuardrailCustomWord
- GuardrailImageBlock
- GuardrailImageCoverage

- GuardrailImageSource
- GuardrailInvocationMetrics
- GuardrailManagedWord
- GuardrailOutputContent
- GuardrailPiiEntityFilter
- GuardrailRegexFilter
- GuardrailSensitiveInformationPolicyAssessment
- GuardrailStreamConfiguration
- GuardrailTextBlock
- GuardrailTextCharactersCoverage
- GuardrailTopic
- GuardrailTopicPolicyAssessment
- GuardrailTraceAssessment
- GuardrailUsage
- GuardrailWordPolicyAssessment
- ImageBlock
- ImageBlockDelta
- ImageBlockStart
- ImageSource
- InferenceConfiguration
- InvokeModelTokensRequest
- InvokeModelWithBidirectionalStreamInput
- InvokeModelWithBidirectionalStreamOutput
- JsonSchemaDefinition
- Message
- MessageStartEvent
- MessageStopEvent
- OutputConfig
- OutputFormat
- OutputFormatStructure

- PayloadPart
- PerformanceConfiguration
- PromptRouterTrace
- PromptVariableValues
- ReasoningContentBlock
- ReasoningContentBlockDelta
- ReasoningTextBlock
- ResponseStream
- S3Location
- SearchResultBlock
- SearchResultContentBlock
- SearchResultLocation
- ServiceTier
- SpecificToolChoice
- SystemContentBlock
- SystemTool
- Tag
- TokenUsage
- Tool
- ToolChoice
- ToolConfiguration
- ToolInputSchema
- ToolResultBlock
- ToolResultBlockDelta
- ToolResultBlockStart
- ToolResultContentBlock
- ToolSpecification
- ToolUseBlock
- ToolUseBlockDelta
- ToolUseBlockStart

- VideoBlock
- VideoSource
- WebLocation

AnyToolChoice

Service: Amazon Bedrock Runtime

The model must request at least one tool (no text is generated). For example, `{"any" : {}}`. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AppliedGuardrailDetails

Service: Amazon Bedrock Runtime

Details about the specific guardrail that was applied during this assessment, including its identifier, version, ARN, origin, and ownership information.

Contents

guardrailArn

The ARN of the guardrail that was applied.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+)`

Required: No

guardrailId

The unique ID of the guardrail that was applied.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `([a-z0-9]+)`

Required: No

guardrailOrigin

The origin of how the guardrail was applied. This can be either requested at the API level or enforced at the account or organization level as a default guardrail.

Type: Array of strings

Valid Values: `REQUEST | ACCOUNT_ENFORCED | ORGANIZATION_ENFORCED`

Required: No

guardrailOwnership

The ownership type of the guardrail, indicating whether it is owned by the requesting account or is a cross-account guardrail shared from another AWS account.

Type: String

Valid Values: `SELF | CROSS_ACCOUNT`

Required: No

guardrailVersion

The version of the guardrail that was applied.

Type: String

Pattern: `(|([1-9][0-9]{0,7})|(DRAFT))`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInvokeOutputDataConfig

Service: Amazon Bedrock Runtime

Asynchronous invocation output data settings.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

s3OutputDataConfig

A storage location for the output data in an S3 bucket

Type: AsyncInvokeS3OutputDataConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInvokeS3OutputDataConfig

Service: Amazon Bedrock Runtime

Asynchronous invocation output data settings.

Contents

s3Uri

An object URI starting with `s3://`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?`

Required: Yes

bucketOwner

If the bucket belongs to another AWS account, specify that account's ID.

Type: String

Pattern: `[0-9]{12}`

Required: No

kmsKeyId

A KMS encryption key ID.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AsyncInvokeSummary

Service: Amazon Bedrock Runtime

A summary of an asynchronous invocation.

Contents

invocationArn

The invocation's ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:async-invoke/[a-z0-9]{12}`

Required: Yes

modelArn

The invoked model's ARN.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `arn:[a-z0-9\-]+:bedrock:[a-z0-9\-]*:[0-9]*:(provisioned-model|foundation-model)/.+`

Required: Yes

outputDataConfig

The invocation's output data settings.

Type: AsyncInvokeOutputDataConfig object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

submitTime

When the invocation was submitted.

Type: Timestamp

Required: Yes

clientRequestToken

The invocation's idempotency token.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[!-~]*`

Required: No

endTime

When the invocation ended.

Type: Timestamp

Required: No

failureMessage

An error message.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Required: No

lastModifiedTime

When the invocation was last modified.

Type: Timestamp

Required: No

status

The invocation's status.

Type: String

Valid Values: `InProgress | Completed | Failed`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AudioBlock

Service: Amazon Bedrock Runtime

An audio content block that contains audio data in various supported formats.

Contents

format

The format of the audio data, such as MP3, WAV, FLAC, or other supported audio formats.

Type: String

Valid Values: `mp3 | opus | wav | aac | flac | mp4 | ogg | mkv | mka | x-aac | m4a | mpeg | mpga | pcm | webm`

Required: Yes

source

The source of the audio data, which can be provided as raw bytes or an S3 location.

Type: AudioSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

error

Error information if the audio block could not be processed or contains invalid data.

Type: ErrorBlock object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AudioSource

Service: Amazon Bedrock Runtime

The source of audio data, which can be provided either as raw bytes or a reference to an S3 location.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

Audio data encoded in base64.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

s3Location

A reference to audio data stored in an Amazon S3 bucket. To see which models support S3 uploads, see Supported models and features for Converse.

Type: S3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

AutoToolChoice

Service: Amazon Bedrock Runtime

The Model automatically decides if a tool should be called or whether to generate text instead. For example, `{"auto" : {}}`. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BidirectionalInputPayloadPart

Service: Amazon Bedrock Runtime

Payload content for the bidirectional input. The input is an audio stream.

Contents

bytes

The audio content for the bidirectional input.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

BidirectionalOutputPayloadPart

Service: Amazon Bedrock Runtime

Output from the bidirectional stream. The output is speech and a text transcription.

Contents

bytes

The speech output of the bidirectional stream.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CacheDetail

Service: Amazon Bedrock Runtime

Cache creation metrics for a specific TTL duration

Contents

inputTokens

Number of tokens written to cache with this TTL (cache creation tokens)

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

ttl

TTL duration for these cached tokens

Type: String

Valid Values: `5m | 1h`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CachePointBlock

Service: Amazon Bedrock Runtime

Defines a section of content to be cached for reuse in subsequent API calls.

Contents

type

Specifies the type of cache point within the CachePointBlock.

Type: String

Valid Values: `default`

Required: Yes

ttl

Optional TTL duration for cache entries. When specified, enables extended TTL caching with the specified duration. When omitted, uses `type` value for caching behavior.

Type: String

Valid Values: `5m | 1h`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Citation

Service: Amazon Bedrock Runtime

Contains information about a citation that references a specific source document. Citations provide traceability between the model's generated response and the source documents that informed that response.

Contents

location

The precise location within the source document where the cited content can be found, including character positions, page numbers, or chunk identifiers.

Type: CitationLocation object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

source

The source from the original search result that provided the cited content.

Type: String

Required: No

sourceContent

The specific content from the source document that was referenced or cited in the generated response.

Type: Array of CitationSourceContent objects

Required: No

title

The title or identifier of the source document being cited.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationGeneratedContent

Service: Amazon Bedrock Runtime

Contains the generated text content that corresponds to or is supported by a citation from a source document.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

The text content that was generated by the model and is supported by the associated citation.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationLocation

Service: Amazon Bedrock Runtime

Specifies the precise location within a source document where cited content can be found. This can include character-level positions, page numbers, or document chunks depending on the document type and indexing method.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

documentChar

The character-level location within the document where the cited content is found.

Type: DocumentCharLocation object

Required: No

documentChunk

The chunk-level location within the document where the cited content is found, typically used for documents that have been segmented into logical chunks.

Type: DocumentChunkLocation object

Required: No

documentPage

The page-level location within the document where the cited content is found.

Type: DocumentPageLocation object

Required: No

searchResultLocation

The search result location where the cited content is found, including the search result index and block positions within the content array.

Type: SearchResultLocation object

Required: No

web

The web URL that was cited for this reference.

Type: WebLocation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationsConfig

Service: Amazon Bedrock Runtime

Configuration settings for enabling and controlling document citations in Converse API responses. When enabled, the model can include citation information that links generated content back to specific source documents.

Contents

enabled

Specifies whether citations from the selected document should be used in the model's response. When set to true, the model can generate citations that reference the source documents used to inform the response.

Type: Boolean

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationsContentBlock

Service: Amazon Bedrock Runtime

A content block that contains both generated text and associated citation information. This block type is returned when document citations are enabled, providing traceability between the generated content and the source documents that informed the response.

Contents

citations

An array of citations that reference the source documents used to generate the associated content.

Type: Array of Citation objects

Required: No

content

The generated content that is supported by the associated citations.

Type: Array of CitationGeneratedContent objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationsDelta

Service: Amazon Bedrock Runtime

Contains incremental updates to citation information during streaming responses. This allows clients to build up citation data progressively as the response is generated.

Contents

location

Specifies the precise location within a source document where cited content can be found. This can include character-level positions, page numbers, or document chunks depending on the document type and indexing method.

Type: CitationLocation object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

source

The source from the original search result that provided the cited content.

Type: String

Required: No

sourceContent

The specific content from the source document that was referenced or cited in the generated response.

Type: Array of CitationSourceContentDelta objects

Required: No

title

The title or identifier of the source document being cited.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationSourceContent

Service: Amazon Bedrock Runtime

Contains the actual text content from a source document that is being cited or referenced in the model's response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

The text content from the source document that is being cited.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CitationSourceContentDelta

Service: Amazon Bedrock Runtime

Contains incremental updates to the source content text during streaming responses, allowing clients to build up the cited content progressively.

Contents

text

An incremental update to the text content from the source document that is being cited.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlock

Service: Amazon Bedrock Runtime

A block of content for a message that you pass to, or receive from, a model with the Converse or ConverseStream API operations.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

audio

An audio content block containing audio data in the conversation.

Type: AudioBlock object

Required: No

cachePoint

CachePoint to include in the message.

Type: CachePointBlock object

Required: No

citationsContent

A content block that contains both generated text and associated citation information, providing traceability between the response and source documents.

Type: CitationsContentBlock object

Required: No

document

A document to include in the message.

Type: DocumentBlock object

Required: No

guardContent

Contains the content to assess with the guardrail. If you don't specify `guardContent` in a call to the Converse API, the guardrail (if passed in the Converse API) assesses the entire message.

For more information, see Use a guardrail with the Converse API.

Type: GuardrailConverseContentBlock object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

image

Image to include in the message.

> **Note**
>
> This field is only supported by Anthropic Claude 3 models.

Type: ImageBlock object

Required: No

reasoningContent

Contains content regarding the reasoning that is carried out by the model. Reasoning refers to a Chain of Thought (CoT) that the model generates to enhance the accuracy of its final response.

Type: ReasoningContentBlock object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

searchResult

Search result to include in the message.

Type: SearchResultBlock object

Required: No

text

Text to include in the message.

Type: String

Required: No

toolResult

The result for a tool request that a model makes.

Type: ToolResultBlock object

Required: No

toolUse

Information about a tool use request from a model.

Type: ToolUseBlock object

Required: No

video

Video to include in the message.

Type: VideoBlock object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlockDelta

Service: Amazon Bedrock Runtime

A block of content in a streaming response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

citation

Incremental citation information that is streamed as part of the response generation process.

Type: CitationsDelta object

Required: No

image

A streaming delta event containing incremental image data.

Type: ImageBlockDelta object

Required: No

reasoningContent

Contains content regarding the reasoning that is carried out by the model. Reasoning refers to a Chain of Thought (CoT) that the model generates to enhance the accuracy of its final response.

Type: ReasoningContentBlockDelta object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

text

The content text.

Type: String

Required: No

toolResult

An incremental update that contains the results from a tool call.

Type: Array of ToolResultBlockDelta objects

Required: No

toolUse

Information about a tool that the model is requesting to use.

Type: ToolUseBlockDelta object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlockDeltaEvent

Service: Amazon Bedrock Runtime

The content block delta event.

Contents

contentBlockIndex

The block index for a content block delta event.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

delta

The delta for a content block delta event.

Type: ContentBlockDelta object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlockStart

Service: Amazon Bedrock Runtime

Content block start information.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

image

The initial event indicating the start of a streaming image block.

Type: ImageBlockStart object

Required: No

toolResult

The

Type: ToolResultBlockStart object

Required: No

toolUse

Information about a tool that the model is requesting to use.

Type: ToolUseBlockStart object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlockStartEvent

Service: Amazon Bedrock Runtime

Content block start event.

Contents

contentBlockIndex

The index for a content block start event.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

start

Start information about a content block start event.

Type: ContentBlockStart object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ContentBlockStopEvent

Service: Amazon Bedrock Runtime

A content block stop event.

Contents

contentBlockIndex

The index for a content block.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseMetrics

Service: Amazon Bedrock Runtime

Metrics for a call to Converse.

Contents

latencyMs

The latency of the call to `Converse`, in milliseconds.

Type: Long

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseOutput

Service: Amazon Bedrock Runtime

The output from a call to Converse.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

message

The message that the model generates.

Type: Message object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseStreamMetadataEvent

Service: Amazon Bedrock Runtime

A conversation stream metadata event.

Contents

metrics

The metrics for the conversation stream metadata event.

Type: ConverseStreamMetrics object

Required: Yes

usage

Usage information for the conversation stream event.

Type: TokenUsage object

Required: Yes

performanceConfig

Model performance configuration metadata for the conversation stream event.

Type: PerformanceConfiguration object

Required: No

serviceTier

Specifies the processing tier configuration used for serving the request.

Type: ServiceTier object

Required: No

trace

The trace object in the response from ConverseStream that contains information about the guardrail behavior.

Type: ConverseStreamTrace object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseStreamMetrics

Service: Amazon Bedrock Runtime

Metrics for the stream.

Contents

latencyMs

The latency for the streaming request, in milliseconds.

Type: Long

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseStreamOutput

Service: Amazon Bedrock Runtime

The messages output stream

Contents

contentBlockDelta

The messages output content block delta.

Type: ContentBlockDeltaEvent object

Required: No

contentBlockStart

Start information for a content block.

Type: ContentBlockStartEvent object

Required: No

contentBlockStop

Stop information for a content block.

Type: ContentBlockStopEvent object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

messageStart

Message start information.

Type: MessageStartEvent object

Required: No

messageStop

Message stop information.

Type: MessageStopEvent object

Required: No

metadata

Metadata for the converse output stream.

Type: ConverseStreamMetadataEvent object

Required: No

modelStreamErrorException

A streaming error occurred. Retry your request.

Type: Exception

HTTP Status Code: 424

Required: No

serviceUnavailableException

The service isn't currently available. For troubleshooting this error, see ServiceUnavailable in the Amazon Bedrock User Guide

Type: Exception

HTTP Status Code: 503

Required: No

throttlingException

Your request was denied due to exceeding the account quotas for *Amazon Bedrock*. For troubleshooting this error, see ThrottlingException in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

The input fails to satisfy the constraints specified by *Amazon Bedrock*. For troubleshooting this error, see ValidationError in the Amazon Bedrock User Guide.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseStreamTrace

Service: Amazon Bedrock Runtime

The trace object in a response from ConverseStream.

Contents

guardrail

The guardrail trace object.

Type: GuardrailTraceAssessment object

Required: No

promptRouter

The request's prompt router.

Type: PromptRouterTrace object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseTokensRequest

Service: Amazon Bedrock Runtime

The inputs from a `Converse` API request for token counting.

This structure mirrors the input format for the `Converse` operation, allowing you to count tokens for conversation-based inference requests.

Contents

additionalModelRequestFields

The additionalModelRequestFields of Converse input request to count tokens for. Use this field when you want to pass additional parameters that the model supports.

Type: JSON value

Required: No

messages

An array of messages to count tokens for.

Type: Array of Message objects

Required: No

system

The system content blocks to count tokens for. System content provides instructions or context to the model about how it should behave or respond. The token count will include any system content provided.

Type: Array of SystemContentBlock objects

Required: No

toolConfig

The toolConfig of Converse input request to count tokens for. Configuration information for the tools that the model can use when generating a response.

Type: ToolConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ConverseTrace

Service: Amazon Bedrock Runtime

The trace object in a response from Converse.

Contents

guardrail

The guardrail trace object.

Type: GuardrailTraceAssessment object

Required: No

promptRouter

The request's prompt router.

Type: PromptRouterTrace object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CountTokensInput

Service: Amazon Bedrock Runtime

The input value for token counting. The value should be either an `InvokeModel` or `Converse` request body.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

converse

A `Converse` request for which to count tokens. Use this field when you want to count tokens for a conversation-based input that would be sent to the `Converse` operation.

Type: ConverseTokensRequest object

Required: No

invokeModel

An `InvokeModel` request for which to count tokens. Use this field when you want to count tokens for a raw text input that would be sent to the `InvokeModel` operation.

Type: InvokeModelTokensRequest object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentBlock

Service: Amazon Bedrock Runtime

A document to include in a message.

Contents

name

A name for the document. The name can only contain the following characters:

- Alphanumeric characters
- Whitespace characters (no more than one in a row)
- Hyphens
- Parentheses
- Square brackets

> **Note**
>
> This field is vulnerable to prompt injections, because the model might inadvertently interpret it as instructions. Therefore, we recommend that you specify a neutral name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 200.

Required: Yes

source

Contains the content of the document.

Type: DocumentSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

citations

Configuration settings that control how citations should be generated for this specific document.

Type: CitationsConfig object

Required: No

context

Contextual information about how the document should be processed or interpreted by the model when generating citations.

Type: String

Required: No

format

The format of a document, or its extension.

Type: String

Valid Values: `pdf | csv | doc | docx | xls | xlsx | html | txt | md`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentCharLocation

Service: Amazon Bedrock Runtime

Specifies a character-level location within a document, providing precise positioning information for cited content using start and end character indices.

Contents

documentIndex

The index of the document within the array of documents provided in the request.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

end

The ending character position of the cited content within the document.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

start

The starting character position of the cited content within the document.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentChunkLocation

Service: Amazon Bedrock Runtime

Specifies a chunk-level location within a document, providing positioning information for cited content using logical document segments or chunks.

Contents

documentIndex

The index of the document within the array of documents provided in the request.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

end

The ending chunk identifier or index of the cited content within the document.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

start

The starting chunk identifier or index of the cited content within the document.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentContentBlock

Service: Amazon Bedrock Runtime

Contains the actual content of a document that can be processed by the model and potentially cited in the response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

The text content of the document.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentPageLocation

Service: Amazon Bedrock Runtime

Specifies a page-level location within a document, providing positioning information for cited content using page numbers.

Contents

documentIndex

The index of the document within the array of documents provided in the request.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

end

The ending page number of the cited content within the document.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

start

The starting page number of the cited content within the document.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

DocumentSource

Service: Amazon Bedrock Runtime

Contains the content of a document.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

The raw bytes for the document. If you use an AWS SDK, you don't need to encode the bytes in base64.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

content

The structured content of the document source, which may include various content blocks such as text, images, or other document elements.

Type: Array of DocumentContentBlock objects

Required: No

s3Location

The location of a document object in an Amazon S3 bucket. To see which models support S3 uploads, see Supported models and features for Converse.

Type: S3Location object

Required: No

text

The text content of the document source.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ErrorBlock

Service: Amazon Bedrock Runtime

A block containing error information when content processing fails.

Contents

message

A human-readable error message describing what went wrong during content processing.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAssessment

Service: Amazon Bedrock Runtime

A behavior assessment of the guardrail policies used in a call to the Converse API.

Contents

appliedGuardrailDetails

Details about the specific guardrail that was applied during this assessment, including its identifier, version, ARN, origin, and ownership information.

Type: AppliedGuardrailDetails object

Required: No

automatedReasoningPolicy

The automated reasoning policy assessment results, including logical validation findings for the input content.

Type: GuardrailAutomatedReasoningPolicyAssessment object

Required: No

contentPolicy

The content policy.

Type: GuardrailContentPolicyAssessment object

Required: No

contextualGroundingPolicy

The contextual grounding policy used for the guardrail assessment.

Type: GuardrailContextualGroundingPolicyAssessment object

Required: No

invocationMetrics

The invocation metrics for the guardrail assessment.

Type: GuardrailInvocationMetrics object

Required: No

sensitiveInformationPolicy

The sensitive information policy.

Type: GuardrailSensitiveInformationPolicyAssessment object

Required: No

topicPolicy

The topic policy.

Type: GuardrailTopicPolicyAssessment object

Required: No

wordPolicy

The word policy.

Type: GuardrailWordPolicyAssessment object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningFinding

Service: Amazon Bedrock Runtime

Represents a logical validation result from automated reasoning policy evaluation. The finding indicates whether claims in the input are logically valid, invalid, satisfiable, impossible, or have other logical issues.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

impossible

Contains the result when the automated reasoning evaluation determines that no valid logical conclusions can be drawn due to contradictions in the premises or policy rules themselves.

Type: GuardrailAutomatedReasoningImpossibleFinding object

Required: No

invalid

Contains the result when the automated reasoning evaluation determines that the claims in the input are logically invalid and contradict the established premises or policy rules.

Type: GuardrailAutomatedReasoningInvalidFinding object

Required: No

noTranslations

Contains the result when the automated reasoning evaluation cannot extract any relevant logical information from the input that can be validated against the policy rules.

Type: GuardrailAutomatedReasoningNoTranslationsFinding object

Required: No

satisfiable

Contains the result when the automated reasoning evaluation determines that the claims in the input could be either true or false depending on additional assumptions not provided in the input context.

Type: GuardrailAutomatedReasoningSatisfiableFinding object

Required: No

tooComplex

Contains the result when the automated reasoning evaluation cannot process the input due to its complexity or volume exceeding the system's processing capacity for logical analysis.

Type: GuardrailAutomatedReasoningTooComplexFinding object

Required: No

translationAmbiguous

Contains the result when the automated reasoning evaluation detects that the input has multiple valid logical interpretations, requiring additional context or clarification to proceed with validation.

Type: GuardrailAutomatedReasoningTranslationAmbiguousFinding object

Required: No

valid

Contains the result when the automated reasoning evaluation determines that the claims in the input are logically valid and definitively true based on the provided premises and policy rules.

Type: GuardrailAutomatedReasoningValidFinding object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningImpossibleFinding

Service: Amazon Bedrock Runtime

Indicates that no valid claims can be made due to logical contradictions in the premises or rules.

Contents

contradictingRules

The automated reasoning policy rules that contradict the claims and/or premises in the input.

Type: Array of GuardrailAutomatedReasoningRule objects

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: GuardrailAutomatedReasoningLogicWarning object

Required: No

translation

The logical translation of the input that this finding evaluates.

Type: GuardrailAutomatedReasoningTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningInputTextReference

Service: Amazon Bedrock Runtime

References a portion of the original input text that corresponds to logical elements.

Contents

text

The specific text from the original input that this reference points to.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningInvalidFinding

Service: Amazon Bedrock Runtime

Indicates that the claims are logically false and contradictory to the established rules or premises.

Contents

contradictingRules

The automated reasoning policy rules that contradict the claims in the input.

Type: Array of GuardrailAutomatedReasoningRule objects

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: GuardrailAutomatedReasoningLogicWarning object

Required: No

translation

The logical translation of the input that this finding invalidates.

Type: GuardrailAutomatedReasoningTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningLogicWarning

Service: Amazon Bedrock Runtime

Identifies logical issues in the translated statements that exist independent of any policy rules, such as statements that are always true or always false.

Contents

claims

The logical statements that are validated while assuming the policy and premises.

Type: Array of GuardrailAutomatedReasoningStatement objects

Required: No

premises

The logical statements that serve as premises under which the claims are validated.

Type: Array of GuardrailAutomatedReasoningStatement objects

Required: No

type

The category of the detected logical issue, such as statements that are always true or always false.

Type: String

Valid Values: `ALWAYS_FALSE | ALWAYS_TRUE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

GuardrailAutomatedReasoningNoTranslationsFinding

Service: Amazon Bedrock Runtime

Indicates that no relevant logical information could be extracted from the input for validation.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningPolicyAssessment

Service: Amazon Bedrock Runtime

Contains the results of automated reasoning policy evaluation, including logical findings about the validity of claims made in the input content.

Contents

findings

List of logical validation results produced by evaluating the input content against automated reasoning policies.

Type: Array of GuardrailAutomatedReasoningFinding objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningRule

Service: Amazon Bedrock Runtime

References a specific automated reasoning policy rule that was applied during evaluation.

Contents

identifier

The unique identifier of the automated reasoning rule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 12.

Pattern: `[a-z0-9]{12}`

Required: No

policyVersionArn

The ARN of the automated reasoning policy version that contains this rule.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:automated-reasoning-policy/[a-z0-9]{12}(:([1-9][0-9]{0,11}))?`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningSatisfiableFinding

Service: Amazon Bedrock Runtime

Indicates that the claims could be either true or false depending on additional assumptions not provided in the input.

Contents

claimsFalseScenario

An example scenario demonstrating how the claims could be logically false.

Type: GuardrailAutomatedReasoningScenario object

Required: No

claimsTrueScenario

An example scenario demonstrating how the claims could be logically true.

Type: GuardrailAutomatedReasoningScenario object

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: GuardrailAutomatedReasoningLogicWarning object

Required: No

translation

The logical translation of the input that this finding evaluates.

Type: GuardrailAutomatedReasoningTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningScenario

Service: Amazon Bedrock Runtime

Represents a logical scenario where claims can be evaluated as true or false, containing specific logical assignments.

Contents

statements

List of logical assignments and statements that define this scenario.

Type: Array of GuardrailAutomatedReasoningStatement objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningStatement

Service: Amazon Bedrock Runtime

A logical statement that includes both formal logic representation and natural language explanation.

Contents

logic

The formal logical representation of the statement.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: No

naturalLanguage

The natural language explanation of the logical statement.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningTooComplexFinding

Service: Amazon Bedrock Runtime

Indicates that the input exceeds the processing capacity due to the volume or complexity of the logical information.

Contents

The members of this exception structure are context-dependent.

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningTranslation

Service: Amazon Bedrock Runtime

Contains the logical translation of natural language input into formal logical statements, including premises, claims, and confidence scores.

Contents

claims

The logical statements that are being validated against the premises and policy rules.

Type: Array of GuardrailAutomatedReasoningStatement objects

Required: No

confidence

A confidence score between 0 and 1 indicating how certain the system is about the logical translation.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

premises

The logical statements that serve as the foundation or assumptions for the claims.

Type: Array of GuardrailAutomatedReasoningStatement objects

Required: No

untranslatedClaims

References to portions of the original input text that correspond to the claims but could not be fully translated.

Type: Array of GuardrailAutomatedReasoningInputTextReference objects

Required: No

untranslatedPremises

References to portions of the original input text that correspond to the premises but could not be fully translated.

Type: Array of GuardrailAutomatedReasoningInputTextReference objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningTranslationAmbiguousFinding

Service: Amazon Bedrock Runtime

Indicates that the input has multiple valid logical interpretations, requiring additional context or clarification.

Contents

differenceScenarios

Scenarios showing how the different translation options differ in meaning.

Type: Array of GuardrailAutomatedReasoningScenario objects

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Required: No

options

Different logical interpretations that were detected during translation of the input.

Type: Array of GuardrailAutomatedReasoningTranslationOption objects

Array Members: Minimum number of 0 items. Maximum number of 2 items.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningTranslationOption

Service: Amazon Bedrock Runtime

Represents one possible logical interpretation of ambiguous input content.

Contents

translations

Example translations that provide this possible interpretation of the input.

Type: Array of GuardrailAutomatedReasoningTranslation objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailAutomatedReasoningValidFinding

Service: Amazon Bedrock Runtime

Indicates that the claims are definitively true and logically implied by the premises, with no possible alternative interpretations.

Contents

claimsTrueScenario

An example scenario demonstrating how the claims are logically true.

Type: GuardrailAutomatedReasoningScenario object

Required: No

logicWarning

Indication of a logic issue with the translation without needing to consider the automated reasoning policy rules.

Type: GuardrailAutomatedReasoningLogicWarning object

Required: No

supportingRules

The automated reasoning policy rules that support why this result is considered valid.

Type: Array of GuardrailAutomatedReasoningRule objects

Required: No

translation

The logical translation of the input that this finding validates.

Type: GuardrailAutomatedReasoningTranslation object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksConfig

Service: Amazon Bedrock Runtime

The configuration for inline guardrail checks. Specify one or more check types to run against the messages.

Contents

contentFilter

The content filter check configuration.

Type: GuardrailChecksContentFilterConfig object

Required: No

promptAttack

The prompt attack check configuration.

Type: GuardrailChecksPromptAttackConfig object

Required: No

sensitiveInformation

The sensitive information check configuration.

Type: GuardrailChecksSensitiveInformationConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksContentBlock

Service: Amazon Bedrock Runtime

A content block within a message to evaluate.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

The text content to evaluate.

Type: String

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksContentFilterCategoryConfig

Service: Amazon Bedrock Runtime

The configuration for a single content filter category to evaluate.

Contents

category

The content filter category to evaluate.

Type: String

Valid Values: `VIOLENCE | HATE | SEXUAL | MISCONDUCT | INSULTS`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksContentFilterConfig

Service: Amazon Bedrock Runtime

The configuration for the content filter check, specifying which categories to evaluate.

Contents

categories

The content filter categories to evaluate.

Type: Array of GuardrailChecksContentFilterCategoryConfig objects

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksContentFilterResult

Service: Amazon Bedrock Runtime

The content filter check results.

Contents

results

The per-category content filter results.

Type: Array of GuardrailChecksContentFilterResultEntry objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksContentFilterResultEntry

Service: Amazon Bedrock Runtime

The evaluation result for a single content filter category.

Contents

category

The content filter category that was evaluated.

Type: String

Valid Values: `VIOLENCE | HATE | SEXUAL | MISCONDUCT | INSULTS`

Required: Yes

severityScore

The severity score for the category, ranging from 0.0 to 1.0. Higher values indicate greater severity.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksContentFilterUsage

Service: Amazon Bedrock Runtime

The text unit usage for the content filter check.

Contents

textUnits

The number of text units consumed by the content filter check.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksMessage

Service: Amazon Bedrock Runtime

A message to evaluate against guardrail checks, containing a role and content blocks.

Contents

content

The content blocks for the message.

Type: Array of GuardrailChecksContentBlock objects

Array Members: Minimum number of 1 item.

Required: Yes

role

The role of the message sender.

Type: String

Valid Values: `user | assistant | system`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksPromptAttackCategoryConfig

Service: Amazon Bedrock Runtime

The configuration for a single prompt attack category to evaluate.

Contents

category

The prompt attack category to evaluate.

Type: String

Valid Values: `JAILBREAK | PROMPT_INJECTION | PROMPT_LEAKAGE`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksPromptAttackConfig

Service: Amazon Bedrock Runtime

The configuration for the prompt attack check, specifying which categories to evaluate.

Contents

categories

The prompt attack categories to evaluate.

Type: Array of GuardrailChecksPromptAttackCategoryConfig objects

Array Members: Minimum number of 1 item. Maximum number of 3 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksPromptAttackResult

Service: Amazon Bedrock Runtime

The prompt attack check results.

Contents

results

The per-category prompt attack results.

Type: Array of GuardrailChecksPromptAttackResultEntry objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksPromptAttackResultEntry

Service: Amazon Bedrock Runtime

The evaluation result for a single prompt attack category.

Contents

category

The prompt attack category that was evaluated.

Type: String

Valid Values: `JAILBREAK | PROMPT_INJECTION | PROMPT_LEAKAGE`

Required: Yes

severityScore

The severity score for the category, ranging from 0.0 to 1.0. Higher values indicate greater severity.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksPromptAttackUsage

Service: Amazon Bedrock Runtime

The text unit usage for the prompt attack check.

Contents

textUnits

The number of text units consumed by the prompt attack check.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksResults

Service: Amazon Bedrock Runtime

The results from the guardrail checks evaluation, organized by check type.

Contents

contentFilter

The content filter check results.

Type: GuardrailChecksContentFilterResult object

Required: No

promptAttack

The prompt attack check results.

Type: GuardrailChecksPromptAttackResult object

Required: No

sensitiveInformation

The sensitive information check results.

Type: GuardrailChecksSensitiveInformationResult object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksSensitiveInformationConfig

Service: Amazon Bedrock Runtime

The configuration for the sensitive information check, specifying which entity types to detect.

Contents

entities

The sensitive information entity types to detect.

Type: Array of GuardrailChecksSensitiveInformationEntityConfig objects

Array Members: Minimum number of 1 item. Maximum number of 31 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksSensitiveInformationEntityConfig

Service: Amazon Bedrock Runtime

The configuration for a single sensitive information entity type to detect.

Contents

type

The PII entity type to detect.

Type: String

Valid Values: `ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksSensitiveInformationResult

Service: Amazon Bedrock Runtime

The sensitive information check results.

Contents

results

The detected sensitive information entities.

Type: Array of GuardrailChecksSensitiveInformationResultEntry objects

Required: Yes

truncated

Specifies whether the results were truncated because the number of detected entities exceeded the maximum limit.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksSensitiveInformationResultEntry

Service: Amazon Bedrock Runtime

The detection result for a single sensitive information entity found in the evaluated messages.

Contents

beginOffset

The start character offset of the detected entity within the content block.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

confidenceScore

The confidence score for the detection, ranging from 0.0 to 1.0. Higher values indicate greater confidence.

Type: Double

Valid Range: Minimum value of 0.0. Maximum value of 1.0.

Required: Yes

contentIndex

The zero-based index of the content block within the message where the entity was detected.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

endOffset

The end character offset of the detected entity within the content block.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

messageIndex

The zero-based index of the message in the input messages array where the entity was detected.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

type

The PII entity type that was detected.

Type: String

Valid Values: `ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksSensitiveInformationUsage

Service: Amazon Bedrock Runtime

The text unit usage for the sensitive information check.

Contents

textUnits

The number of text units consumed by the sensitive information check.

Type: Integer

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailChecksUsageResults

Service: Amazon Bedrock Runtime

The text unit usage for the guardrail checks evaluation, organized by check type.

Contents

contentFilter

The text unit usage for the content filter check.

Type: GuardrailChecksContentFilterUsage object

Required: No

promptAttack

The text unit usage for the prompt attack check.

Type: GuardrailChecksPromptAttackUsage object

Required: No

sensitiveInformation

The text unit usage for the sensitive information check.

Type: GuardrailChecksSensitiveInformationUsage object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConfiguration

Service: Amazon Bedrock Runtime

Configuration information for a guardrail that you use with the Converse operation.

Contents

guardrailIdentifier

The identifier for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(|([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: No

guardrailVersion

The version of the guardrail.

Type: String

Pattern: `(|([1-9][0-9]{0,7})|(DRAFT))`

Required: No

trace

The trace behavior for the guardrail.

Type: String

Valid Values: `enabled | disabled | enabled_full`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentBlock

Service: Amazon Bedrock Runtime

The content block to be evaluated by the guardrail.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

image

Image within guardrail content block to be evaluated by the guardrail.

Type: GuardrailImageBlock object

Required: No

text

Text within content block to be evaluated by the guardrail.

Type: GuardrailTextBlock object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentFilter

Service: Amazon Bedrock Runtime

The content filter for a guardrail.

Contents

action

The guardrail action.

Type: String

Valid Values: `BLOCKED | NONE`

Required: Yes

confidence

The guardrail confidence.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: Yes

type

The guardrail type.

Type: String

Valid Values: `INSULTS | HATE | SEXUAL | VIOLENCE | MISCONDUCT | PROMPT_ATTACK`

Required: Yes

detected

Indicates whether content that breaches the guardrail configuration is detected.

Type: Boolean

Required: No

filterStrength

The filter strength setting for the guardrail content filter.

Type: String

Valid Values: `NONE | LOW | MEDIUM | HIGH`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContentPolicyAssessment

Service: Amazon Bedrock Runtime

An assessment of a content policy for a guardrail.

Contents

filters

The content policy filters.

Type: Array of GuardrailContentFilter objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContextualGroundingFilter

Service: Amazon Bedrock Runtime

The details for the guardrails contextual grounding filter.

Contents

action

The action performed by the guardrails contextual grounding filter.

Type: String

Valid Values: `BLOCKED | NONE`

Required: Yes

score

The score generated by contextual grounding filter.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: Yes

threshold

The threshold used by contextual grounding filter to determine whether the content is grounded or not.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

Required: Yes

type

The contextual grounding filter type.

Type: String

Valid Values: `GROUNDING | RELEVANCE`

Required: Yes

detected

Indicates whether content that fails the contextual grounding evaluation (grounding or relevance score less than the corresponding threshold) was detected.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailContextualGroundingPolicyAssessment

Service: Amazon Bedrock Runtime

The policy assessment details for the guardrails contextual grounding filter.

Contents

filters

The filter details for the guardrails contextual grounding filter.

Type: Array of GuardrailContextualGroundingFilter objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConverseContentBlock

Service: Amazon Bedrock Runtime

A content block for selective guarding with the Converse or ConverseStream API operations.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

image

Image within converse content block to be evaluated by the guardrail.

Type: GuardrailConverseImageBlock object

Required: No

text

The text to guard.

Type: GuardrailConverseTextBlock object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConverseImageBlock

Service: Amazon Bedrock Runtime

An image block that contains images that you want to assess with a guardrail.

Contents

format

The format details for the image type of the guardrail converse image block.

Type: String

Valid Values: `png | jpeg`

Required: Yes

source

The image source (image bytes) of the guardrail converse image block.

Type: GuardrailConverseImageSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConverseImageSource

Service: Amazon Bedrock Runtime

The image source (image bytes) of the guardrail converse image source.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

The raw image bytes for the image.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailConverseTextBlock

Service: Amazon Bedrock Runtime

A text block that contains text that you want to assess with a guardrail. For more information, see GuardrailConverseContentBlock.

Contents

text

The text that you want to guard.

Type: String

Required: Yes

qualifiers

The qualifier details for the guardrails contextual grounding filter.

Type: Array of strings

Valid Values: `grounding_source | query | guard_content`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailCoverage

Service: Amazon Bedrock Runtime

The action of the guardrail coverage details.

Contents

images

The guardrail coverage for images (the number of images that guardrails guarded).

Type: GuardrailImageCoverage object

Required: No

textCharacters

The text characters of the guardrail coverage details.

Type: GuardrailTextCharactersCoverage object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailCustomWord

Service: Amazon Bedrock Runtime

A custom word configured in a guardrail.

Contents

action

The action for the custom word.

Type: String

Valid Values: `BLOCKED | NONE`

Required: Yes

match

The match for the custom word.

Type: String

Required: Yes

detected

Indicates whether custom word content that breaches the guardrail configuration is detected.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailImageBlock

Service: Amazon Bedrock Runtime

Contain an image which user wants guarded. This block is accepted by the guardrails independent API.

Contents

format

The format details for the file type of the image blocked by the guardrail.

Type: String

Valid Values: `png | jpeg`

Required: Yes

source

The image source (image bytes) details of the image blocked by the guardrail.

Type: GuardrailImageSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailImageCoverage

Service: Amazon Bedrock Runtime

The details of the guardrail image coverage.

Contents

guarded

The count (integer) of images guardrails guarded.

Type: Integer

Required: No

total

Represents the total number of images (integer) that were in the request (guarded and unguarded).

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailImageSource

Service: Amazon Bedrock Runtime

The image source (image bytes) of the guardrail image source. Object used in independent api.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

The bytes details of the guardrail image source. Object used in independent api.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailInvocationMetrics

Service: Amazon Bedrock Runtime

The invocation metrics for the guardrail.

Contents

guardrailCoverage

The coverage details for the guardrail invocation metrics.

Type: GuardrailCoverage object

Required: No

guardrailProcessingLatency

The processing latency details for the guardrail invocation metrics.

Type: Long

Required: No

usage

The usage details for the guardrail invocation metrics.

Type: GuardrailUsage object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailManagedWord

Service: Amazon Bedrock Runtime

A managed word configured in a guardrail.

Contents

action

The action for the managed word.

Type: String

Valid Values: `BLOCKED | NONE`

Required: Yes

match

The match for the managed word.

Type: String

Required: Yes

type

The type for the managed word.

Type: String

Valid Values: `PROFANITY`

Required: Yes

detected

Indicates whether managed word content that breaches the guardrail configuration is detected.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailOutputContent

Service: Amazon Bedrock Runtime

The output content produced by the guardrail.

Contents

text

The specific text for the output content produced by the guardrail.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailPiiEntityFilter

Service: Amazon Bedrock Runtime

A Personally Identifiable Information (PII) entity configured in a guardrail.

Contents

action

The PII entity filter action.

Type: String

Valid Values: `ANONYMIZED | BLOCKED | NONE`

Required: Yes

match

The PII entity filter match.

Type: String

Required: Yes

type

The PII entity filter type.

Type: String

Valid Values: `ADDRESS | AGE | AWS_ACCESS_KEY | AWS_SECRET_KEY | CA_HEALTH_NUMBER | CA_SOCIAL_INSURANCE_NUMBER | CREDIT_DEBIT_CARD_CVV | CREDIT_DEBIT_CARD_EXPIRY | CREDIT_DEBIT_CARD_NUMBER | DRIVER_ID | EMAIL | INTERNATIONAL_BANK_ACCOUNT_NUMBER | IP_ADDRESS | LICENSE_PLATE | MAC_ADDRESS | NAME | PASSWORD | PHONE | PIN | SWIFT_CODE | UK_NATIONAL_HEALTH_SERVICE_NUMBER | UK_NATIONAL_INSURANCE_NUMBER | UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER | URL | USERNAME | US_BANK_ACCOUNT_NUMBER | US_BANK_ROUTING_NUMBER | US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER | US_PASSPORT_NUMBER | US_SOCIAL_SECURITY_NUMBER | VEHICLE_IDENTIFICATION_NUMBER`

Required: Yes

detected

Indicates whether personally identifiable information (PII) that breaches the guardrail configuration is detected.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailRegexFilter

Service: Amazon Bedrock Runtime

A Regex filter configured in a guardrail.

Contents

action

The region filter action.

Type: String

Valid Values: `ANONYMIZED | BLOCKED | NONE`

Required: Yes

detected

Indicates whether custom regex entities that breach the guardrail configuration are detected.

Type: Boolean

Required: No

match

The regesx filter match.

Type: String

Required: No

name

The regex filter name.

Type: String

Required: No

regex

The regex query.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailSensitiveInformationPolicyAssessment

Service: Amazon Bedrock Runtime

The assessment for a Personally Identifiable Information (PII) policy.

Contents

piiEntities

The PII entities in the assessment.

Type: Array of GuardrailPiiEntityFilter objects

Required: Yes

regexes

The regex queries in the assessment.

Type: Array of GuardrailRegexFilter objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailStreamConfiguration

Service: Amazon Bedrock Runtime

Configuration information for a guardrail that you use with the ConverseStream action.

Contents

guardrailIdentifier

The identifier for the guardrail.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `(|([a-z0-9]+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:guardrail/[a-z0-9]+))`

Required: No

guardrailVersion

The version of the guardrail.

Type: String

Pattern: `(|([1-9][0-9]{0,7})|(DRAFT))`

Required: No

streamProcessingMode

The processing mode.

The processing mode. For more information, see Configure streaming response behavior.

Type: String

Valid Values: `sync | async`

Required: No

trace

The trace behavior for the guardrail.

Type: String

Valid Values: `enabled | disabled | enabled_full`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTextBlock

Service: Amazon Bedrock Runtime

The text block to be evaluated by the guardrail.

Contents

text

The input text details to be evaluated by the guardrail.

Type: String

Required: Yes

qualifiers

The qualifiers describing the text block.

Type: Array of strings

Valid Values: `grounding_source | query | guard_content`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTextCharactersCoverage

Service: Amazon Bedrock Runtime

The guardrail coverage for the text characters.

Contents

guarded

The text characters that were guarded by the guardrail coverage.

Type: Integer

Required: No

total

The total text characters by the guardrail coverage.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopic

Service: Amazon Bedrock Runtime

Information about a topic guardrail.

Contents

action

The action the guardrail should take when it intervenes on a topic.

Type: String

Valid Values: `BLOCKED | NONE`

Required: Yes

name

The name for the guardrail.

Type: String

Required: Yes

type

The type behavior that the guardrail should perform when the model detects the topic.

Type: String

Valid Values: `DENY`

Required: Yes

detected

Indicates whether topic content that breaches the guardrail configuration is detected.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTopicPolicyAssessment

Service: Amazon Bedrock Runtime

A behavior assessment of a topic policy.

Contents

topics

The topics in the assessment.

Type: Array of GuardrailTopic objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailTraceAssessment

Service: Amazon Bedrock Runtime

A Top level guardrail trace object. For more information, see ConverseTrace.

Contents

actionReason

Provides the reason for the action taken when harmful content is detected.

Type: String

Required: No

inputAssessment

The input assessment.

Type: String to GuardrailAssessment object map

Required: No

modelOutput

The output from the model.

Type: Array of strings

Required: No

outputAssessments

the output assessments.

Type: String to array of GuardrailAssessment objects map

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailUsage

Service: Amazon Bedrock Runtime

The details on the use of the guardrail.

Contents

contentPolicyUnits

The content policy units processed by the guardrail.

Type: Integer

Required: Yes

contextualGroundingPolicyUnits

The contextual grounding policy units processed by the guardrail.

Type: Integer

Required: Yes

sensitiveInformationPolicyFreeUnits

The sensitive information policy free units processed by the guardrail.

Type: Integer

Required: Yes

sensitiveInformationPolicyUnits

The sensitive information policy units processed by the guardrail.

Type: Integer

Required: Yes

topicPolicyUnits

The topic policy units processed by the guardrail.

Type: Integer

Required: Yes

wordPolicyUnits

The word policy units processed by the guardrail.

Type: Integer

Required: Yes

automatedReasoningPolicies

The number of automated reasoning policies that were processed during the guardrail evaluation.

Type: Integer

Required: No

automatedReasoningPolicyUnits

The number of text units processed by the automated reasoning policy.

Type: Integer

Required: No

contentPolicyImageUnits

The content policy image units processed by the guardrail.

Type: Integer

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

GuardrailWordPolicyAssessment

Service: Amazon Bedrock Runtime

The word policy assessment.

Contents

customWords

Custom words in the assessment.

Type: Array of GuardrailCustomWord objects

Required: Yes

managedWordLists

Managed word lists in the assessment.

Type: Array of GuardrailManagedWord objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageBlock

Service: Amazon Bedrock Runtime

Image content for a message.

Contents

format

The format of the image.

Type: String

Valid Values: `png | jpeg | gif | webp`

Required: Yes

source

The source for the image.

Type: ImageSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

error

Error information if the image block could not be processed or contains invalid data.

Type: ErrorBlock object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageBlockDelta

Service: Amazon Bedrock Runtime

A streaming delta event that contains incremental image data during streaming responses.

Contents

error

Error information if this image delta could not be processed.

Type: ErrorBlock object

Required: No

source

The incremental image source data for this delta event.

Type: ImageSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageBlockStart

Service: Amazon Bedrock Runtime

The initial event in a streaming image block that indicates the start of image content.

Contents

format

The format of the image data that will be streamed in subsequent delta events.

Type: String

Valid Values: `png | jpeg | gif | webp`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ImageSource

Service: Amazon Bedrock Runtime

The source for an image.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

The raw image bytes for the image. If you use an AWS SDK, you don't need to encode the image bytes in base64.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

s3Location

The location of an image object in an Amazon S3 bucket. To see which models support S3 uploads, see Supported models and features for Converse.

Type: S3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

InferenceConfiguration

Service: Amazon Bedrock Runtime

Base inference parameters to pass to a model in a call to Converse or ConverseStream. For more information, see Inference parameters for foundation models.

If you need to pass additional parameters that the model supports, use the `additionalModelRequestFields` request field in the call to `Converse` or `ConverseStream`. For more information, see Model parameters.

Contents

maxTokens

The maximum number of tokens to allow in the generated response. The default value is the maximum allowed value for the model that you are using. For more information, see Inference parameters for foundation models.

Type: Integer

Valid Range: Minimum value of 1.

Required: No

stopSequences

A list of stop sequences. A stop sequence is a sequence of characters that causes the model to stop generating the response.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 2500 items.

Length Constraints: Minimum length of 1.

Required: No

temperature

The likelihood of the model selecting higher-probability options while generating a response. A lower value makes the model more likely to choose higher-probability options, while a higher value makes the model more likely to choose lower-probability options.

The default value is the default value for the model that you are using. For more information, see Inference parameters for foundation models.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

topP

The percentage of most-likely candidates that the model considers for the next token. For example, if you choose a value of 0.8 for `topP`, the model selects from the top 80% of the probability distribution of tokens that could be next in the sequence.

The default value is the default value for the model that you are using. For more information, see Inference parameters for foundation models.

Type: Float

Valid Range: Minimum value of 0. Maximum value of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvokeModelTokensRequest

Service: Amazon Bedrock Runtime

The body of an `InvokeModel` API request for token counting. This structure mirrors the input format for the `InvokeModel` operation, allowing you to count tokens for raw text inference requests.

Contents

body

The request body to count tokens for, formatted according to the model's expected input format. To learn about the input format for different models, see Model inference parameters and responses.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvokeModelWithBidirectionalStreamInput

Service: Amazon Bedrock Runtime

Payload content, the speech chunk, for the bidirectional input of the invocation step.

Contents

chunk

The audio chunk that is used as input for the invocation step.

Type: BidirectionalInputPayloadPart object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InvokeModelWithBidirectionalStreamOutput

Service: Amazon Bedrock Runtime

Output from the bidirectional stream that was used for model invocation.

Contents

chunk

The speech chunk that was provided as output from the invocation step.

Type: BidirectionalOutputPayloadPart object

Required: No

internalServerException

The request encountered an unknown internal error.

Type: Exception

HTTP Status Code: 500

Required: No

modelStreamErrorException

The request encountered an error with the model stream.

Type: Exception

HTTP Status Code: 424

Required: No

modelTimeoutException

The connection was closed because a request was not received within the timeout period.

Type: Exception

HTTP Status Code: 408

Required: No

serviceUnavailableException

The request has failed due to a temporary failure of the server.

Type: Exception

HTTP Status Code: 503

Required: No

throttlingException

The request was denied due to request throttling.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

The input fails to satisfy the constraints specified by an AWS service.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JsonSchemaDefinition

Service: Amazon Bedrock Runtime

JSON schema structured output format options.

Contents

schema

The JSON schema to constrain the model's output. For more information, see JSON Schema Reference.

Type: String

Required: Yes

description

A description of the JSON schema.

Type: String

Required: No

name

The name of the JSON schema.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Message

Service: Amazon Bedrock Runtime

A message input, or returned from, a call to Converse or ConverseStream.

Contents

content

The message content. Note the following restrictions:

- You can include up to 20 images. Each image's size, height, and width must be no more than 3.75 MB, 8000 px, and 8000 px, respectively.
- You can include up to five documents. Each document's size must be no more than 4.5 MB.
- If you include a `ContentBlock` with a `document` field in the array, you must also include a `ContentBlock` with a `text` field.
- You can only include images and documents if the `role` is `user`.

Type: Array of ContentBlock objects

Required: Yes

role

The role that the message plays in the message.

Type: String

Valid Values: `user | assistant | system`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MessageStartEvent

Service: Amazon Bedrock Runtime

The start of a message.

Contents

role

The role for the message.

Type: String

Valid Values: `user | assistant | system`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MessageStopEvent

Service: Amazon Bedrock Runtime

The stop event for a message.

Contents

stopReason

The reason why the model stopped generating output.

Type: String

Valid Values: `end_turn | tool_use | max_tokens | stop_sequence | guardrail_intervened | content_filtered | malformed_model_output | malformed_tool_use | model_context_window_exceeded`

Required: Yes

additionalModelResponseFields

The additional model response fields.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputConfig

Service: Amazon Bedrock Runtime

Output configuration for a model response in a call to Converse or ConverseStream.

Contents

textFormat

Structured output parameters to control the model's text response.

Type: OutputFormat object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputFormat

Service: Amazon Bedrock Runtime

Structured output parameters to control the model's response.

Contents

structure

The structure that the model's output must adhere to.

Type: OutputFormatStructure object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

type

The type of structured output format.

Type: String

Valid Values: `json_schema`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputFormatStructure

Service: Amazon Bedrock Runtime

The structure that the model's output must adhere to.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

jsonSchema

A JSON schema structure that the model's output must adhere to.

Type: JsonSchemaDefinition object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PayloadPart

Service: Amazon Bedrock Runtime

Payload content included in the response.

Contents

bytes

Base64-encoded bytes of payload data.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PerformanceConfiguration

Service: Amazon Bedrock Runtime

Performance settings for a model.

Contents

latency

To use a latency-optimized version of the model, set to `optimized`.

Type: String

Valid Values: `standard | optimized`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptRouterTrace

Service: Amazon Bedrock Runtime

A prompt router trace.

Contents

invokedModelId

The ID of the invoked model.

Type: String

Pattern: `(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})|(arn:aws(|-us-gov|-cn|-iso|-iso-b):bedrock:(|[0-9a-z-]{1,20}):(|[0-9]{12}):inference-profile/[a-zA-Z0-9-:.]+)`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PromptVariableValues

Service: Amazon Bedrock Runtime

Contains a map of variables in a prompt from Prompt management to an object containing the values to fill in for them when running model invocation. For more information, see How Prompt management works.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

text

The text value that the variable maps to.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReasoningContentBlock

Service: Amazon Bedrock Runtime

Contains content regarding the reasoning that is carried out by the model with respect to the content in the content block. Reasoning refers to a Chain of Thought (CoT) that the model generates to enhance the accuracy of its final response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

reasoningText

The reasoning that the model used to return the output.

Type: ReasoningTextBlock object

Required: No

redactedContent

The content in the reasoning that was encrypted by the model provider for safety reasons. The encryption doesn't affect the quality of responses.

Type: Base64-encoded binary data object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReasoningContentBlockDelta

Service: Amazon Bedrock Runtime

Contains content regarding the reasoning that is carried out by the model with respect to the content in the content block. Reasoning refers to a Chain of Thought (CoT) that the model generates to enhance the accuracy of its final response.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

redactedContent

The content in the reasoning that was encrypted by the model provider for safety reasons. The encryption doesn't affect the quality of responses.

Type: Base64-encoded binary data object

Required: No

signature

A token that verifies that the reasoning text was generated by the model. If you pass a reasoning block back to the API in a multi-turn conversation, include the text and its signature unmodified.

Type: String

Required: No

text

The reasoning that the model used to return the output.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ReasoningTextBlock

Service: Amazon Bedrock Runtime

Contains the reasoning that the model used to return the output.

Contents

text

The reasoning that the model used to return the output.

Type: String

Required: Yes

signature

A token that verifies that the reasoning text was generated by the model. If you pass a reasoning block back to the API in a multi-turn conversation, include the text and its signature unmodified.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResponseStream

Service: Amazon Bedrock Runtime

Definition of content in the response stream.

Contents

chunk

Content included in the response.

Type: PayloadPart object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception

HTTP Status Code: 500

Required: No

modelStreamErrorException

An error occurred while streaming the response. Retry your request.

Type: Exception

HTTP Status Code: 424

Required: No

modelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

Type: Exception

HTTP Status Code: 408

Required: No

serviceUnavailableException

The service isn't available. Try again later.

Type: Exception

HTTP Status Code: 503

Required: No

throttlingException

Your request was throttled because of service-wide limitations. Resubmit your request later or in a different region. You can also purchase Provisioned Throughput to increase the rate or number of tokens you can process.

Type: Exception

HTTP Status Code: 429

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception

HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Location

Service: Amazon Bedrock Runtime

A storage location in an Amazon S3 bucket.

Contents

uri

An object URI starting with `s3://`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?`

Required: Yes

bucketOwner

If the bucket belongs to another AWS account, specify that account's ID.

Type: String

Pattern: `[0-9]{12}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SearchResultBlock

Service: Amazon Bedrock Runtime

A search result block that enables natural citations with proper source attribution for retrieved content.

> **Note**
>
> This field is only supported by Anthropic Claude Opus 4.1, Opus 4, Sonnet 4.5, Sonnet 4, Sonnet 3.7, and 3.5 Haiku models.

Contents

content

An array of search result content block.

Type: Array of SearchResultContentBlock objects

Required: Yes

source

The source URL or identifier for the content.

Type: String

Required: Yes

title

A descriptive title for the search result.

Type: String

Required: Yes

citations

Configuration setting for citations

Type: CitationsConfig object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SearchResultContentBlock

Service: Amazon Bedrock Runtime

A block within a search result that contains the content.

Contents

text

The actual text content

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SearchResultLocation

Service: Amazon Bedrock Runtime

Specifies a search result location within the content array, providing positioning information for cited content using search result index and block positions.

Contents

end

The ending position in the content array where the cited content ends.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

searchResultIndex

The index of the search result content block where the cited content is found.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

start

The starting position in the content array where the cited content begins.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ServiceTier

Service: Amazon Bedrock Runtime

Specifies the processing tier configuration used for serving the request.

Contents

type

Specifies the processing tier type used for serving the request.

Type: String

Valid Values: `priority | default | flex | reserved`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpecificToolChoice

Service: Amazon Bedrock Runtime

The model must request a specific tool. For example, `{"tool" : {"name" : "Your tool name"}}`. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide

> **Note**
>
> This field is only supported by Anthropic Claude 3 models.

Contents

name

The name of the tool that the model must request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_-]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SystemContentBlock

Service: Amazon Bedrock Runtime

Contains configurations for instructions to provide the model for how to handle input. To learn more, see Using the Converse API.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

cachePoint

CachePoint to include in the system prompt.

Type: CachePointBlock object

Required: No

guardContent

A content block to assess with the guardrail. Use with the Converse or ConverseStream API operations.

For more information, see Use a guardrail with the Converse API.

Type: GuardrailConverseContentBlock object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

text

A system prompt for the model.

Type: String

Length Constraints: Minimum length of 1.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SystemTool

Service: Amazon Bedrock Runtime

Specifies a system-defined tool for the model to use. *System-defined tools* are tools that are created and provided by the model provider.

Contents

name

The name of the system-defined tool that you want to call.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_-]+`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tag

Service: Amazon Bedrock Runtime

A tag.

Contents

key

The tag's key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

value

The tag's value.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `[a-zA-Z0-9\s._:/=+@-]*`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TokenUsage

Service: Amazon Bedrock Runtime

The tokens used in a message API inference call.

Contents

inputTokens

The number of tokens sent in the request to the model.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

outputTokens

The number of tokens that the model generated for the request.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

totalTokens

The total of input tokens and tokens generated by the model.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

cacheDetails

Detailed breakdown of cache writes by TTL. Empty if no cache creation occurred. Sorted by TTL duration (1h before 5m).

Type: Array of CacheDetail objects

Required: No

cacheReadInputTokens

The number of input tokens read from the cache for the request.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

cacheWriteInputTokens

The number of input tokens written to the cache for the request.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tool

Service: Amazon Bedrock Runtime

Information about a tool that you can use with the Converse API. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

cachePoint

CachePoint to include in the tool configuration.

Type: CachePointBlock object

Required: No

systemTool

Specifies the system-defined tool that you want use.

Type: SystemTool object

Required: No

toolSpec

The specfication for the tool.

Type: ToolSpecification object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolChoice

Service: Amazon Bedrock Runtime

Determines which tools the model should request in a call to `Converse` or `ConverseStream`. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

any

The model must request at least one tool (no text is generated).

Type: AnyToolChoice object

Required: No

auto

(Default). The Model automatically decides if a tool should be called or whether to generate text instead.

Type: AutoToolChoice object

Required: No

tool

The Model must request the specified tool. Only supported by Anthropic Claude 3 and Amazon Nova models.

Type: SpecificToolChoice object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolConfiguration

Service: Amazon Bedrock Runtime

Configuration information for the tools that you pass to a model. For more information, see Tool use (function calling) in the Amazon Bedrock User Guide.

Contents

tools

An array of tools that you want to pass to a model.

Type: Array of Tool objects

Array Members: Minimum number of 1 item.

Required: Yes

toolChoice

If supported by model, forces the model to request a tool.

Type: ToolChoice object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolInputSchema

Service: Amazon Bedrock Runtime

The schema for the tool. The top level schema type must be `object`. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

json

The JSON schema for the tool. For more information, see JSON Schema Reference.

Type: JSON value

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolResultBlock

Service: Amazon Bedrock Runtime

A tool result block that contains the results for a tool request that the model previously made. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

content

The content for tool result content block.

Type: Array of ToolResultContentBlock objects

Required: Yes

toolUseId

The ID of the tool request that this is the result for.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_.:-]+`

Required: Yes

status

The status for the tool result content block.

> **Note**
>
> This field is only supported by Amazon Nova and Anthropic Claude 3 and 4 models.

Type: String

Valid Values: `success | error`

Required: No

type

The type for the tool result content block.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolResultBlockDelta

Service: Amazon Bedrock Runtime

Contains incremental updates to tool results information during streaming responses. This allows clients to build up tool results data progressively as the response is generated.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

json

The JSON schema for the tool result content block. see JSON Schema Reference.

Type: JSON value

Required: No

text

The reasoning the model used to return the output.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolResultBlockStart

Service: Amazon Bedrock Runtime

The start of a tool result block. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

toolUseId

The ID of the tool that was used to generate this tool result block.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_.:-]+`

Required: Yes

status

The status of the tool result block.

Type: String

Valid Values: `success | error`

Required: No

type

The type for the tool that was used to generate this tool result block.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolResultContentBlock

Service: Amazon Bedrock Runtime

The tool result content block. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

document

A tool result that is a document.

Type: DocumentBlock object

Required: No

image

A tool result that is an image.

> **Note**
>
> This field is only supported by Amazon Nova and Anthropic Claude 3 and 4 models.

Type: ImageBlock object

Required: No

json

A tool result that is JSON format data.

Type: JSON value

Required: No

searchResult

A tool result that is a search result.

Type: SearchResultBlock object

Required: No

text

A tool result that is text.

Type: String

Required: No

video

A tool result that is video.

Type: VideoBlock object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolSpecification

Service: Amazon Bedrock Runtime

The specification for the tool. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

inputSchema

The input schema for the tool in JSON format.

Type: ToolInputSchema object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

name

The name for the tool.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_-]+`

Required: Yes

description

The description for the tool.

Type: String

Length Constraints: Minimum length of 1.

Required: No

strict

Flag to enable structured output enforcement on a tool usage response.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolUseBlock

Service: Amazon Bedrock Runtime

A tool use content block. Contains information about a tool that the model is requesting be run., The model uses the result from the tool to generate a response. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

input

The input to pass to the tool.

Type: JSON value

Required: Yes

name

The name of the tool that the model wants to use.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_-]+`

Required: Yes

toolUseId

The ID for the tool request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_.:-]+`

Required: Yes

type

The type for the tool request.

Type: String

Valid Values: `server_tool_use`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolUseBlockDelta

Service: Amazon Bedrock Runtime

The delta for a tool use block.

Contents

input

The input for a requested tool.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ToolUseBlockStart

Service: Amazon Bedrock Runtime

The start of a tool use block. For more information, see Call a tool with the Converse API in the Amazon Bedrock User Guide.

Contents

name

The name of the tool that the model is requesting to use.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_-]+`

Required: Yes

toolUseId

The ID for the tool request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9_.:-]+`

Required: Yes

type

The type for the tool request.

Type: String

Valid Values: `server_tool_use`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoBlock

Service: Amazon Bedrock Runtime

A video block.

Contents

format

The block's format.

Type: String

Valid Values: `mkv | mov | mp4 | webm | flv | mpeg | mpg | wmv | three_gp`

Required: Yes

source

The block's source.

Type: VideoSource object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VideoSource

Service: Amazon Bedrock Runtime

A video source. You can upload a smaller video as a base64-encoded string as long as the encoded file is less than 25MB. You can also transfer videos up to 1GB in size from an S3 bucket.

Contents

> **Important**
>
> This data type is a UNION, so only one of the following members can be specified when used or returned.

bytes

Video content encoded in base64.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 1.

Required: No

s3Location

The location of a video object in an Amazon S3 bucket. To see which models support S3 uploads, see Supported models and features for Converse.

Type: S3Location object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

WebLocation

Service: Amazon Bedrock Runtime

Provides the URL and domain information for the website that was cited when performing a web search.

Contents

domain

The domain that was cited when performing a web search.

Type: String

Required: No

url

The URL that was cited when performing a web search.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Common Parameters

The following list contains the parameters that all actions use for signing Signature Version 4 requests with a query string. Any action-specific parameters are listed in the topic for that action. For more information about Signature Version 4, see Signing AWS API requests in the *IAM User Guide*.

X-Amz-Algorithm

The hash algorithm that you used to create the request signature.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Valid Values: `AWS4-HMAC-SHA256`

Required: Conditional

X-Amz-Credential

The credential scope value, which is a string that includes your access key, the date, the region you are targeting, the service you are requesting, and a termination string ("aws4_request"). The value is expressed in the following format: *access_key*/*YYYYMMDD*/*region*/*service*/aws4_request.

For more information, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-Date

The date that is used to create the signature. The format must be ISO 8601 basic format (YYYYMMDD'T'HHMMSS'Z'). For example, the following date time is a valid X-Amz-Date value: `20120325T120000Z`.

Condition: X-Amz-Date is optional for all requests; it can be used to override the date used for signing requests. If the Date header is specified in the ISO 8601 basic format, X-Amz-Date is not required. When X-Amz-Date is used, it always overrides the value of the Date header. For more information, see Elements of an AWS API request signature in the *IAM User Guide*.

Type: string

Required: Conditional

X-Amz-Security-Token

The temporary security token that was obtained through a call to AWS Security Token Service (AWS STS). For a list of services that support temporary security credentials from AWS STS, see AWS services that work with IAM in the *IAM User Guide*.

Condition: If you're using temporary security credentials from AWS STS, you must include the security token.

Type: string

Required: Conditional

X-Amz-Signature

Specifies the hex-encoded signature that was calculated from the string to sign and the derived signing key.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-SignedHeaders

Specifies all the HTTP headers that were included as part of the canonical request. For more information about specifying signed headers, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

Common Error Types

This section lists common error types that this AWS service may return. Not all services return all error types listed here. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You don't have permission to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 403

ExpiredTokenException

The security token included in the request has expired. Request a new security token and try again.

HTTP Status Code: 403

IncompleteSignature

The request signature doesn't conform to AWS standards. Verify that you're using valid AWS credentials and that your request is properly formatted. If you're using an SDK, ensure it's up to date.

HTTP Status Code: 403

InternalFailure

The request can't be processed right now because of an internal server issue. Try again later. If the problem persists, contact AWS Support.

HTTP Status Code: 500

MalformedHttpRequestException

The request body can't be processed. This typically happens when the request body can't be decompressed using the specified content encoding algorithm. Verify that the content encoding header matches the compression format used.

HTTP Status Code: 400

NotAuthorized

You don't have permissions to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 401

OptInRequired

Your AWS account needs a subscription for this service. Verify that you've enabled the service in your account.

HTTP Status Code: 403

RequestAbortedException

The request was aborted before a response could be returned. This typically happens when the client closes the connection.

HTTP Status Code: 400

RequestEntityTooLargeException

The request entity is too large. Reduce the size of the request body and try again.

HTTP Status Code: 413

RequestTimeoutException

The request timed out. The server didn't receive the complete request within the expected time frame. Try again.

HTTP Status Code: 408

ServiceUnavailable

The service is temporarily unavailable. Try again later.

HTTP Status Code: 503

ThrottlingException

Your request rate is too high. The AWS SDKs automatically retry requests that receive this exception. Reduce the frequency of requests.

HTTP Status Code: 400

UnknownOperationException

The action or operation isn't recognized. Verify that the action name is spelled correctly and that it's supported by the API version you're using.

HTTP Status Code: 404

UnrecognizedClientException

The X.509 certificate or AWS access key ID you provided doesn't exist in our records. Verify that you're using valid credentials and that they haven't expired.

HTTP Status Code: 403

ValidationError

The input doesn't meet the required format or constraints. Check that all required parameters are included and that values are valid.

HTTP Status Code: 400