



User Guide

Amazon Bedrock



Amazon Bedrock: User Guide

Copyright © 2024 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.

Table of Contents

What is Amazon Bedrock?	1
What can I do with Amazon Bedrock?	1
How do I get started with Amazon Bedrock?	2
Amazon Bedrock pricing	3
Key terminology	3
Getting started	6
Request access to an Amazon Bedrock foundation model	9
(Optional tutorials) Explore Amazon Bedrock features through the console or API	10
Getting started in the console	10
Explore the text playground	11
Explore the image playground	11
Getting started with the API	12
Get credentials to grant programmatic access	13
Attach Amazon Bedrock permissions to a user or role	17
Request access to Amazon Bedrock models	18
Try making API calls to Amazon Bedrock	18
Run examples with the AWS CLI	18
Run examples with the AWS SDK for Python (Boto3)	20
Run examples with a SageMaker AI notebook	25
Working with AWS SDKs	29
Access foundation models	31
Grant permissions to request access to foundation models	31
Add or remove access to foundation models	34
Foundation model information	37
Get model information	39
Supported foundation models	40
Model support by Region	71
Feature support by Region	83
Model support by feature	86
Model inference parameters and responses	99
Amazon Nova models	100
Amazon Titan models	101
Anthropic Claude models	160
Cohere models	189

AI21 Labs models	212
Meta Llama models	222
Mistral AI models	228
Stability AI models	247
Custom model hyperparameters	274
Amazon Nova	275
Amazon Titan text models	276
Amazon Titan Image Generator G1 models	279
Amazon Titan Multimodal Embeddings G1	280
Anthropic Claude 3 models	282
Cohere Command models	284
Meta Llama 2 models	286
Meta Llama 3.1 models	288
Model lifecycle	289
On-Demand, Provisioned Throughput, and model customization	290
Legacy versions	290
Amazon Bedrock Marketplace	294
Set up Amazon Bedrock Marketplace	295
End-to-end workflow	301
Discover a model	310
Subscribe to a model	310
Deploy a model	311
Bring your own endpoint	314
Call the endpoint	314
Manage your endpoints	315
Model compatibility	316
Submit prompts and generate responses with model inference	324
Influence response generation with inference parameters	326
Randomness and diversity	327
Length	329
Supported Regions and models	329
Prerequisites	330
Generate responses in the console using playgrounds	332
Optimize model inference for latency	336
Generate responses using the API	337
Submit a single prompt	340

Carry out a conversation with Converse	344
Use a tool to complete a model response	376
Call a tool with the Converse API	377
Converse API tool use examples	382
Use a computer use tool to complete a model response	391
Example code	393
Example response	395
Prompt caching	395
How it works	396
Supported models, regions, and limits	397
Getting started	399
Process multiple prompts with batch inference	405
Supported Regions and models	405
Prerequisites	410
Permissions	410
Set up data	412
[Optional] Set up a VPC	415
Create a job	419
Monitor jobs	422
Stop a job	424
View the results of a job	424
Code examples	426
Set up a model invocation resource using inference profiles	428
Supported Regions and models	430
Supported cross-region inference profiles	430
Supported Regions and models for application inference profiles	434
Prerequisites	436
Create an application inference profile	438
Modify the tags for an application inference profile	440
View information about an inference profile	440
Use an inference profile in model invocation	441
Delete an application inference profile	443
Prompt engineering concepts	444
What is a prompt?	445
Components of a prompt	446
Few-shot prompting vs. zero-shot prompting	447

Prompt template	449
Maintain recall over inference requests	450
What is prompt engineering?	451
Intelligent prompt routing	452
Considerations and limitations	453
Design a prompt	454
Provide simple, clear, and complete instructions	455
Place the question or instruction at the end of the prompt for best results	456
Use separator characters for API calls	456
Use output indicators	457
Best practices for good generalization	461
Optimize prompts for text models on Amazon Bedrock—when the basics aren't good enough	461
Control the model response with inference parameters	465
Prompt templates and examples for Amazon Bedrock text models	465
Text classification	466
Question-answer, without context	469
Question-answer, with context	472
Summarization	477
Text generation	479
Code generation	481
Mathematics	484
Reasoning/logical thinking	485
Entity extraction	486
Chain-of-thought reasoning	488
Construct and store reusable prompts with Prompt management	490
Key definitions	491
Supported Regions and models	491
Prerequisites	492
Create a prompt	495
View information about prompts	500
Modify a prompt	501
Test a prompt	502
Optimize a prompt	504
Supported Regions and models	505
Submit a prompt for optimization	505

Deploy to your application using versions	508
Create a version	509
View information about versions	510
Compare versions	511
Delete a version	511
Delete a prompt	512
Run code samples	513
Stop harmful content in models using Amazon Bedrock Guardrails	519
.....	521
How charges are calculated for using Amazon Bedrock Guardrails	522
Supported regions and models for Amazon Bedrock Guardrails	522
Components of a guardrail	524
Content filters	525
Filter classification and blocking levels	527
Filter strength	527
Prompt attacks	528
Denied topics	530
Sensitive information filters	531
Word filters	537
Contextual grounding check	538
Block images with image content filter	545
Prerequisites for using guardrails	552
Create a guardrail	553
Permissions for Amazon Bedrock Guardrails	562
Permissions to create and manage guardrails for the policy role	563
Permissions you need to invoke guardrails to filter content	563
(Optional) Create a customer managed key for your guardrail for additional security	564
Test a guardrail	566
View information about your guardrails	575
Modify a guardrail	579
Delete a guardrail	581
Deploy your guardrail	582
Create a version of a guardrail	583
View information about guardrail versions	584
Delete a guardrail version	587
Use guardrails for your use case	588

Use the inference operations	590
Evaluate the performance of Amazon Bedrock resources	615
Supported Regions and models	616
Automatic model evaluation jobs	619
Prerequisites	619
Model evaluation task types	622
Prompt datasets	629
Create job	634
List job	637
Stop job	639
Delete job	640
Human-based model evaluation jobs	642
Creating your first model evaluation that uses human workers	643
Custom prompt datasets (human)	646
Human-based model evaluation jobs	647
List model evaluation jobs	653
Stop job	654
Delete job	656
Manage a work team for human evaluations	658
LLM as a judge model evaluation jobs	660
Creating job	660
Prompt datasets	663
Evaluator prompts	664
Create job	742
List job	747
Stop job	748
Knowledge base evaluation jobs	749
Prerequisites	750
Prompt dataset for knowledge base evaluation	753
Evaluator prompts	755
Create job	827
List job	832
Stop a knowledge base evaluation job	833
Knowledge base evaluation of retrieval or generation	834
Reports and metrics for knowledge base evaluation	838
Delete a knowledge base evaluation job	842

CORS requirements	842
Reports and metrics for model evaluation	843
Review metrics for an automated model evaluation job	844
Review a human model evaluation job	847
Understand Amazon S3 output from a model evaluation job	853
Data management and encryption in Amazon Bedrock evaluation job	860
Key policy requirements	861
IAM policy requirements	864
Data encryption for knowledge base evaluation jobs	866
Management events	874
Retrieve data and generate responses with Amazon Bedrock Knowledge Bases	876
How knowledge bases work	877
Turning data into a knowledge base	880
Retrieving information from data sources	884
Customizing your knowledge base	885
Supported models and regions	894
Supported models for vector embeddings	897
Supported models and Regions for parsing	897
Supported models and Regions for reranking results during query	898
Chat with your document with zero setup	898
Build a knowledge base by connecting to a data source	899
Prerequisites	900
Create a knowledge base	912
Sync a data source	966
Ingest changes directly into a knowledge base	968
View information about a data source	985
Modify a data source	987
Delete a data source	990
Build a knowledge base by connecting to a structured data store	991
Prerequisites	992
Create a knowledge base	1005
Sync a structured data store	1016
Build a knowledge base with an Amazon Kendra GenAI index	1017
Create a knowledge base	1017
Build a knowledge base with graphs	1020
Test your knowledge base with queries and responses	1021

Query a knowledge base and retrieve data	1022
Query a knowledge base and generate responses	1027
Generate a query for structured data	1034
Configure and customize queries and responses	1035
Deploy your knowledge base for your application	1061
View information about a knowledge base	1063
Modify a knowledge base	1064
Delete a knowledge base	1065
Improve the relevance of query responses with a reranker model	1068
Supported Regions and models	1069
Permissions	1070
Use a reranker model	1075
Automate tasks in your application using AI agents	1079
How Amazon Bedrock Agents work	1080
Build-time configuration	1080
Runtime process	1082
Supported regions	1084
Build and modify agents for your application	1085
Configure your agent using conversational builder	1087
Configure an inline agent at runtime	1090
Create and configure agent manually	1098
View information about an agent	1105
Modify an agent	1106
Delete an agent	1108
Use action groups to define actions for your agent	1109
Define actions in the action group	1110
Handle fulfillment of the action	1122
Add an action group to your agent	1138
View information about an action group	1145
Modify an action group	1147
Delete an action group	1148
Use multi-agent collaboration for complex tasks	1149
Supported regions and models for multi-agent collaboration	1150
Create multi-agent collaboration	1151
Disassociate collaborator agent	1158
Disable a multi-agent collaboration	1159

Configure agent to request information from user	1161
Enable user input	1161
Disable user input	1163
Augment response generation for your agent with knowledge base	1164
View information about an agent-knowledge base association	1166
Modify an agent-knowledge base association	1167
Disassociate a knowledge base from an agent	1168
Retain conversational context using memory	1169
Enable agent memory	1171
View memory sessions	1172
Delete session summaries	1174
Disable agent memory	1176
Enable memory summarization log delivery	1176
Generate, run, and test code with code interpretation	1177
Enable code interpretation	1179
Test code interpretation	1180
Disable code interpretation	1184
Implement safeguards for your application	1186
Provision additional throughput	1186
Test and troubleshoot agent behavior	1186
Track agent's step-by-step reasoning process using trace	1194
Customize agent for your use case	1208
Customize agent orchestration	1209
Control agent session context	1314
Optimize performance for agents using a single knowledge base	1320
Working with models not yet optimized	1322
Deploy and integrate agent into your application	1322
View information about versions of agents in Amazon Bedrock	1326
Delete a version of an agent in Amazon Bedrock	1327
View information about aliases of agents in Amazon Bedrock	1328
Edit an alias of an agent in Amazon Bedrock	1329
Delete an alias of an agent in Amazon Bedrock	1330
Build a generative AI workflow with Amazon Bedrock Flows	1332
How it works	1334
Key definitions	1334
Define inputs with expressions	1336

Node types in flow	1338
Example flows	1360
Supported regions and models	1366
Prerequisites	1367
Create a flow	1369
View information about flows	1374
Modify a flow	1375
Include guardrails in your flow	1376
Test a flow	1377
Track each step in your flow by viewing its trace	1379
Deploy to your application using versions and aliases	1381
Create a version	1382
View information about versions	1383
Delete a version	1384
Create an alias	1385
View information about aliases	1386
Modify an alias	1387
Delete an alias	1388
Invoke a Lambda function in a different AWS account	1389
Converse with a flow	1389
How to process a multi-turn conversation in a flow	1390
Creating and running an example flow	1394
Run code samples	1399
Delete a flow	1406
Customize a model for your use case	1408
Supported regions and models	1409
Guidelines for model customization	1410
Amazon Nova models	1410
Amazon Titan Text Premier	1411
Prerequisites for model customization	1412
Prepare the datasets	1413
[Optional] Protect your model customization jobs using a VPC	1423
Submit a model customization job	1427
Monitor your model customization job	1430
Analyze model customization job results	1431
Stop a model customization job	1434