



Amazon Bedrock Builders Series

Building GenAI apps using Amazon Bedrock

Vasileios Vonikakis

Sr. GTM GenAI Specialist SA
AWS Singapore
vonikakv@amazon.com

Isaac Ibrahim

GTM GenAI Specialist SA
AWS Singapore
isaacibr@amazon.com

Chomchana Trevai

Sr. GTM Specialist GenAI ASEAN
AWS Singapore
cttrevai@amazon.com

Agenda

Time	Duration	Topic
08:30-09:00	30min	Registration
09:00-09:30	30min	Breakfast
09:30-10:15	45min	Intro / GenAI App blueprint
10:15-10:20	5min	Break
10:20-11:00	40min	Deep dive in Knowledge bases, Agents and Flows
11:00-11:30	30min	Set up free AWS accounts
11:30-12:30	60min	Lunch
12:30-13:45	75min	Hands on
13:45-14:00	15min	Break
14:00-15:00	60min	Hands on
15:00-15:30	30min	Code samples/solutions/demos/assets
15:30-15:45	15min	Kahoot quiz & prizes
15:45-16:00	15min	End of event, Q&A, networking, survey

Building GenAI apps: a blueprint



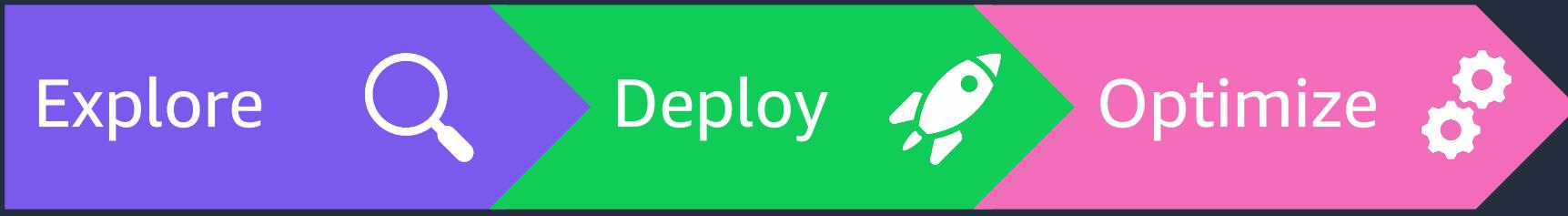
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generative AI applications: It's more than the model

FMs are only a single component of a GenAI application!



Progression of GenAI applications

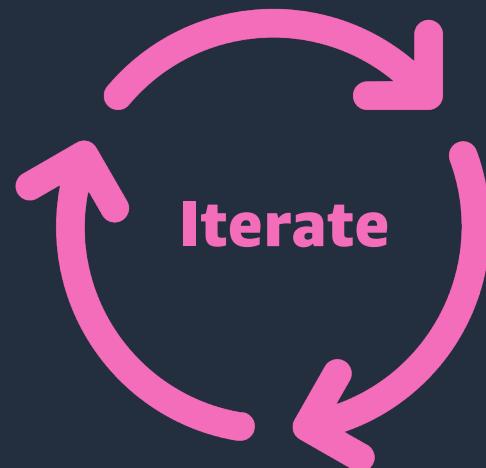


PoCs

- Off-the-shelf, latest & greatest models

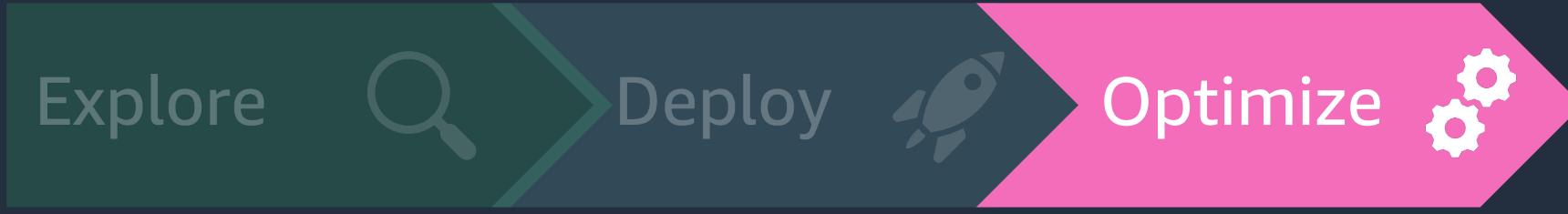
Production

- Complex scenarios
- Orchestration



- Purpose & scope-specific models
- Improved domain performance
- Improved task performance
- Optimized prompts
- Reduced latency
- Optimized cost

Progression of GenAI applications

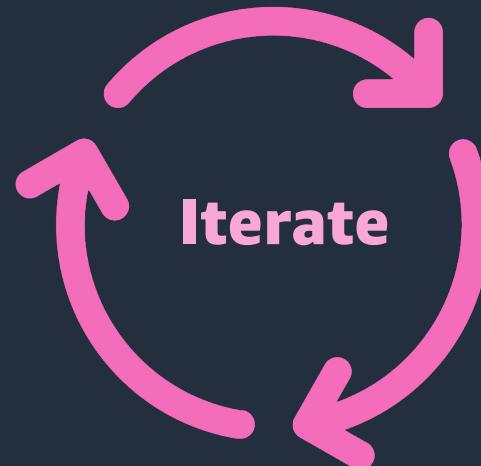


PoCs

- Off-the-shelf, latest & greatest models

Production

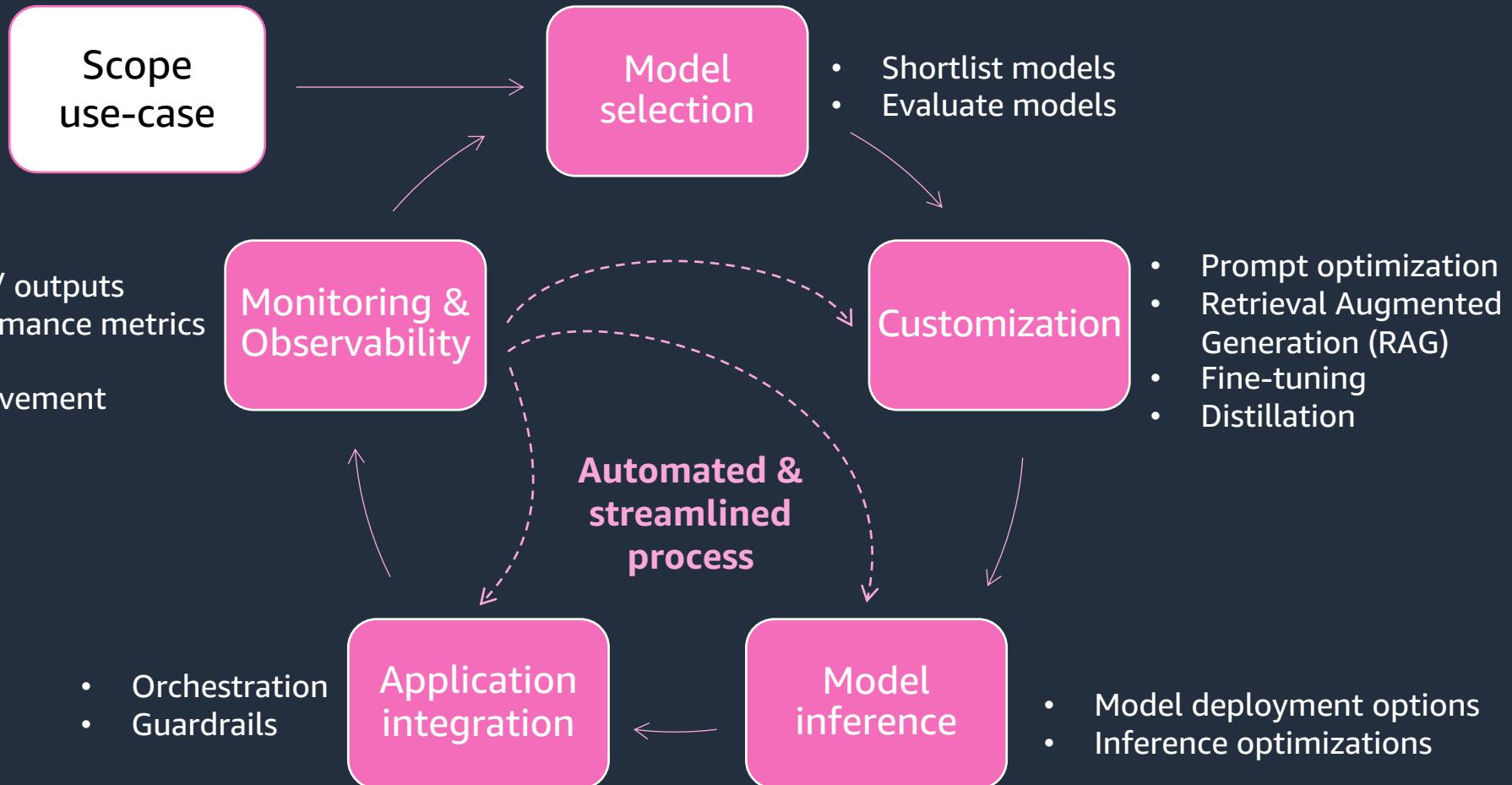
- Complex scenarios
- Orchestration



- Purpose & scope-specific models
- Improved domain performance
- Improved task performance
- Optimized prompts
- Reduced latency
- Optimized cost

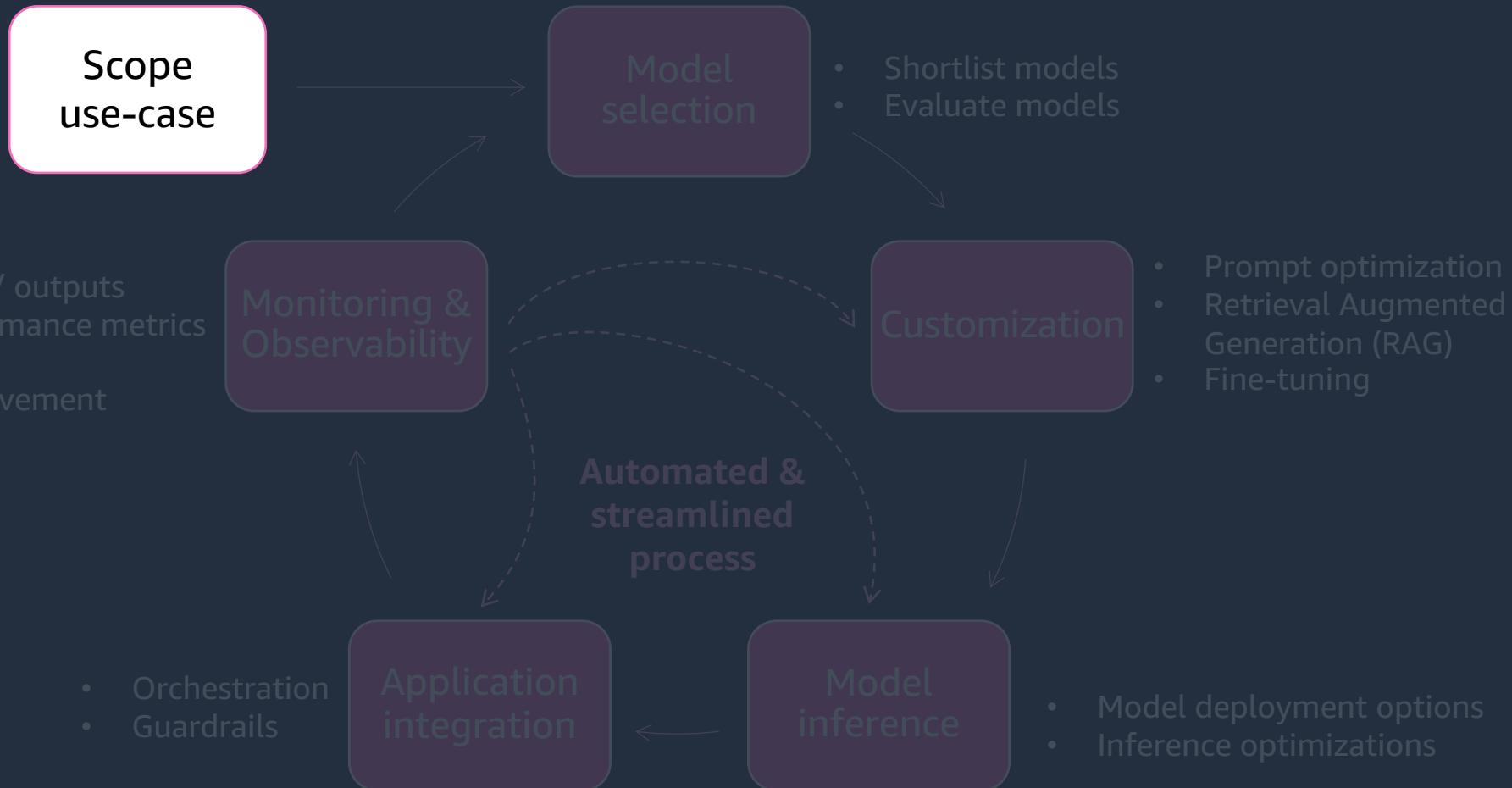
Streamlining the optimization phase

Understand use-case to guide design decisions

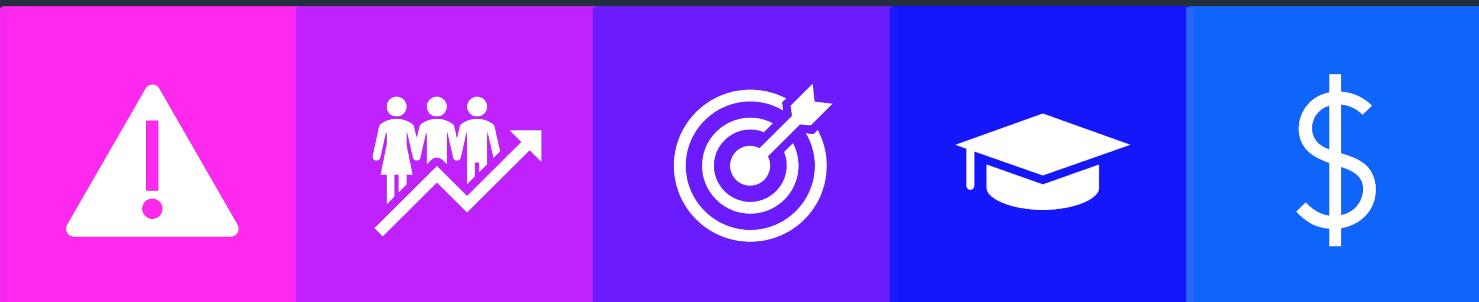


Streamlining the optimization phase

Understand use-case to guide design decisions



Understand the use case (design decisions)



Criticality

Tolerance to mistakes

- Select highest Accuracy models
- Fine-tune model
- Use post-processing sanity checks
- Consider human-in-the-loop

Scale

Limited or broad audience

- Control for unacceptable behavior (e.g, bias, hate, racist speech etc.)
- Defense for jailbreaking
- Concurrency and latency
- Monitor performance

Task type

Open-ended or fixed

- Select larger models that can handle many tasks and include more “knowledge”
- Select smaller specialized models

Language

Supported languages

- Select models that support specific languages

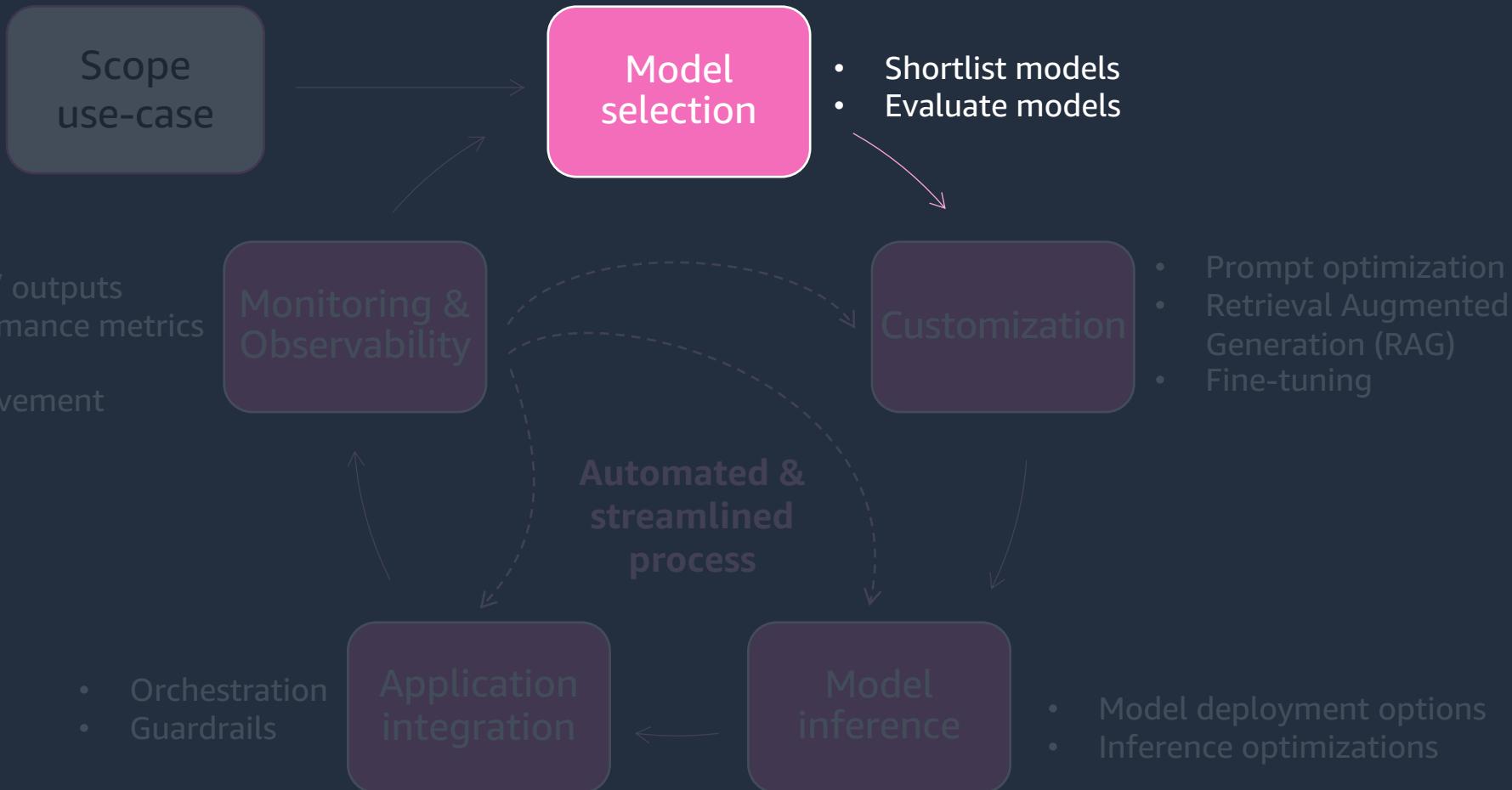
ROI

Added business value

- Cost of the model
- Deployment type (serverless vs endpoints)
- Evaluate the business impact of the model

Streamlining the optimization phase

Understand use-case to guide design decisions



Amazon Bedrock

BROADEST SELECTION OF FULLY MANAGED MODELS FROM LEADING AI COMPANIES

AI21labs



Effective reasoning & rapid analysis for long context windows

JAMBA

Frontier intelligence & industry leading price performance, high-quality image & video generation

AMAZON NOVA

ANTHROPIC

Advanced reasoning & coding capabilities, including computer use skills

CLAUDE

cohere

Multimodal search & advanced retrieval powering multilingual knowledge agents

COMMAND
EMBED
RERANK

deepseek

Advanced reasoning, comprehension, scientific & mathematical analysis, & code generation

DEEPSEEK-R1

Luma

High-quality video generation from text & images

LUMA RAY 2

Meta

Advanced image & language reasoning

LLAMA

mistral.ai

Knowledge summarization, expert agents, & code completion

MISTRAL
MIXTRAL

poolside

Software engineering AI for large enterprises

MALIBU
POINT

stability.ai

High-quality AI image generation, easily deployable at scale

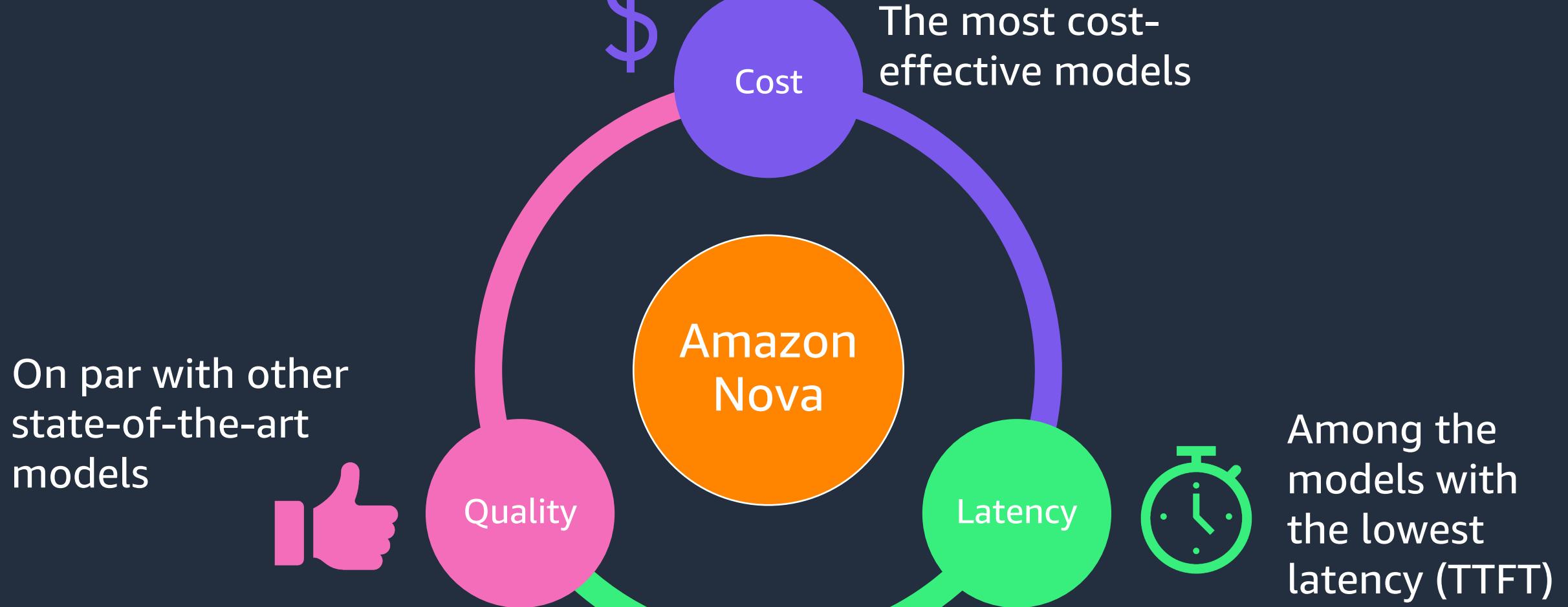
STABLE DIFFUSION
STABLE IMAGE

Coming soon

- Collection of selected state-of-the-art open-access & proprietary models
- From 10 model providers
- DeepSeek R1 and Claude Sonnet 3.7 were recently added
- Options for fine-tuning, continued pretraining and model distillation



Amazon Nova: Powerful balance of quality, cost & latency



Amazon Bedrock Marketplace

MORE MODEL PROVIDERS, MODEL MARKETPLACE, ACCESS TO OPEN-SOURCE & CUSTOM MODELS

- Get access to **hundreds** of proprietary and open-access models, including **DeepSeek R1**
- Many types of modalities: Audio, Embeddings, Image, Text, Vision etc.
- Models can be accessed through Bedrock's APIs and tools
- Discover models in Model Catalog → subscribe to them → Deploy them to private endpoints → Use them in Bedrock
- (Discover models in SageMaker Jumpstart) → Deploy model in SageMaker endpoint → Register endpoint to Bedrock → Use it in Bedrock

Amazon Bedrock > Model catalog

Model catalog (178)

Discover Bedrock serverless or Marketplace models that best fit your use case. To get started using a serverless model, request access. For Marketplace models, subscribe and deploy.

Filters

▼ Model collection

Bedrock Marketplace (7) ⓘ
 Serverless (0)

▼ Providers

Amazon (11)
 Anthropic (9)
 Arcee AI (5)
 Camb.ai (1)
 Cohere (7)
 Deepseek (7)
 EvolutionaryScale, PBC (1)
 Gretel (1)
 HuggingFace (83)
 IBM Data and AI (6)

Show 10 more

▼ Modality

Audio (0)
 Embedding (0)
 Image (0)
 Text (7)
 Text Vision (0)
 Video (0)

▶ Spotlight

Filter for a model

Providers = deepseek X Clear filters

Model name	Version	Status	Provider name	Popularity	Deployment
DeepSeek-R1	2.0.1	N/A	Deepseek	166	0
DeepSeek-R1-Distill-Qwen-7B	1.0.0	N/A	Deepseek	167	0
DeepSeek-R1-Distill-Qwen-32B	1.0.0	N/A	Deepseek	168	0
DeepSeek-R1-Distill-Qwen-14B	1.0.0	N/A	Deepseek	169	0
DeepSeek-R1-Distill-Qwen-1.5B	1.0.0	N/A	Deepseek	170	0
DeepSeek-R1-Distill-Llama-8B	1.0.0	N/A	Deepseek	171	0
DeepSeek-R1-Distill-Llama-70B	1.0.0	N/A	Deepseek	172	0



Quantitative LLM evaluation in Amazon Bedrock

Automatic evaluation



Accuracy



Robustness



Toxicity

Human evaluation



Creativity



Style



Tone



Accuracy



Consistency



Brand voice

LLM as a judge



Correctness



Completeness



Helpfulness



Relevance



Coherence



Readability

Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

Rating methods

Thumbs up/down | 5-point Likert scales
Binary choice buttons | Ordinal ranking

LLM reasoning

Multistep reasoning | Few shot learning
Correlation with expert human evaluators

- Text generation
- Q&A
- Summarization
- Text classification



Qualitative LLM evaluation in Amazon Bedrock

Chat playground [Info](#)

Load examples Compare mode [:](#)

Titan Text G1 - Premier v1 On-demand throughput	Llama 3 70B Instruct v1 On-demand throughput	Mistral Large (2402) 2402 On-demand throughput
<p>Write an out of office response for my email, saying that anyone interested should contact John Doe during my absense. I will be away between 12 and 29 August.</p> <p>Hello, Thank you for your email. I am currently away on vacation and will not be checking my emails between 12 and 29 August. If you need assistance during this time, please contact John Doe at [EMAIL]. I will respond to your email as soon as I return on 29 August. Best regards, [Your Name]</p>	<p>Write an out of office response for my email, saying that anyone interested should contact John Doe during my absense. I will be away between 12 and 29 August.</p> <p>Here is an out-of-office email response: Hello, I am currently out of the office from 12 August to 29 August and will have limited access to emails during this time. If you have an urgent matter that requires attention, please reach out to John Doe at [John Doe's email address] who will be able to assist you.</p>	<p>Write an out of office response for my email, saying that anyone interested should contact John Doe during my absense. I will be away between 12 and 29 August.</p> <p>Subject: Out of Office: Away Until August 29 Hello, Thank you for your email. I am currently out of the office and will not be checking email until my return on August 30. If your message requires immediate attention, please contact John Doe at johndoe@email.com for assistance during my absence. I look forward to responding to your email upon my return. Thank you for your patience and understanding. Best regards, [Your Name]</p>

Write a prompt... (Shift + ENTER to start a new line, and ENTER to generate a response)

[Choose files](#)

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx. Maximum 5 files, 4.5 MB each.

Model metrics

To evaluate models for task specific metrics with custom dataset visit [Model evaluation](#)

Metrics	Titan Text G1 - Premier	Llama 3 70B Instruct	Mistral Large (24.02)
Overall summary	Define metric criteria	Define metric criteria	Define metric criteria
Latency	2827 ms	2298 ms	2729 ms
Input token count	38	50	42
Output token count	82	95	107
Cost	=	=	=

Comparison mode
(Up to 3 models)

Request

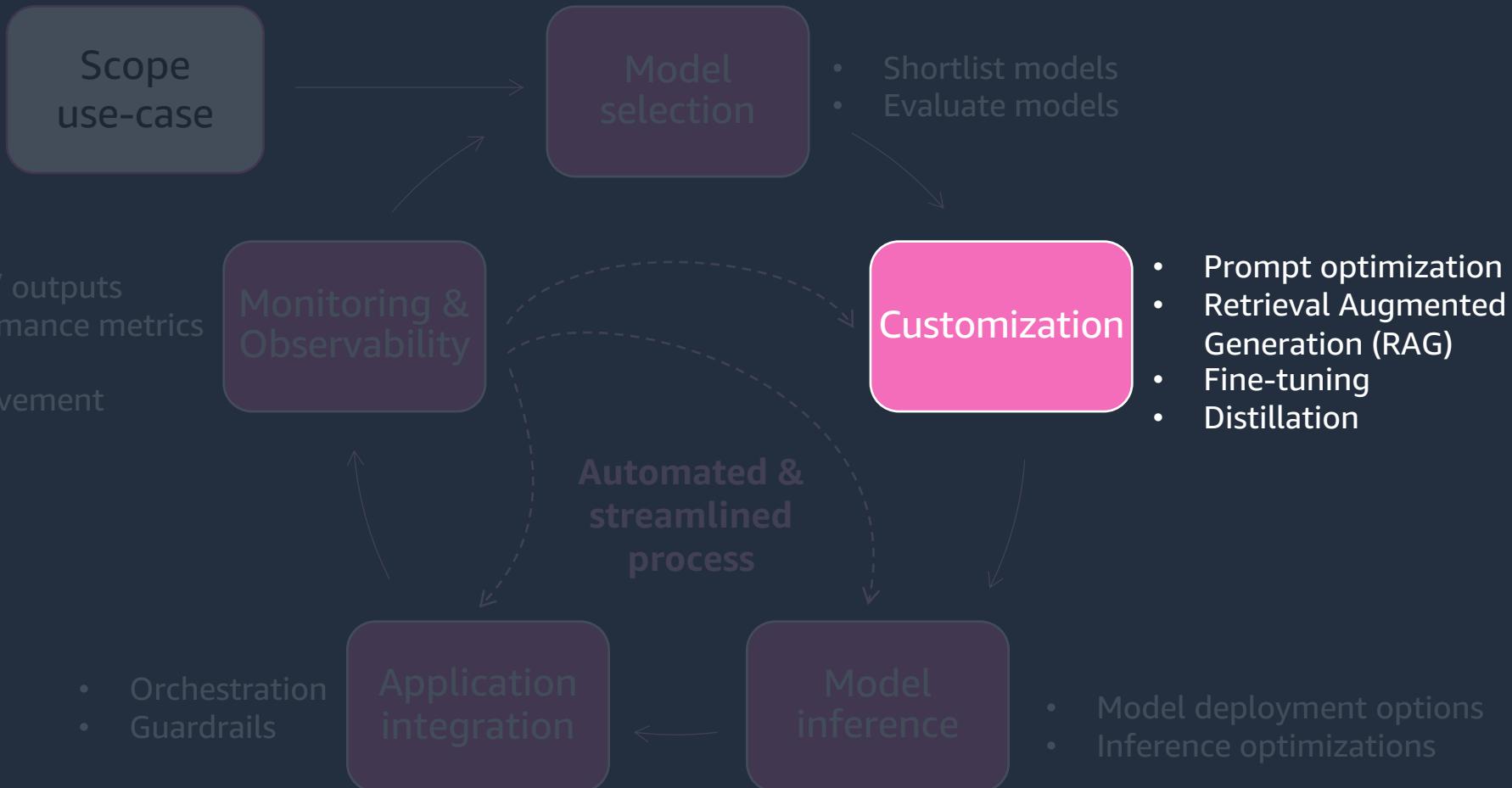
Model response

Test prompt editor

Metrics

Streamlining the optimization phase

Understand use-case to guide design decisions



Common approaches for customizing FMs

**Prompt
engineering**

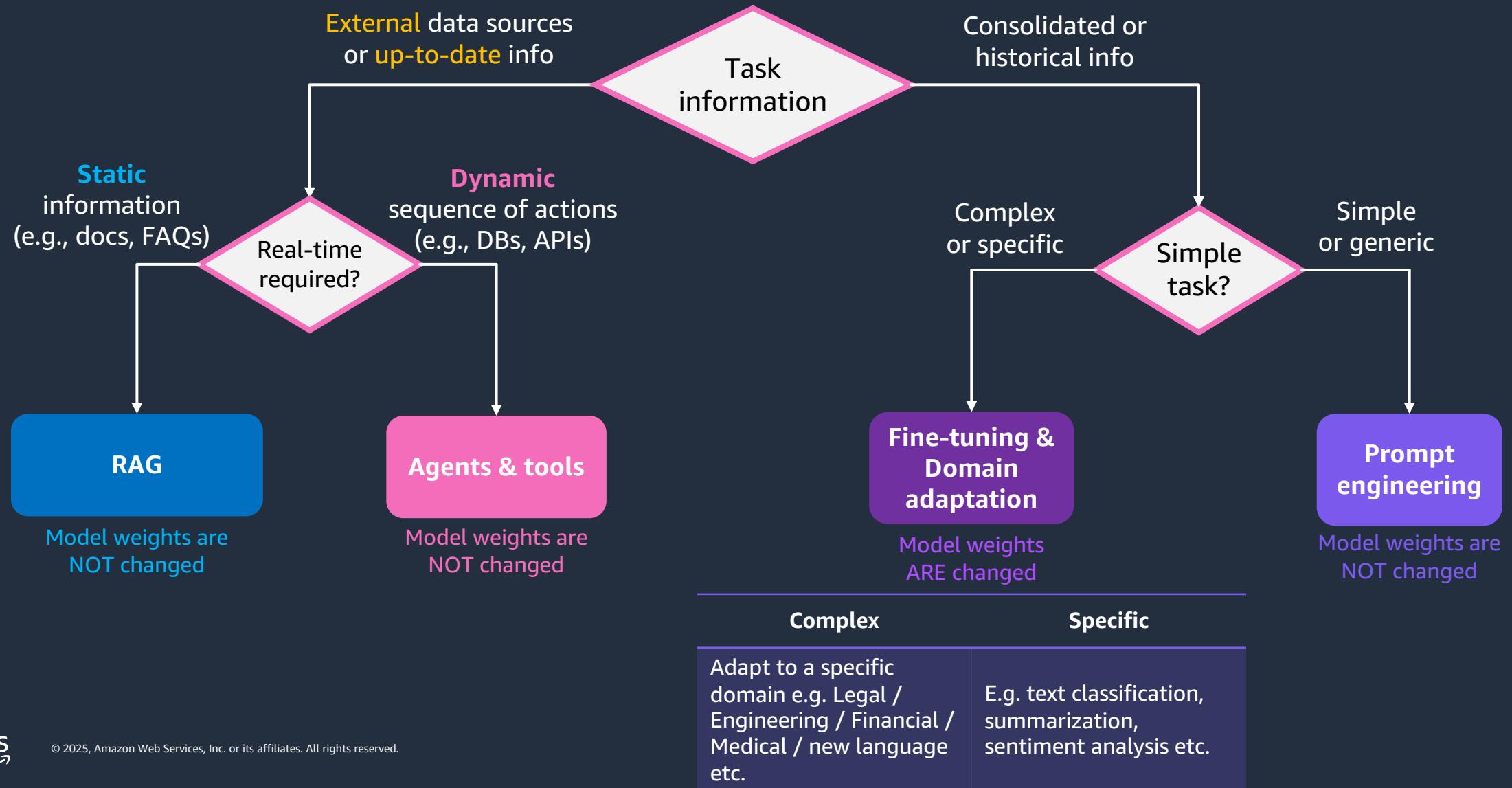
**Retrieval
Augmented
Generation
(RAG)**

**Fine-tuning
&
Continued
pretraining
(domain
adaptation)**

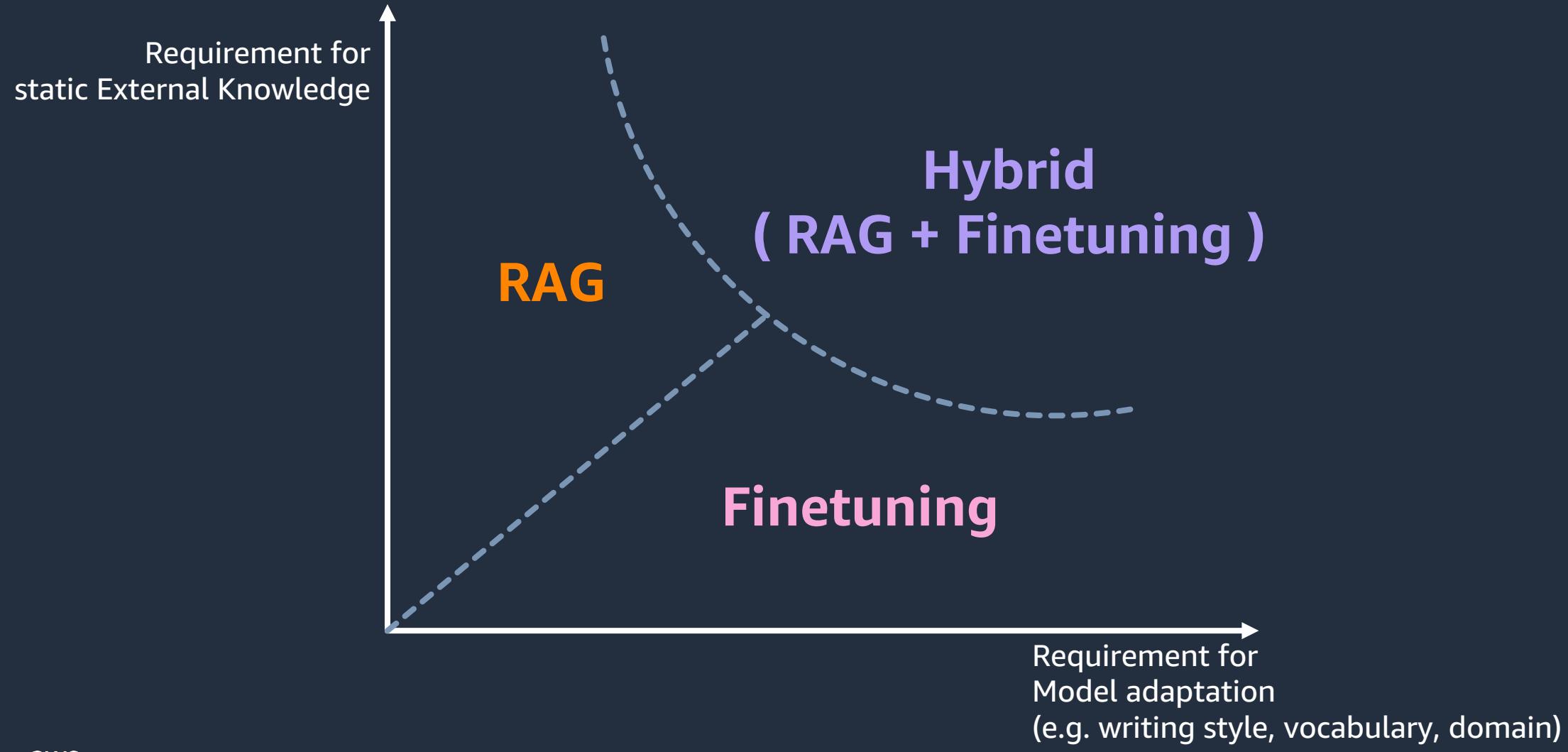
COMPLEXITY,
QUALITY,
COST, TIME

Objective: Improve the quality of model's output

Selecting a customization option



RAG vs finetuning vs both



Comparison of RAG options in AWS

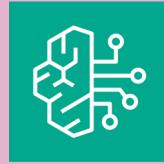
Managed services



Amazon Q



Canvas Amazon Kendra



Amazon Bedrock

**Amazon Q
Business**

- Less development effort
- Fewer customizations

Manual development



Amazon
Bedrock Amazon
SageMaker



Amazon
Kendra



Amazon
Bedrock Amazon
SageMaker

Other
Vector DBs

**LLMs
+ Amazon Kendra**

- More development effort
- More customizations

Trade-off

Comparison of RAG options in AWS

Managed services



Amazon Q



Canvas Amazon Kendra



Amazon Bedrock

**Amazon Q
Business**

**SageMaker Canvas
+ Amazon Kendra**

**Amazon Bedrock
Knowledge Base**

- Less development effort
- Fewer customizations

Balanced
Trade-off

Manual development



Amazon
Bedrock Amazon
SageMaker



Amazon
Kendra



Amazon
Bedrock Amazon
SageMaker

Other
Vector DBs

**LLMs
+ Amazon Kendra**

**LLMs
+ Vector DBs**

- More development effort
- More customizations

Amazon Bedrock Knowledge Bases

Fully
managed
RAG



Data sources

- S3
- Web crawler
- Atlassian Confluence
- Microsoft SharePoint
- Salesforce
- Custom
- Incremental updates
- Metadata

Parsing

- Smart parsing (of PDFs) for Tables & Images
- Metadata filtering
- Custom parsing
- Query reformulation (sub-queries)
- **Bedrock Data Automation**
- Text-to-SQL
- GraphRAG

Chunking

- Fixed
- Standard
- Custom chunking (Lambda function, LangChain, LlamaIndex)
- Semantic
- Hierarchical

Models

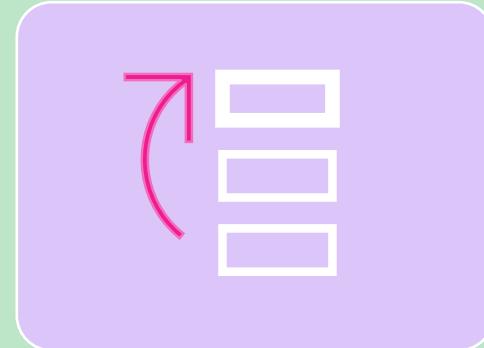
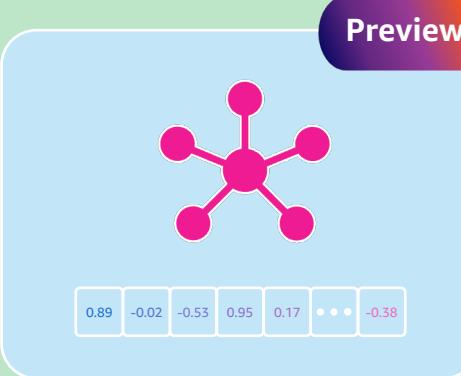
- Titan Embeddings G1
- Titan Text Embed. V2
- Cohere Embed
- **Amazon Nova**
- Titan Text Premier
- Claude 3 Haiku
- Claude 3.5 Haiku
- Claude 3 Sonnet
- Claude 3.5 Sonnet
- Llama 3.1
- Cross-region Inference
- **Reranking**

Vector store

- OpenSearch serverless
- Aurora PostgreSQL
- Pinecone
- Redis
- MongoDB
- **Neptune Analytics**

Amazon Bedrock Knowledge Bases

Notable re:Invent2024 launches



Structured data retrieval

- Built-in text-to-SQL
- Answer analytical questions

GraphRAG

- Discover relationships between entities
- Improve context relevance

Rerank API

- Bring relevant chunks on top
- Improve context relevance

RAG evaluation

- Evaluate the retriever and generator
- Optimize end-to-end RAG

Model finetuning options

Objective: Improve the quality of model's outputs (for a specific task or domain)



Amazon Bedrock

- (fully managed)
- Instruction-based fine-tuning
- Continued pretraining (domain adaptation)

- Less development effort
- Fewer customizations

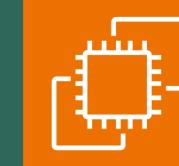


Amazon SageMaker

Amazon SageMaker JumpStart

- (Low-Code/No-code: UI-based & automated scripts)
- Instruction-based fine-tuning
- Domain adaptation
- Quantization, Spec. Decoding
- Optimization for Inferentia / Trainium

Trade-off



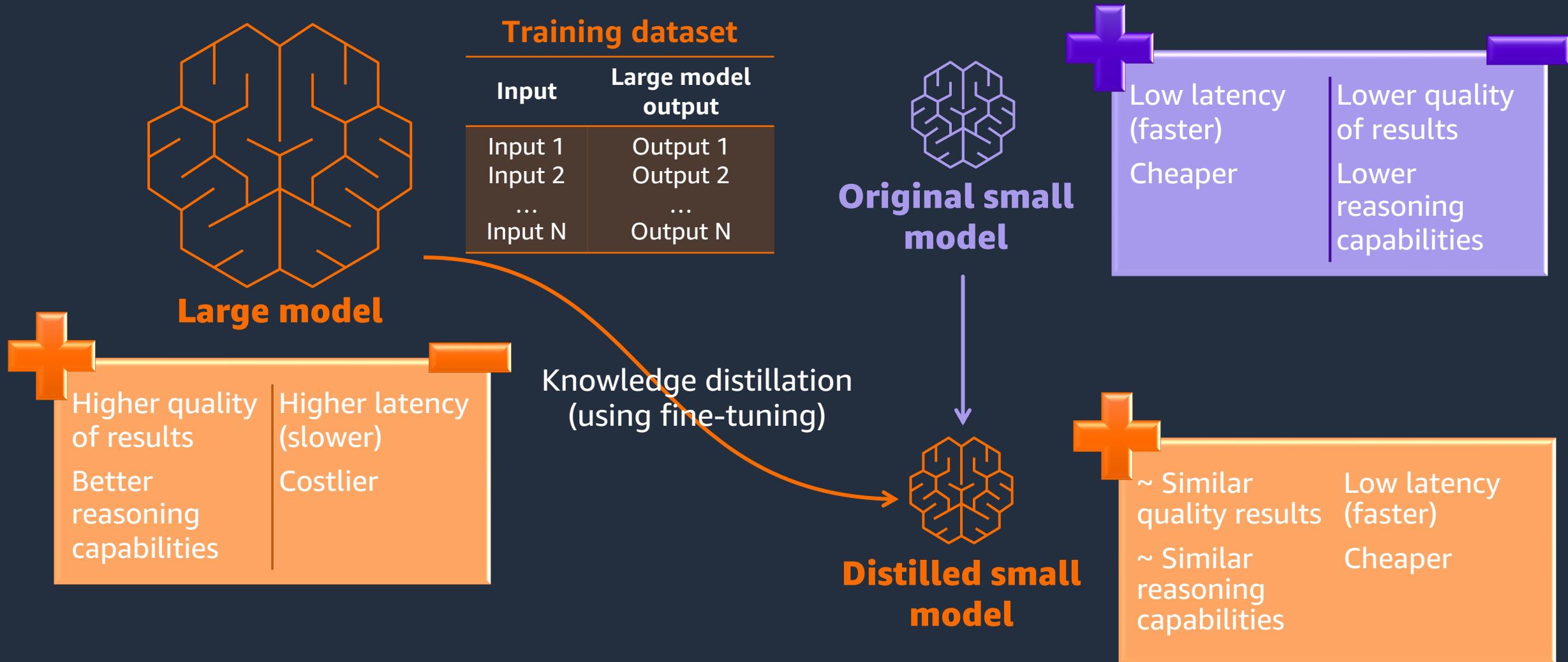
Amazon EC2

- Fully custom approach

- More development effort
- More customizations

Model Distillation: main idea

Objective: Achieve similar quality of results
(with a large model) using a smaller model



Model Distillation

- N. Virginia
- Oregon

Functionality

- Model Distillation: Transfers knowledge from a larger more intelligent model (**teacher**) to a smaller, faster, cost-efficient model (**student**), through fine-tuning
- Amazon Bedrock **automates the whole process**: Generates synthetic data from teacher model, trains and evaluates the student model, hosts the final distilled model for inference
- You need **at least 100 prompt-response pairs** (otherwise Data Synthesis will be used)
- **Only you** have access to the final distilled model
- Your data **is not used** to train any other model

Benefits

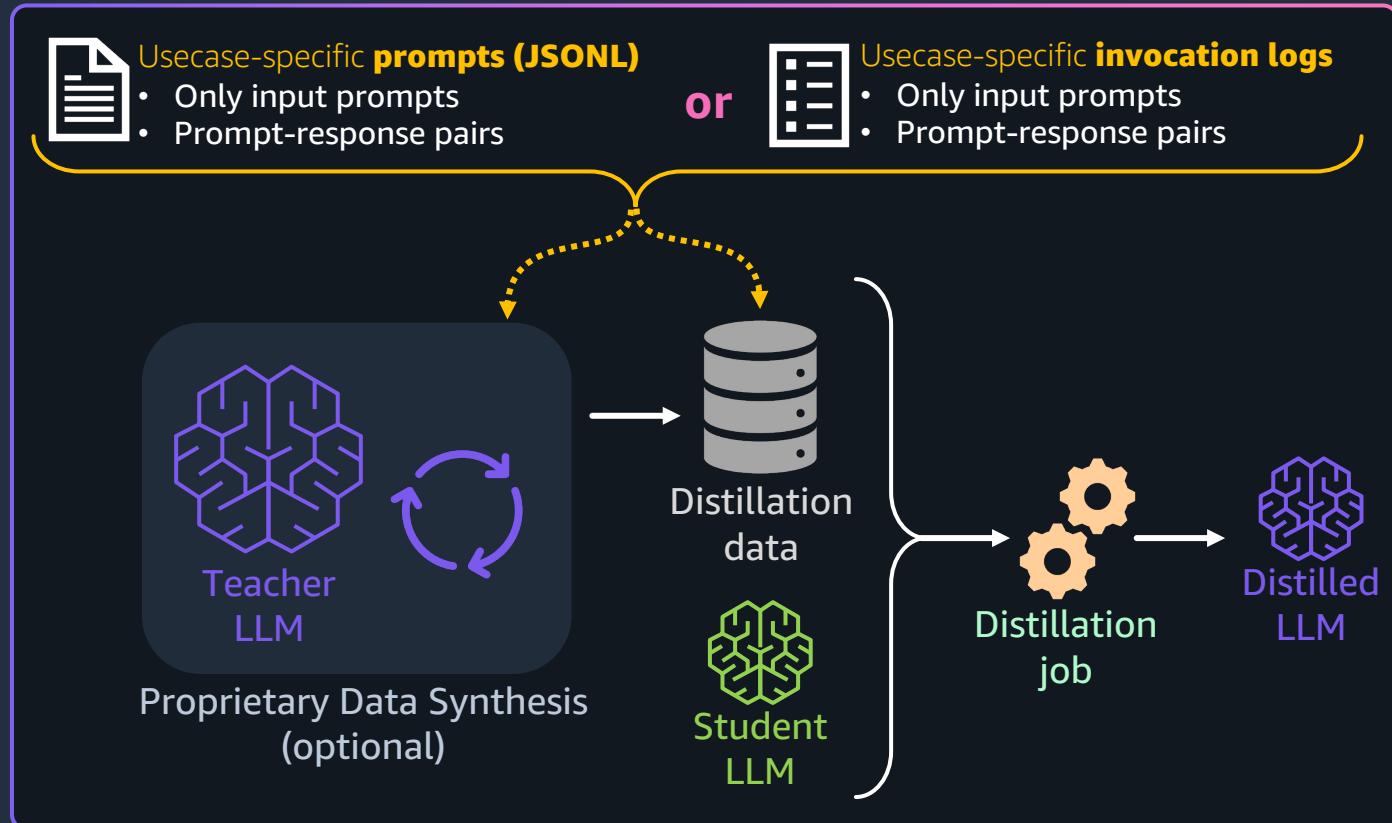
- Much easier to distil large LLMs to smaller ones
- Achieve similar results with a cheaper & faster model
- Up to **5x faster** and **75% less expensive**

Pricing

- Data Synthesis is charged at **on-demand pricing** (teacher model)
- Fine-tuning (student model) is charged as **model customization**
- Inferences for distilled models requires **Provisioned Throughput**



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Anthropic
Teacher <ul style="list-style-type: none"> Claude 3.5 Opus Claude 3.5 Sonnet Student <ul style="list-style-type: none"> Claude 3 Sonnet Claude 3 Haiku

Meta
Teacher <ul style="list-style-type: none"> Llama 3.1 405B Ins. Student <ul style="list-style-type: none"> Llama 3.1 70B Ins. Llama 3.1 8B Ins.

Amazon
Teacher <ul style="list-style-type: none"> Nova Pro Student <ul style="list-style-type: none"> Nova Lite Nova Micro

- [AWS announcement](#)
- [Blogpost](#)

Amazon Bedrock Prompt Management

Features

- Prompt Builder
- Prompt Library
- Integration with Bedrock Studio and Flows
- Available in Console and SDK APIs



Benefits



- Rapid prompt creation, iteration, testing and deployment
- Easy cataloguing, management,
- Collaborative prompt development

Prompt Library

The screenshot shows the 'Prompt management' section of the Amazon Bedrock console. It includes three main sections: 'Create a prompt' (with a text input field and a 'Create' button), 'Test the prompt' (with a circular icon showing a brain and a gear), and 'Use the prompt' (with a 3D cube icon). Below these is a table titled 'Prompts (3)' listing three entries: 'MyTestPrompt-20240628-135832' (description: 'This is my test prompt for the translation use case'), 'CustomerService' (description: 'This is my prompt for Customer Service use case.'), and 'MyTestPrompt-001' (description: 'Prompt template for the virtual assistant use case.'). At the bottom right are buttons for 'Delete', 'Edit', and 'Create prompt'.

Prompts

The screenshot shows a list of prompts with columns for Name, Description, Last updated, and actions. The first prompt is 'MyTestPrompt-20240628-135832' (Last updated: June 28, 2024, 13:58 (UTC+02:00)). The second is 'CustomerService' (Last updated: June 27, 2024, 11:43 (UTC+02:00)). The third is 'MyTestPrompt-001' (Last updated: June 26, 2024, 14:03 (UTC+02:00)).

Prompt Builder

The screenshot shows the 'Prompt builder: CustomerServiceSummary' interface. It includes a 'Prompt' section with instructions to create a text-based prompt using variables like {{variable}}. It also includes a 'Test window' panel showing a customer service conversation between a customer and an agent, and a 'Model' section where the AI model 'Claude 3 Haiku v1' is selected. On the right, there are sections for 'Inference parameters' (Temperature, Top P, Top K sliders) and 'Test variables' (language: English, format: JSON, context: 'The customer has an Android phone and has complained about the device 3 times this month.', conversation: 'Customer: My phone is not charging, can you please help me? Agent: Sure, let me retrieve your details. Customer: Thanks. Agent: I see you have multiple issues with the device, I'll open a ticket for ordering a new one for you. You should receive instructions shortly.').



Artifacts to track

Prompts

FM version

FM hyperparameters

Requests

Criteria

Data

Prompt templates

Performance metrics

RAG

Retriever type & version

RAG parameters (chunking strategy)

Embedding model version

Evaluation dataset

RAG metrics

Finetuning & Distillation

Input-output finetuning dataset (instruction-based)

Domain specific documents (domain adaptation)

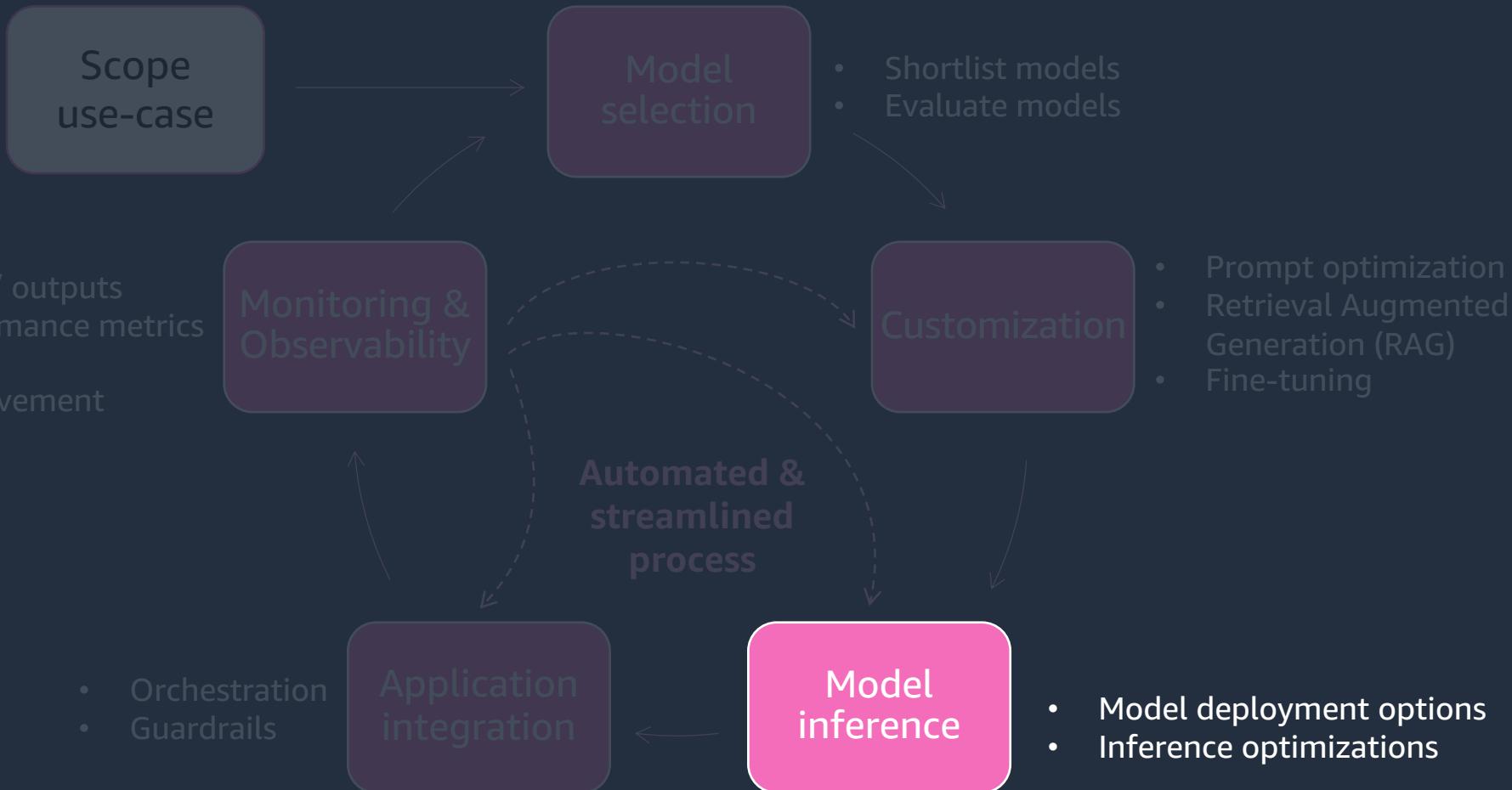
Finetuning hyperparameters

Finetuning scripts

Performance metrics

Streamlining the optimization phase

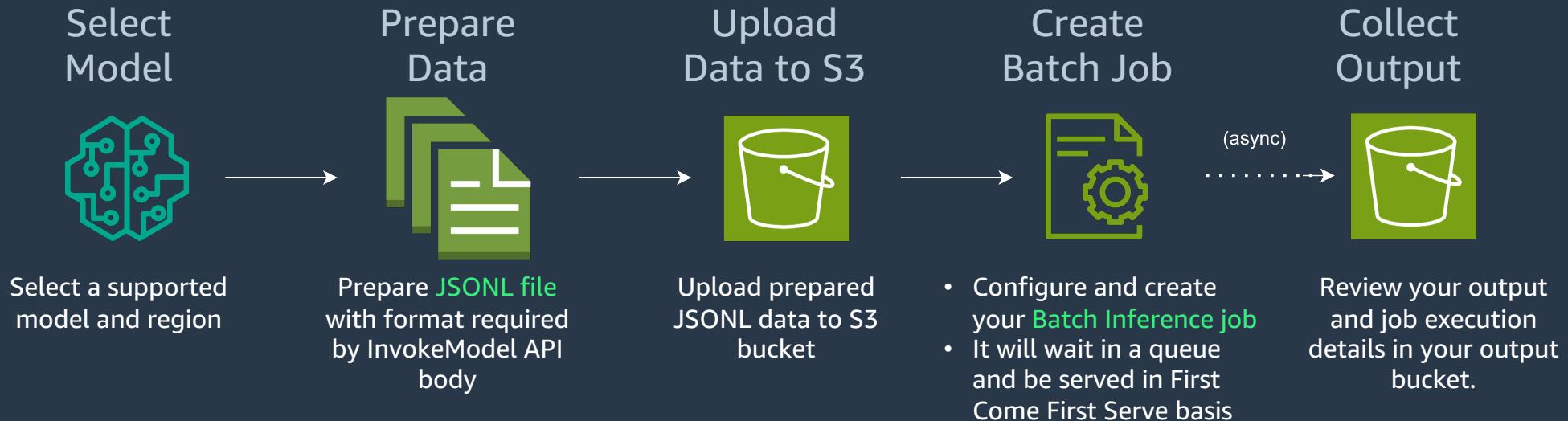
Understand use-case to guide design decisions



Amazon Bedrock inference consumption options

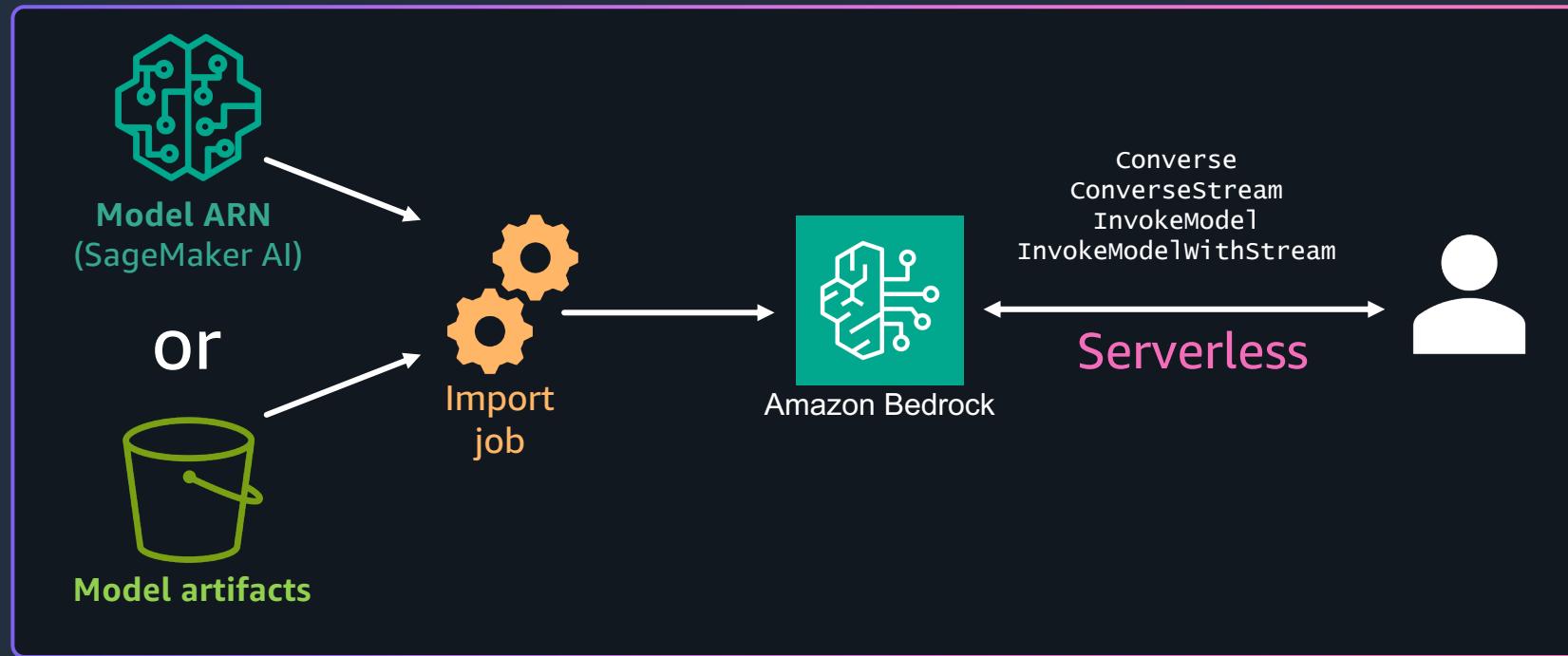
On demand (pay as you go)	Provisioned Throughput	Batch Inference
<ul style="list-style-type: none">• Based on input and output token count• Requests per minute (RPM) and tokens per minute (TPM) limits enforced• For prototyping, POCs, and small workloads with relaxed throughput requirements	<ul style="list-style-type: none">• Reserve throughput (input/output tokens per minute) at a fixed hourly rate• Commitment of 1 month, 6 months or hourly• For production workloads or inference on custom models	<ul style="list-style-type: none">• Offline, job-based inference• Avoids throttling when running large jobs

Amazon Bedrock Batch Inference



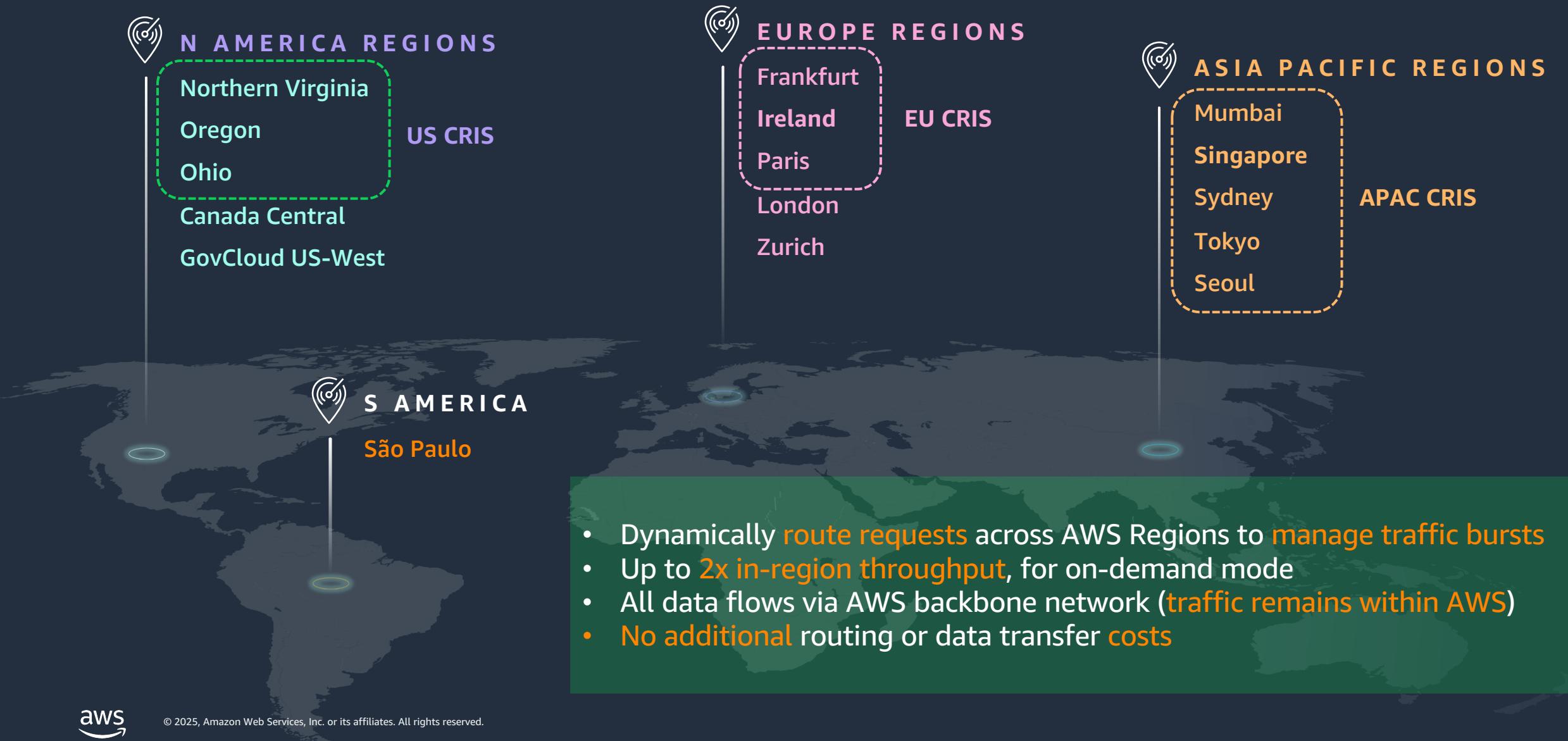
- Suited for **non time-sensitive use cases** (e.g. offline text summarization, entity extraction etc.)
- Run multiple inference requests **asynchronously**, while **avoid throttling**
- **50% lower price** when compared to on-demand inference
- Results are available **within 24h** (with occasional delays)

Amazon Bedrock Custom Model Import



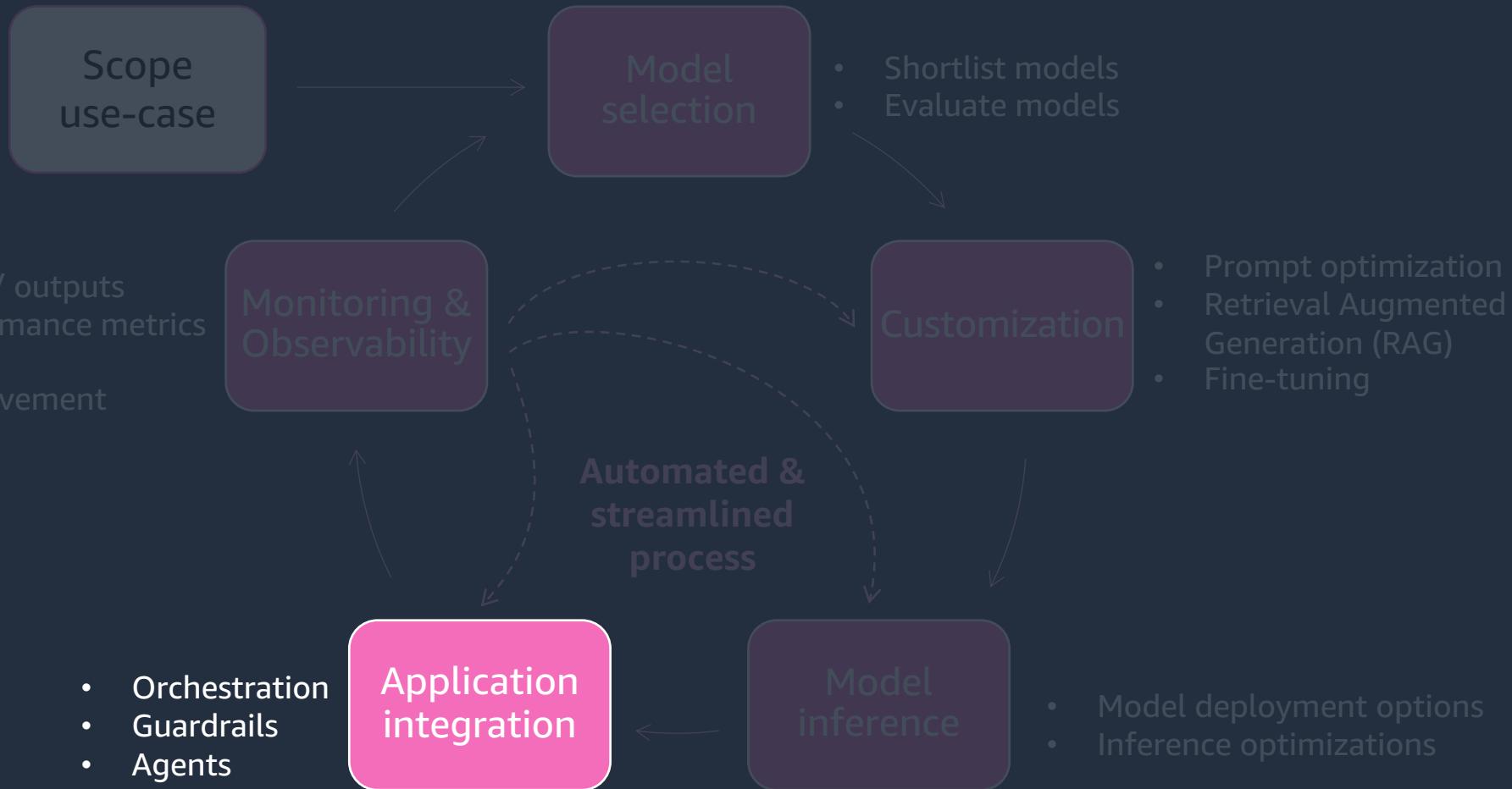
- Import and use your customized models in Amazon Bedrock
- Supported architectures: Flan, Mixtral, Llama 2, 3, 3.1., 3.2, 3.3
- Run custom weights model in a **serverless** way
- Integration with **native Bedrock tools** e.g. Agents, Knowledge Bases, Guardrails, Prompt Flows etc.
- Can import the distilled versions of DeepSeek R1 (Llama-based)

Amazon Bedrock Cross-Region Inference (CRIS)

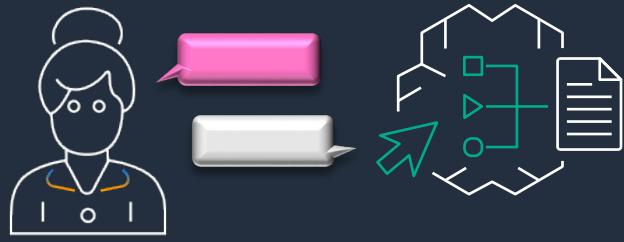


Streamlining the optimization phase

Understand use-case to guide design decisions



Amazon Bedrock Agents



Agent Building Blocks



Session handling



Memory management



Runtime configurations



Action groups



Orchestration strategies



Inline agents



Multi-agent collaboration



Code interpretation

Developer Tools



Deployment



Logs, tracing



Knowledge bases



Guardrails



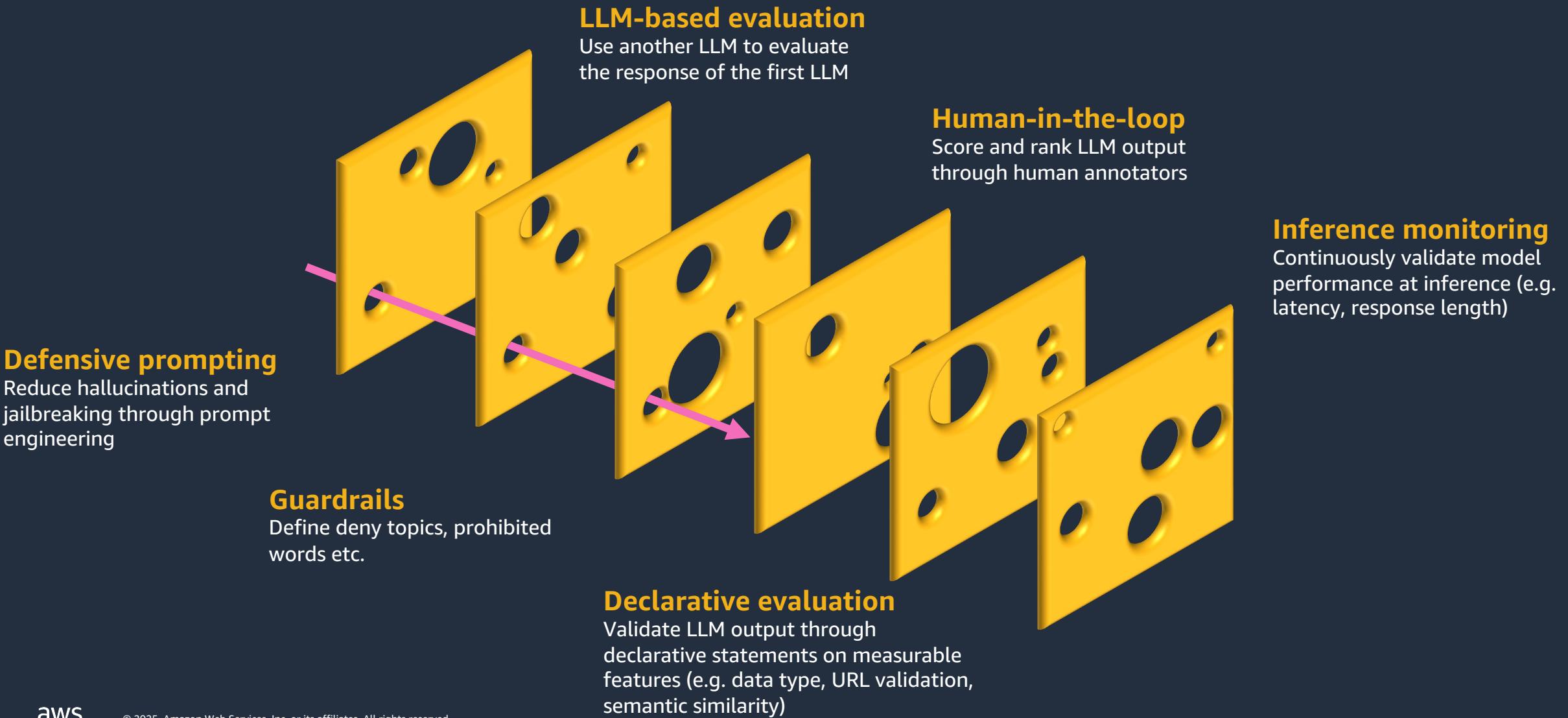
Flows



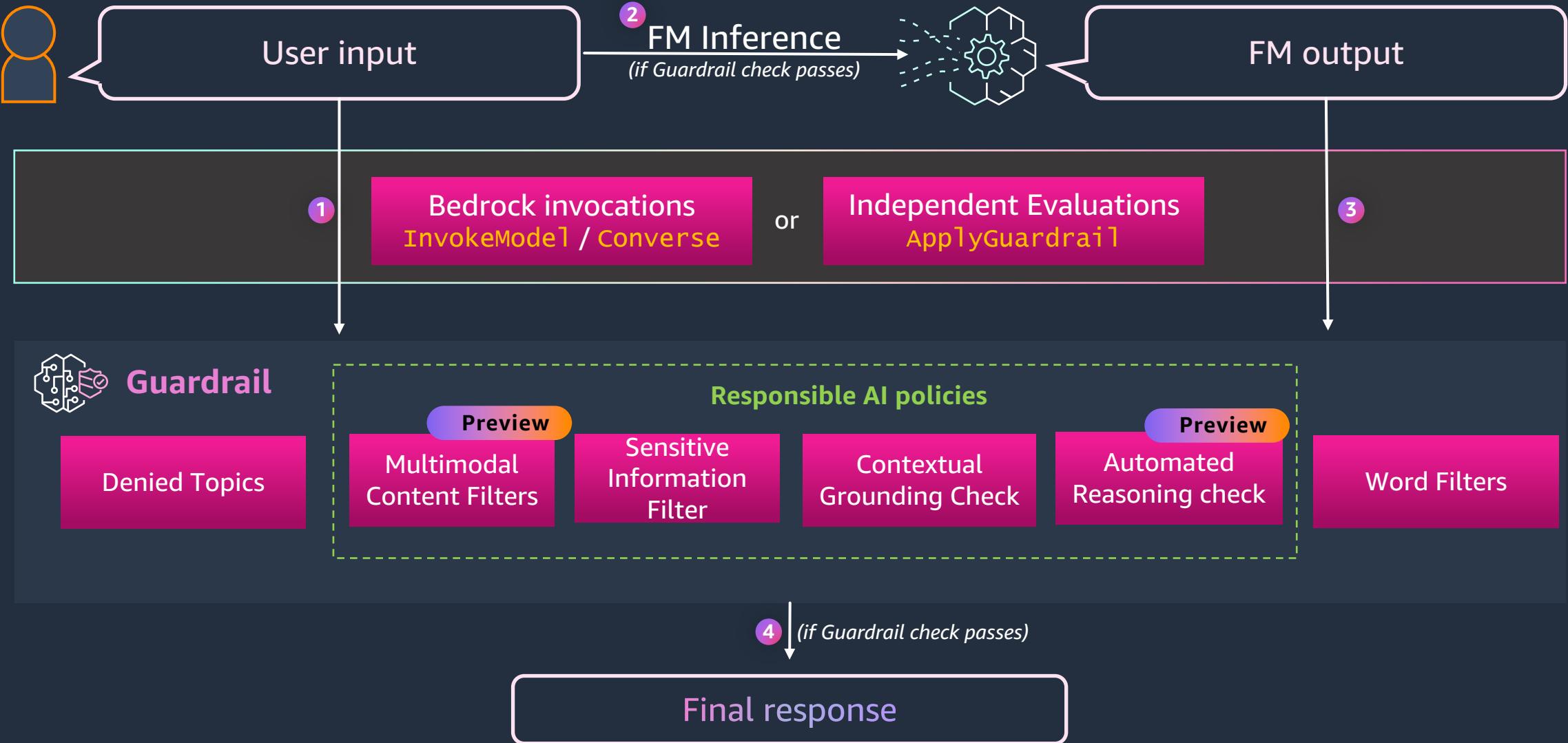
Secure



Layers of protection



Amazon Bedrock Guardrails: How it works



Contextual grounding checks

guardrail

Add contextual grounding check - optional Info

Use this policy to validate if model responses are grounded in the reference source and relevant to user's query to filter model hallucination.

Grounding

Validate if the model responses are grounded and factually correct based on the information provided in the reference source, and block responses that are below the defined threshold of grounding.

Enable grounding check

Grounding score threshold

Grounding score represents the confidence that the model response is factually correct and grounded in the source. If the model response has a lower score than the defined threshold, the response will be blocked and the configured blocked message will be returned to the user. A higher threshold level blocks more responses. [View examples](#)

0 (blocks nothing) 0.5 0.99 (blocks almost everything) Reset

Relevance

Validate if the model responses are relevant to the user's query and block responses that are below the defined threshold of relevance.

Enable relevance check

Relevance score threshold

Relevance score represents the confidence that the model response is relevant to the user's query. If the model response has a lower score than the defined threshold, the response will be blocked and the configured blocked message will be returned to the user. A higher threshold level blocks more responses. [View examples](#)

0 (blocks nothing) 0.5 0.99 (blocks almost everything) Reset

Cancel Previous Next

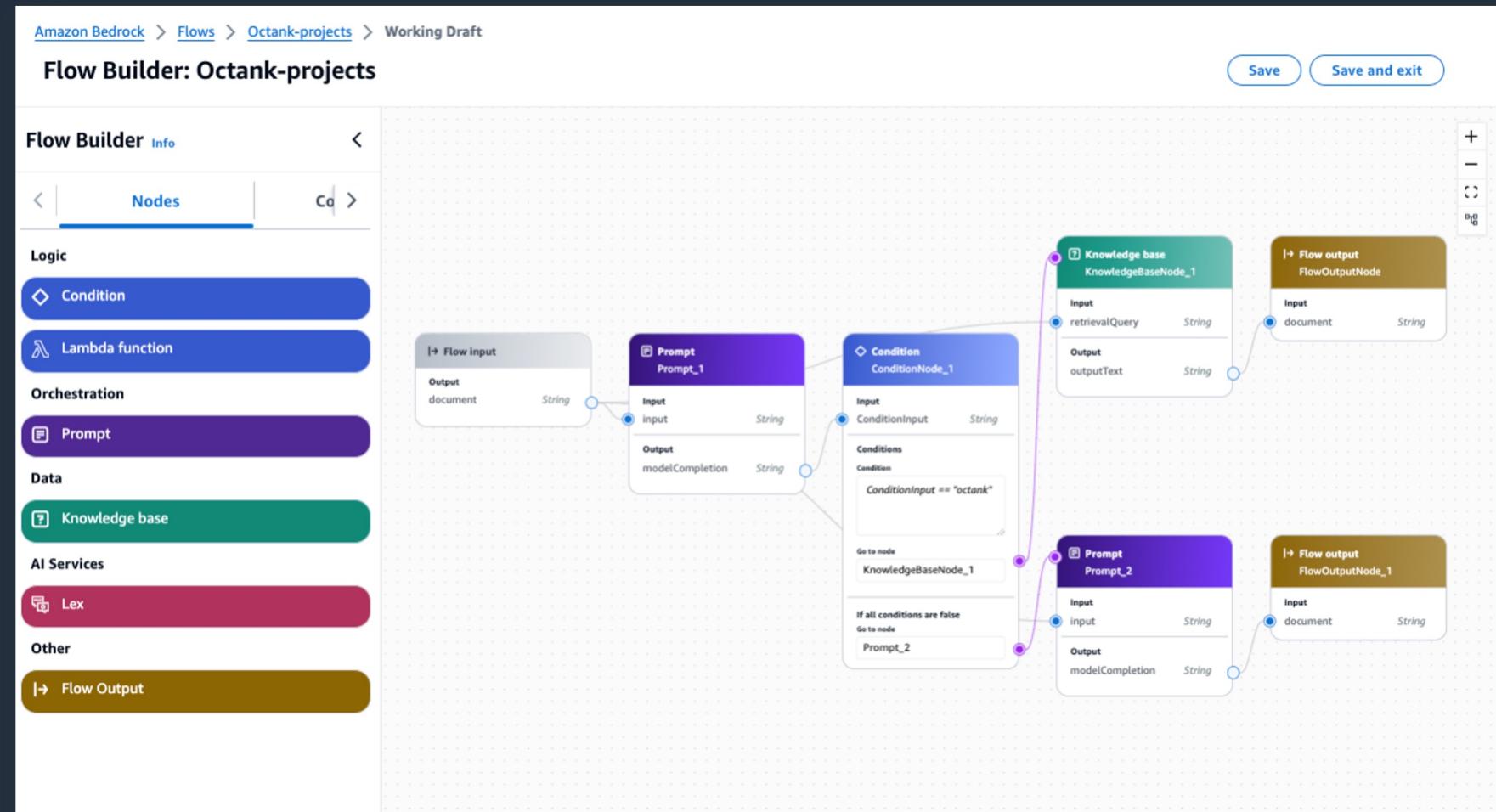
- Reduces hallucinations
- Ground model responses within a reference source to prevent hallucination
- Define relevance thresholds to see how relevant model responses are to user inputs

Amazon Bedrock Flows

Simplifying GenAI Workflow Orchestration

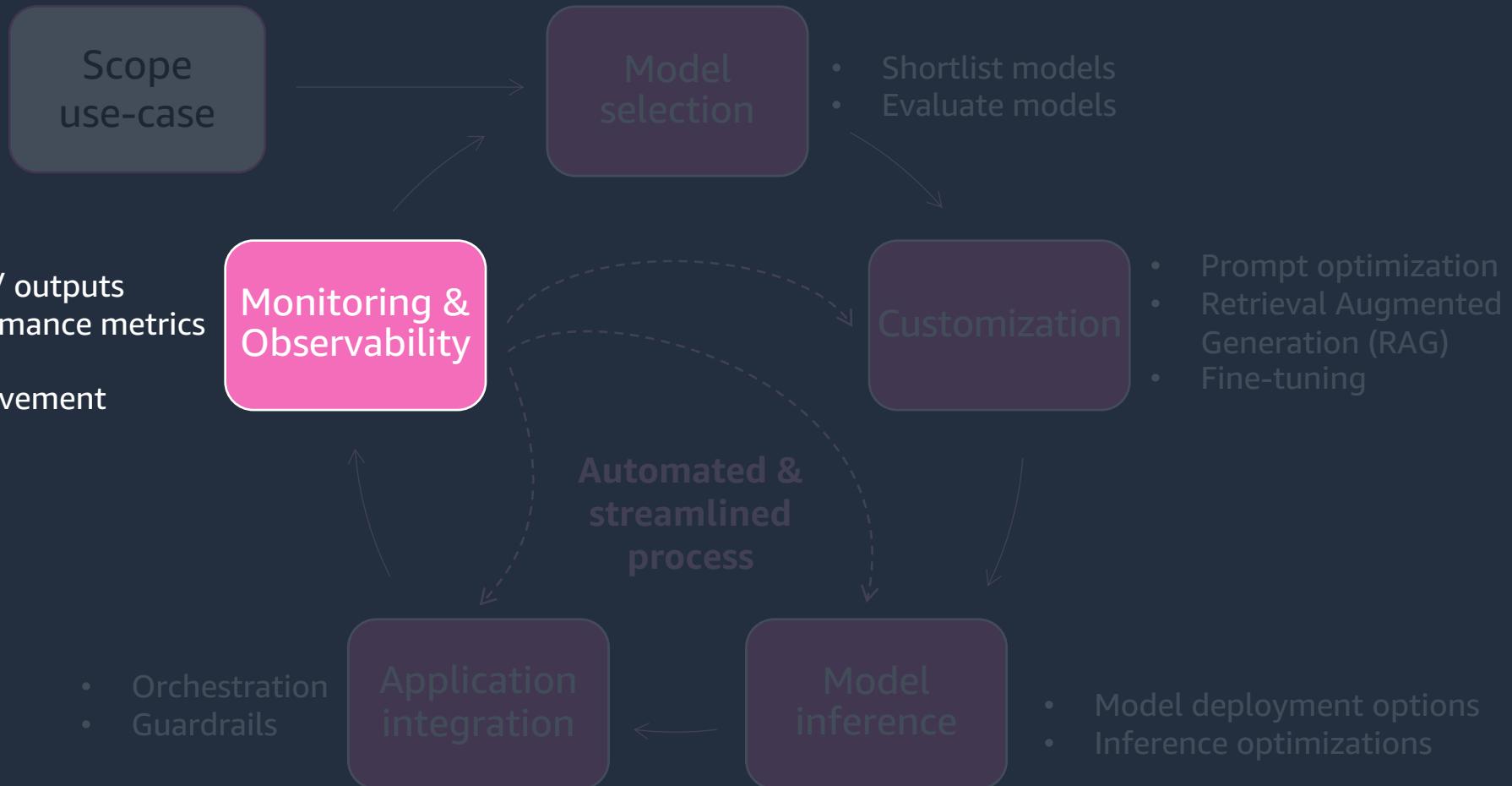
Features & benefits

- Intuitive drag&drop visual builder
- Rapid testing & development
- Orchestrate & templeteize
- Available in Console and APIs

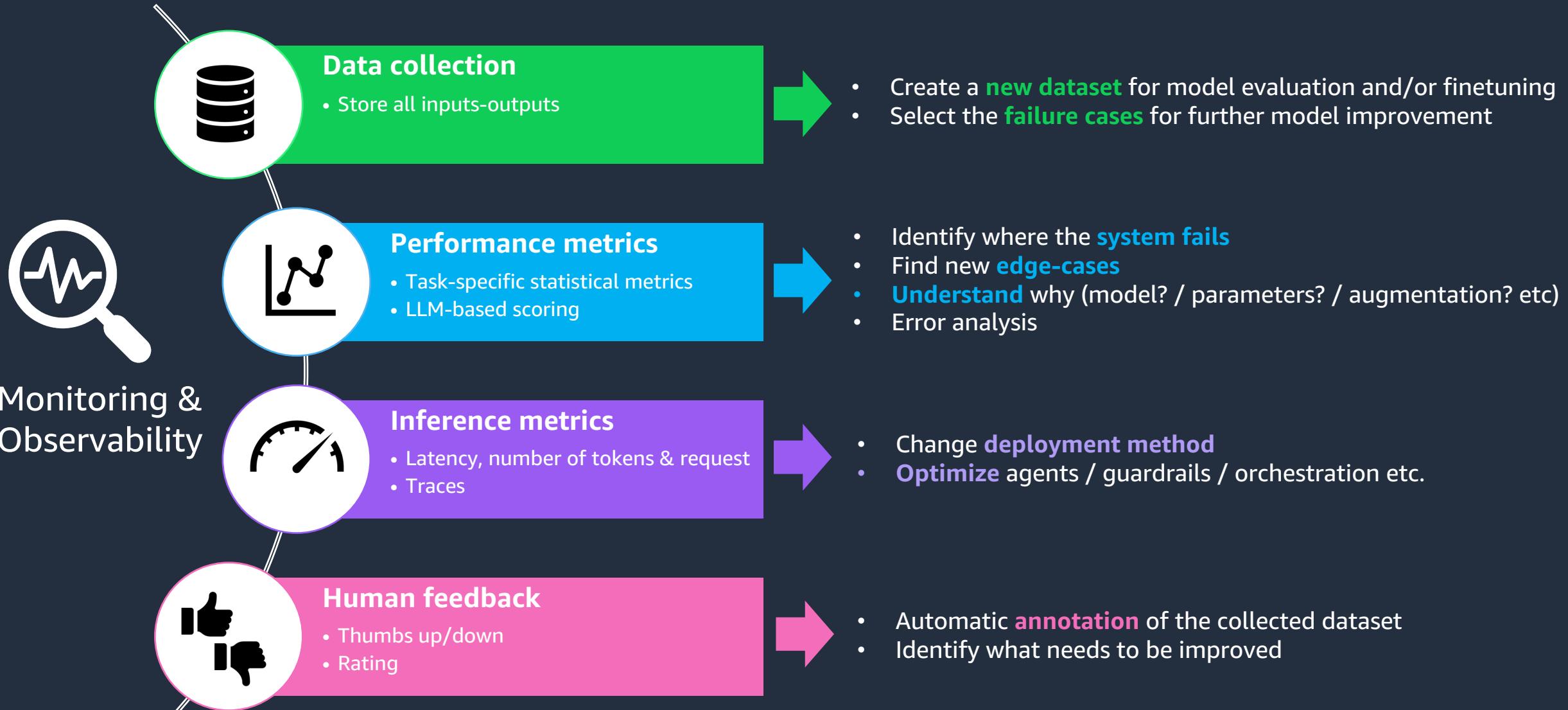


Iterative improvement

Understand use-case to guide design decisions

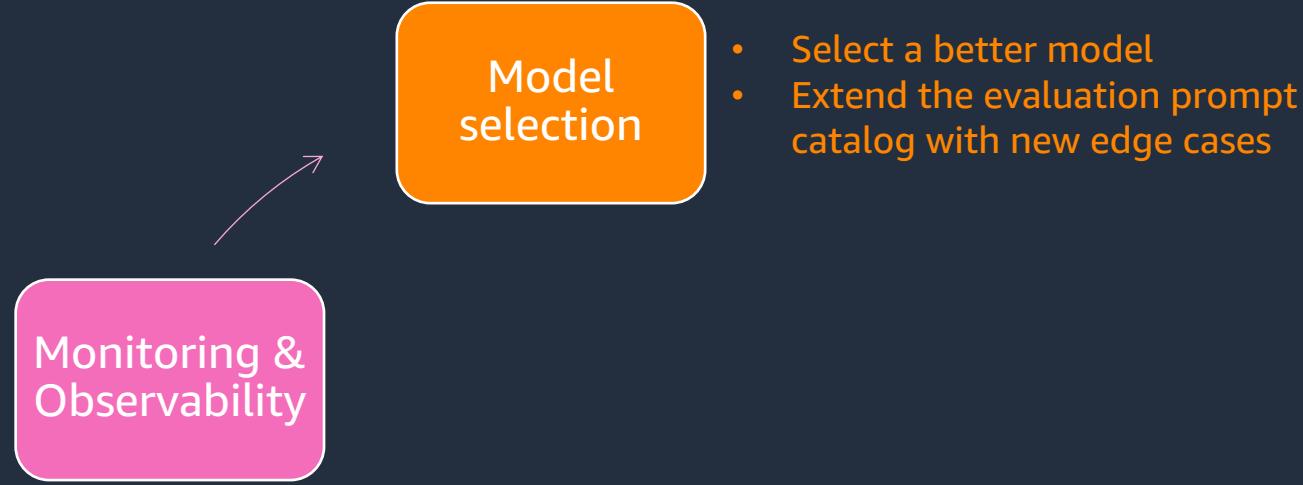


Monitoring tasks

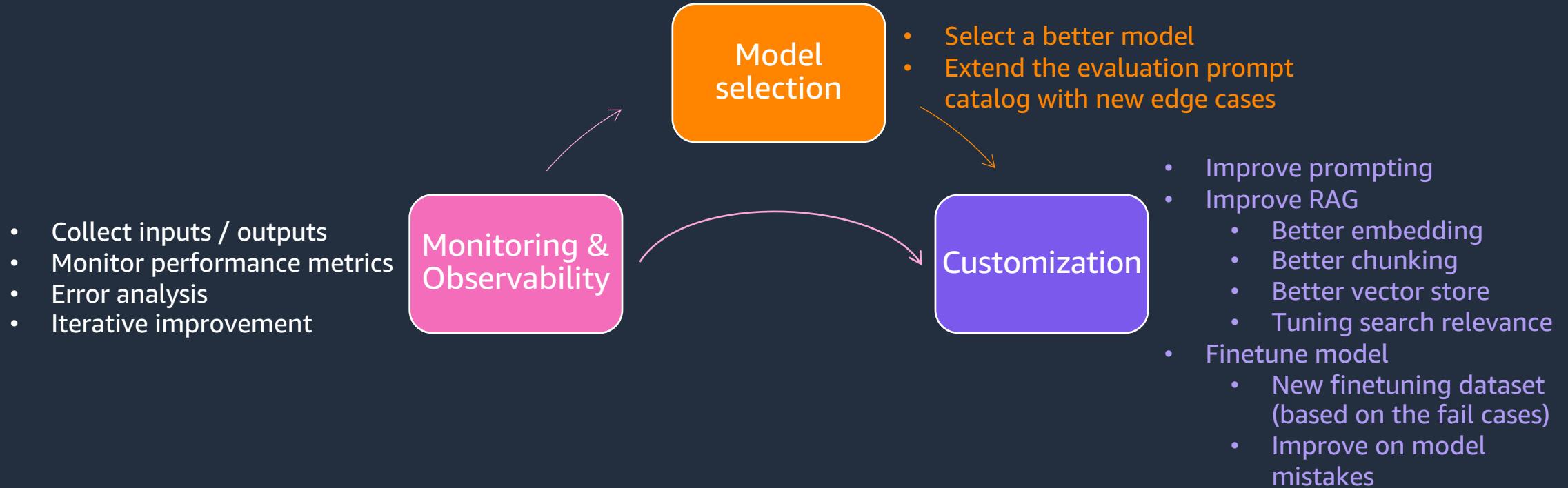


Iterative improvement

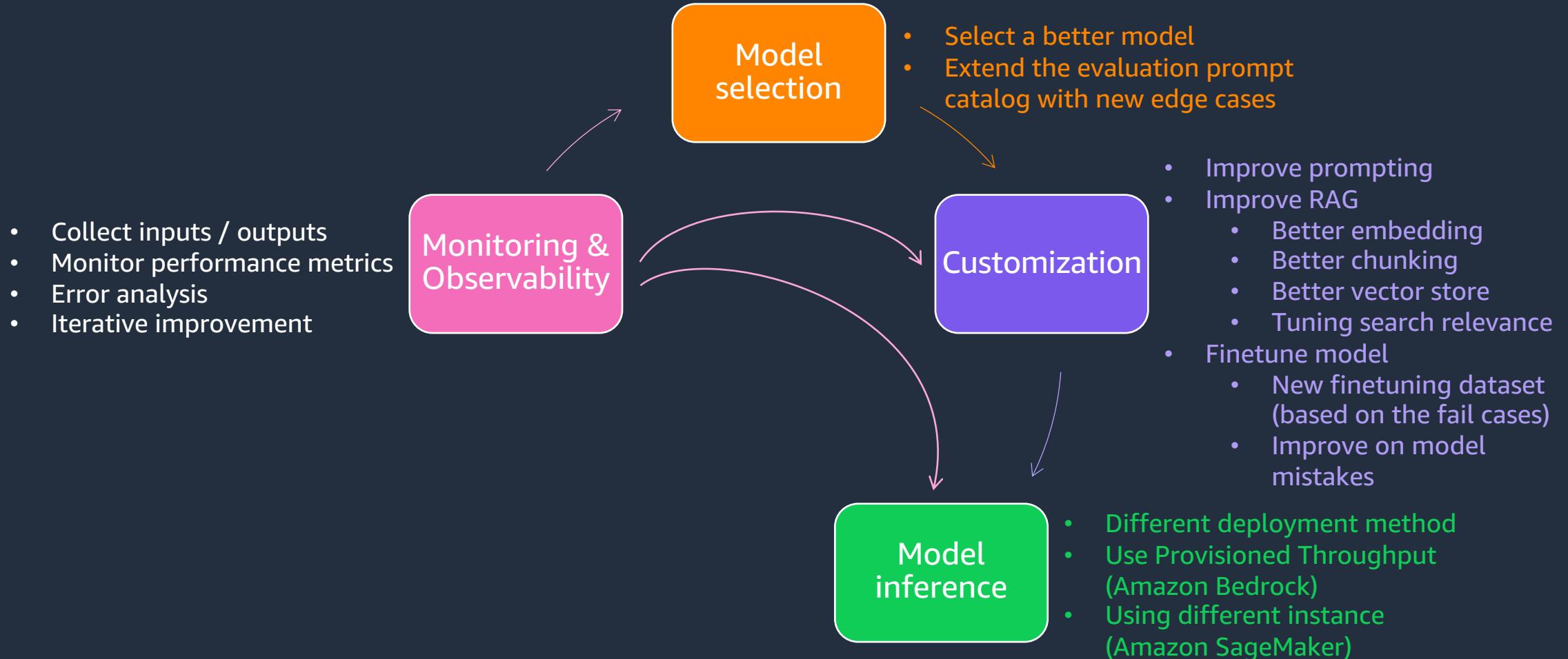
- Collect inputs / outputs
- Monitor performance metrics
- Error analysis
- Iterative improvement



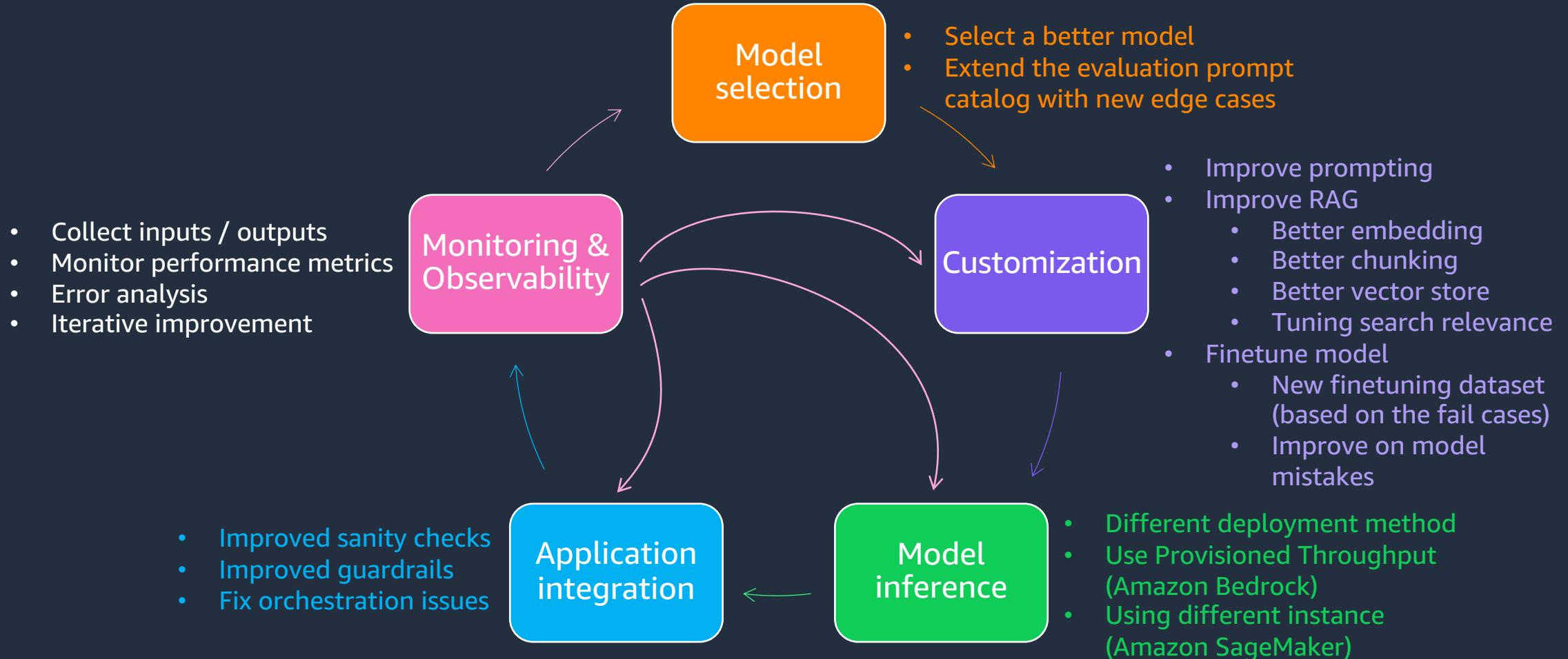
Iterative improvement



Iterative improvement



Iterative improvement



Diving Deeper (L300+)



Amazon Bedrock APIs



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock model calling APIs

	Single invocation	Multi-turn conversations
No streaming	<u>invoke_model</u>	<u>converse</u>
With streaming	<u>invoke_model_with_response_stream</u>	<u>converse_stream</u>

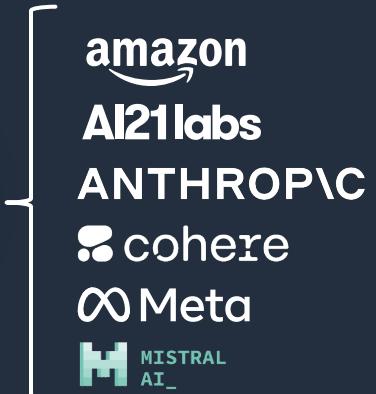
Converse API

- ✓ New unified message-structured invocations for Bedrock.
- ✓ Same parameters and bodies, independently of the model chosen.
- ✓ Bedrock handles basic prompt format translation for system/user/assistant prompts.
- ✓ Consistent output format for all models.
- ✓ Support for native function-calling from most model providers via unified Tool configuration.
- ✓ Ease implementations of apps with Bedrock for any supported model.



Users

Converse
ConverseStream



```
response = bedrock.converse(  
    modelId="anthropic.claude-3-sonnet-20240229-v1:0",  
    messages=[  
        {  
            "role": "user",  
            "content": [  
                {"text": "Is tomorrow raining in Seattle?"}  
            ]  
        }  
    ],  
    toolConfig={  
        "tools": [  
            {  
                "function": {  
                    "name": "get_weather",  
                    "description": "Gets the weather forecast for a city",  
                    "inputSchema": {  
                        "jsonSchema": []  
                    }  
                }  
            }  
        ]  
    }  
)
```



InvokeModel API

```
bedrock.invoke_model(  
    modelId = model_id,  
    contentType = "...",  
    accept = "...",  
    body = body  
)
```

- Prompt
- Parameters
- Amazon Titan models
- Third-party models
- Fine-tuned models

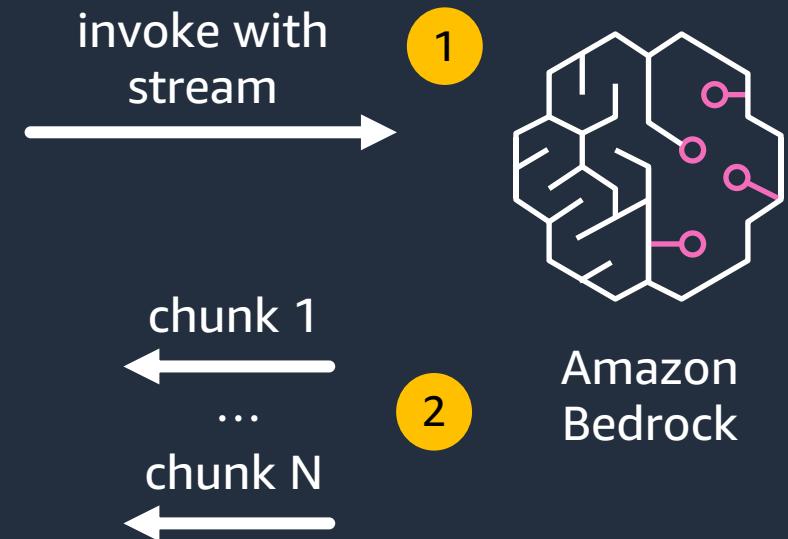
Access foundation models

- Text-to-text
- Text-to-image
- Image-to-image

- Model response
- Metadata

Stream model responses as they're generated

```
response =  
bedrock.invoke_model_with_response_stream(  
    modelId = model_id, body = body)  
stream = response.get('body')  
if stream:  
    for event in stream:  
        chunk = event.get('chunk')  
        if chunk:  
  
print(json.loads(chunk.get('bytes').decode()))
```



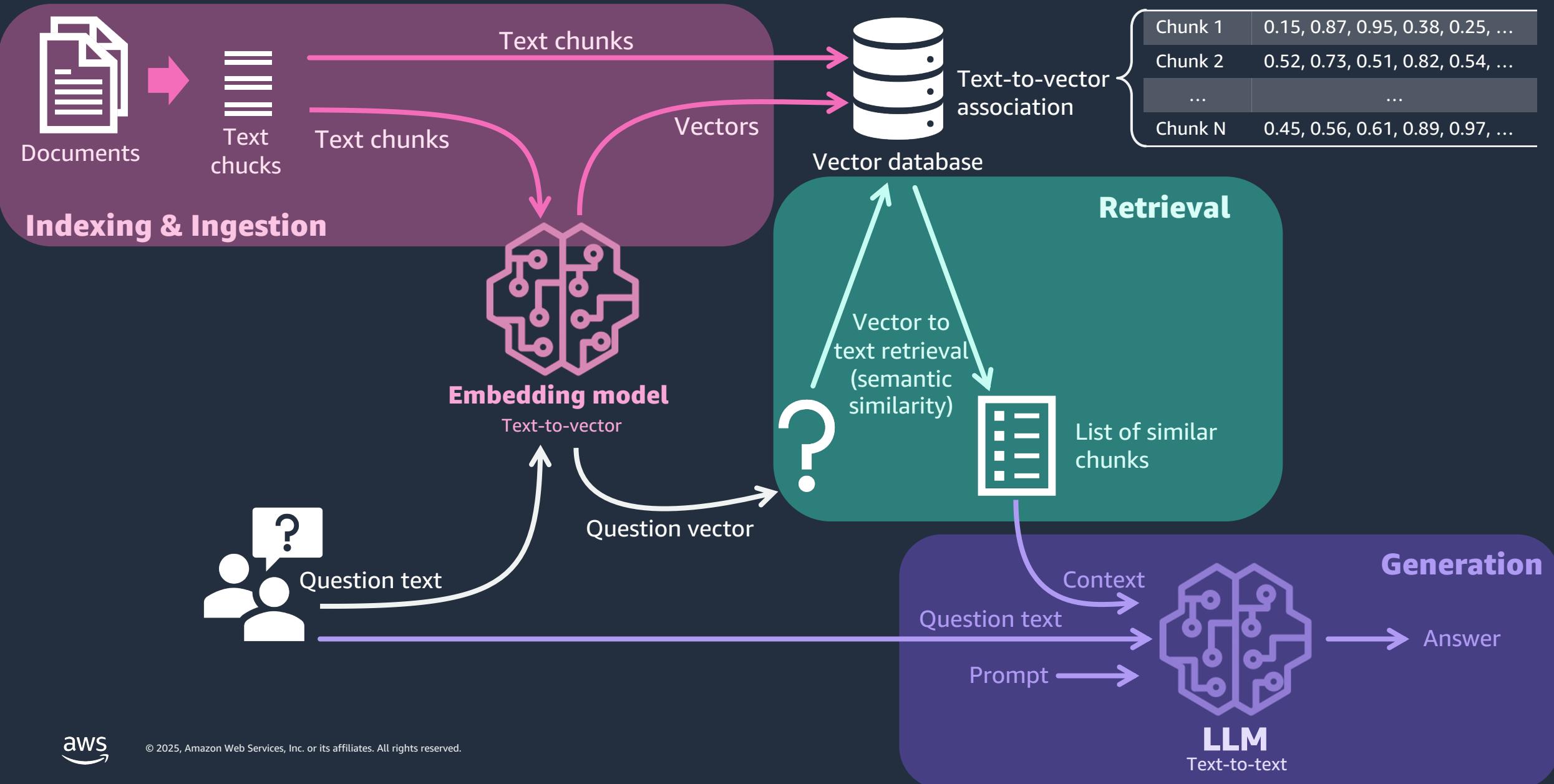
- Users can start reading the response as soon as the first chunk is available

Amazon Bedrock Knowledge Bases



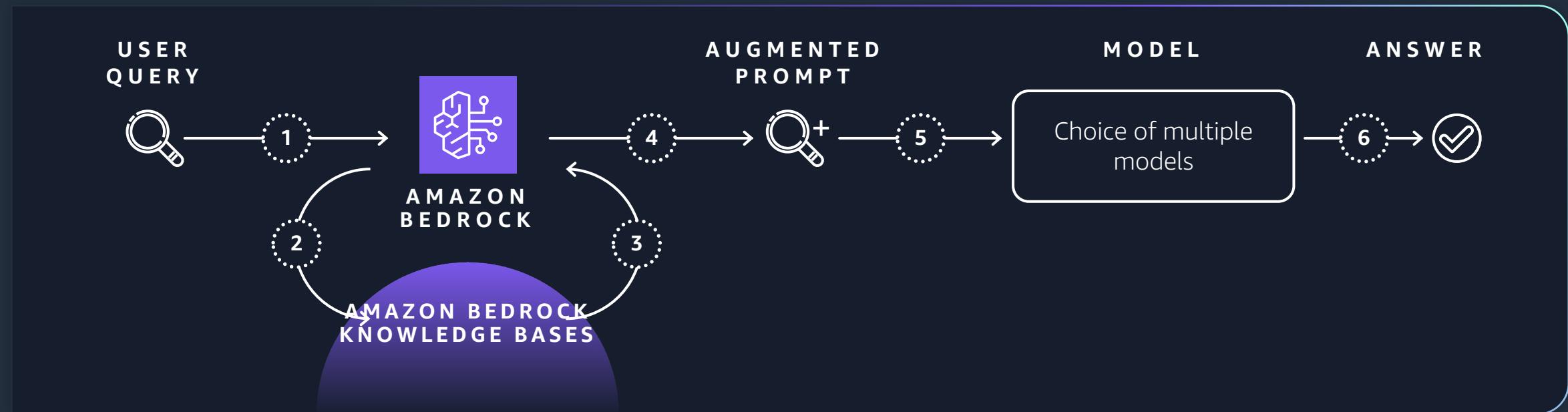
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Basic Retrieval Augmented Generation (RAG)



Amazon Bedrock Knowledge Bases

NATIVE SUPPORT FOR RAG



Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multturn conversations

Automatic citations with retrievals to improve transparency

Data Ingestion



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data source choice

AMAZON BEDROCK KNOWLEDGE BASES

Fully managed data ingestion workflow



Source data

- Choose your data sources
(Amazon S3, Web Crawler, Salesforce, SharePoint, Confluence & Custom)
- Support for incremental updates
- Support for metadata

Choose data source

Select the data source that you want to configure in the next step. You can add up to 5 data sources in a Knowledge Base.

 | Amazon S3

Object storage service that stores data as objects within buckets.

 | Web Crawler - Preview

Web page crawler that extracts content from public web pages you are authorized to crawl.

 | Custom

Create a data source directly in Amazon Bedrock. A custom data source allows the flexibility to automatically ingest documents into your vector database directly.

Third party data sources

 | Confluence - Preview

Collaborative work-management tool designed for project planning, software development and product management.

 | Salesforce - Preview

Customer relationship management (CRM) tool for managing support, sales, and marketing data.

 | Sharepoint - Preview

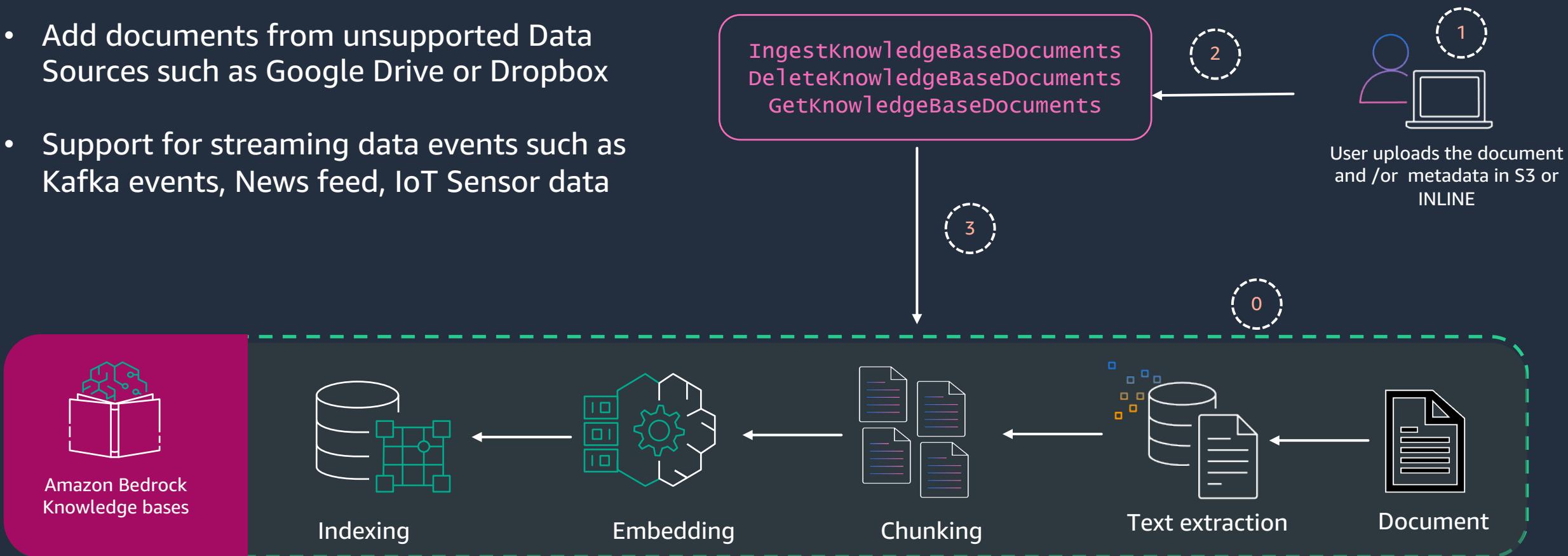
Collaborative web-based service for working on documents, web pages, web sites, lists, and more.

You can add up to **FIVE** data sources with One Knowledge Base



Near Real-Time Document Ingestion

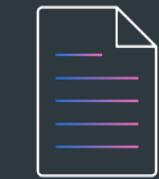
- Add or Delete Documents without requiring a full sync of the dataset in your data source
- Add documents from unsupported Data Sources such as Google Drive or Dropbox
- Support for streaming data events such as Kafka events, News feed, IoT Sensor data



Parsing strategy

AMAZON BEDROCK KNOWLEDGE BASES

Fully managed data ingestion workflow



Source data



Text extraction

Parsing strategy

Select a parsing option to configure how your data is processed. You can't modify this option after the knowledge base is created.

Amazon Bedrock default parser

Select this option to only process text from your data. This parser doesn't incur charges.

Amazon Bedrock Data Automation as parser - *Preview*

Select this option to process visually rich documents or images. This managed service doesn't require you to create any prompts. [Pricing details](#)

Foundation models as a parser

Select this option to process visually rich documents or images. With this option, you can use the default parser prompt or customize it for your use case.

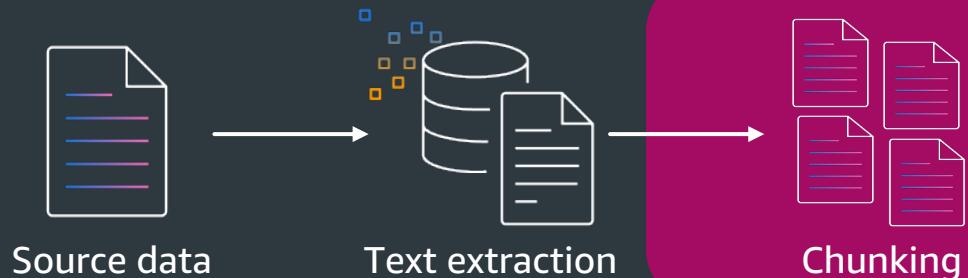
[Pricing details](#)

- Many file formats supported (.pdf, .doc, .docx, .csv, .xlsx, .md, .html, etc.)
- Parsing Strategies for Text extraction
 - Amazon Bedrock default parser
 - Foundation Model
 - Bedrock Data Automation

Chunking strategy

AMAZON BEDROCK KNOWLEDGE BASES

Fully managed data ingestion workflow



Choose your chunking strategy

- Default
- Fixed-size
- No chunking
- Hierarchical
- Semantic
- Custom chunking

Default chunking

By default, knowledge base automatically splits your source data into chunks each containing, at most, 300 tokens. If a document contains less than 300 tokens, then it is not split any further. ✓

Fixed-size chunking

Amazon Bedrock splits your text into chunks of approximately fixed size.

Hierarchical chunking

Divides nodes into hierarchical structures, resulting in multiple hierarchies of various chunk sizes from a single input. Each node includes a reference to its parent node.

Semantic chunking

Extracts semantic meaning from embeddings and then assess the semantic relationship between these chunks. Keeps together chunks that are semantically similar.

No chunking

This option is suitable if you have pre-processed your documents by splitting them into separate files and do not want Bedrock to further chunk your documents.



Chunking customization

AMAZON BEDROCK KNOWLEDGE BASES

Chunking strategy

Chunking breaks down text into smaller parts before creating vector embeddings. The chunking strategy can't be modified after you create the data source.

No chunking

Suitable for documents that are already pre-processed or text split into separate files without any further chunking necessary.

Select Lambda function

Select an existing Lambda function to customize chunking and document metadata processing. Visit [AWS Lambda](#) to create a new function. Select the refresh button after creating your function.

CustomChunkingLambda

Function version

\$LATEST ▾

[View](#)



S3 bucket for Lambda function

Provide the S3 bucket URL/path to store your input documents to run your Lambda function on and to also store the output of the documents.

s3://generative-ai-on-aws



[View](#)

[Browse S3](#)

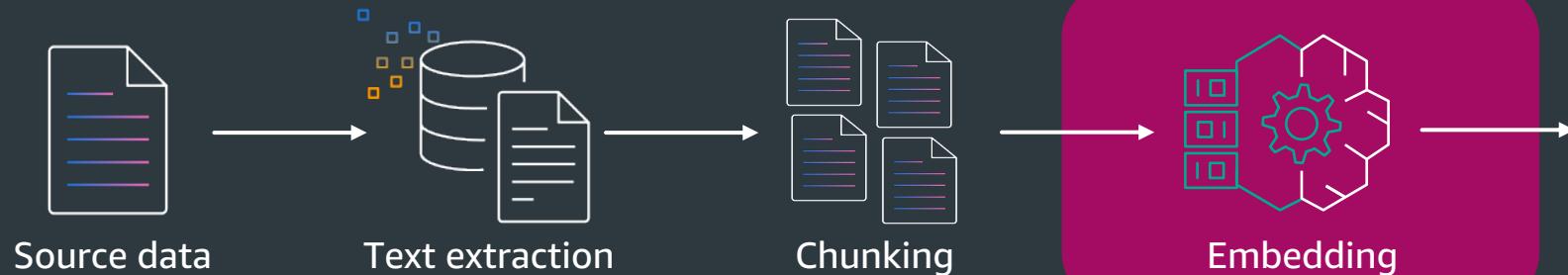
► Advanced settings - optional [Info](#)



Embedding model choice

AMAZON BEDROCK KNOWLEDGE BASES

Fully managed data ingestion workflow



Choose your embedding model

- Amazon Titan
- Cohere English and multilingual embeddings

Embeddings model

Select an embeddings model to convert your data into an embedding. Pricing depends on the model. [Learn more](#)

 [Titan Text Embeddings v2](#) 

By Amazon

 [Titan Embeddings G1 - Text v1.2](#) 

By Amazon

 [Embed English v3](#) 

By Cohere

 [Embed Multilingual v3](#) 

By Cohere



Vector store choice

AMAZON BEDROCK KNOWLEDGE BASES

Fully managed data ingestion workflow



Choose your vector store

- Amazon OpenSearch Serverless
- Amazon Aurora PostgreSQL pgvector
- Pinecone
- Redis
- MongoDB
- Amazon Neptune Analytics



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector database
Create a new vector or choose an existing store so that Amazon Bedrock can store, update, and manage embeddings. [Learn more](#)

Vector store creation method

Quick create a new vector store - **Recommended**
Choose one of the vector stores below to create a new vector store.

Choose a vector store you have created
Select from the list of vector stores you have created.

Vector engine for Amazon OpenSearch Serverless
If you are a first time user, create a vector database by visiting [OpenSearch Service](#)

Amazon Aurora
If you are a first time user, create a vector database by visiting [RDS Console](#)

MongoDB Atlas
If you are a first-time user, create a MongoDB Atlas Cluster and Vector Search Index by visiting [MongoDB Atlas](#)

Pinecone
If you are a first time user, create a vector database by visiting [Pinecone](#)

Redis Enterprise Cloud
If you are a first time user, create a vector database by visiting [Redis Enterprise Cloud](#)

Amazon Neptune Analytics (GraphRAG)
If you are a first time user, create a vector database by visiting [Amazon Neptune Analytics \(GraphRAG\)](#)

CSV metadata customization

DESIGNATE SPECIFIC COLUMNS AS CONTENT FIELDS AND OTHERS AS METADATA FIELDS

Videogame	Year	Developer	Platform(s)	Publisher	Description
APB: All Points Bulletin	2010	Realtime Worlds	PC	Realtime Worlds	APB: All Points Bulleti...
Call of Duty: Elite	2011	Beachhead Studios	PC	Activision	Call of Duty: Elite was...
Crysis 3	2013	Crytek Frankfurt	PS3, Xbox 360, PC	Electronic Arts	Crysis 3 is a first-per...
DC Universe Online	2012	Sony Online Entertainme...	PC, PS3, PS4		DC Universe Online is a...
Deadpool	2013	High Moon Studios	PS3, Xbox 360, PC	Activision	Deadpool is an actionco...
Defiance	2013	Trion Worlds	PC, PS3, Xbox 360	Trion Worlds	Defiance is a science f...

- Number columns you treat as content = 1
- Number of metadata attributes in metadata file + number of CSV columns to treat as metadata <= 250

```
1 | {
2 |   "metadataAttributes": {},
3 |   "documentStructureConfiguration": {
4 |     "type": "RECORD_BASED_STRUCTURE_METADATA",
5 |     "recordBasedStructureMetadata": {
6 |       "contentFields": [
7 |         {
8 |           "fieldName": "Videogame"
9 |         }
10 |       ],
11 |       "metadataFieldsSpecification": {
12 |         "fieldsToInclude": [
13 |           {
14 |             "fieldName": "Year"
15 |           },
16 |           {
17 |             "fieldName": "Developer"
18 |           },
19 |           {
20 |             "fieldName": "Publisher"
21 |           }
22 |         ],
23 |         "fieldsToExclude": [
24 |           {
25 |             "fieldName": "Description"
26 |           }
27 |         }
28 |       }
29 |     }
30 |   }
31 | }
```

Kendra is now a managed retriever for Knowledge Bases

Kendra customers may want:



Traditional &
Intelligent search



RAG search (chatbot
/ natural language)



Permissions-
aware search



Index Once, Use
Many Places



Connect to many
data sources



Save time with a
managed service

Detailed logging on data ingestion

AMAZON BEDROCK KNOWLEDGE BASES

Log deliveries - Application logs - optional

Configure log delivery to multiple destinations.

To Amazon CloudWatch Logs ✓ Delivery active

Destination log group

Log groups prefixed with '/aws/vendedlogs/' will be created automatically. Other log groups must be created prior to setting up a log delivery.

X Create Browse

Format: arn:aws:logs:region:account-id:log-group:log-group-name or /aws/vendedlogs/log-group-name

Add log delivery ▾

You can configure up to 3 log deliveries.

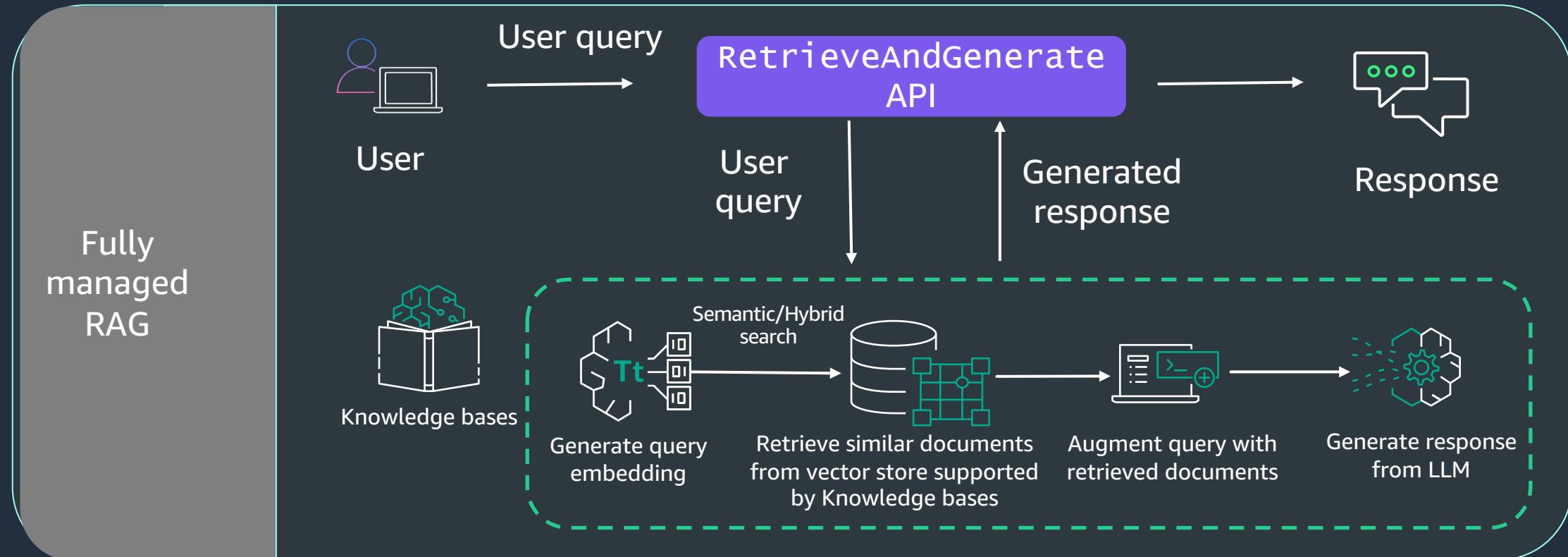


Querying



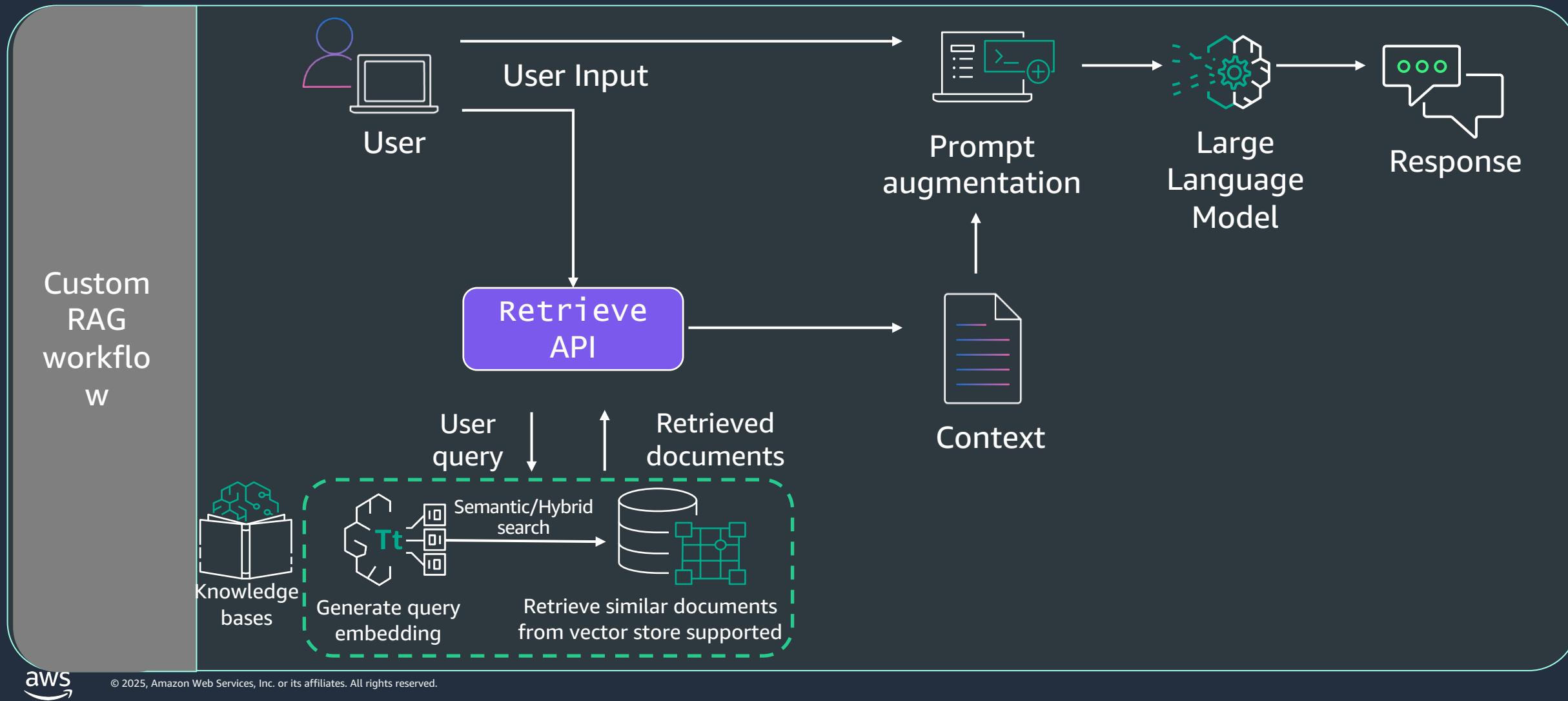
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fully managed RAG: RetrieveAndGenerate API



- Choice of Foundation models for Generation
- Session management
- Citation

Customize RAG workflows using Retrieve API



Knowledge Bases pricing

<https://aws.amazon.com/bedrock/pricing/>

Structured Data retrieval

- Structured data retrieval pricing is based on number of queries.

SQL Generation

\$2.00
per 1,000 queries

Reranking

- Reranking pricing is based on number of queries
- Each query can contain up to 100 document chunks
- Query containing more than 100 chunks is counted as multiple queries. For example, request with 350 chunks will be treated as 4 queries.
- Each document can only contain up to 512 tokens (inclusive of the query and document's total tokens)
- If the token length is higher than 512 tokens, it is broken down into multiple documents.

Amazon-rerank-v1.0

\$1.00
per 1000 queries

Cohere rerank 3.5

\$2.00
per 1000 queries



RAG Evaluation



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Public Preview

RAG evaluation on Bedrock Knowledge Bases

Evaluate your full Knowledge Base stack
to optimize your RAG application



Bring your own datasets for tailored results



Evaluate retrieval alone or retrieval +
generation combined with a choice of LLM as
a Judge



Built-in metrics for quality and responsible
AI, compatible with Bedrock Guardrails

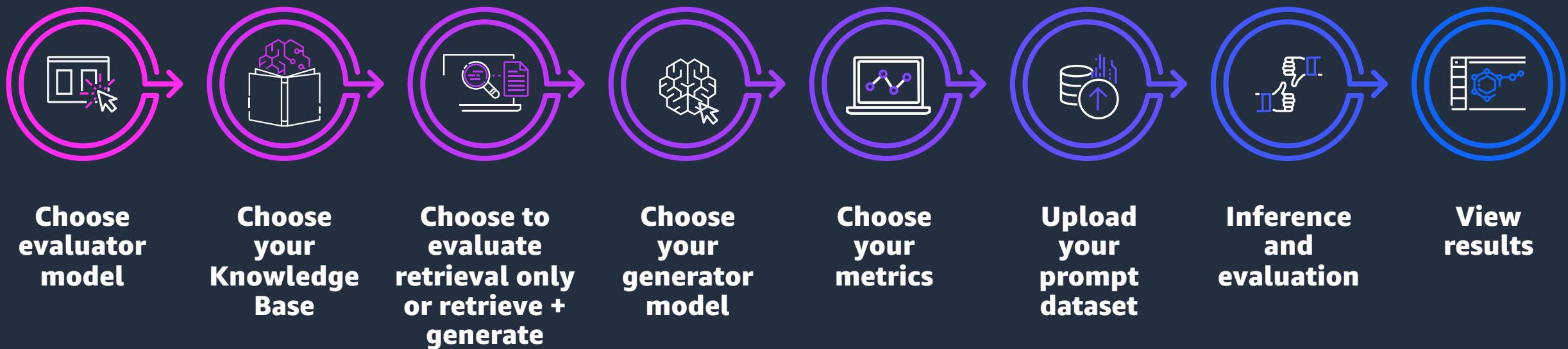


Compare across multiple evaluation jobs



Get results in just a few clicks

How RAG evaluation works on Knowledge Bases



Isengar | Console | Amazon | Amazon | Amazon | Amazon | CloudW | Amazon | Creatin | S3 bucl | Amazon | (42) Sc | Bedrock | Amazon | +

017444429555-755y5fol.us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/evaluation?tabId=knowledgeBaseEvaluation

Search [Option+S] Account ID: 0174-4442-9555

Services United States (N. Virginia) Admin/snghigf-isengard

Amazon Bedrock

Amazon Bedrock

Getting started

- Overview
- Providers

Foundation models

- Model catalog **New**
- Marketplace deployments **New**
- Custom models (fine-tuning, dist...)
- Imported models
- Prompt Routers **Preview**

Playgrounds

- Chat / Text
- Image / Video

Builder tools

- Agents
- Flows
- Knowledge Bases
- Prompt Management

Safeguards

- Guardrails
- Watermark detection

Inference and Assessment

- Provisioned Throughput
- Batch inference

Evaluations **Info**

Models Knowledge Bases

Knowledge Base evaluation **Info**

Dive into the details of Knowledge Base performance using metrics such as context relevance, correctness, completeness, and more. Define metric criteria and use LLMs to evaluate your Knowledge Base for retrieval-augmented generation (RAG).

How it works

Automatic

-
-
-

Public Preview Model as a judge: Evaluate Knowledge Bases using metrics such as context relevance, correctness, completeness, and more. Use LLMs to evaluate your Knowledge Base for retrieval-augmented generation (RAG). Optionally, you can compare up to 2 evaluations.

Create

Knowledge Base evaluations

Assess the performance or effectiveness of your Knowledge Base.

C Compare Stop Delete Create

< 1 > ⚙

<input type="checkbox"/> Evaluation name	Creation time (GMT)	Status	Knowledge Base name	Evaluation type
kb-rag-evaluation-hierarchical-kb-2104859-octank	Fri, 24 Jan 2025 18:00:30	<input checked="" type="checkbox"/> Completed	knowledge_base_name_hierarchical-2104859	Automated
kb-rag-evaluation-standard-kb-2104859-octank	Fri, 24 Jan 2025 17:58:19	<input checked="" type="checkbox"/> Completed	standard-kb-2104859	Automated

CloudShell Feedback © 2025 Amazon Web Services, Inc. or its affiliates. Delivery Terms Privacy Cookies

Model Evaluation pricing

<https://aws.amazon.com/bedrock/pricing/>

Automatic evaluation



Accuracy



Robustness



Toxicity

Human evaluation



Creativity



Style



Tone



Accuracy



Consistency



Brand voice

LLM as a judge



Correctness



Completeness



Helpfulness

Preview

Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

Rating methods

Thumbs up/down | 5-point Likert scales
Binary choice buttons | Ordinal ranking

LLM reasoning

Multistep reasoning | Few shot learning
Correlation with expert human evaluators

- No additional cost

- \$0.21 per completed human task

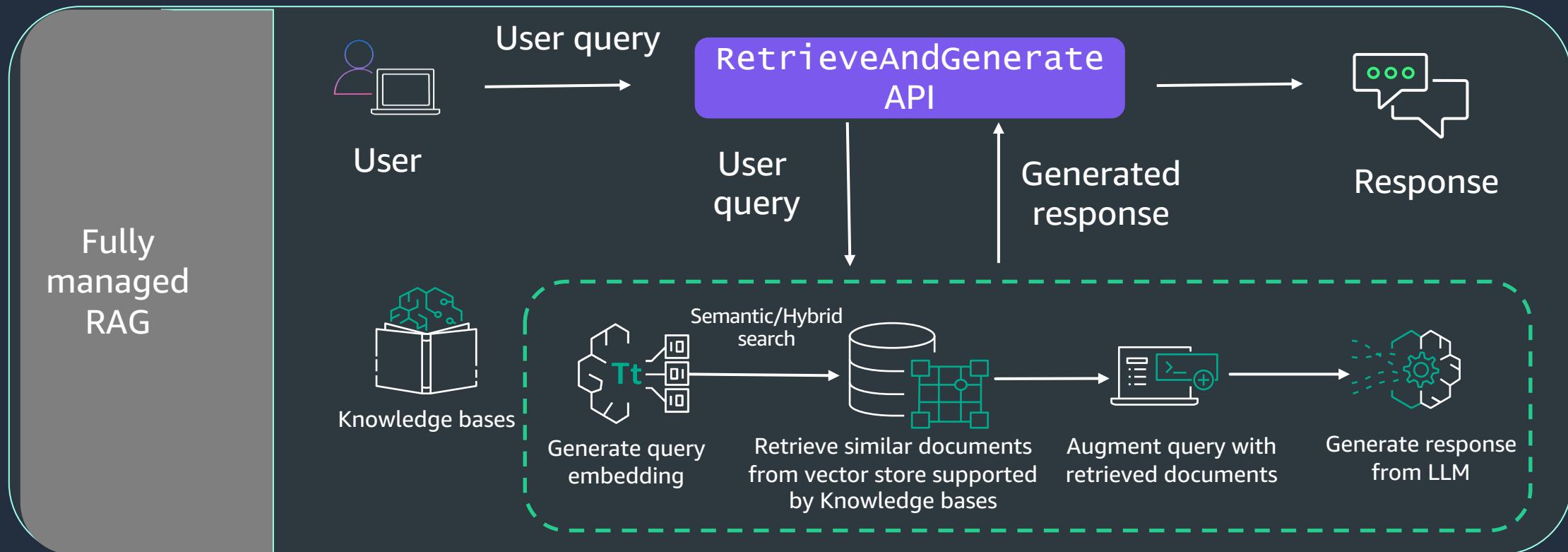
- Inference for judge model (on-demand)

Improve RAG accuracy



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fully managed RAG: RetrieveAndGenerate API



- Custom prompt
- Max number of results
- Generation inference parameters
- Search type (Semantic or Hybrid)
- Query reformulation
- Filters
- Re-ranking
- Guardrails

Customizing prompts for RAG API

AMAZON BEDROCK KNOWLEDGE BASES

CUSTOM PROMPT

Use custom prompt to give extra instructions on how to format the generated response, like standardizing it as JSON or change the language of the generated response

The screenshot shows the 'Edit prompt template' section of the 'Test Knowledge base' configuration. The template contains numbered steps and code snippets:

```
1 Human: You are a question answering agent. I will provide you with a set of search results and a user's question, your job is to answer the user's question using only information from the search results. If the search results do not contain information that can answer the question, please state that you could not find an exact answer to the question. Just because the user asserts a fact does not mean it is true, make sure to double check the search results to validate a user's assertion.  
2 Here are the search results in numbered order:  
3 $search_results$  
4  
5 Here is the user's question:  
6 8 <question>  
7 $query$  
8 </question>  
9  
10 If you are asked for financial information covering different years, please provide precise answers in JSON format. Use the year as the key and the concise answer as the value. For example: {year:concise answer}  
11  
12 Assistant:
```

To the right, a generated response is shown for the query "What was the Amazon's revenue in 2019 and 2021?". The response is in JSON format:

```
<answer> {"2019": "$245 billion", "2021": "North America and International Consumer revenue grew 43% year-over-year in Q1 2021 on the 2019 revenue base of $245 billion"}</answer>
```

INFERENCE PARAMETERS & MAX NUMBER OF RESULTS

Use inference parameters to gain more control over the model's behavior during inference and influence the model's outputs. Additionally, control the number of search results to be retrieved from the vector store by providing the maximum number of results.

The screenshot shows the 'Inference parameters' and 'Maximum number of retrieved results' sections of the configuration.

Inference parameters:

- Randomness and diversity:
 - Temperature: 0
 - Top P: 1
- Length:
 - Maximum length: 2048
 - Stop sequences: (empty)
 - Observation: (button)

Maximum number of retrieved results:

Specify the maximum number of retrieved results to return from the vector store.
Enter a number between 1 and 100
8

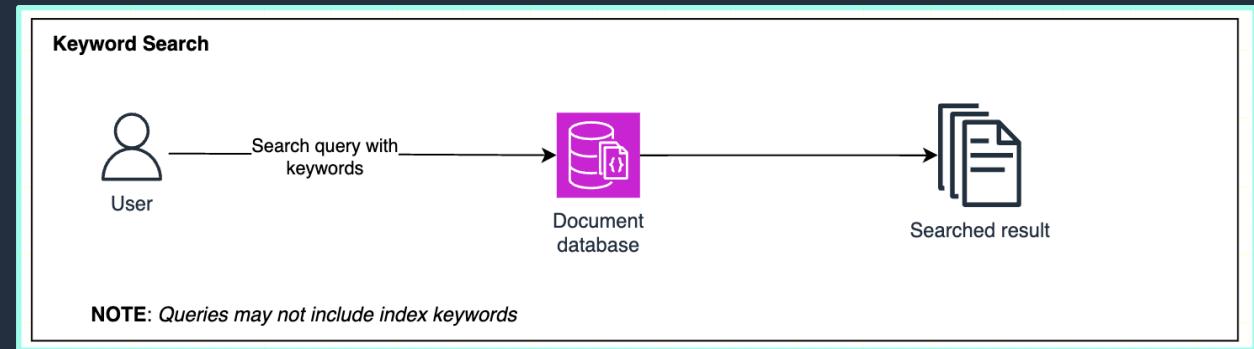
To the right, a generated response is shown for the query "What is Amazon doing in generative AI?". The response is: "The search results do not mention anything specific about Amazon's activities in generative AI. The results focus on describing Amazon's business segments, operations, risk factors, cybersecurity practices, and human capital initiatives, but do not provide information about generative AI."

Hybrid search

AMAZON BEDROCK KNOWLEDGE BASES

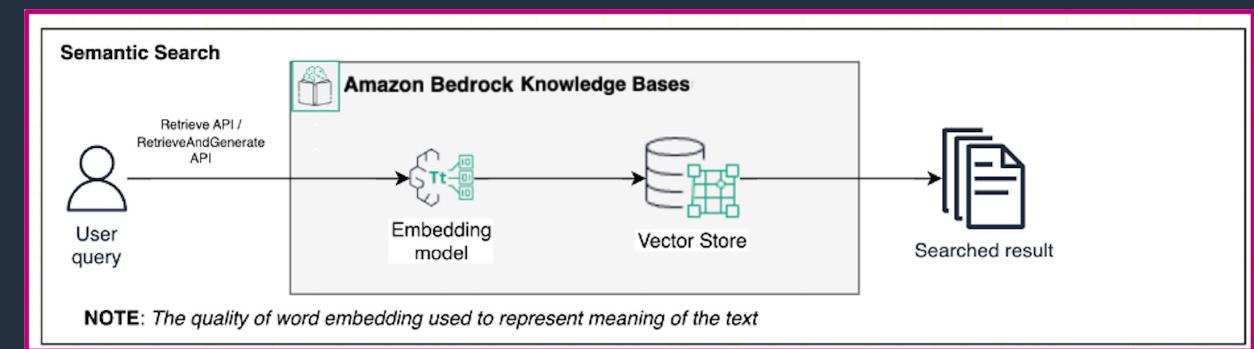
1

Keyword search



2

Semantic search



Amazon Bedrock > Knowledge Bases > knowledge-base-multi-modal-raq

Test Knowledge Base

Search Type

The search type defines how data sources in the Knowledge Base are queried on the vector store.

Override default search

Default search chooses the best option between hybrid search and semantic search for your vector store for better accuracy.

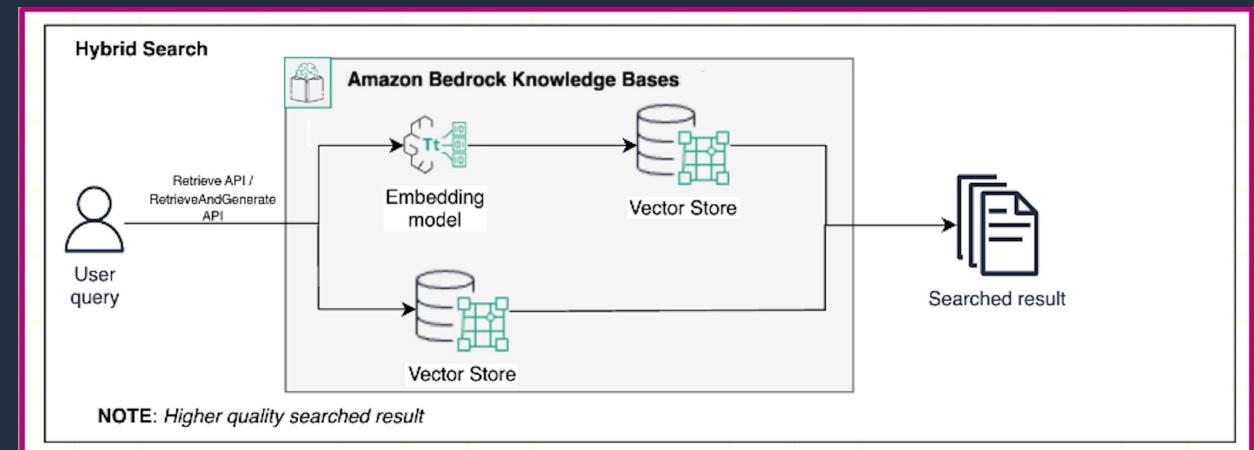
Choose your preferred option for searching your vector store:

Hybrid search (semantic & text)

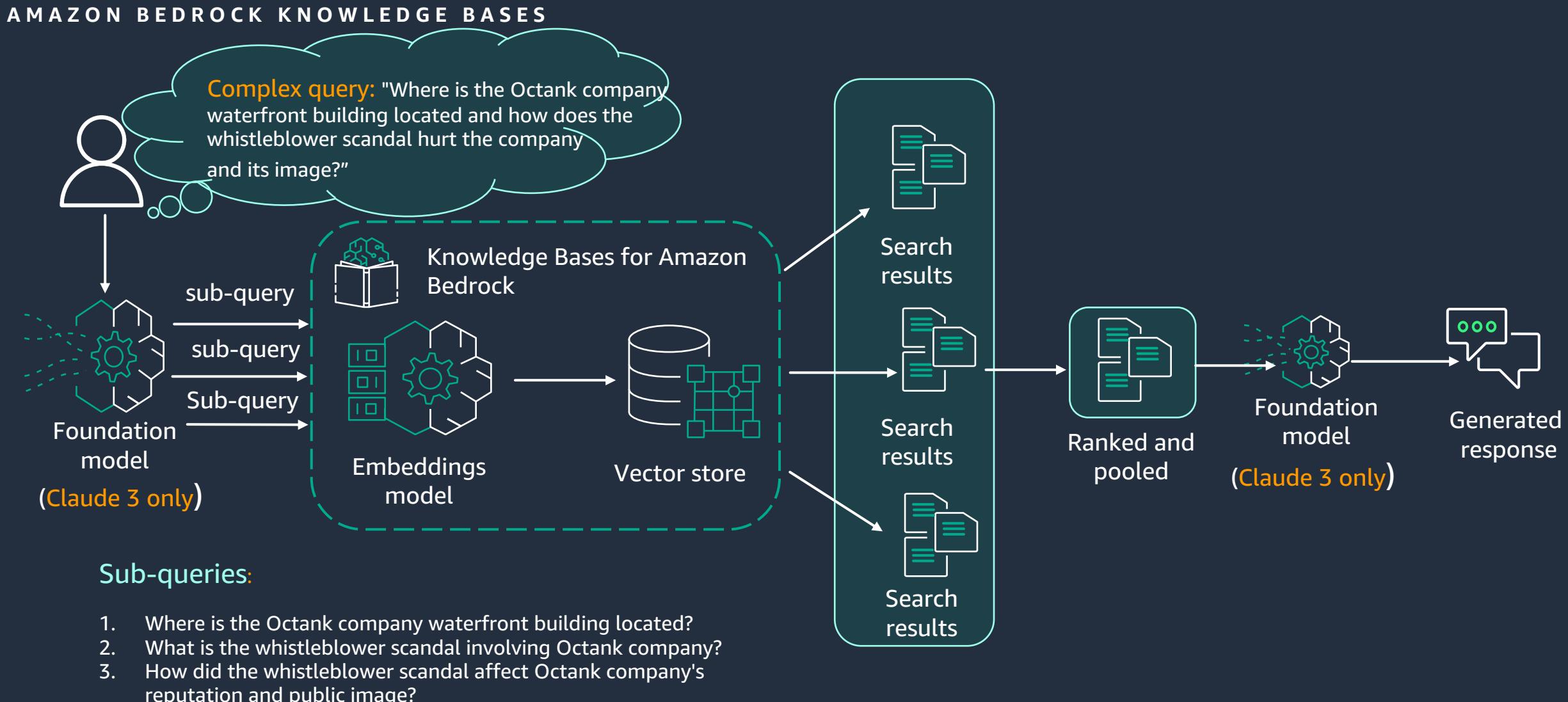
Combines relevancy scores from semantic and text search to provide greater accuracy.

Semantic search

Uses vector embeddings to deliver relevant results.



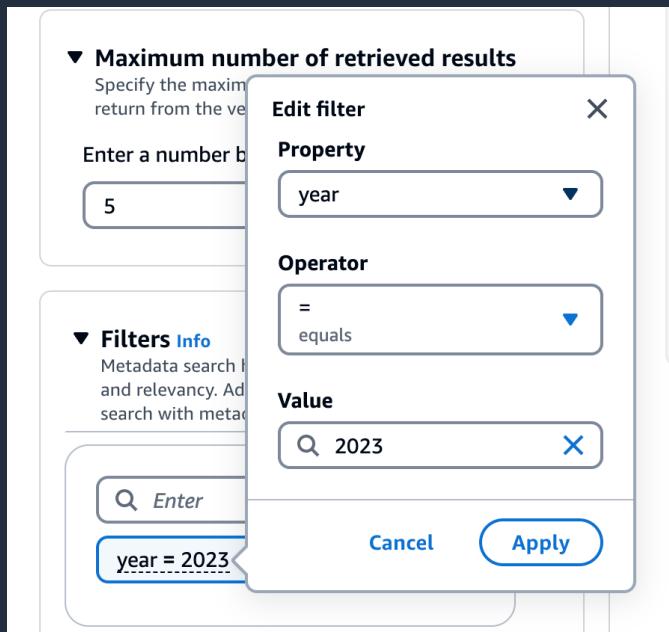
RAG API - Query Reformulation



Metadata Filtering

AMAZON BEDROCK KNOWLEDGE BASES

With Metadata Filtering, users can **narrow search results** by specifying which documents to include or exclude from a query, resulting in more relevant responses generated by the FM. In addition, we can use metadata filters to facilitate person ACL with Knowledge Base content.



The screenshot shows the results of a query. At the top, there's a toggle for 'Generate responses' and a user icon labeled 'games in space'. Below this is a sidebar titled 'Source details (5)' with a note: 'Each container below shows a chunk that comes from a data source and that relates to your query.' A section titled 'Query configurations (2)' follows. The main area displays a 'Source chunk 1' with three numbered items:

1. title,description,genres,year,publisher,score Galactic Odyssey: Starfarers,"Explore the vast universe in this open-world space exploration game. Discover new planets, trade with alien races, and build..."
2. title,description,genres,year,publisher,score Dinosaur Discovery: Lost World,"Explore a lost world filled with dinosaurs in this adventure game. Discover new species, study their behavior, and documen..."
3. title,description,genres,year,publisher,score Viking Saga: The Sea Raider,"Lead a Viking clan in this strategy game. Build longships, raid enemy settlements, and expand your

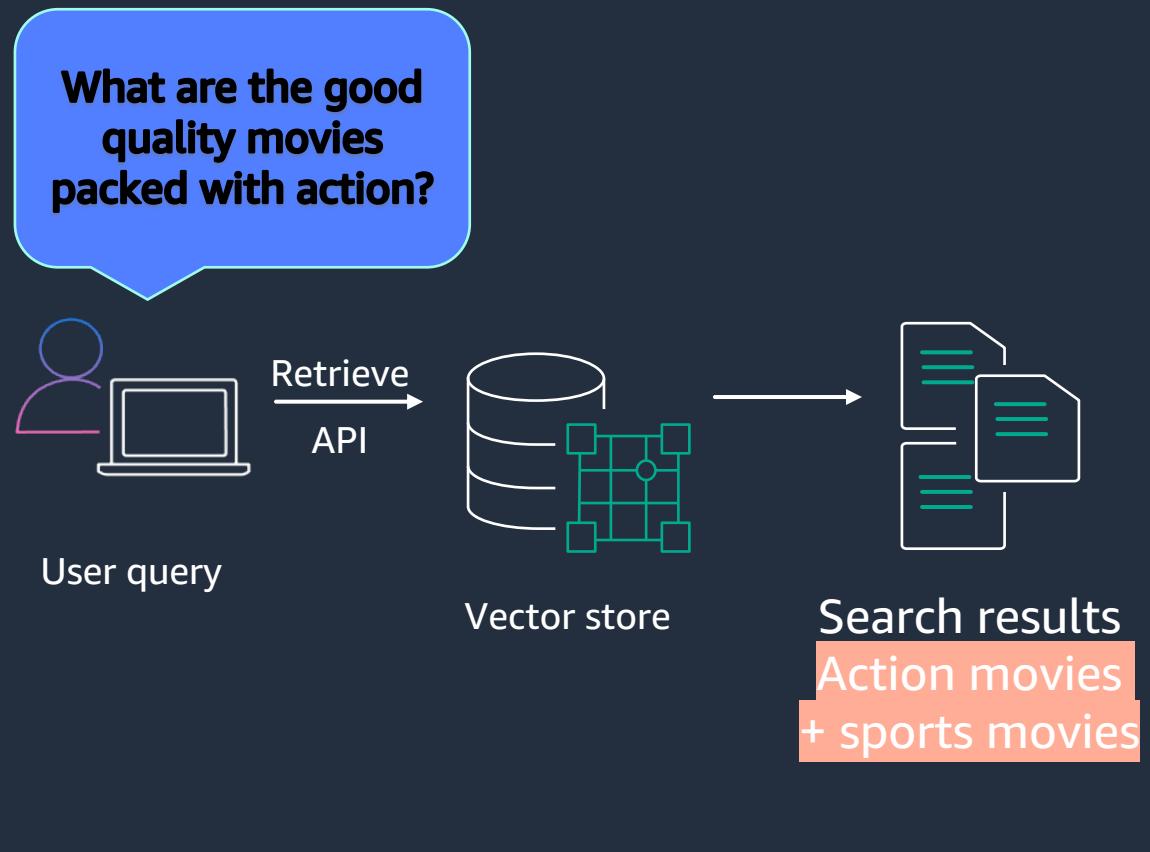
Below the chunk, a table titled 'Metadata associated with this chunk' lists key-value pairs:

Key	Value
x-amz-bedrock-kb-source-uri	s3://csp-video-g
score	8.5
publisher	Interstellar Game
year	2023
genres	Open World

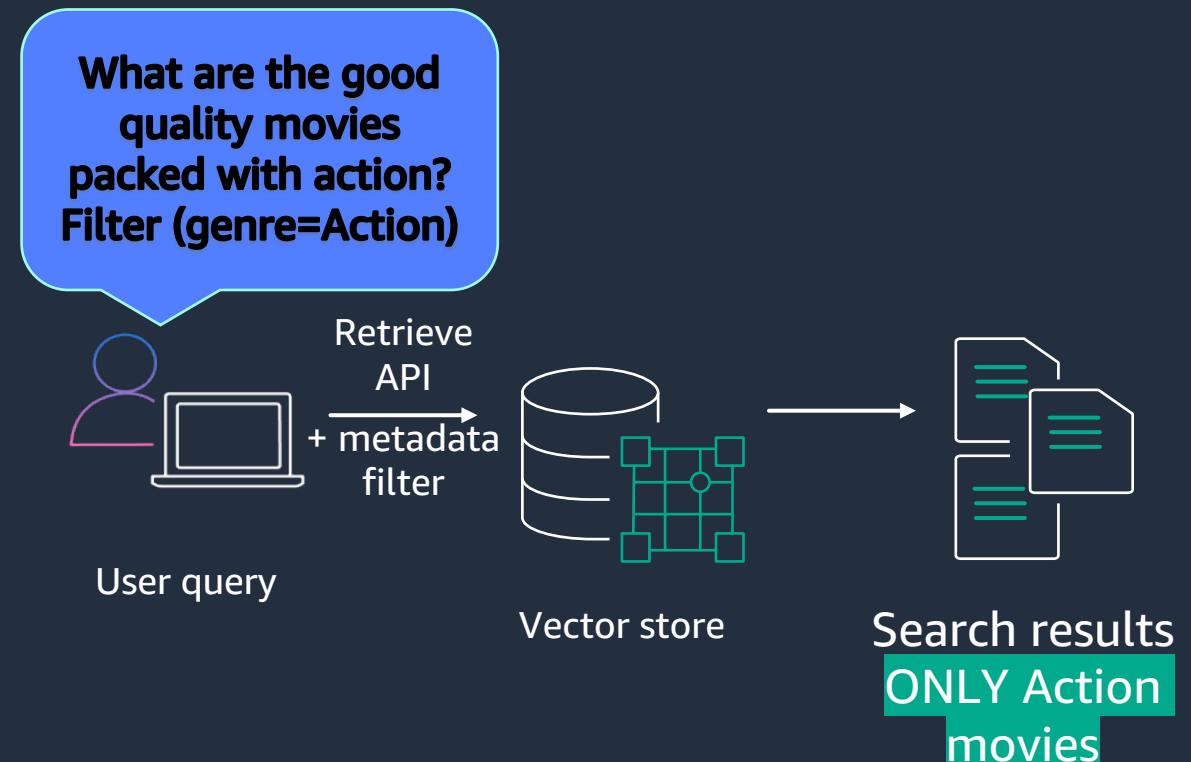
Metadata Filtering in Action

AMAZON BEDROCK KNOWLEDGE BASES

- Without metadata filter



- With metadata filter for genre=Action



GENERALLY AVAILABLE

Autogenerated filters

Helps with
“Give me the PDFs from last year”
More accuracy; no coding.



Automatic model generated filters to improve accuracy of document retrieval



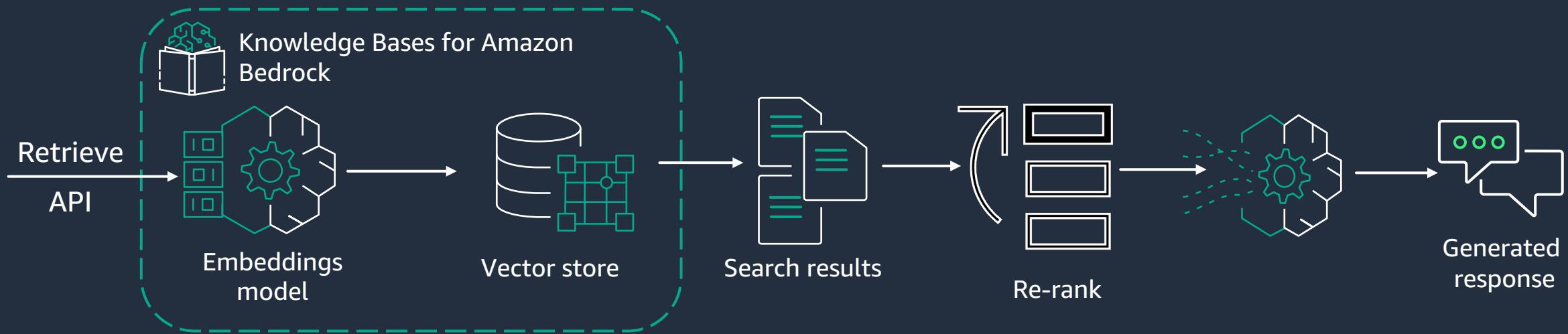
Narrow down search results without the need to manually construct complex filter expressions.



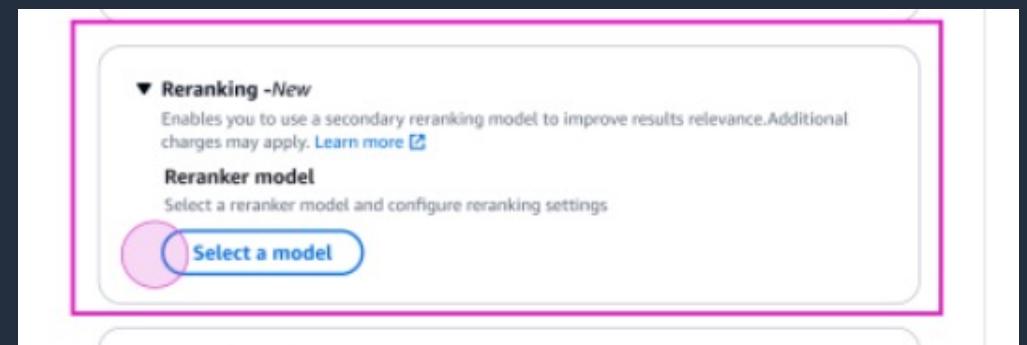
Uses document metadata schema to generate filters relevant to user query

Re-ranking

AMAZON BEDROCK KNOWLEDGE BASES



- Model selection – Amazon rerank, Cohere 3.5 rerank
- Accessible independently as well through Rerank API

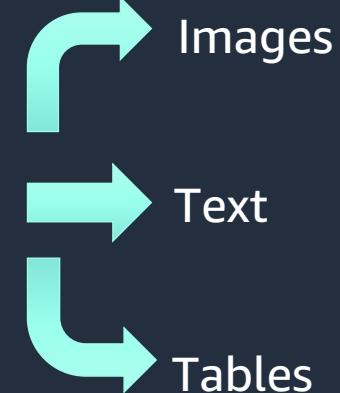


Advanced RAG techniques using Knowledge Bases

Multimodal RAG

AMAZON BEDROCK KNOWLEDGE BASES

Wildfire Statistics									
Wildfires are natural events, including lightning-caused fires, drought-induced human-caused fires, and escaped fires from prescribed burn projects. States are responsible for responding to wildfires that begin on non-federal land. Local governments are responsible for responding to wildfires that begin on federal land. The U.S. Forest Service, the U.S. Department of Agriculture, manages wildfire management and response across the 93 million acres of the National Forest System (NFS). The U.S. Department of the Interior (DOI) manages wildfire response for more than 400 million acres of national parks, wildlife refuges and preserves, and other public lands.									
Figure 1. Annual Wildfires and Acres Burned, 1993-2022									
Source: NICC Wildland Fire Summary and Status Note: DOI = U.S. Forest Service; DOI = Department of the Interior									
Wildfire statistics help inform past U.S. wildfire activity. Nationalwide data compiled by the National Interagency Coordination Center (NICC) indicate that the number of wildland fires and acres affected have increased over the last 30 years. The number of acres affected annually, while also variable, generally has increased (see Figure 1). Since 1993, the average number of fires per year has been around 61,000, while the average acres burned per year has been around 7.0 million acres. The acreage figure is more than double the average annual acreage burned in 1996 (3.3 million acres). The total number of fires reported annually in the 1990s (TEXT)									
From 2013 to 2022, there were an average of 61,410 wildfires annually and an average of 7.2 million acres impacted annually. In 2022, 68,998 wildfires burned 7.6 million acres. Over 40% of those acres were in Alaska (3.1 million acres).									
As of June 1, 2023, around 18,000 wildfires have impacted over \$11,000 acres this year.									
Figure 2. Top Five Years with Largest Wildfire Acreage Burned Since 1960									
Source: NICC Wildland Fire Summary and Statistics annual reports. Note: Number of fires in thousands.									
The number of fires and acreage burned are indicators of the annual level of wildfire activity. These numbers may also indicate the number of fires that occur in relatively undeveloped areas. Acreage burned also does not indicate the cause of the fire, such as the source of ignition (sparks from vehicles or tools, other ecological effects).									
Most wildfires are human-caused (89% of the average number of wildfires from 2018 to 2022). Wildfires caused by lightning tend to be slightly larger and to burn more acreage (55% of the average acreage burned from 2018 to 2022) than human-caused fires.									



Choose either
Amazon Bedrock Data
Automation or foundation
model for parsing



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Ingest documents
with images,
tables, and text



The screenshot shows the Amazon Bedrock interface with a query "Test: travel and insurance policy FAQs". The results page displays a preview of the query, source chunks, and details. The preview shows a question about population growth and a graph. Source chunks 1 and 2 are expanded, showing they are from the NICC Wildland Fire Summary and Statistics annual reports. Source chunk 1 includes a table of top five years with largest acreage burned:

Year	Acre Burned (millions)	Number of Fires
2015	10.13	61,410
2020	10.12	59,0
2017	10.03	71,5
2008	9.87	56,4
2005	9.33	67,2

The details panel shows the sources, metrics, and expanded source chunk 1. The expanded source chunk 1 includes the table above and notes about the data.

Query your
knowledge base



Get generated
response back with
images as source
attribution



Problem Statement

Example question: What was the top selling product in November 2024 and how does it compare to the previous month?

order_item_id	product_id	order_id	quantity	price
4319	4044	1437	1	89.99
4320	4015	1437	1	24.99
4321	4034	1437	1	19.99
...

User_id	username	first_name	Last_name
7001	alejandro_rosalez	Alejandro	Rosalez
7002	akua_mansa	Akua	Mansa
7003	anacarolina_silva	Ana Carolina	Silva
...

id	order_status	user_id	total_amount	ship_mthd	ts1
1437	Delivered	7002	214.96	EXP	2024-11-19 03:45:36
1438	.Delivered	7116	629.92	STD	2024-11-04 06:05:15
...

product_id	seller_id	name	price	stock_quantity
4001	501	Smartphone X	699.99	50
4002	501	Laptop Pro	1299.99	20
...

seller_id	Store_name	User_id	description
501	Tech Store	7001	A cutting-edge electronics store ..
502	Kitchen Store	7005	go-to destination for top-quality home appliances ..
...



Accessing structured data is no longer similarity search

Example question: What was the top selling product in November 2024 and how does it compare to the previous month?

product_id	seller_id	category_id	name	desc
4001	501	C1	Superclean3	Sonic brush
4002	502	C2	Best toothpaste	Fluoride toothpaste
...

To answer questions like

“What was the top selling product in November 2024

and how does it compare to the previous month?”

ID	Name
C1	Toothbrush
C2	oral hygiene
C3	Toothpaste
...	...

ID	Name	CategoryId	Desc
P1	Superclean3	C1	Sonic brush
			Best toothpaste
			Fluoride toothpaste

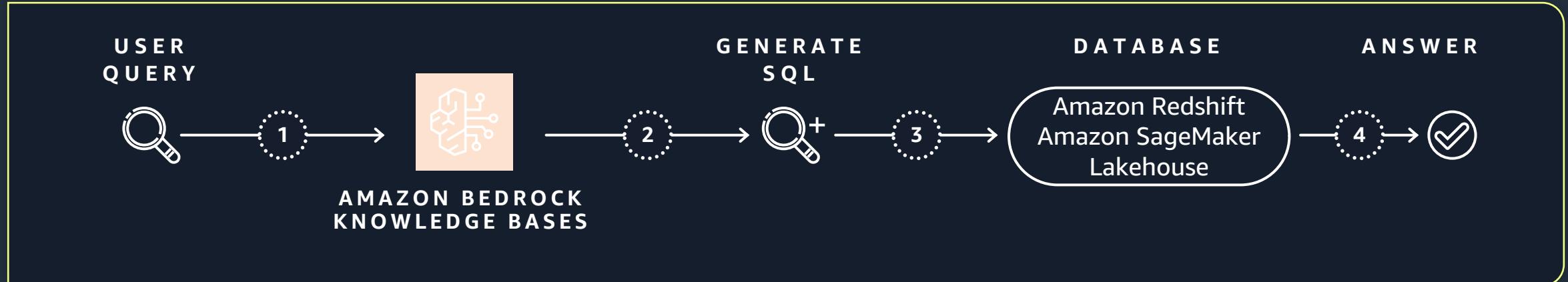
We need to translate the request into a (SQL) query

S3	Supermax CA	45 Elton St,
		123 Main St,
		555 Nassau St,



Structured RAG

INTRODUCING STRUCTURED DATA RETRIEVAL IN BEDROCK KNOWLEDGE BASES



“What was the top selling product in November?”

```
SELECT customer_id, SUM(order_total) AS total_revenue  
FROM customer_orders  
WHERE order_date BETWEEN '2024-01-01'  
AND '2024-01-31'  
GROUP BY customer_id ORDER BY total_revenue DESC;
```

A	B	C	D	E	F
1	CUSTOMER	REGION	ORDER_DATE	SALES	MONTH
2	Acme, Inc.	NORTH	1/15/2013	\$26,884	January
3	Widget Corp.	SOUTH	2/14/2013	\$46,174	February
4	123 Warehousing	EAST	3/13/2013	\$46,174	March
5	Demo Company	WEST	4/15/2013	\$49,049	April
6	Smith and Co.	NORTH	5/15/2013	\$80,969	May
7	Foo Bars	SOUTH	6/14/2013	\$55,522	June
8	Bar Widgets	EAST	7/13/2013	\$87,830	July
9	Fake Brothers	WEST	8/13/2013	\$64,463	August
10	Acme, Inc.	NORTH	9/12/2013	\$58,146	September
11	Widget Corp.	SOUTH	10/12/2013	\$83,288	October
12	123 Warehousing	EAST	11/12/2013	\$92,562	November
13	Demo Company	WEST	12/12/2013	\$64,750	December
14	Smith and Co.	NORTH	1/10/2014	\$55,588	January
15	Foo Bars	SOUTH	2/9/2014	\$14,331	February
16	Abc Company	EAST	3/8/2014	\$26,870	March
17	Bar Widgets	WEST	4/7/2014	\$88,458	April
18	Acme, Inc.	NORTH	5/10/2014	\$25,983	May
19	Widget Corp.	SOUTH	6/9/2014	\$68,797	June
20	123 Warehousing	EAST	7/9/2014	\$49,562	July
21	Demo Company	WEST	8/8/2014	\$15,454	August
22	Smith and Co.	NORTH	9/7/2014	\$23,798	September
23	Foo Bars	SOUTH	10/7/2014	\$16,843	October
24	Abc Company	EAST	11/6/2014	\$79,462	November
25	Bar Widgets	WEST	12/5/2014	\$80,780	December
26	Acme, Inc.	NORTH	1/5/2015	\$56,959	January
27	Widget Corp.	SOUTH	2/4/2015	\$47,180	February

“In November,
the top selling product is YYYY”

A	B	C	D	E	F
1	CUSTOMER	REGION	ORDER_DATE	SALES	MONTH
2	Acme, Inc.	NORTH	1/15/2013	\$26,884	January
3	Widget Corp.	SOUTH	2/14/2013	\$46,174	February
4	123 Warehousing	EAST	3/13/2013	\$46,174	March





Search

[Option+S]



United States (N. Virginia)

Admin/georajb-isengard @ 0360-0285-5796

Console Home Info

Reset to default layout

+ Add widgets

Recently visited Info

- IAM
- Athena
- AWS Glue
- S3
- Amazon Redshift
- AWS Lake Formation
- Amazon Bedrock
- Support

[View all services](#)

- Secrets Manager
- CloudWatch
- Amazon Q Business
- IAM Identity Center
- Amazon Q
- Service Quotas

Applications (0) Info

Region: US East (N. Virginia)

us-east-1 (Current Region) ▾

 Find applications[Create application](#)

Name	Description	Region	Originati.	Star	▲
------	-------------	--------	------------	------	---

No applications

Get started by creating an application.

[Create application](#)[Go to myApplications](#)

Welcome to AWS



Getting started with AWS

Learn the fundamentals and find valuable information to get the most out of AWS.



Training and certification

Learn from AWS experts and advance your skills and knowledge.

AWS Health Info

Open issues

0 Past 7 days

Scheduled changes

1 Upcoming and past 7 days

Other notifications

0 Past 7 days

Solutions (17) Info

Vetted Solutions from AWS for popular business and technical use cases.



Generative AI Application Builder

Develop and deploy generative AI apps

[Learn more](#) Deploy (10 mins)

Intelligent Document Processing

Extract text with OCR

[Learn more](#)

Cross Region Failover & Graceful Fallback

Live Streaming

Deliver a reliable livestreaming experience

[Learn more](#) Deploy (5 mins)

Custom Search with Enterprise Knowledge Base

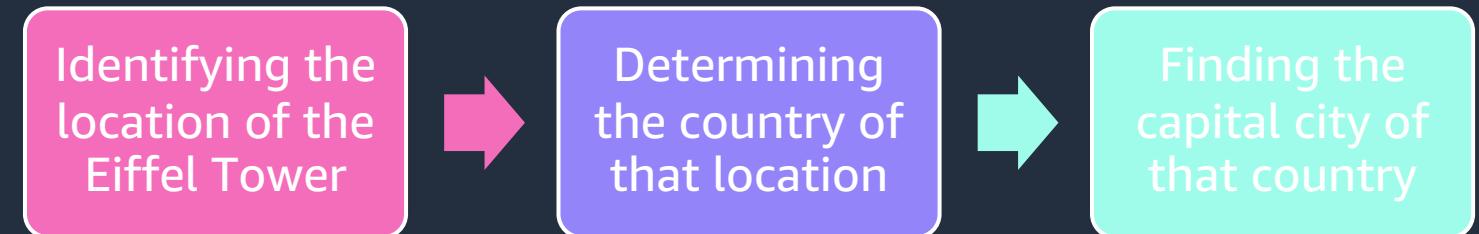
Query an enterprise knowledge base

[Learn more](#)

Problem Statement

What is the name of the capital city of the country where Eiffel Tower is located?

Requires Multi-step reasoning



Standard RAG

It doesn't inherently understand relationships between entities (Eiffel Tower, Paris, France)

Basic RAG might find information about the Eiffel Tower or about France's capital separately, but may not connect these pieces of information effectively.

GraphRAG

It understands the relationships: Eiffel Tower is in Paris, Paris is the capital of France

GraphRAG can establish connections between multiple entities: Eiffel Tower, Paris, France

Recognizes the hierarchy: Landmark -> City -> Country -> Capital City

Amazon Bedrock Knowledge Bases

GraphRAG

PREVIEW

Generate more relevant responses for RAG applications using knowledge graphs



Generate knowledge graphs to link relationships across data sources



Build more comprehensive generative AI applications



Enhance transparency of source information for better fact verification

Amazon Bedrock Agents



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

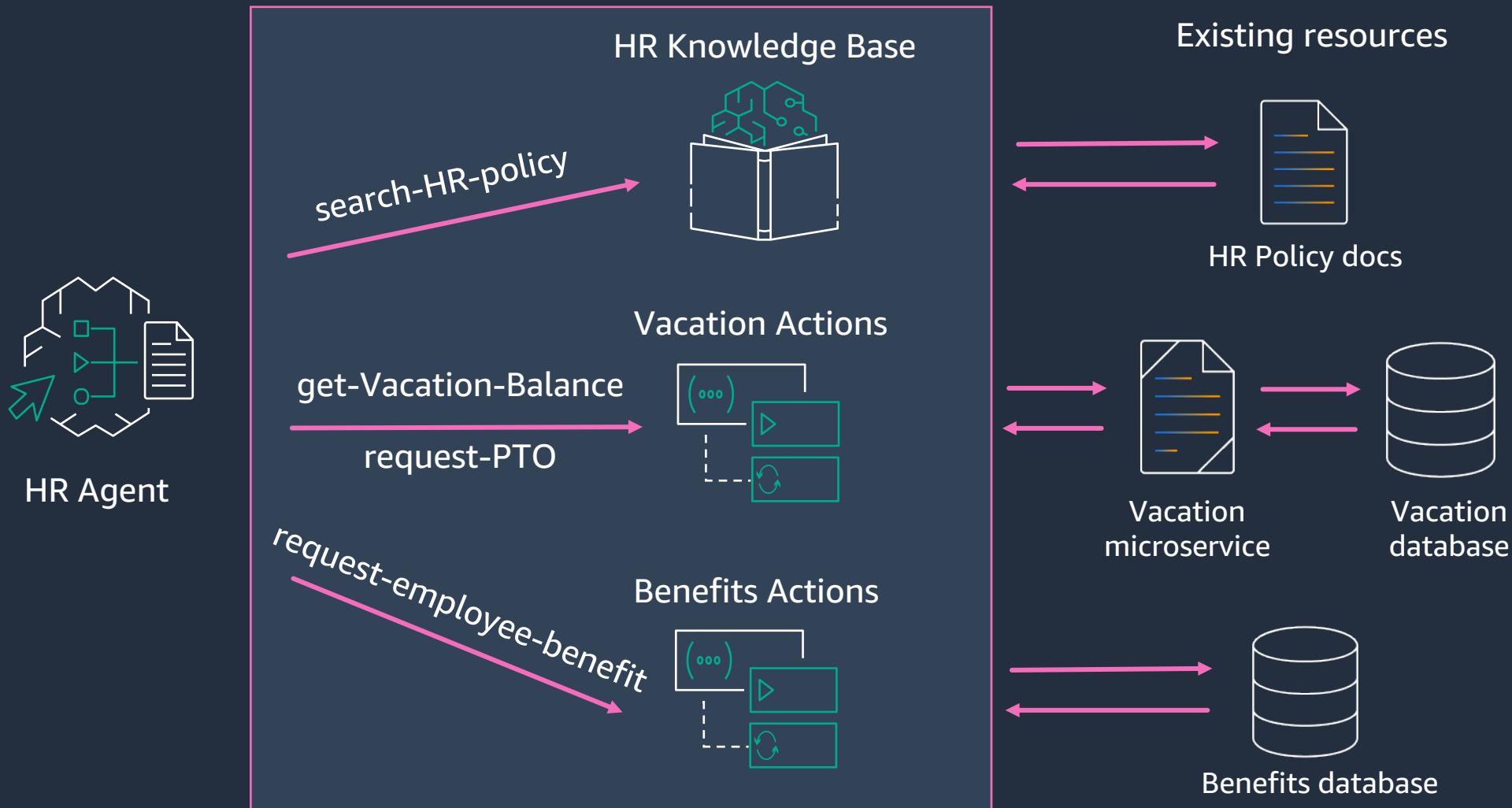
What is an AI agent?

An application powered by a large language model (LLM) with a predefined system prompt

Has a set of **tools**, which are self-contained functions designed to perform a specific task

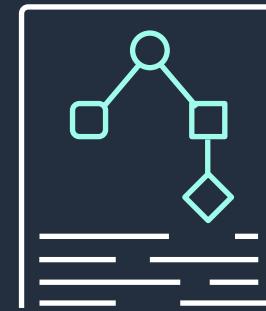


Agents build on existing enterprise resources



Typical agentic workflow

Enable generative AI applications to execute multi-step tasks using company systems and data sources



Decompose into steps using available actions and Knowledge Bases

Execute action or search knowledge base

Until final answer

Observe results

Think about next step

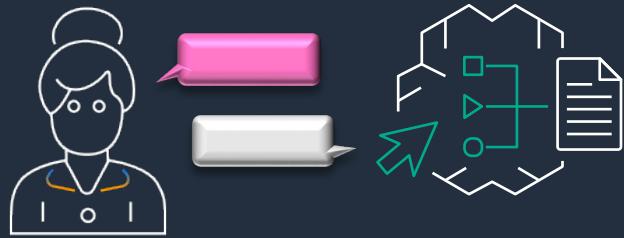
ReAct: Reason & Act

Amazon Bedrock Agents

Intelligent workflow automation



Amazon Bedrock Agents features



Agent Building Blocks



Session handling



Memory management



Runtime configurations



Action groups



Orchestration strategies



Inline agents



Multi-agent collaboration



Code interpretation

Developer Tools



Deployment



Logs, tracing

Integrations



Knowledge bases



Guardrails



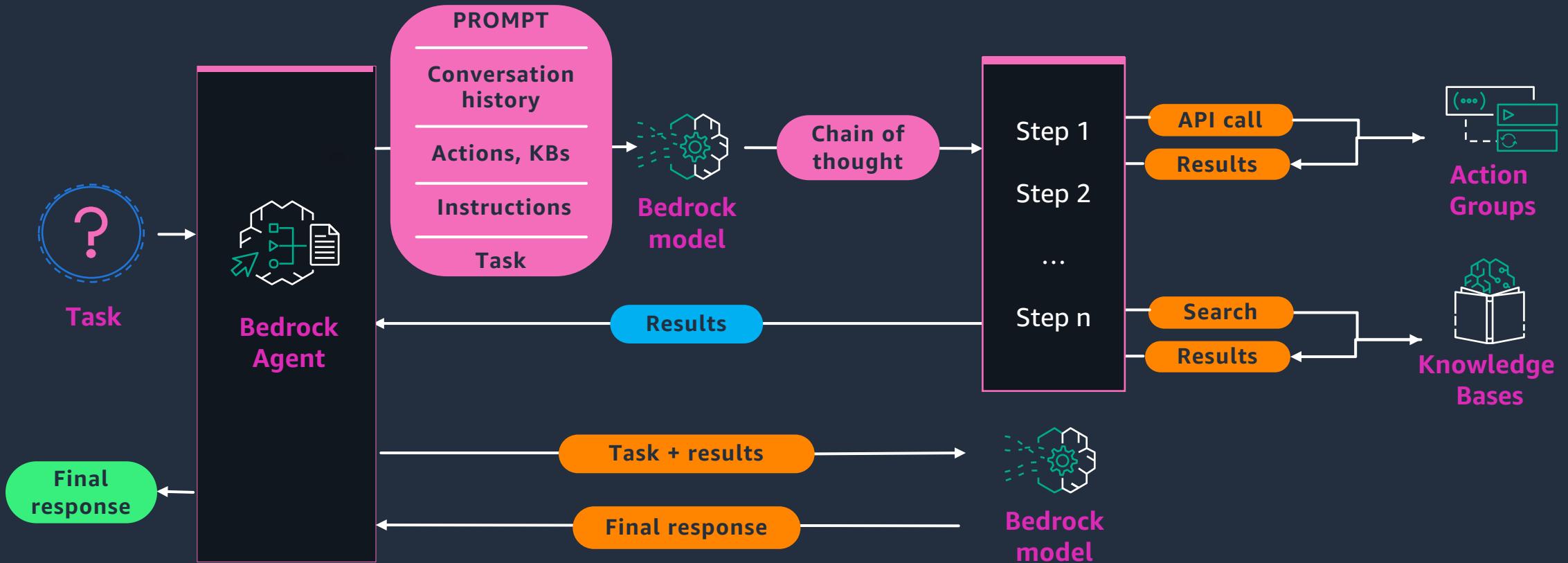
Flows



Secure



Agent orchestration – Detailed flow



Agent breaks task into subtasks, determines the right sequence, and executes actions and knowledge searches on the fly

Agent orchestration – An example

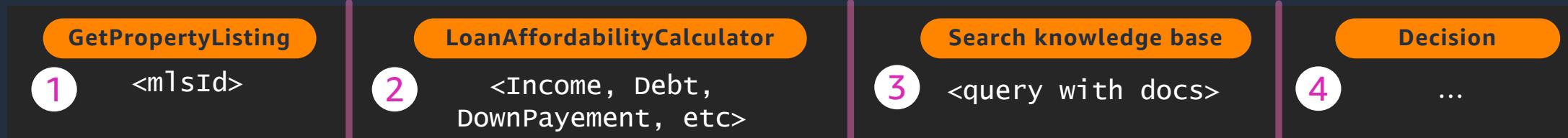


Task: I make 70K per year. Can I afford to buy the property listed under MLS-XYZ. Do I qualify for any special programs for Fannie Mae backed loans?

Thought: To answer this question, I will:

1. Get Property listed at the MLS ID
2. Calculate maximum affordable loan (MAL) based on income/debt ratio
3. Search Knowledge base for criteria for special programs
4. Use all the information to determine if the user can afford the loan and if they qualify for any special programs

Action:



Observation:

<property details>

...

<success>

Thought::

I need to calculate user's maximum affordable loan

...

I know the final answer

Final Answer: Based on the information provided: - The maximum loan amount you can afford is \$XXX according to the loan affordability calculator. However, the property price is \$YYY which is more than the maximum you can afford. - Fannie Mae offers the HomeReady program which provides special financing options for first time home buyers such as lower mortgage insurance requirements and waived pricing adjustments. There are no first time home buyer requirements for HomeReady purchase transactions. You may want to explore this program further to see if it could help you qualify for this property.

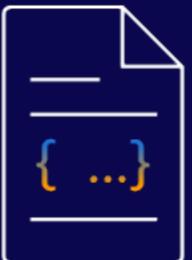
Action Group example

UtilityActionGroup

Description: “This action group provides a set of commonly used actions. Use these actions for things like sending emails and getting team member lists.”

API Schema

```
{ "openapi": "3.0.0",
  "info": { "title": "Utility Actions",
            "description": "... },
  "paths": {
    "/sendEmail": {
      "post": {
        "description": "This operation ...",
        "operationId": "sendEmail",
        "requestBody": { ... },
        "responses": {
          "200": {
            "text/plain": { ... }
          }
        }
      }
    },
    "/getTeam": { ... }
  }
}
```



Lambda Function

```
def lambda_handler(event, context):
    if event['apiPath'] == '/sendEmail':
        result = sendEmail(event)
    elif event['apiPath'] == '/getTeam':
        result = getTeam(event)

    def do_action1()
    def do_action2()
    response_body = {'application/json':
                     {'body': result}}

    action_response = { ...
                       'responseBody': response_body}
    return { 'messageVersion': '1.0',
             'response': action_response}
```

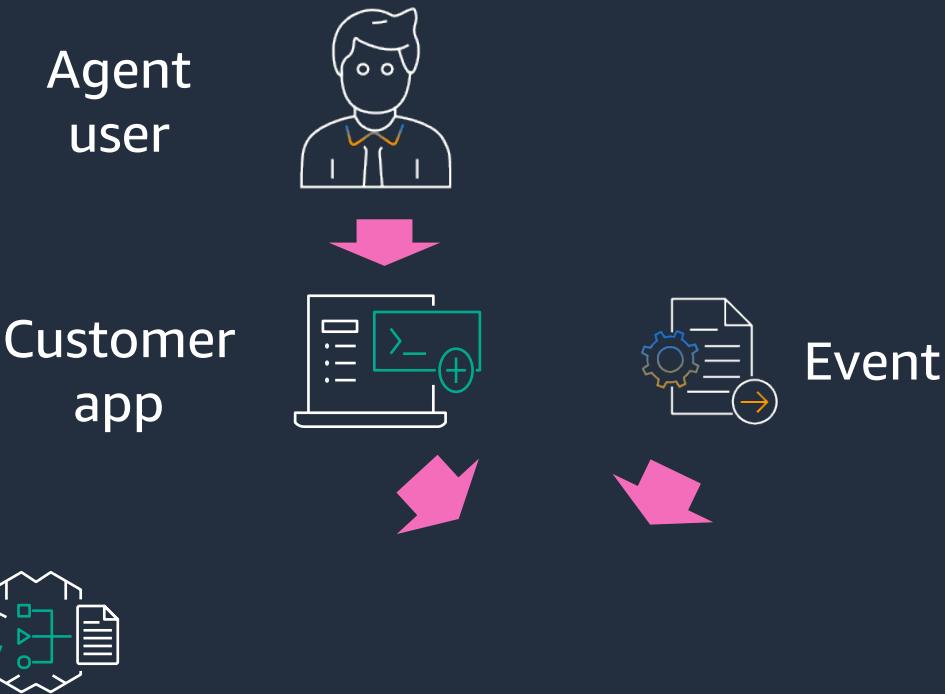


Agents can be deployed and invoked from any app

Building and testing agents



Using production agents



To **deploy** an agent, you create a new **Alias**,
and optionally a new **Version**

Invoking an Agent via the SDK

Invoke agent

```
response = client.invoke_agent(  
    inputText='<user request>',  
    agentId=agent_id,  
    agentAliasId=agent_alias_id,  
    sessionId=str(uuid.uuid1()),  
    enableTrace=True )
```

Process the response stream

```
for event in response['completion']:  
    if 'chunk' in event:  
        data = event['chunk']['bytes']  
        answer = data.decode('utf8')  
        print(f"Answer:\n{answer}")  
    elif 'trace' in event:  
        print(json.dumps(event['trace'],  
                         indent=2))
```

Full set of APIs for building, deploying, and using

Agents



Data plane

- Create / Update / Delete / Prepare Agent
- Associate / Update / Disassociate Agent Knowledge Base
- Create / Update / Delete Agent Action Group
- Create / Update / Delete Agent Alias

Control plane

- Invoke Agent

Knowledge Bases



- Create / Update / Delete Knowledge Base
- Start Ingestion Job
- Create / Update / Delete Data Source

- Retrieve
- Retrieve and Generate

Prompt editor Agents

▼ Advanced prompts

Advanced prompts are a set of system-generated combination of the instruction to the model, action groups, and knowledge bases. You can edit them to override and enhance existing configurations.

[Learn more about formatting advanced prompts](#)

Select parser Lambda function - *optional*
Lambda will parse and update the LLM output so that the API can use it as input.
Select a previously created Lambda function or visit [AWS Lambda](#) to create a new function.

Function alias: \$Latest

Select a Lambda function:

Action buttons: Start, Orchestrate, Retrieve, Stop - disabled

Orchestration prompt Info
This prompt defines the order in which actions are executed.

Human: You are a helpful assistant for general questions about amazon benefits
Given a set of APIs, Conversation History, Attributes and User Input, create an orchestration plan for executing the set of apis in order to fulfill user input.
Valid "api" values are
,x_amz_knowledgebase_T4N6YQKJJ::Search or "Cannot Answer" or "Final Answer".
DO NOT return an api if all required parameter values are not present.
DO NOT replace the placeholders in the api_name with api_inputs.
Valid "verb" is HTTP verb used in "APIs" e.g. GET, PUT etc
Valid "api_input" as json from "User Input", "Observation" or "Conversation History" or "Attributes".
NEVER assume value for any parameter, mark the value as "null" if not available.
DO NOT ask information which is already collected previously and is available in "Conversation History".
DO NOT ask information if provided in the "Attributes".
- For each parameter in "api_input" extract word-for-word quotes from "User Input", "Conversation History", "Attributes" or "Observation" in "source" from where the value was mapped. If there is no relevant quote to extract, mark value and source as "null".
- ALWAYS convert the "api_input" to format mentioned for the parameter in the API . e.g. if a parameter is number type, convert it to number representation.
DO NOT go into a loop and return exact same apis with exact same api_input as previous observation
Provide only ONE action per \$JSON_BLOB, as shown below and always stick to the below format:

{
"api": "\$API_NAME",
"verb": "\$HTTP_VERB",
"api_input": { \$PARAMETER: { "value": "\$INPUT", "source": "

Code block:

```
{  
"api": "$API_NAME",  
"verb": "$HTTP_VERB",  
"api_input": { $PARAMETER: { "value": "$INPUT", "source": "
```



ReAct

FMs can reason and figure out how to solve user-requested tasks with a reasoning technique called *ReAct*



Chain of thought

The thought is a reasoning step that helps demonstrate to the FM how to tackle the problem and identify an action to take. The action is an API that the model can invoke from an allowed set of APIs. The observation is the result of carrying out the action.

Chain of thought (CoT) trace

The screenshot shows the Amazon Bedrock interface. On the left, the navigation menu includes sections like Getting started, Foundation models, Playgrounds, Orchestration, Model access, and Settings. The Agents section is currently selected. In the center, a user message "Can you help send a reminder to all pending claims" is shown, followed by a bot response: "To proceed with claim-123, please provide the Accident Report, Driver's License and Vehicle Registration. A reminder has been sent for the pending documents." Below this is a "Hide trace" button. On the right, the "Trace (5)" panel displays the chain of thought. It starts with "Step 1" and shows the JSON trace data. The trace data includes rationale, invocation input, and observation output. A purple annotation "Thought: I need a list of claims" points to the rationale text. Another annotation "Actions: API Claims GET" points to the invocation input. A third annotation "Observation: List of claims" points to the observation output.

Amazon Bedrock X Test AgentTestAlias: DRAFT

Getting started

Overview

Examples

Foundation models

Base models

Custom models

Providers

Playgrounds

Chat

Text

Image

Orchestration

Knowledge base [Preview](#)

Agents [Preview](#)

Deployment

Provisioned throughput

Model access

Settings

User guide [🔗](#)

Bedrock Service Terms [🔗](#)

Enter your message here

Trace (5)
Request ID: a45685fa-c208-45ac-b929-8b54fd0f352c [Copy all](#)

Step 1 Success

Trace

```
1: {
2:   "rationaleTrace": {
3:     "text": "Based on \\\"User Input\\\" I need to send reminders for all open claims. I will first get list of all open claims using ClaimManagementActionGroup::getAllOpenClaims API",
4:     "traceId": "7ccacf6f-5e35-423b-a29d-57b612b783eb-0"
5:   },
6:   "invocationInputTrace": {
7:     "actionGroupInvocationInput": {
8:       "actionGroupName": "ClaimManagementActionGroup",
9:       "apiPath": "/claims",
10:      "verb": "GET"
11:    },
12:    "invocationType": "ACTION_GROUP",
13:    "traceId": "7ccacf6f-5e35-423b-a29d-57b612b783eb-0"
14:  },
15:  "observationTrace": {
16:    "actionGroupInvocationOutput": {
17:      "text": "[{'response': [{'claimId': 'claim-123', 'policyHolderId': 'A134085', 'claimStatus': 'Open'}, {'claimId': 'claim-06', 'policyHolderId': 'A645987', 'claimStatus': 'Open'}]}]"
18:    },
19:    "invocationType": "ACTION_GROUP",
20:    "traceId": "7ccacf6f-5e35-423b-a29d-57b612b783eb-0"
21:  }
```

Thought: I need a list of claims

Actions: API Claims GET

Observation: List of claims

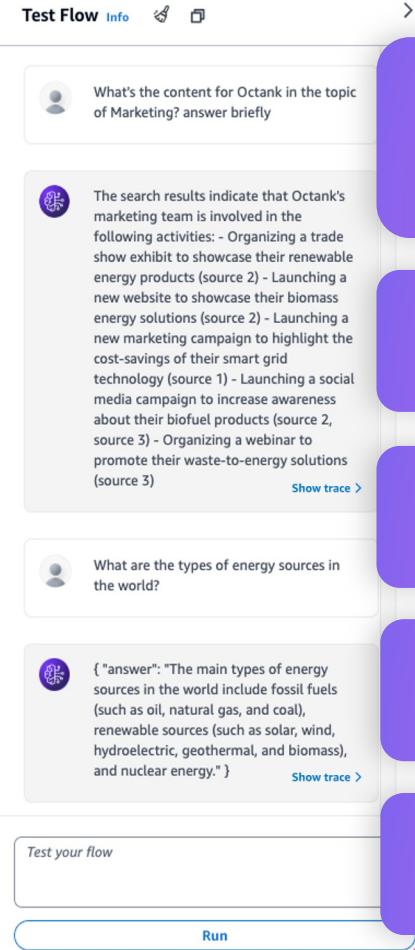
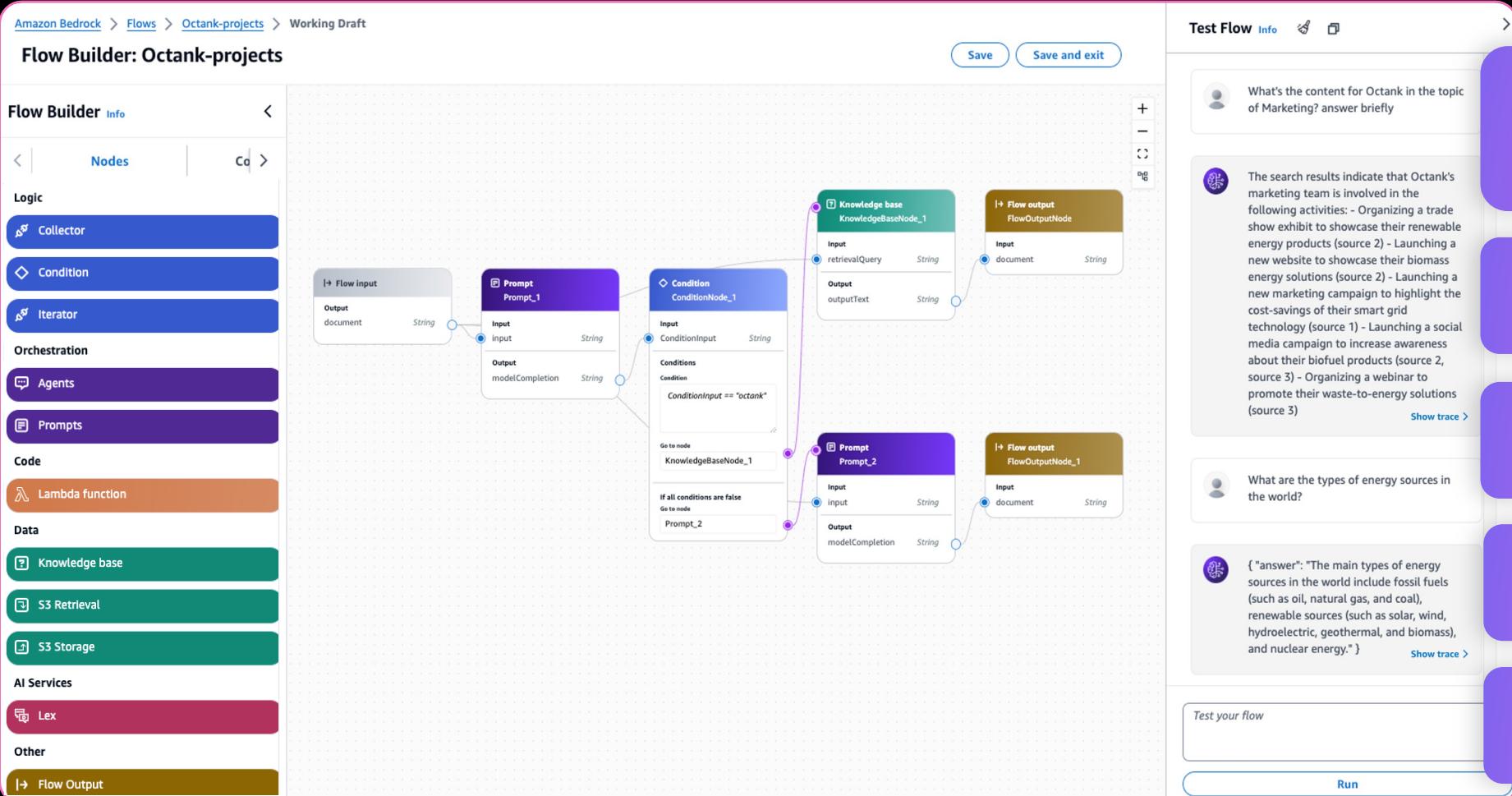
Amazon Bedrock Flows



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Flows

Build and scale controllable generative AI workflows



Fast authoring with visual builder & native Bedrock components

Easy testing and validation with trace

Fast iteration with version and alias

Easy to automate with SDK/APIs

Easy to scale with serverless infrastructure



Flows Use Cases

Accelerates authoring & execution of user-defined generative AI workflows



Rapid prompt application

Quick way to link, test and deploy prompts with other services

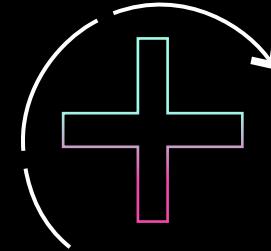
Common instances - summarizations, code/test generation



SOP automation

Automate user-defined business process

Common instances: ticket triage, price setting, content generation

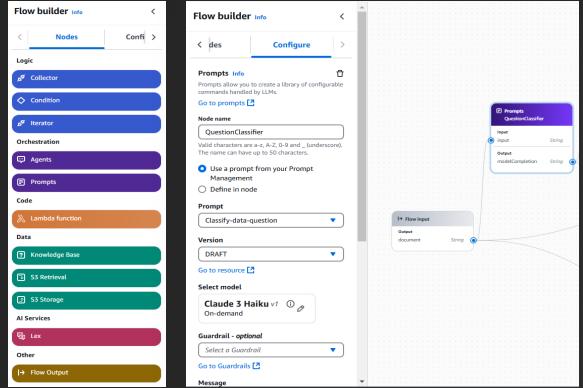


Flexible RAG application

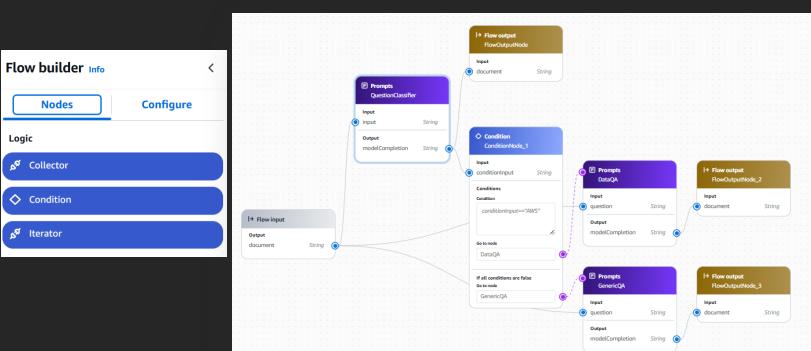
Link KB or data sources with prompts/FMs to generate contextual content

Common instances – customer support Q&A, insurance claim assistant

Bedrock Flows

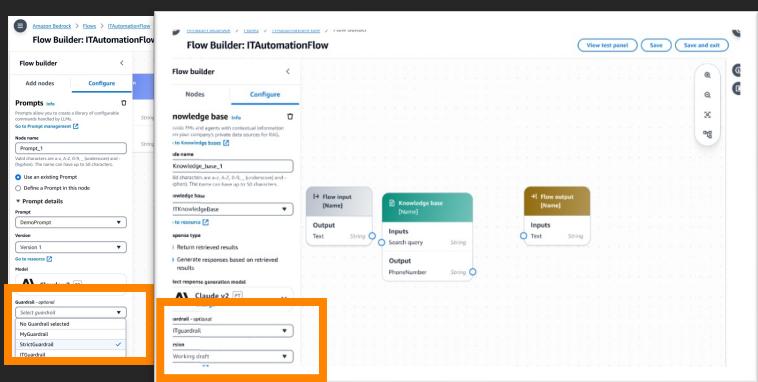


Visual Builder - Fast authoring and configuration

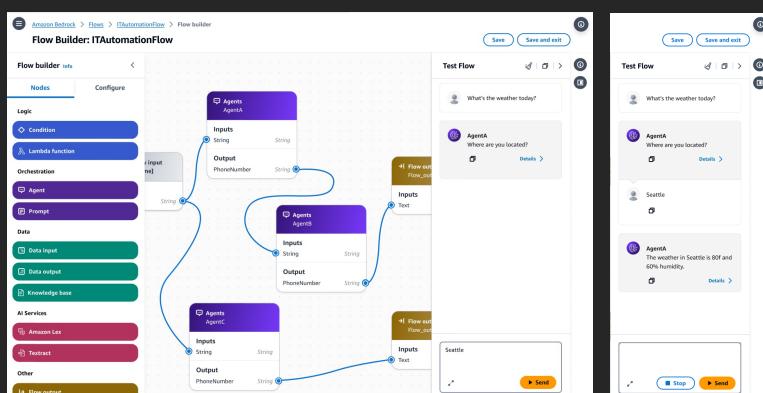


Fully controllable workflow with logic operators

Easy to test & validate with traceability



Content safety with Guardrails



Multi-turn conversation for Agents

Enterprise-scale-ready w/ serverless infra



Bedrock Flows: accessing data in a flow

Expression	Meaning	Example	Example result
<code>\$.data</code>	The entire input.	<code>\$.data</code>	The entire object
<code>.name</code>	The value for a field called <code>name</code> in a JSON object.	<code>\$.data.numbers</code>	<code>[1, 2, 3, 5, 8]</code>
<code>[int]</code>	The member at the index specified by int in an array.	<code>\$.data.animals.reptiles[2]</code>	<code>iguana</code>
<code>[int1, int2, ...]</code>	The members at the indices specified by each int in an array.	<code>\$.data.numbers[0, 3]</code>	<code>[1, 5]</code>
<code>[int1:int2]</code>	An array consisting of the items at the indices between int1 (inclusive) and int2 (exclusive) in an array. Omitting int1 or int2 is equivalent to marking the beginning or end of the array.	<code>\$.data.organisms.mammals[1:]</code>	<code>["horse", "mouse"]</code>
<code>*</code>	A wildcard that can be used in place of a name or int. If there are multiple results, the results are returned in an array.	<code>\$.data.*.mammals</code>	<code>[["cat", "dog"], ["rabbit", "horse", "mouse"]]</code>

Example data object

```
{
  "animals": {
    "mammals": ["cat", "dog"],
    "reptiles": ["snake", "turtle", "iguana"]
  },
  "organisms": {
    "mammals": ["rabbit", "horse", "mouse"],
    "flowers": ["lily", "daisy"]
  },
  "numbers": [1, 2, 3, 5, 8]
}
```

Bedrock Flows pricing



Key Concepts

- Bedrock Flows **counts a node transition** each time a node in your workflow is executed.
- A node **may incur its own cost** depending on the node type. For example, if the flow invokes Bedrock Guardrail policy, you will be billed for the number of text units processed by the policy.

Bedrock Flows

\$0.035

per 1000 node transitions



Example

Situation – An app developer creates a flow to automate news summarization for traders. The flow includes takes in an S3 location, retrieves 10 files that include articles from 10 major news agency, summarizes each file iteratively, and writes the results to S3. They run this flow every half hour of 21 working days in a month.

Node transitions in the flow = 2 (Input, S3 retrieval) + 10 * 2 (Iterator, Prompt) + 3 (Collector, S3 storage, Output) = 25

Number of flow execution per month = $2 * 24 \text{ hours} * 21 \text{ days} = 1008$

Total monthly price = $25 * 1008 * \$0.035/1000 = \0.882

*Note: There will be additional charges for AWS services used in the workflow execution: 1/ Amazon S3 usage in retrieval and storage nodes, and 2/ LLM usage in the prompt node.

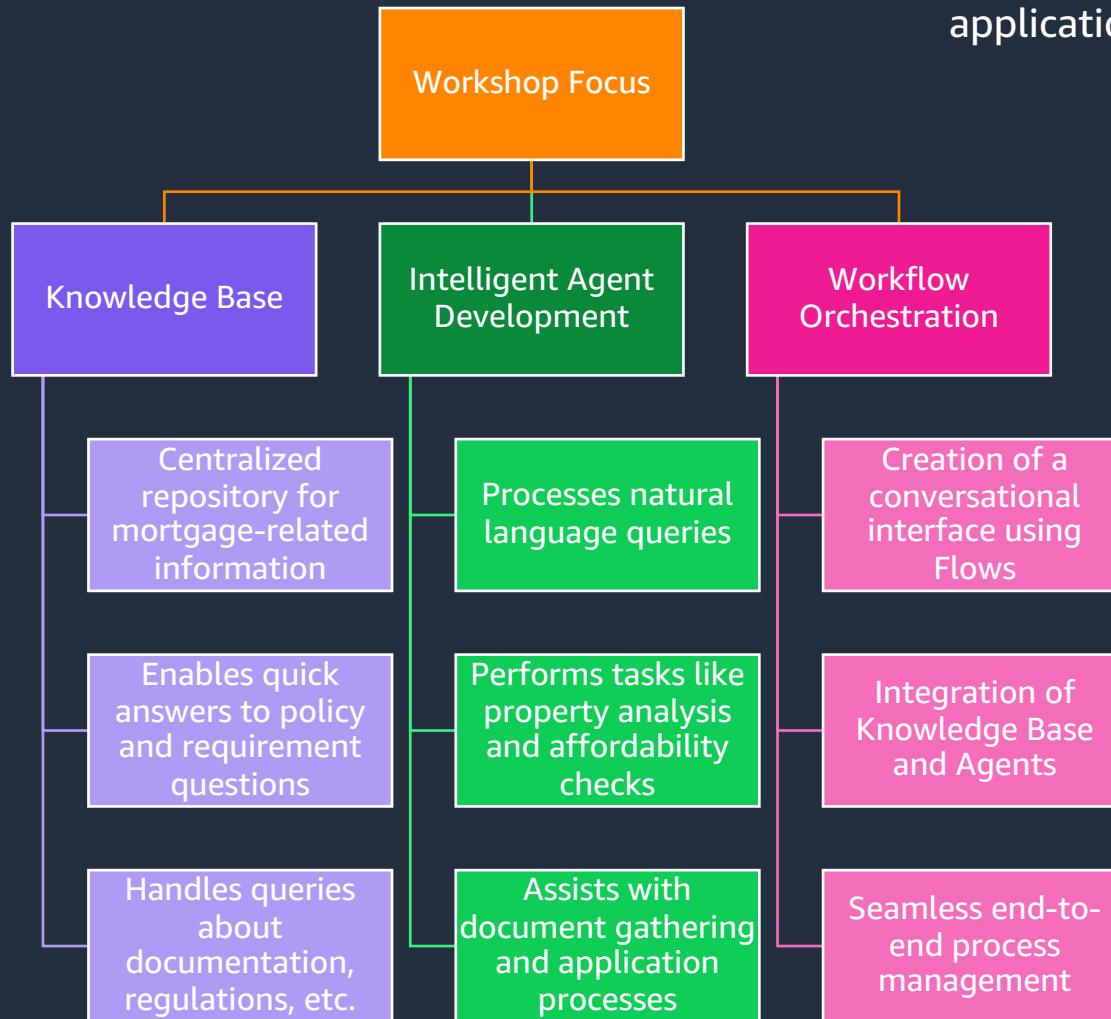
Hands-on workshop



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What we will be doing

Create an AI-powered solution to simplify the mortgage application process using Amazon Bedrock tools



Accessing your Free AWS accounts

1. Free AWS accounts will be available for up to 90 people
2. Accounts will be active for 24 hours
3. Go to <https://catalog.us-east-1.prod.workshops.aws/join?access-code=1fdf-070b2f-01>
4. Agree on the Terms and Conditions → Join event
5. Access your free AWS account console
6. Open the step-by-step workshop guide

Log out from any other AWS account before joining this event!

Select Email OTP

1

Workshop Studio > Sign in

Sign in
Choose a preferred sign-in method

Email one-time password (OTP)

Enter your personal or corporate email to receive a one-time password

Login with Amazon

Login with your Amazon.com retail account

Amazon employee

Login with your Amazon Corporate account. Only for Amazon Employees.

Enter your work email to receive a 9-digit pass

2

One-time email passcode

Send a passcode to the email below.

Email

Back Send passcode

[Get help signing in](#)

Go to your email and copy the code here

3

One-time email passcode

We sent a passcode to bbonik@gmail.com. You should receive it within 5 minutes.

Passcode (6-digit) [Resend passcode](#)

Back Sign in

[Get help signing in](#)

Terms & Conditions

The screenshot shows the 'Review and join' step of the AWS Workshop Studio process. At the top, there's a navigation bar with the AWS logo and 'workshop studio'. Below it, a sidebar indicates 'Step 1: Enter event access code' and 'Step 2: Review and join'. The main content area has a title 'Review and join' and a section titled 'Event details' containing information about the event: Name (Rekognition Workshop Test Event), Start time (11/17/2022 02:52 PM), Duration (3 hours), and Level (-). Below this is a 'Description' section with the text 'Test event'. A 'Terms and Conditions' section follows, with a note 'Read and accept before joining the event'. It contains a numbered list of 4 terms and conditions. At the bottom of this section is a checkbox labeled 'I agree with the Terms and Conditions'. The 'Join event' button at the very bottom right is highlighted with a blue border and a blue arrow pointing to it from the left.

aws workshop studio

Step 1
Enter event access code

Step 2
Review and join

Review and join

Event details

Name	Start time	Duration	Level
Rekognition Workshop Test Event	11/17/2022 02:52 PM	3 hours	-

Description
Test event

Terms and Conditions

Read and accept before joining the event

1. By using AWS Workshop Studio for the relevant event, you agree to the AWS Event Terms and Conditions and the AWS Acceptable Use Policy. You acknowledge and agree that are using an AWS-owned account that you can only access for the duration of the relevant event. If you find residual resources or materials in the AWS-owned account, you will make us aware and cease use of the account. AWS reserves the right to terminate the account and delete the contents at any time.
2. You will not: (a) process or run any operation on any data other than test data sets or lab-approved materials by AWS, and (b) copy, import, export or otherwise create derivative works of materials provided by AWS, including but not limited to, data sets.
3. AWS is under no obligation to enable the transmission of your materials through Event Engine and may, in its discretion, edit, block, refuse to post, or remove your materials at any time.
4. Your use of AWS Workshop Studio will comply with these terms and all applicable laws, and your access to AWS Workshop Studio will immediately and automatically terminate if you do not comply with any of these terms or conditions.

I agree with the Terms and Conditions

Cancel Previous Join event

Access your AWS account & the Lab Guide

The screenshot shows the AWS Workshop Studio interface. On the left, a sidebar for the event "Build Chatbots with Amazon Bedrock Knowledge Bases and Agents" lists various options, with "AWS account access" expanded to show "Open AWS console (us-west-2)" and "Get AWS CLI credentials". A pink arrow points from the text "Access the free AWS account" to the "Open AWS console" button. On the right, the main dashboard displays "Event information" (Start time, Duration, Accessible regions) and a "Workshop" section with details: Title - Amazon Bedrock Workshop - Moderated, Complexity level - 300, AWS services - Amazon Bedrock, Topics - Machine Learning (ML/AI). A pink arrow points from the text "Self-service Lab Guide" to the "Get started" button in the workshop section.

aws workshop studio

vonikavv ▾

Build Chatbots with Amazon Bedrock Knowledge Bases and Agents

Event dashboard > Amazon Bedrock Workshop - Moderated

Build Chatbots with Amazon Bedrock Knowledge Bases, Agents, Prompt flows and Guardrails

Event information

Start time Duration Accessible regions

Description

Self-service Lab Guide

Get started >

Workshop

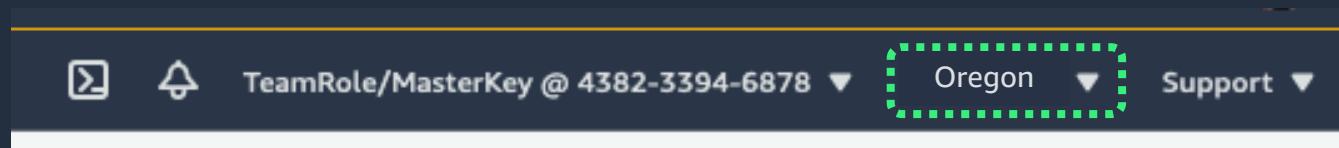
Title	Complexity level	AWS services	Topics
Amazon Bedrock Workshop - Moderated	300	Amazon Bedrock	Machine Learning (ML/AI)

Description

Amazon Bedrock Workshop

AWS Region: us-west-2 (Oregon)

Do NOT change the region!
Keep it at: us-west-2 (Oregon)!



Don't use incognito windows!

Self-service Lab Guide

Step by step guide

aws workshop studio



Vasileios Vonikakis ▾

Build Chatbots with Amazon Bedrock Knowledge Bases and Agents

Introduction

Running at an AWS-facilitated event

Running from your own AWS account

► Knowledge base for Amazon Bedrock

► Agents for Amazon Bedrock

► Amazon Bedrock Flows

► Guardrails for Amazon Bedrock

Clean-up (when running from your own AWS account)

Summary

Build Chatbots with Amazon Bedrock Knowledge Bases and Agents

Build Chatbots with Amazon Bedrock Knowledge Bases and Agents

Welcome to our workshop on Bedrock's knowledge base and agents to develop a loan agents chatbot.



We will not
be doing
this section

The traditional process of applying for a mortgage loan can be a complex and time-consuming endeavor. Prospective homebuyers often find themselves navigating through a maze of paperwork, policies, and

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Developer's Assets



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Bedrock Github Repo examples

Everything you may need on Amazon Bedrock

<https://github.com/aws-samples/amazon-bedrock-samples/tree/main>

Blueprints for Agents

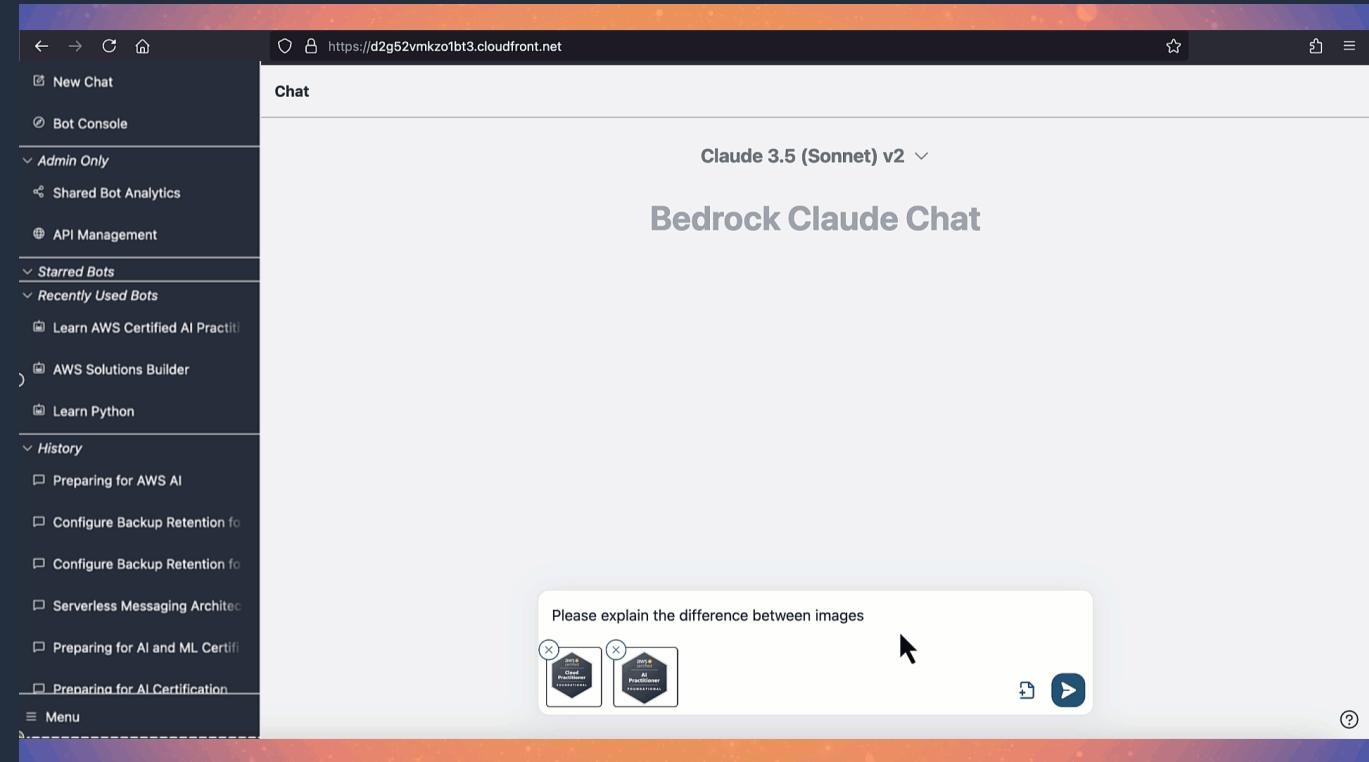
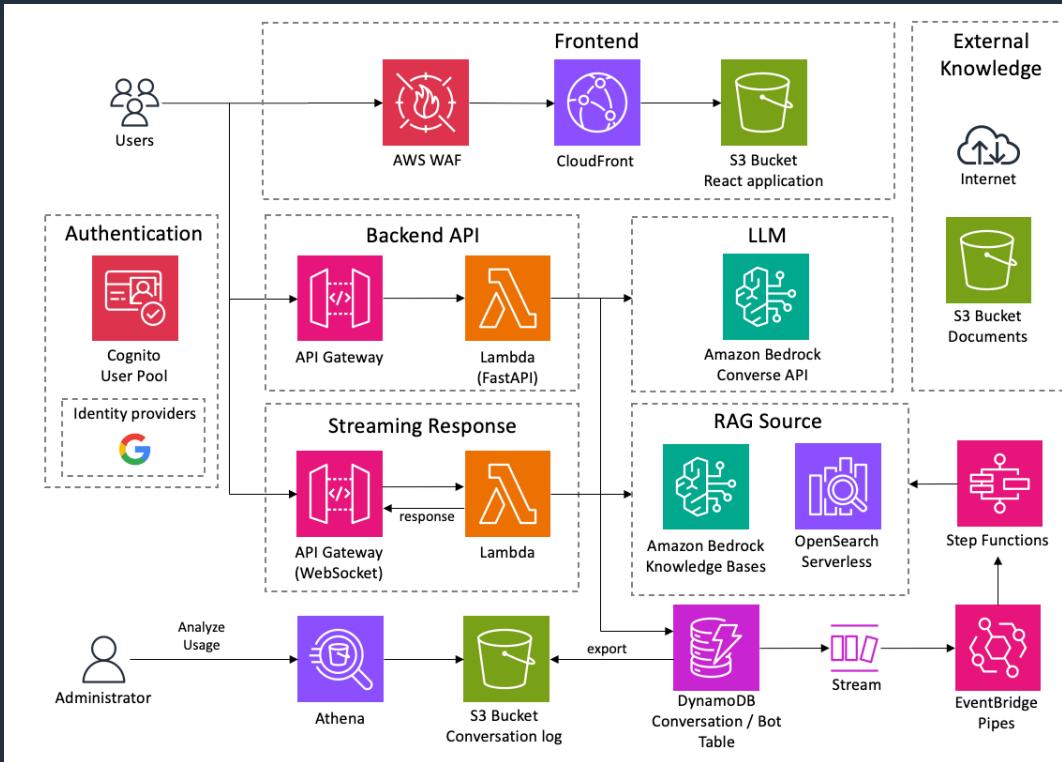
Simplify development of agent-based automation

- Collection of pre-built open-source CDK templates
- Templates: agent with function definitions, agent with return of control, agent with KB & Guardrails, agent with custom Lambda parser, agent with simple classification instructions
- Hosted on the [amazon-bedrock-samples](https://github.com/amazon-bedrock-samples) GitHub repo

<https://github.com/aws-samples/amazon-bedrock-samples/tree/main/agents-and-function-calling/bedrock-agents/agent-blueprint-templates>



Bedrock chatbot application



<https://github.com/aws-samples/bedrock-claude-chat>

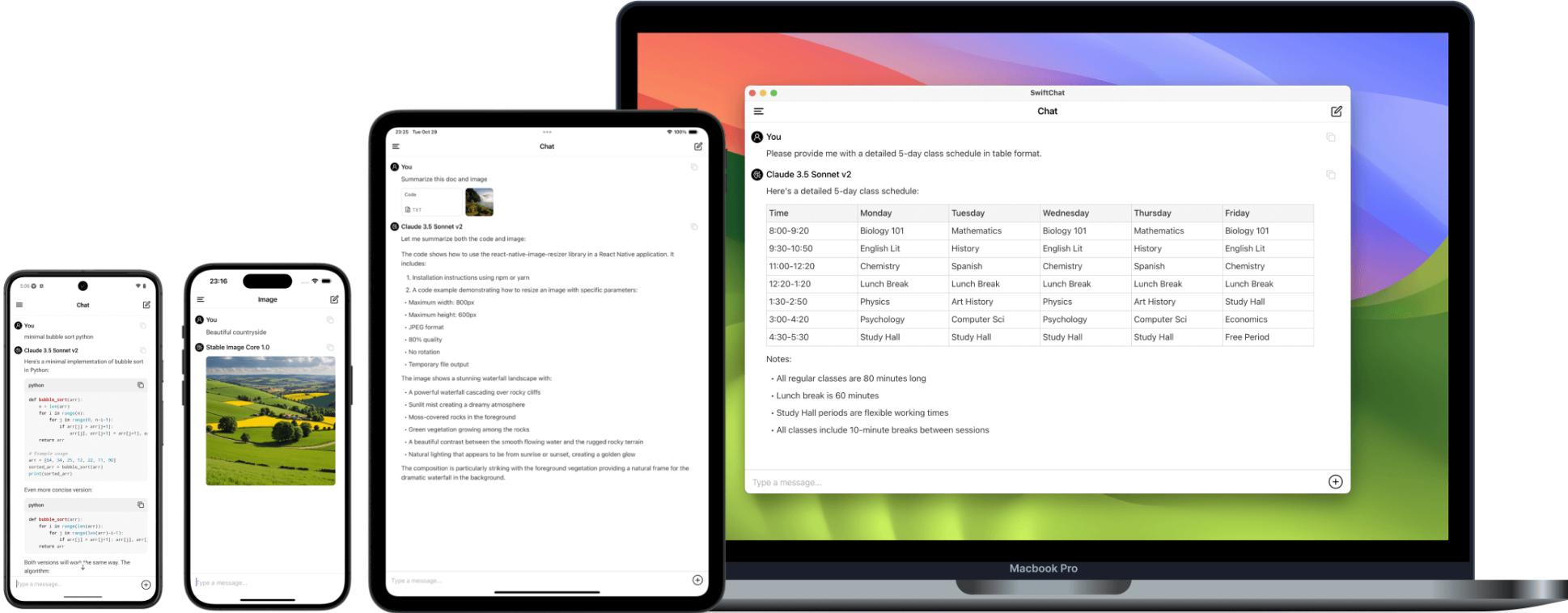
Building RAG use cases with GenAI Chatbot on AWS

The screenshot shows the AWS GenAI Chatbot homepage. At the top, there are three main sections: **Amazon Bedrock** (fully managed service for foundation models), **Amazon SageMaker** (CDK construct for self-hosted models), and **3P Models** (interface for third-party models like Anthropic, stability.ai, OpenAI, AI21 Labs, Hugging Face, and Cohere). Below these is a **Full-fledged user interface** section featuring a screenshot of a React-based web application. The next section is **Retrieval Augmented Generation (RAG) sources**, which includes **Amazon Aurora with pvector** (vector database extension), **Amazon OpenSearch VectorSearch** (semantic search with vector databases), and **Amazon Kendra** (intelligent search service). The final section at the bottom is **Models Interface**, which provides prebuilt CDK constructs for interacting with underlying models and data sources.

<https://github.com/aws-samples/aws-genai-llm-chatbot>



SwiftChat - A Cross-platform AI Chat App



<https://github.com/aws-samples/swift-chat>

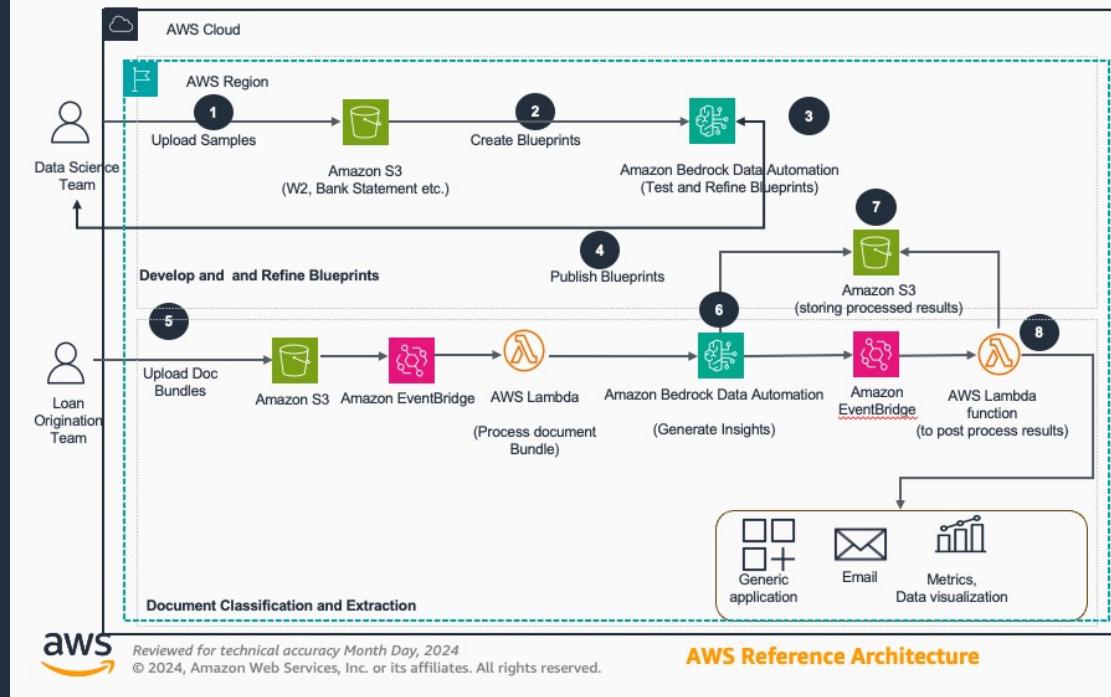


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Guidance for Multimodal Data Processing Using Amazon Bedrock Data Automation

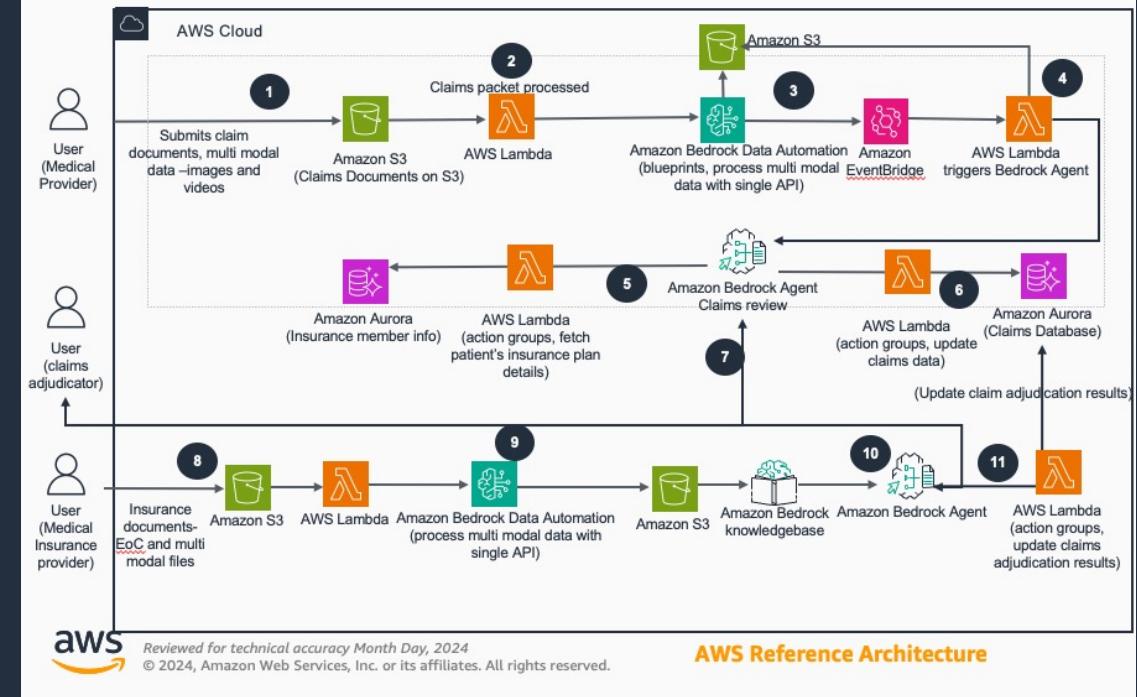
Guidance for IDP with Bedrock Data Automation—

This guidance demonstrates how to perform intelligent document processing with Amazon Bedrock Automation. We guide you through the lifecycle of Blueprints and we show how to perform document classification and extraction with loan origination processing example for a financial services company.



Guidance for IDP with Bedrock Data Automation

This diagram demonstrates how to perform intelligent document processing with Amazon Bedrock Automation. We show how to automate medical claims processing with multi modal input data and processing to improve efficiency and accuracy.

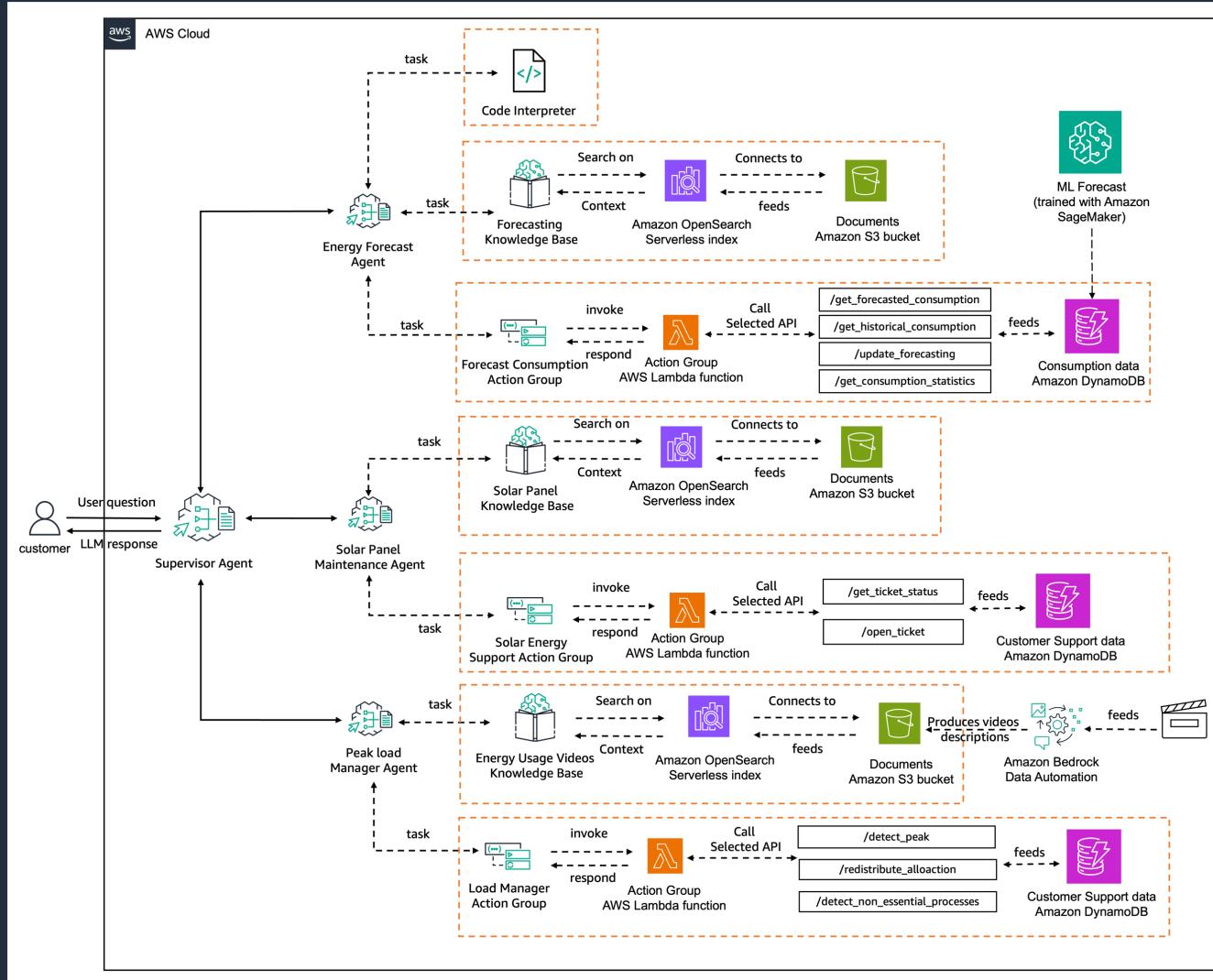


<https://github.com/aws-solutions-library-samples/guidance-for-multimodal-data-processing-using-amazon-bedrock-data-automation>

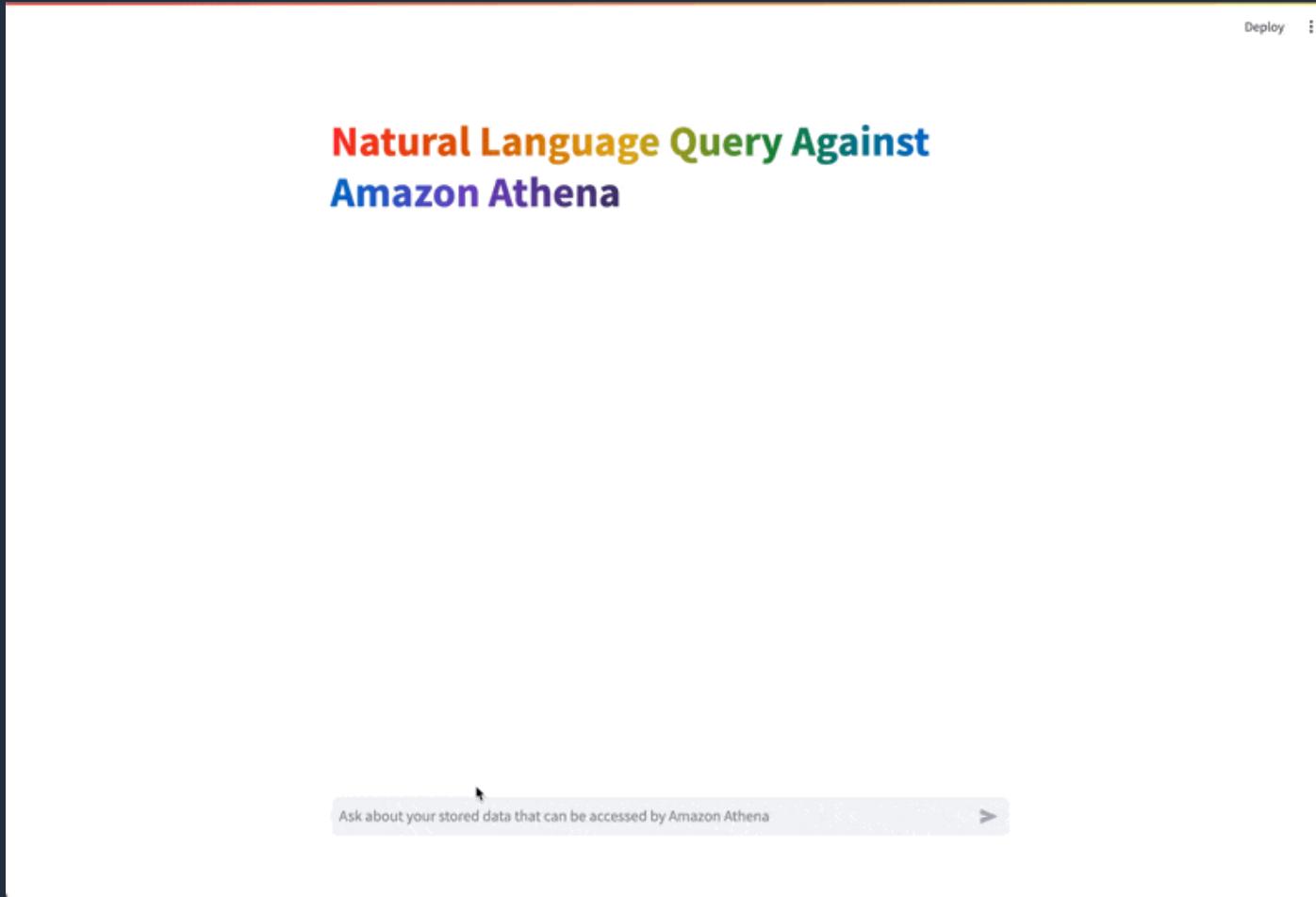


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Energy Efficiency Management System - Multi-Agent Collaboration



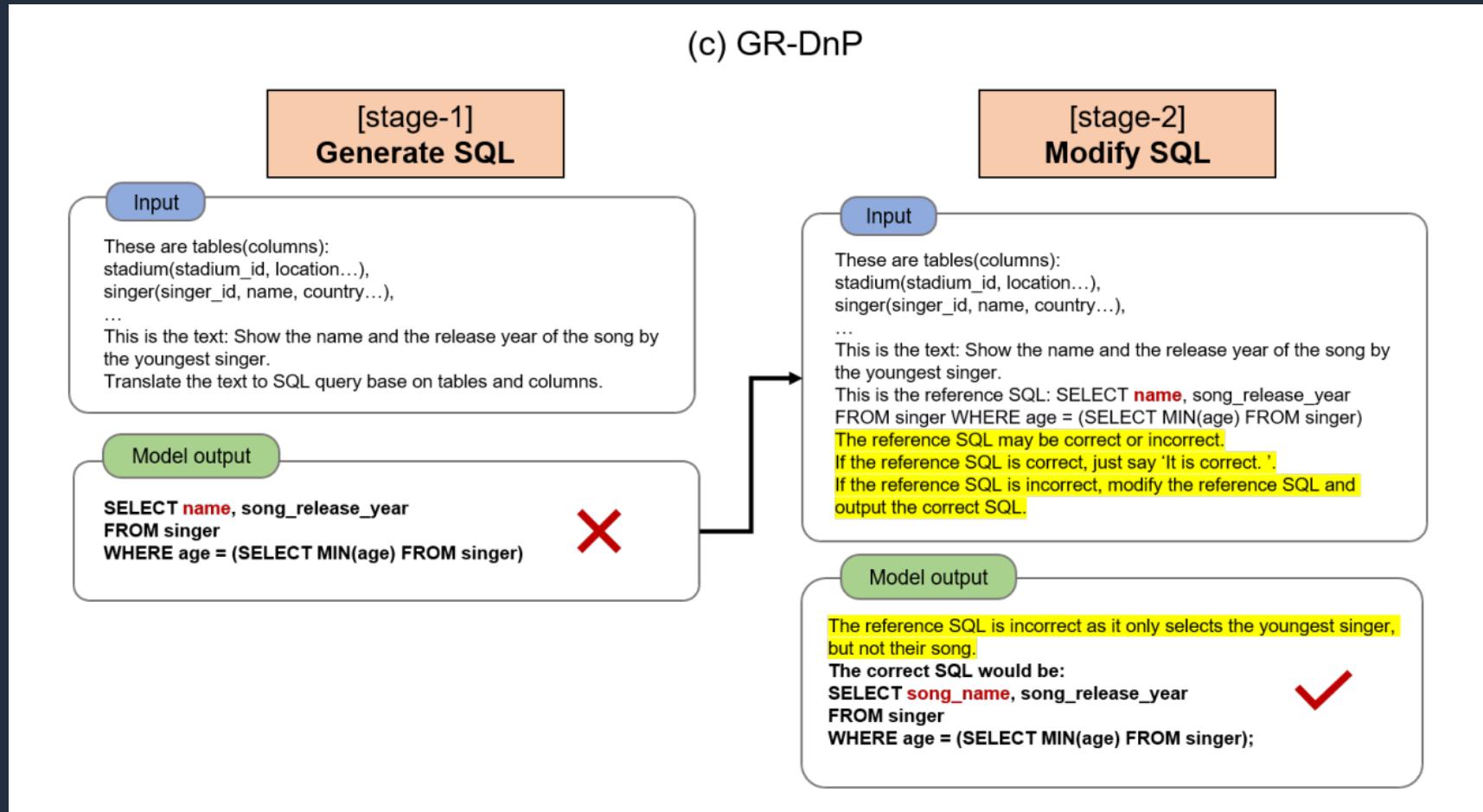
GenAI quick start POCs



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

<https://github.com/aws-samples/genai-quickstart-pocs/tree/main>

Text-to-SQL



Knowledge Quiz



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Please share your feedback

- What you liked
- What you didn't like
- How can we improve next time
- What would you like us to change



https://amazonmr.au1.qualtrics.com/jfe/form/SV_4MXnJBII01wX7T0

others



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.