AWS re:Invent 2023 Generative AI Recap

Michael Lin

Sr. Solutions Architect
Amazon Web Services

Agenda

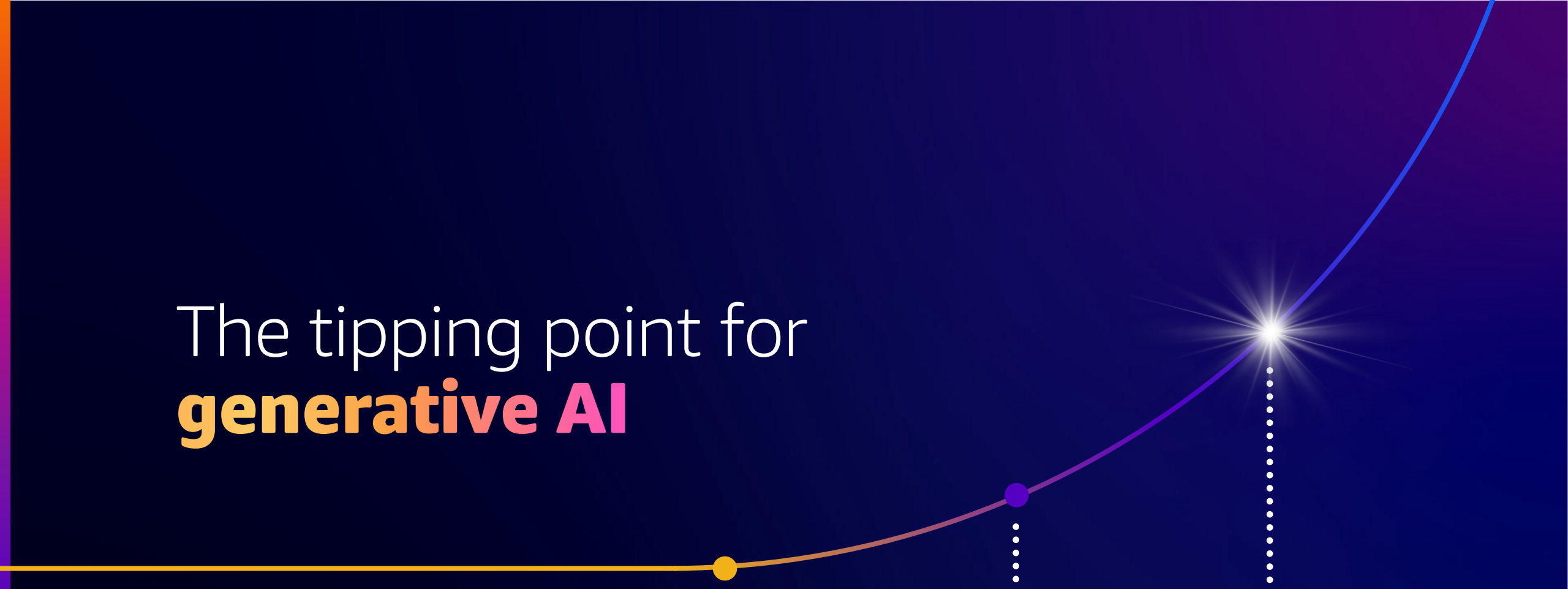
- Bedrock New Models
- SageMaker JumpStart for LLMs
- RAG and Knowledge Base
- Fine-tuning and Pre-training
- Automation and Agent

Agenda

- Responsible AI and Guardrails
- Model Evaluation
- AI Assistant and Amazon Q

Build your first generative AI application with Amazon Bedrock

The tipping point for **generative AI**

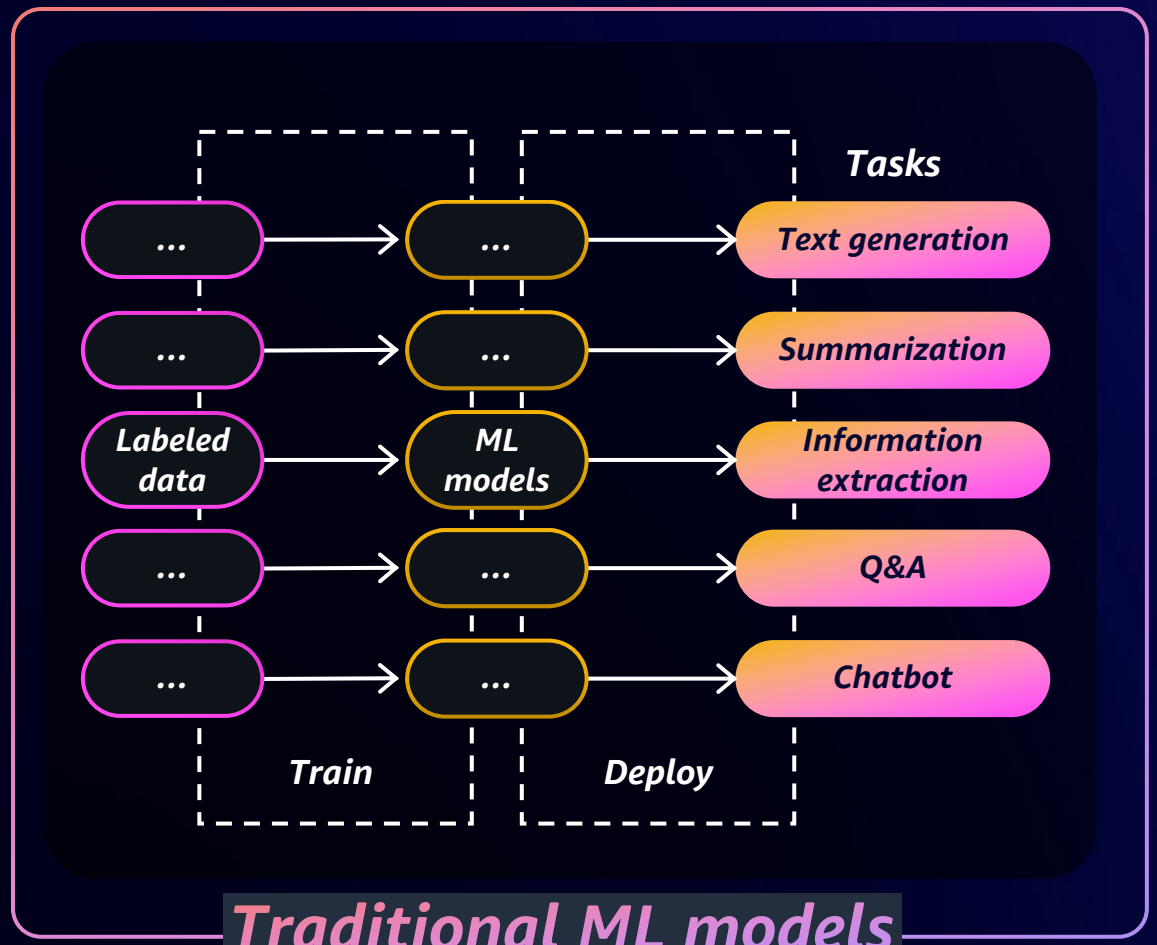


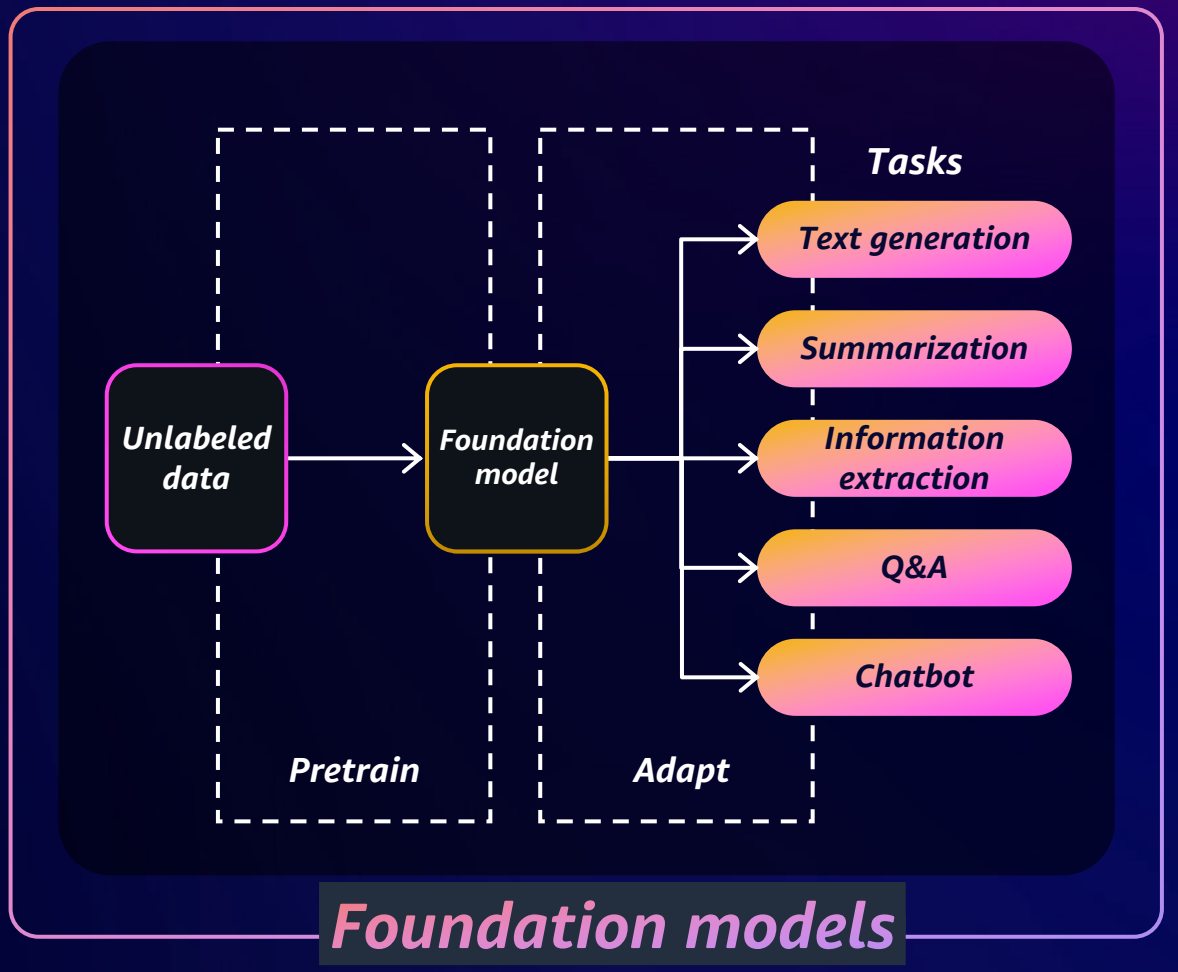
Massive proliferation of data

Availability of scalable compute capacity

Machine learning innovation

Generative AI is powered by foundation models (FMs)





GENERALLY AVAILABLE

Amazon *Bedrock*

The easiest way to build and scale generative AI applications with foundation models

Access a range of leading FMs through a single API

Privately customize FMs using your organization's data

Build agents that execute complex business tasks by dynamically invoking APIs

Extend the power of FMs with your data using Retrieval Augmented Generation (RAG)

Enable data security and compliance

More than **10,000 customers** are using Amazon Bedrock

https://aws.amazon.com/bedrock/testimonials/

Amazon *Bedrock*

Broad choice of foundation models

AI21labs

Jurassic-2

Contextual answers, summarization, paraphrasing

ANTHROPIC

Claude

Summarization, complex reasoning, writing, coding

cohere

Command & Embed

Text generation, search, classification

Meta

Llama 2

Q&A and reading comprehension

stability.ai

Stable Diffusion XL

High-quality images and art

amazon

Amazon Titan

Text summarization, generation, Q&A, search

NOW GENERALLY AVAILABLE

AWS News Blog

Amazon Bedrock now provides access to Meta's Llama 2 Chat 13B model

by Sébastien Stormacq | on 13 NOV 2023 | in Amazon Bedrock, Announcements, Artificial Intelligence, Generative AI, Launch, News | Permalink | Comments | Share

0:00 / 0:00

Voiced by Amazon Polly

Update: November 29, 2023 — Today, we're adding the Llama 2 70B model in Amazon Bedrock, in addition to the already available Llama 2 13B model. As its name implies, the Llama 2 70B model has been trained on larger datasets than the Llama 2 13B model. If you're wondering when to use which model, consider using Llama 13B for smaller-scale tasks such as text classification, sentiment analysis, and language translation, and Llama 2 70B for large-scale tasks such as language modeling, text generation, and dialogue systems. According to Meta, Llama 2 70B's training took 1,720,320 GPU-hours, the equivalent of 196.38 years. Start using the Llama 2 70B model in Amazon Bedrock today. We're excited to see what you build with these models.

https://aws.amazon.com/blogs/aws/amazon-bedrock-now-provides-access-to-llama-2-chat-13b-model/

NOW GENERALLY AVAILABLE

AWS News Blog

Mistral AI models now available on Amazon Bedrock

by Donnie Prakoso | on 01 MAR 2024 | in Amazon Bedrock, Amazon Machine Learning, Announcements, Artificial Intelligence, Generative AI, Launch, News | Permalink | Comments | Share

0:00 / 0:00

Voiced by Amazon Polly

Last week, we announced that Mistral AI models are coming to Amazon Bedrock. In that post, we elaborated on a few reasons why Mistral AI models may be a good fit for you. Mistral AI offers a balance of cost and performance, fast inference speed, transparency and trust, and is accessible to a wide range of users.

Today, we're excited to announce the availability of two high-performing Mistral AI models, Mistral 7B and Mixtral 8x7B, on Amazon Bedrock. Mistral AI is the 7th foundation model provider offering cutting-edge models in Amazon Bedrock, joining other leading AI companies like AI21 Labs, Anthropic, Cohere, Meta, Stability AI, and Amazon. This integration provides you the flexibility to choose optimal high-performing foundation models in Amazon Bedrock.

Mistral 7B is the first foundation model from Mistral AI, supporting English text generation tasks with natural coding capabilities. It is optimized for low latency with a low memory requirement and high throughput for its size. Mixtral 8x7B is a popular, high-quality, sparse Mixture-of-Experts (MoE) model, that is ideal for text summarization, question and answering, text classification, text completion, and code generation.

https://aws.amazon.com/blogs/aws/mistral-ai-models-now-available-on-amazon-bedrock/

CLAUDE SONNET AVAILABLE

AWS News Blog

Anthropic's Claude 3 Sonnet foundation model is now available in Amazon Bedrock

by Channy Yun | on 04 MAR 2024 | in Amazon Bedrock, Artificial Intelligence, Generative AI, Launch, News | Permalink | Comments | Share

0:00 / 0:00

Voiced by Amazon Polly

In September 2023, we announced a strategic collaboration with Anthropic that brought together their respective technology and expertise in safer generative artificial intelligence (AI), to accelerate the development of Anthropic's Claude foundation models (FMs) and make them widely accessible to AWS customers. You can get early access to unique features of Anthropic's Claude model in Amazon Bedrock to reimagine user experiences, reinvent your businesses, and accelerate your generative AI journeys.

In November 2023, Amazon Bedrock provided access to Anthropic's Claude 2.1, which delivers key capabilities to build generative AI for enterprises. Claude 2.1 includes a 200,000 token context window, reduced rates of hallucination, improved accuracy over long documents, system prompts, and a beta tool use feature for function calling and workflow orchestration.

Today, Anthropic announced Claude 3, a new family of state-of-the-art AI models that allows customers to choose the exact combination of intelligence, speed, and cost that suits their business needs. The three models in the family are **Claude 3 Haiku**, the fastest and most compact model for near-instant responsiveness, **Claude 3 Sonnet**, the ideal balanced model between skills and speed, and **Claude 3 Opus**, the most intelligent offering for the top-level performance on highly complex tasks.

https://aws.amazon.com/blogs/aws/anthropics-claude-3-sonnet-foundation-model-is-now-available-in-amazon-bedrock/

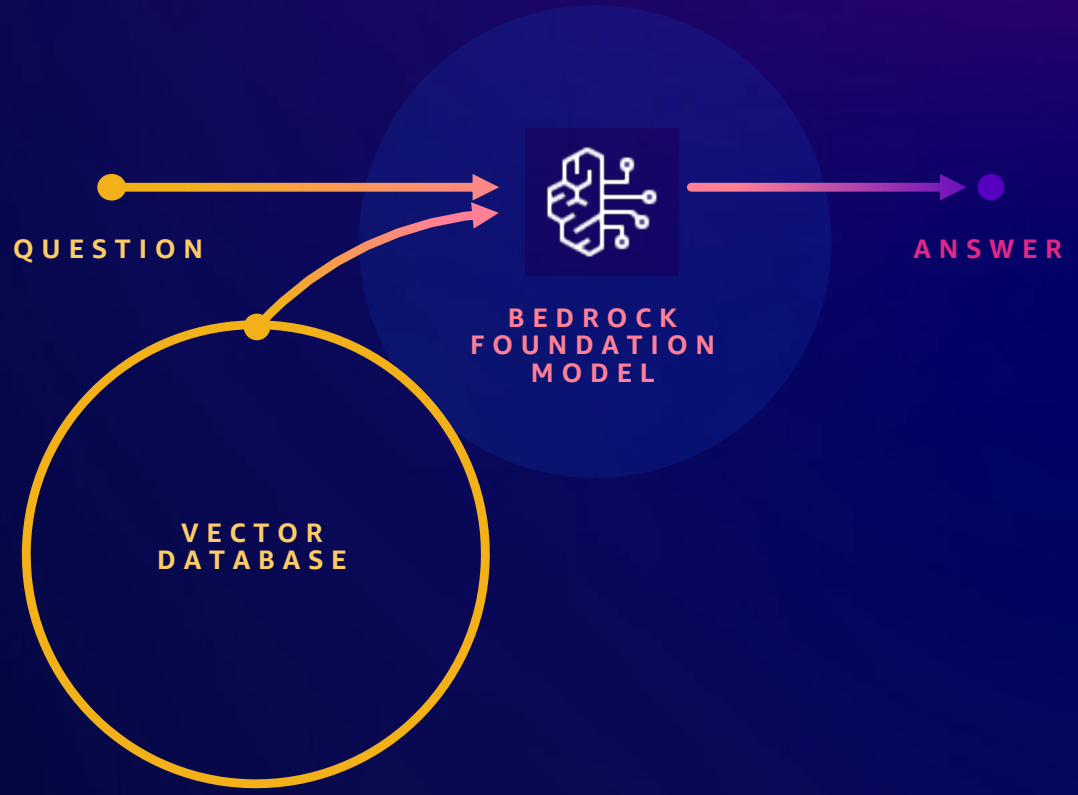
Knowledge base for Amazon Bedrock

USE RETRIEVAL AUGMENTED GENERATION (RAG)

Connect FMs to data sources including vector engine for Amazon OpenSearch Serverless, Pinecone, and Redis Enterprise Cloud

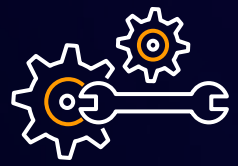
Enable automatic data source detection

Provide citations



Your data is your differentiator

PRIVATELY CUSTOMIZE FOUNDATION MODELS USING YOUR ORGANIZATION'S DATA



Fine-tune

Purpose	Maximizing accuracy for specific tasks
Data need	Small number of labeled examples

Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO COMPLETE TASKS IN JUST A FEW QUICK STEPS

1. *Select your foundation model*
2. *Provide basic instructions*
3. *Select relevant data sources*
4. *Specify available actions*

| Breaks down and orchestrates tasks |

| Securely accesses and retrieves company data |

| Takes action by invoking API calls on your behalf |

| Provides fully managed infrastructure |

aws

Explore text-generation FMs for top use cases with Amazon Bedrock

Amazon Titan Text

TEXT GENERATION MODELS

Titan Text Lite

Price-performant version, ideal for English-language tasks

Highly customizable for fine-tuning tasks such as article summarization and copywriting

- ✓ Max Tokens: 4K
- ✓ Languages: English
- ✓ Fine Tuning: Yes
- ✓ Recommended use-cases: Summarization, fine-tuning, copywriting, among others

Titan Text Express

Ideal for wide range of tasks, such as open-ended text generation and conversational chat

Support within RAG workflows

- ✓ Max Tokens: 8K
- ✓ Languages: 100+ languages
- ✓ Fine Tuning: Yes
- ✓ Recommended use-case: RAG, conversational chat, text generation, COT, code generation, among others

Claude 3 Highlights

Claude 3 Benchmark	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5-shot	**79.0%** 5-shot	**75.2%** 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level knowledge *GPQA, Diamond*	**50.4%** 0-shot CoT	**40.4%** 0-shot CoT	**33.3%** 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	-	-
Undergraduate level knowledge *GSM8K*	**95.0%** 0-shot CoT	**92.3%** 0-shot CoT	**88.9%** 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot CoT	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	**43.1%** 0-shot CoT	**38.9%** 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	**83.5%** 0-shot	**75.1%** 0-shot	74.5% 8-shot	-	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	**73.0%** 0-shot	**75.9%** 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	**78.9** 3-shot	**78.4** 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed Evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	**82.9%** 3-shot CoT	**73.7%** 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Logical reasoning *ARC-Challenge*	**96.4%** 25-shot	**93.2%** 25-shot	**89.2%** 25-shot	96.3% 25-shot	85.2% 25-shot	-	-
Logical reasoning *HellaSwag*	**95.4%** 10-shot	**89.0%** 10-shot	**85.9%** 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84% 10-shot

- **TL;DR**;
  - 🚀 3 different versions (Opus, Sonnet, Haiku)
  - 🥇 Opus (best models) outperforms GPT-4 and Gemini 1.0 Ultra on Text Benchmarks
  - 📚 200k context window
  - 🌅 Vision support (59% on MMMU)
  - ⚡ Sonnet is 2x faster than Claude 2
  - 🤖 Can be used for task automation/agent workflows
  - 🌐 Sonnet available in Amazon Bedrock today (March 4), Opus and Haiku coming soon
- **Cost**:
  - Opus: $15/1M input (0.5x of GPT-4); $75M output (1.25x of GPT-4)
  - Sonnet: $3/1M input (0.33x of GPT-4 Turbo); $15M output (0.5x of GPT-4 Turbo)
  - Haiku: $0.25/1M input (0.5x of GPT-3.5 Turbo); $1.25M output (1.2x of GPT-3.5 Turbo)

https://www.anthropic.com/news/claude-3-family

Explore image generation and search with Foundational Models on Amazon Bedrock

Amazon Titan Multimodal FMs

TWO NEW MULTIMODAL MODELS FROM AMAZON

Amazon Titan Multimodal Embeddings

for enterprise tasks such as image search and similarity

Amazon Titan Image Generator

for text-to-image generation and image editing

Features

- Build multimodal semantic search apps using Amazon Titan Multimodal Embeddings FM
- Content creation at scale with Amazon Titan Image Generator FM
- Comprehensive editing features to edit existing or generated images
- FMs can be customized on your data
- Support responsible use of AI

Amazon Titan Multimodal Embeddings

NOW GENERALLY AVAILABLE

Example

"Blue sneakers without laces"

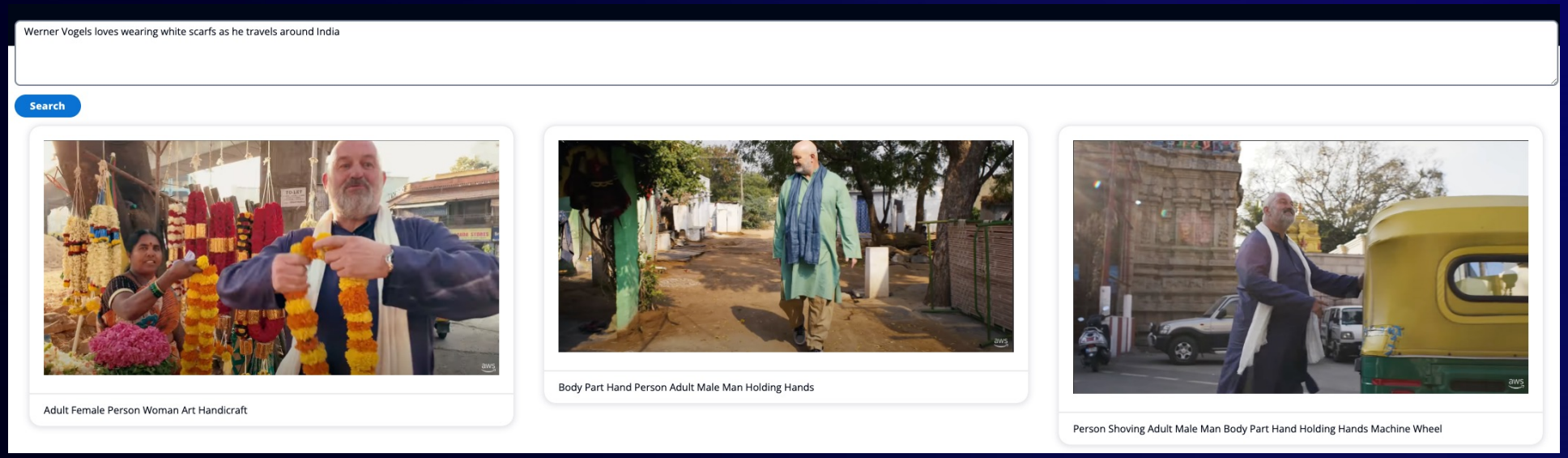
Result 1			Result 2		
10 results			**10 results**		
1		text: No Tie Shoelaces -	1		text_description: Capelli New York Toddlers Unisex Slip On Sneaker
Not applicable			Not applicable		
2		text: Shoe Laces No Tie	2		text_description: Tesla Men's Ultra Lightweight Running Shoes L510 PR2
Not applicable			Not applicable		
3		text: Shoe Laces No Tie	3		text_description: Crocs Men's Norlin Canvas Slip-on Upper: Canvas
Not applicable			Not applicable		
4		text: Shoe Laces No Tie	4		text_description: Bed Stu Men's Bluegill Slip-On Loafer Slip on
Not applicable			Not applicable		

Image Search

AWS Machine Learning Blog

Easily build semantic image search using Amazon Titan

by Mark Watkins and Dan Johns | on 30 NOV 2023 | in Amazon Bedrock, Amazon Comprehend, Amazon OpenSearch Service, Amazon Rekognition, Artificial Intelligence, Customer Solutions, Generative AI, Intermediate (200), Technical How-to, Thought Leadership | Permalink | Comments | Share



https://aws.amazon.com/blogs/machine-learning/easily-build-semantic-image-search-using-amazon-titan/

Amazon Titan Image Generator

NOW IN PREVIEW

Text to Image

GENERATE IMAGES FROM TEXT PROMPTS

A person wearing a hat and dark glasses is running in a forest. The forest is lush green with red and yellow flowers

Customization

FINE-TUNE THE MODEL ON YOUR DATA

Base Model

Customized Model

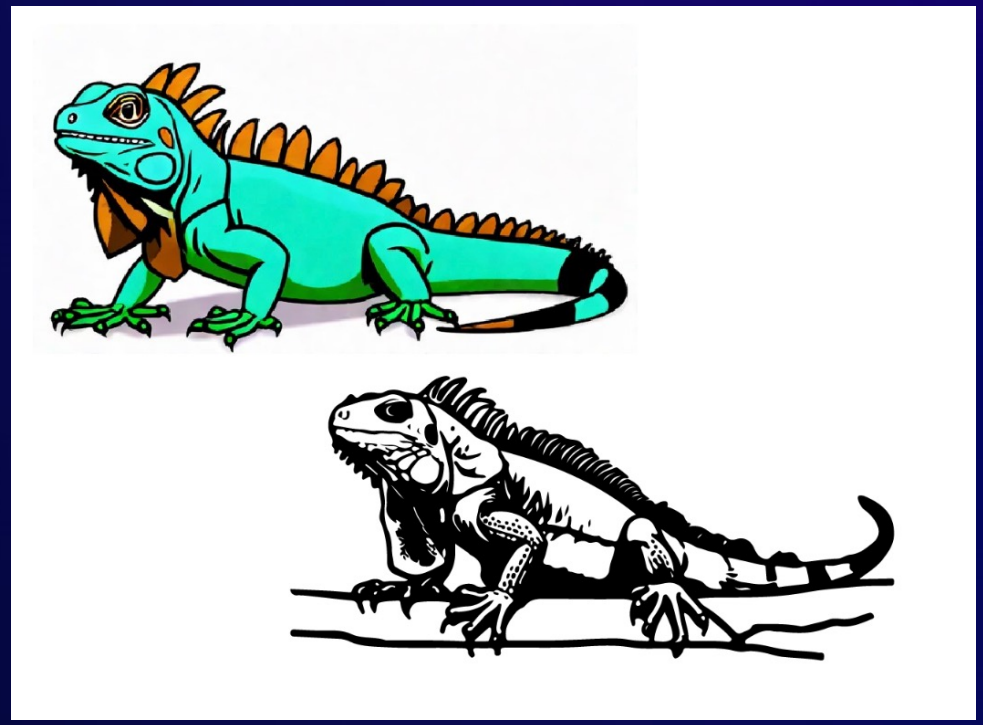
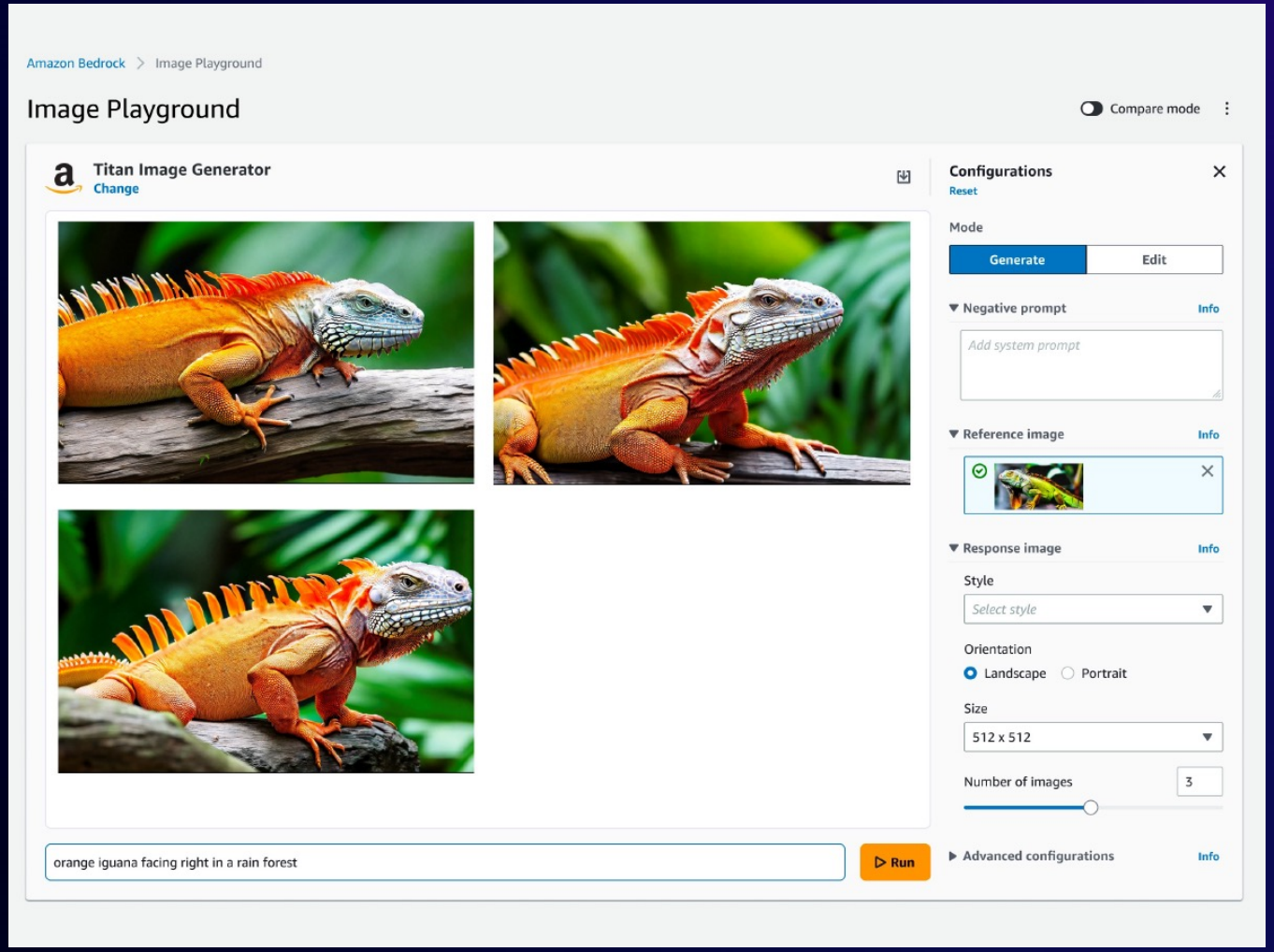


Image Variations

GENERATE VARIATIONS OF AN IMAGE



Automatic editing

MASK-FREE EDITING

Input image

\+

Input prompt

“change flowers to orange color”

→

Generated image

Inpainting

EDIT AN IMAGE

Input image

Inpainting the image with a car

Outpainting

GENERATE DIFFERENT BACKGROUNDS

Input image

Generated images with different backgrounds

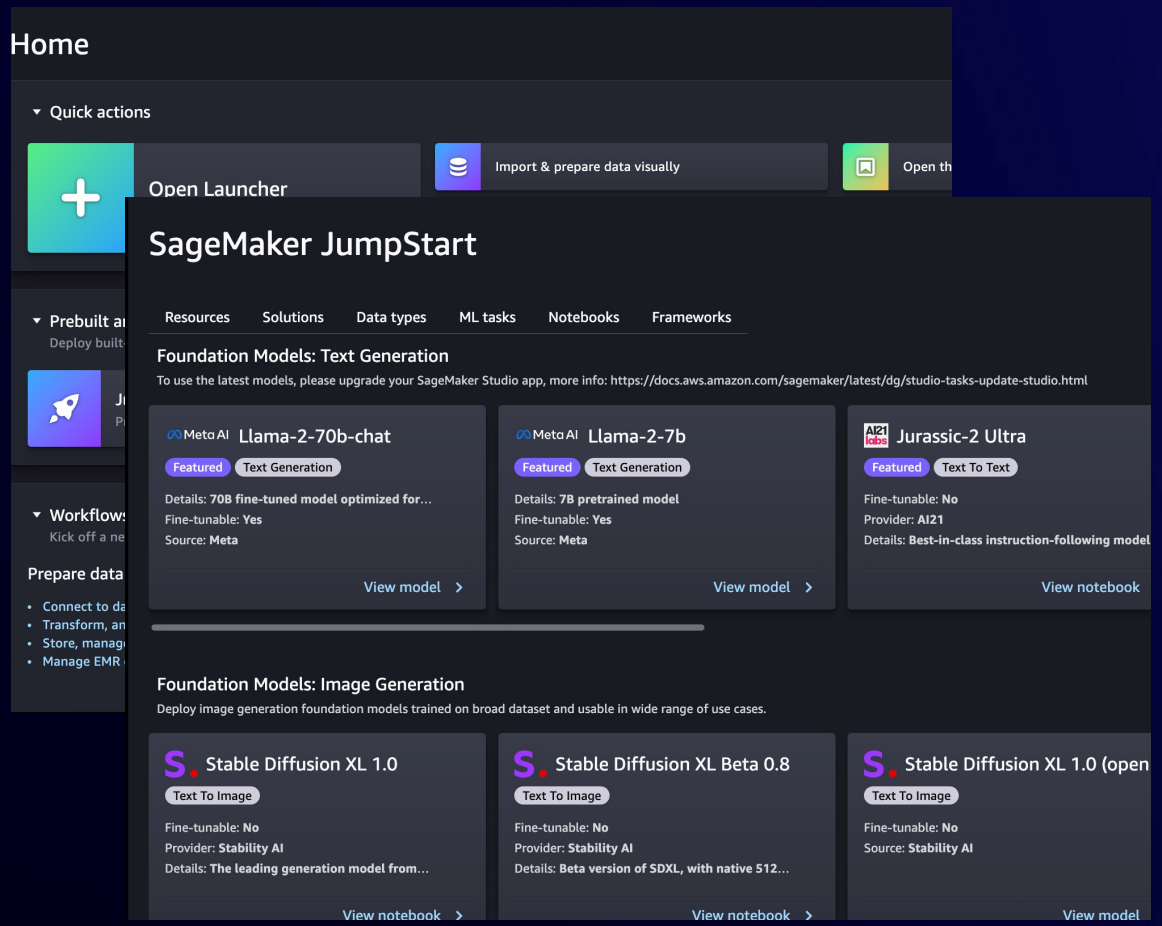
Accelerate FM development with Amazon SageMaker JumpStart

Discover foundation models and deploy with SageMaker's enterprise-ready features

AVAILABLE ON SAGEMAKER JUMPSTART

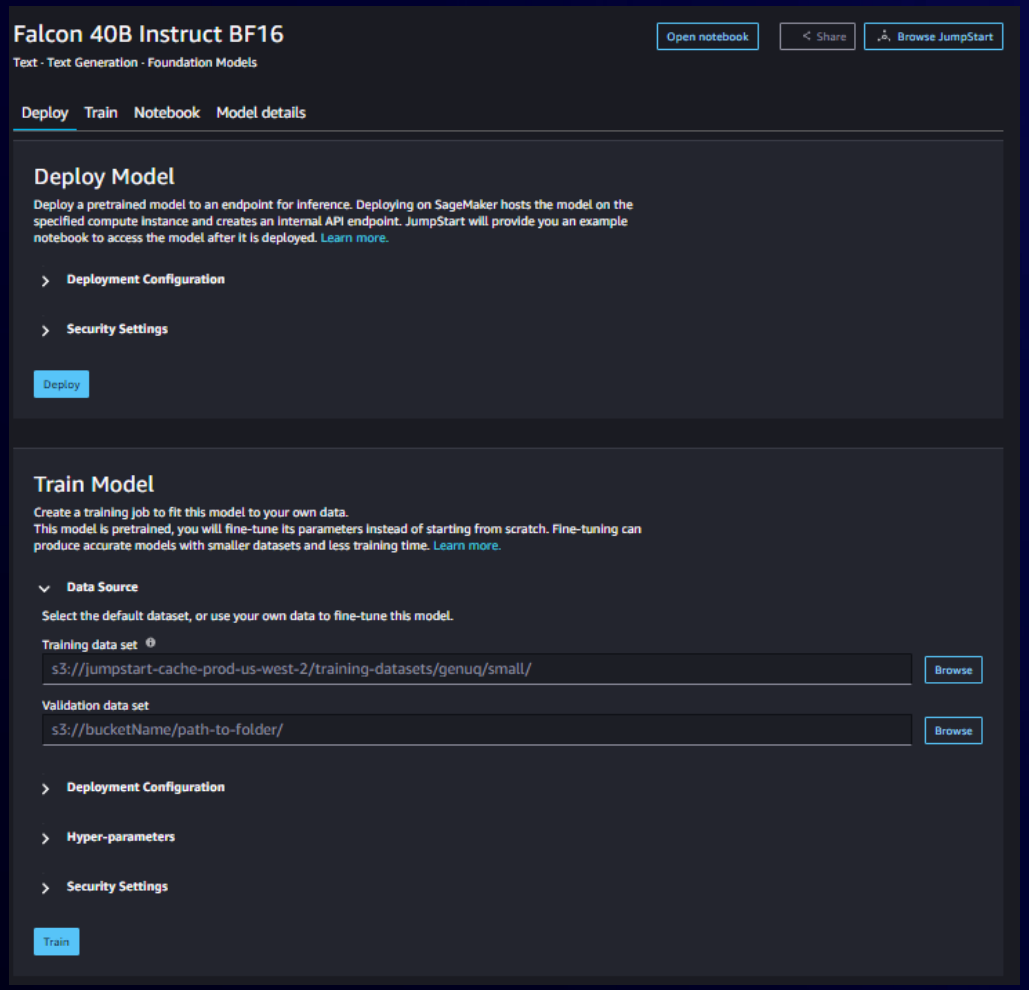
AI21labs	Meta AI	cohere	Hugging Face	stability.ai	LightOn	databricks	alexa
Models	***Models***	***Models***	***Models***	***Models***	***Models***	***Models***	***Models***
Jurassic-2	Llama 2 7B, 13B, 70B	Command	Falcon-7B, 40B,	Stable Diffusion	Lyra-Fr	Dolly	AlexaTM 20B
Ultra, Mid	Code Llama 7B, 13B, 34B	Cohere Light	180B,	XL 1.0	10B, Mini		
	Open LlaMA		Mistral 7B	2.1 base			
			RedPajama	Upscaling			
			MPT-7B	Inpainting			
			BloomZ 176B				
			Flan T-5				
			DistilGPT2				
			GPT NeoXT				
			Bloom				

Discover foundation models from multiple providers



- Browse in SageMaker Studio
- Search for specific model or provider from search bar
- View model-specific information

Review model details and take action



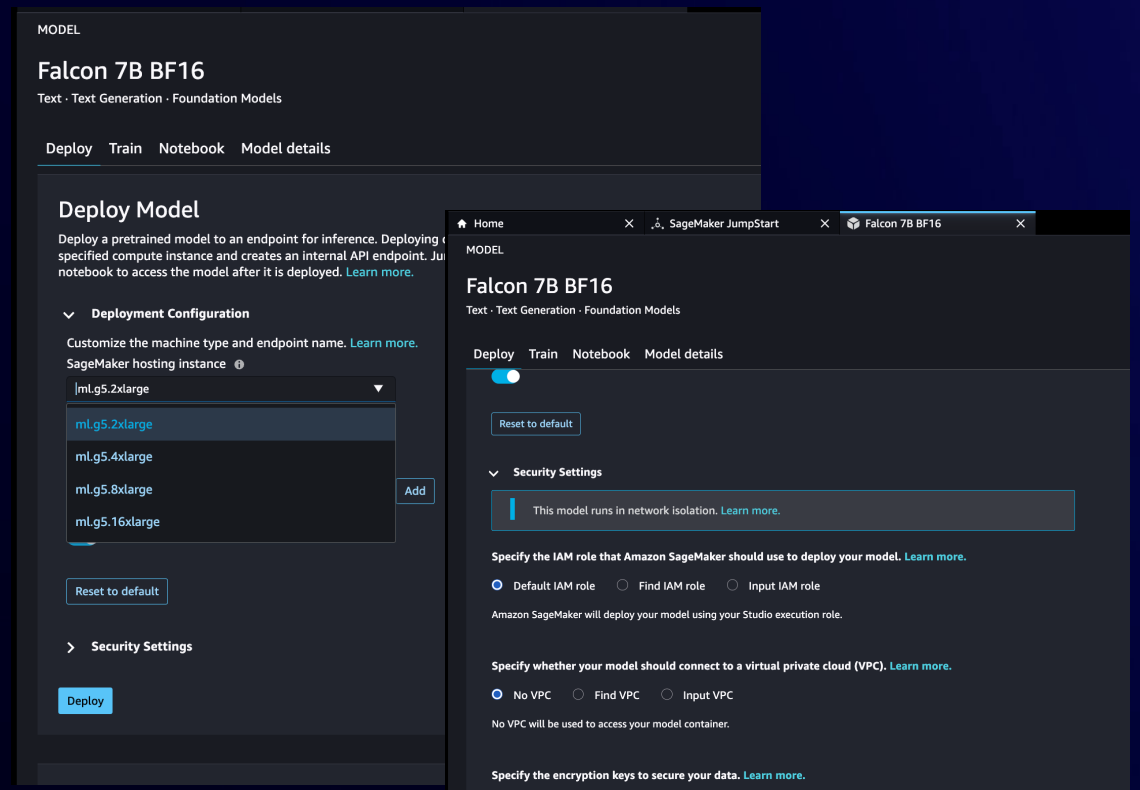
Details from model provider:

- Model size and description
- License info
- Use cases and how-to use model

Take action:

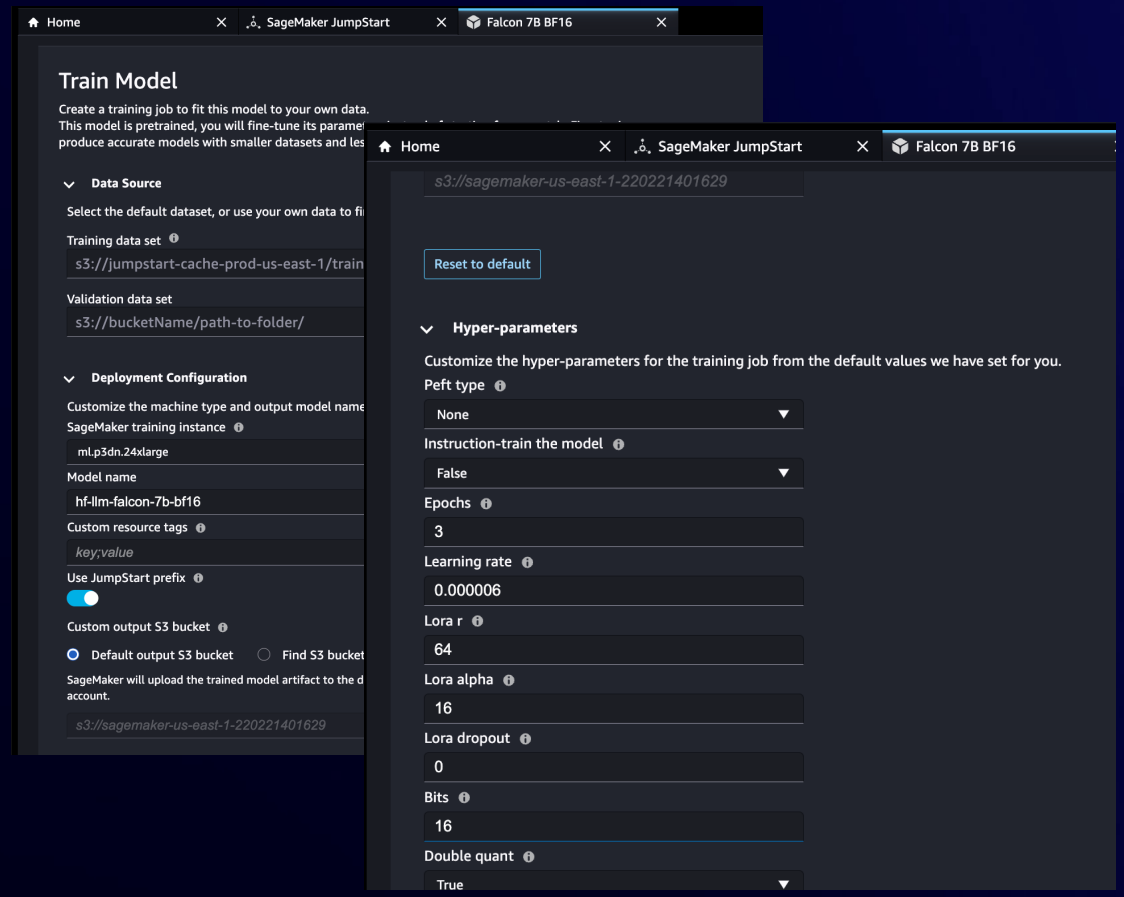
- Deploy
- Train
- View API snippet

Deploy to SageMaker with just a few clicks



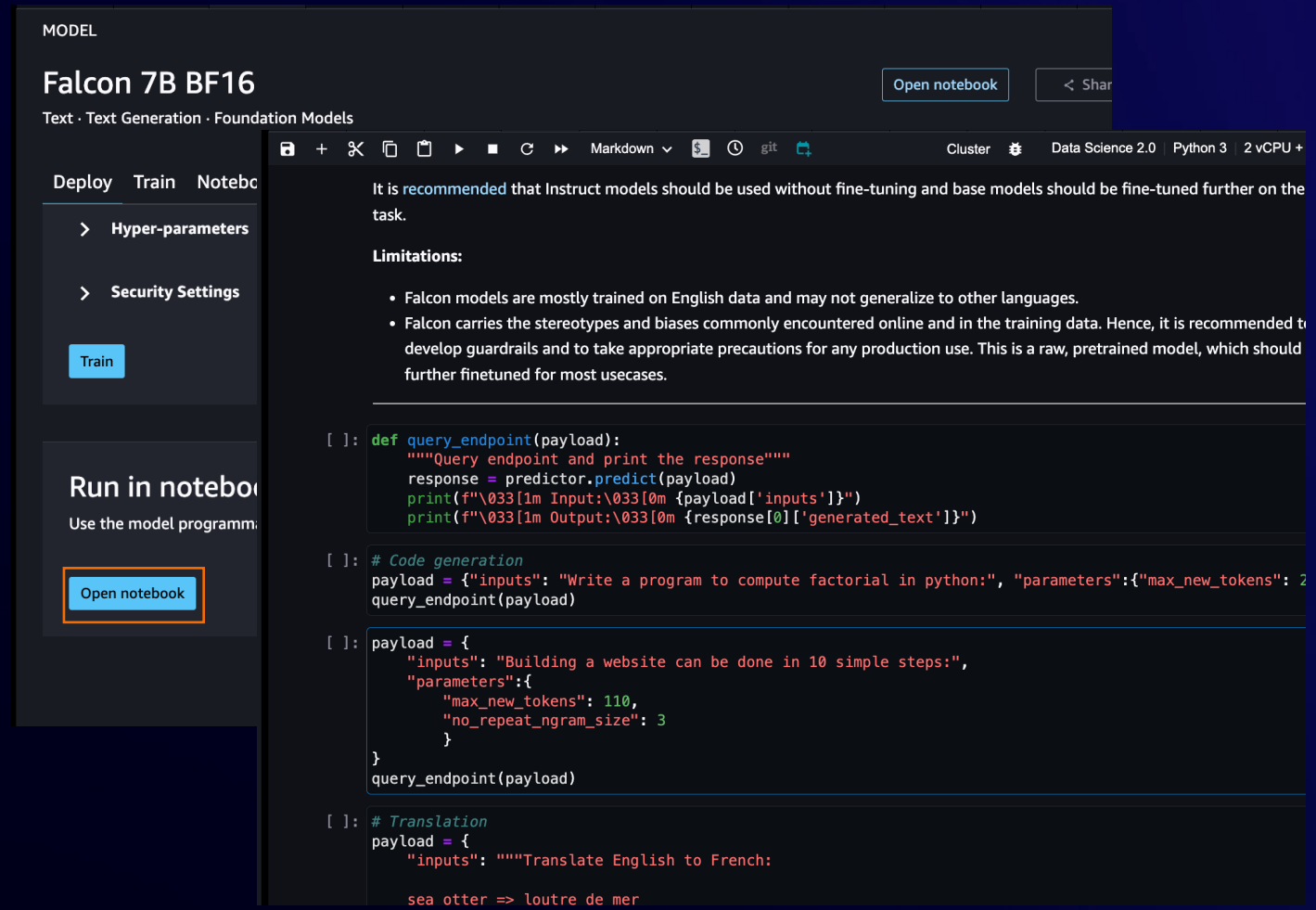
- One-click deploy with defaults
- Configure for cost, throughput, and latency
- Control security and VPC settings

Customize models with your data



- Fine-tune open and closed models
- Store model weights in your S3 bucket
- Control hyper-parameters
- Choose preferred instance type

Scale using the SageMaker SDK



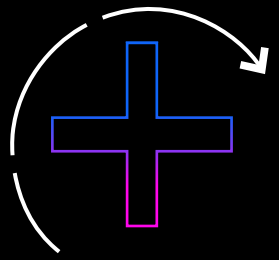
- Automate using the SageMaker APIs
- View example code for each model
- Control deployment parameters

Use RAG to improve responses in generative AI applications

What is Retrieval Augmented Generation?

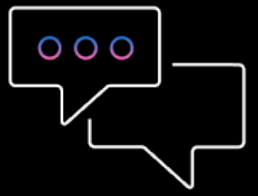
Retrieval

Fetches the relevant content from the external knowledge base or data sources based on a user query



Augmentation

Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model



Generation

Response from the foundation model based on the augmented prompt.

Knowledge Bases for Amazon Bedrock

Gives FMs and agents contextual information from your private data sources for Retrieval Augmented Generation (RAG) to deliver more relevant, accurate, and customized responses.

Fully managed support for end-to-end RAG workflow

Securely connect FMs and agents to data sources

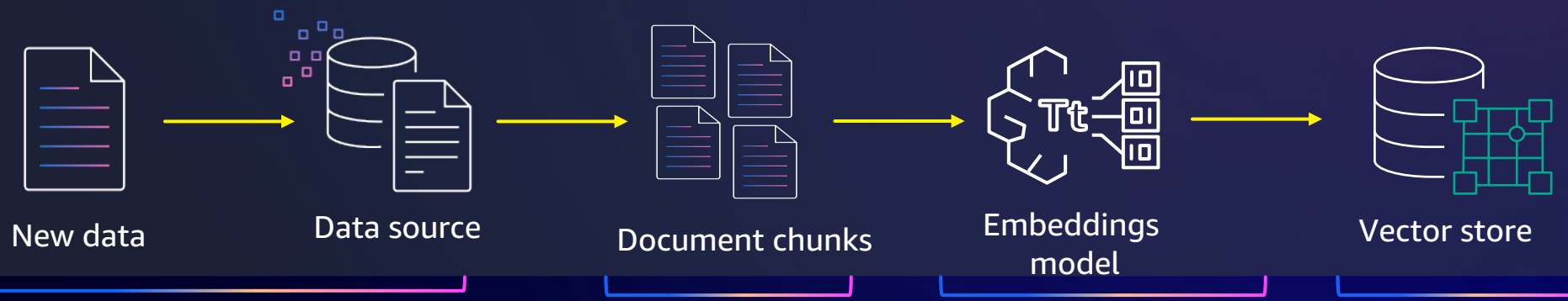
Easily retrieve relevant data and augment prompts

Provide source attribution

Data Ingestion Workflow

KNOWLEDGE BASES FOR AMAZON BEDROCK

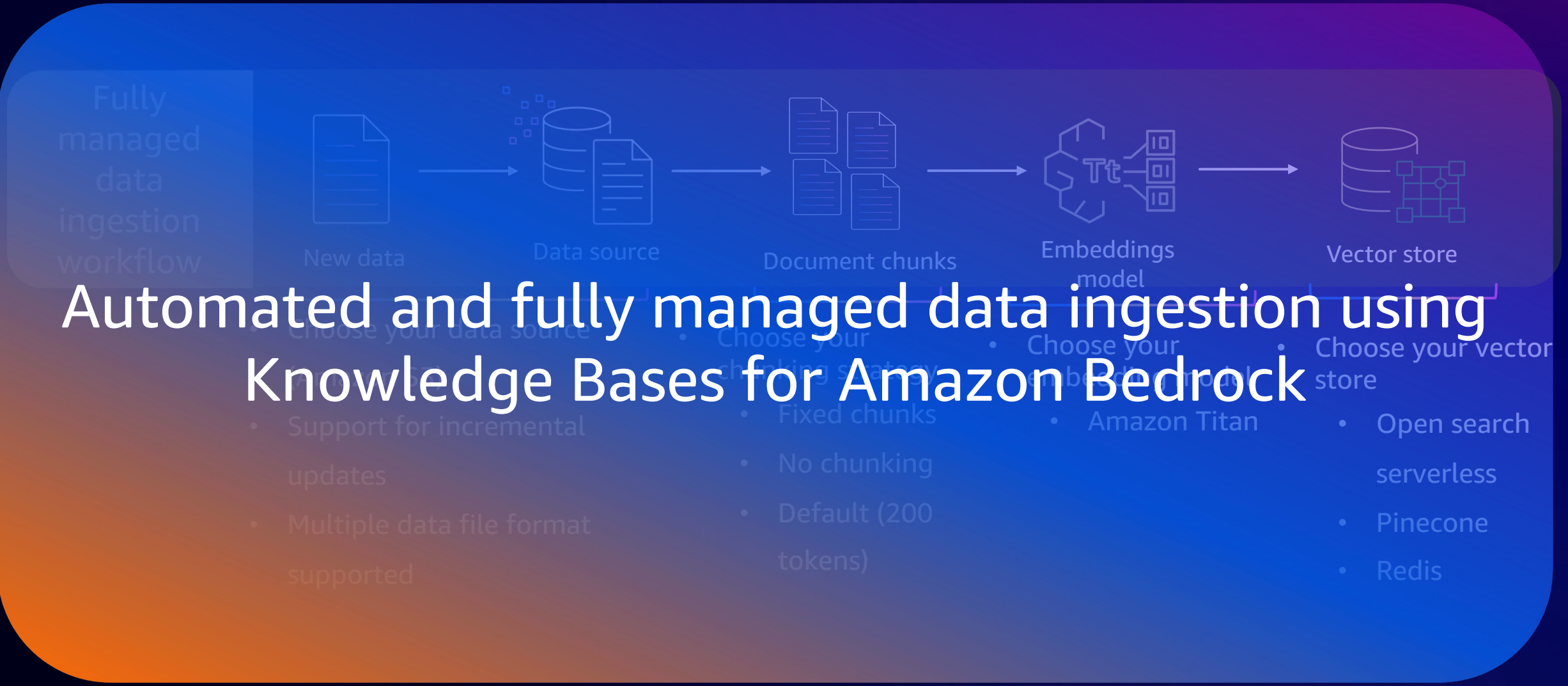
Fully managed data ingestion workflow



- Choose your data source (Amazon S3)
  - Support for incremental updates
  - Multiple data file formats supported
- Choose your chunking strategy
  - Fixed chunks
  - No chunking
  - Default (200 tokens)
- Choose your embedding model
  - Amazon Titan
- Choose your vector store
  - Open search serverless
  - Pinecone
  - Redis

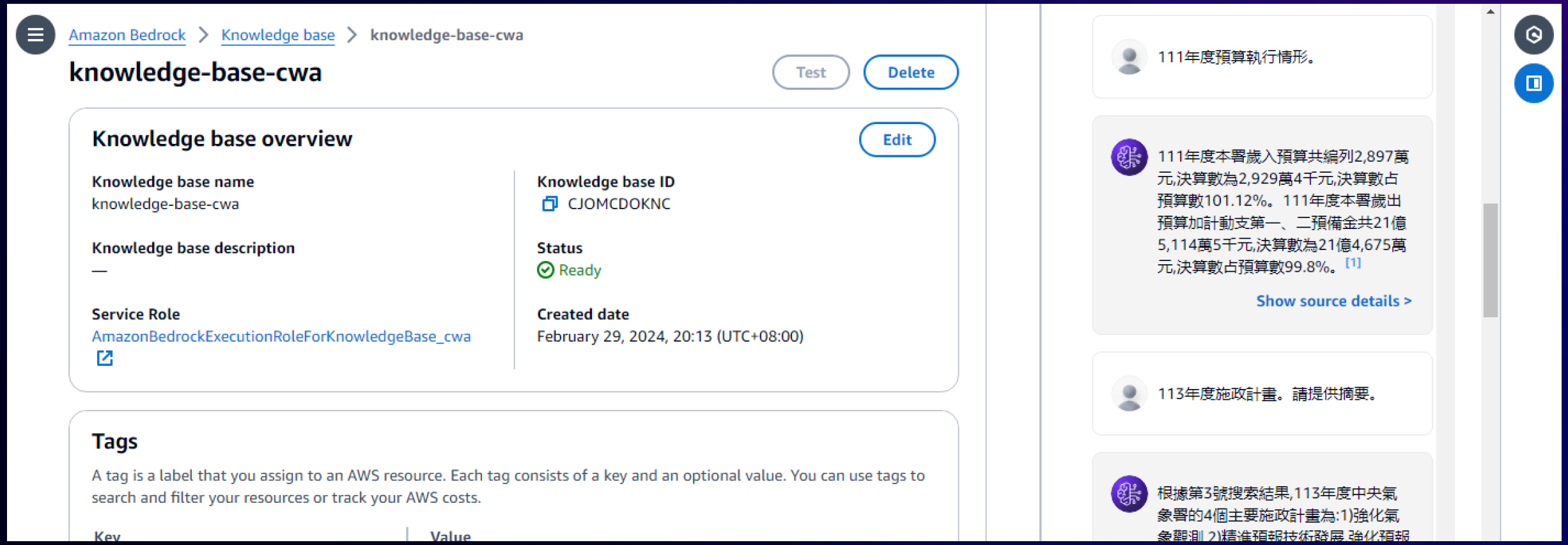
Fully managed data ingestion

KNOWLEDGE BASES FOR AMAZON BEDROCK



Automated and fully managed data ingestion using Knowledge Bases for Amazon Bedrock

KB in Action



Fine-tuning and Continued Pre-training with Amazon Bedrock

Why customize?

Customize to specific business needs

E.g. Healthcare – Understand medical terminology and provide accurate responses related to patient's health

Adapt to domain-specific language

E.g. Finance – Teach financial & accounting terms to provide good analysis for earnings reports

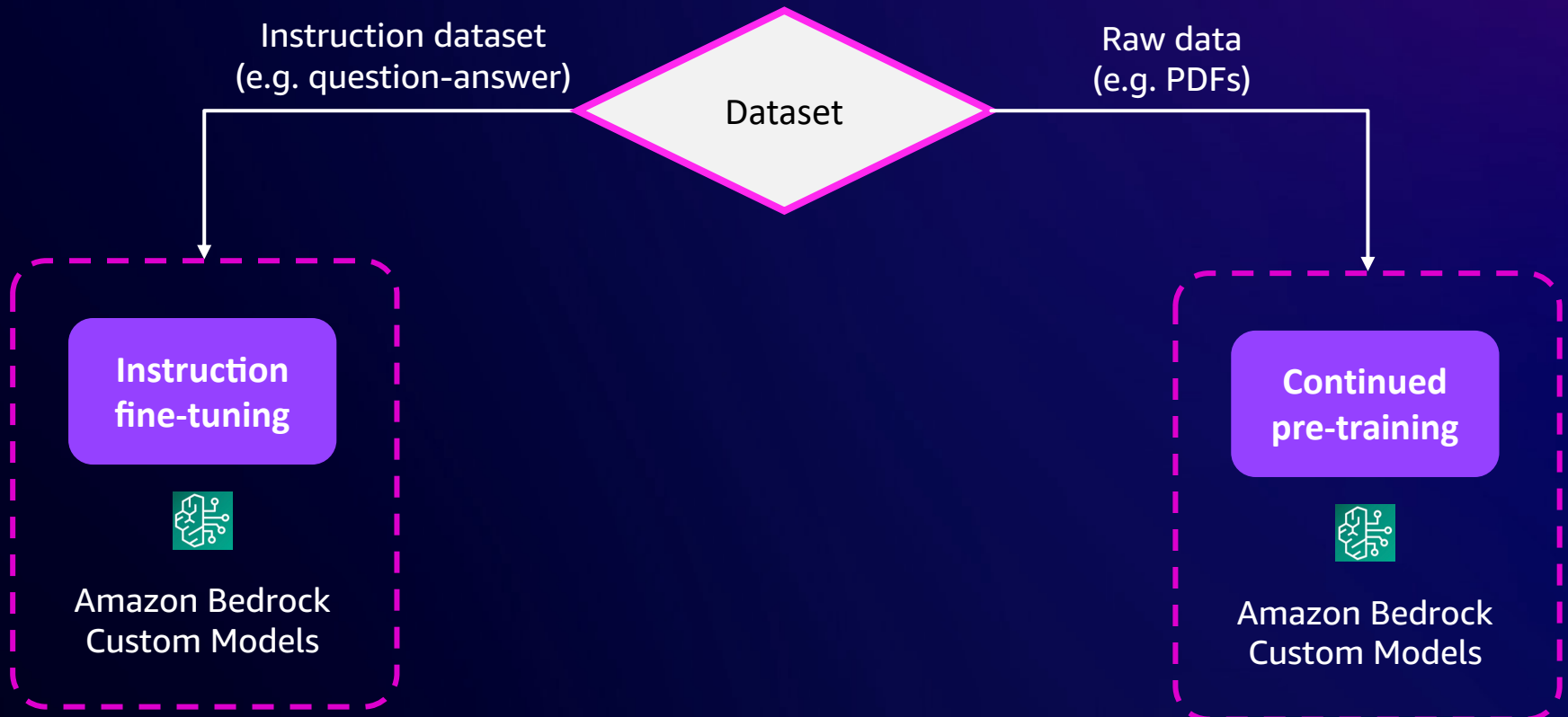
Enhance performance for specific tasks

E.g. Customer Service – Improve ability to understand and respond to customer's inquires and complaints

Improve context-awareness in responses

E.g. Legal Services – Better understand case facts and law to provide useful insights for attorneys

Datasets for instruction fine-tuning and continued pre-training



{"prompt": "<prompt text>", "completion": "<expected generated text>"}
{"prompt": "<prompt text>", "completion": "<expected generated text>"}
{"prompt": "<prompt text>", "completion": "<expected generated text>"}

{"input": "<raw text>"}
{"input": "<raw text>"}
{"input": "<raw text>"}

Amazon Bedrock custom models

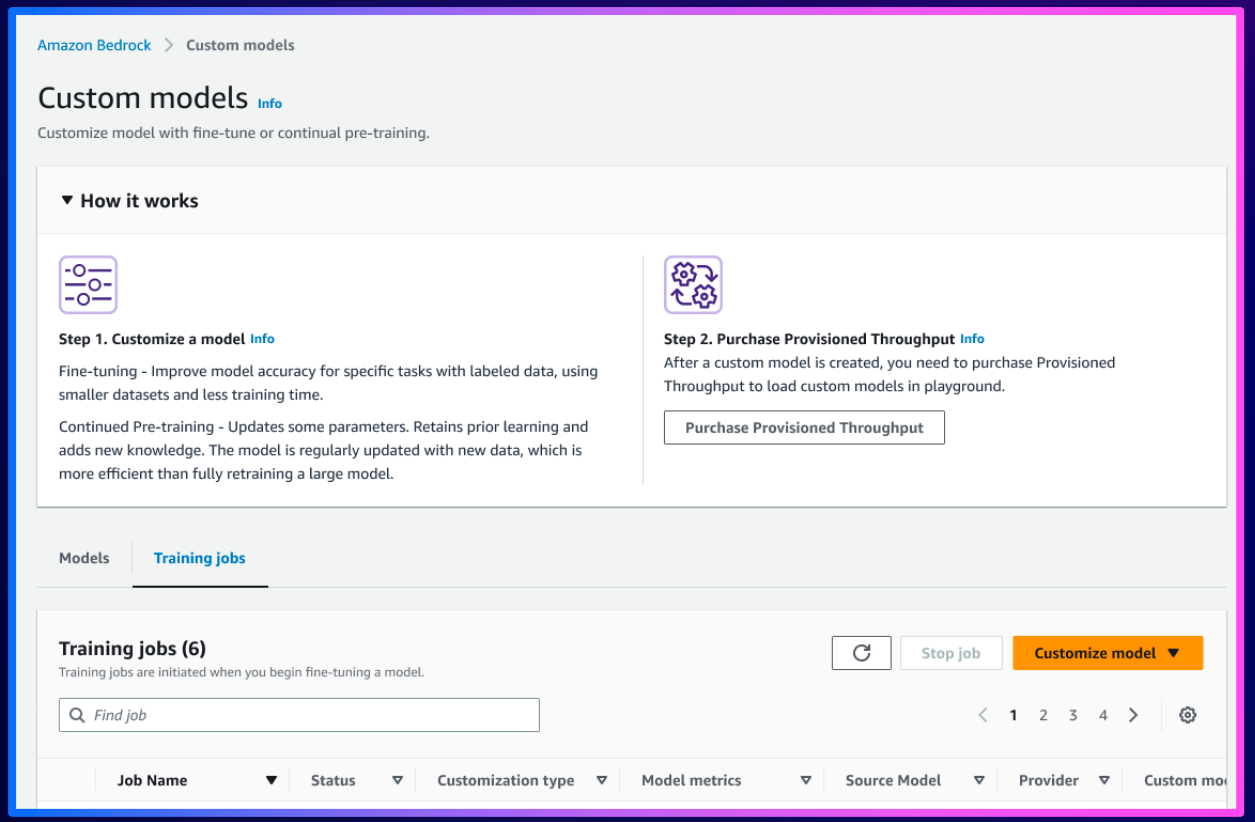
New!

Create custom models using the console or APIs

Maximize accuracy of FMs by providing labeled or raw unlabeled data

Once deployed, custom models are invoked the same way as base models (playground or API)

Customizations now supported for Amazon Titan and some third party FMs



Security and privacy

You are always in control of your data

- ✓ Data **not used** to improve models, and **not shared** with model providers
- ✓ Customer **data remain** in Region
- ✓ Support for **AWS PrivateLink** and **VPC configurations**
- ✓ Integration with **AWS IAM**
- ✓ API monitoring in **AWS CloudTrail**, logging & metrics in **Amazon CloudWatch**
- ✓ Custom models encrypted and stored with **Service or Customer Managed Keys (CMK)** - Only you have access to your models

Fine-Tuning in Action: Text Summarization

```
prompt = """
Summarize the simplest and most interesting part of the following conversation.

#Person1#: Hello. My name is John Sandals, and I've got a reservation.
#Person2#: May I see some identification, sir, please?
#Person1#: Sure. Here you are.
#Person2#: Thank you so much. Have you got a credit card, Mr. Sandals?
#Person1#: I sure do. How about American Express?
#Person2#: Unfortunately, at the present time we take only MasterCard or VISA.
#Person1#: No American Express? Okay, here's my VISA.
#Person2#: Thank you, sir. You'll be in room 507, nonsmoking, with a queen-size bed. Do you approve, sir?
#Person1#: Yeah, that'll be fine.
#Person2#: That's great. This is your key, sir. If you need anything at all, anytime, just dial zero.

Summary:
"""

body = {
    "prompt": prompt,
    "temperature": 0.5,
    "top_p": 0.9,
    "max_gen_len": 512,
}
```

Fine-Tuning in Action: Baseline Completion

```
response = bedrock_runtime.invoke_model(
    modelId="meta.llama2-13b-chat-v1", # compare to chat model
    body=json.dumps(body)
)

response_body = response["body"].read().decode('utf8')
print(json.loads(response_body)["generation"])
```

```
A man named John Sandals checks into a hotel and provides identification and a credit card. The hotel only takes M
asterCard or VISA, so he uses his VISA card. He is given room 507, a nonsmoking room with a queen-size bed.
```

Fine-Tuning in Action: Improved Completion

```
response = bedrock_runtime.invoke_model(
    modelId=provisioned_model_arn, # custom fine-tuned model
    body=json.dumps(body)
)

response_body = response["body"].read().decode('utf8')
print(json.loads(response_body)["generation"])
```

```
John Sandals checks in the hotel with VISA and is assigned room 507, nonsmoking, with a queen-size bed.
```

Simplify generative AI app development with Agents for Amazon Bedrock

Workflow automation challenges

Knowledge workers stretched, need productivity tools

LLM's are powerful, but they can't take actions

Integration of databases and systems is expensive and slow

Building production agents involves complex engineering

Need diverse set of programming languages and interfaces

Agents for Amazon Bedrock

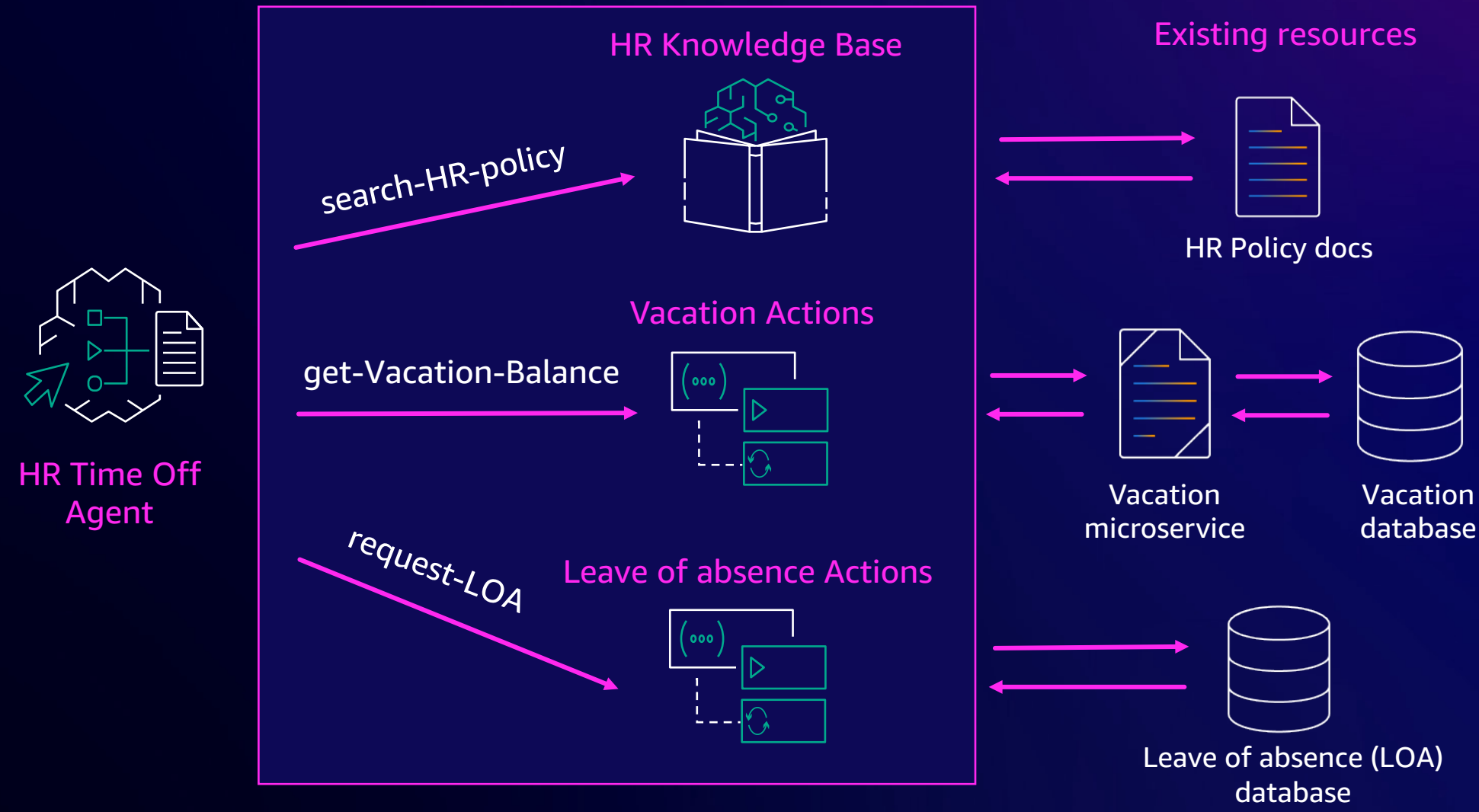
Enable generative AI applications to execute multi-step business tasks using natural language

Generally available

Features

- Uses power of LLM's to prompt and respond using natural language
- Breaks down and orchestrates tasks
- Completes tasks by dynamically invoking APIs
- Securely and privately accesses company data
- Surfaces chain-of-thought trace and underlying agent prompts

Agents build on existing enterprise resources



Agent action groups

Sets of actions made available to your agent to get work done

Each Action Group has 3 key elements

Action Group Description

Overview of actions provided – helps agent know when this action group is relevant

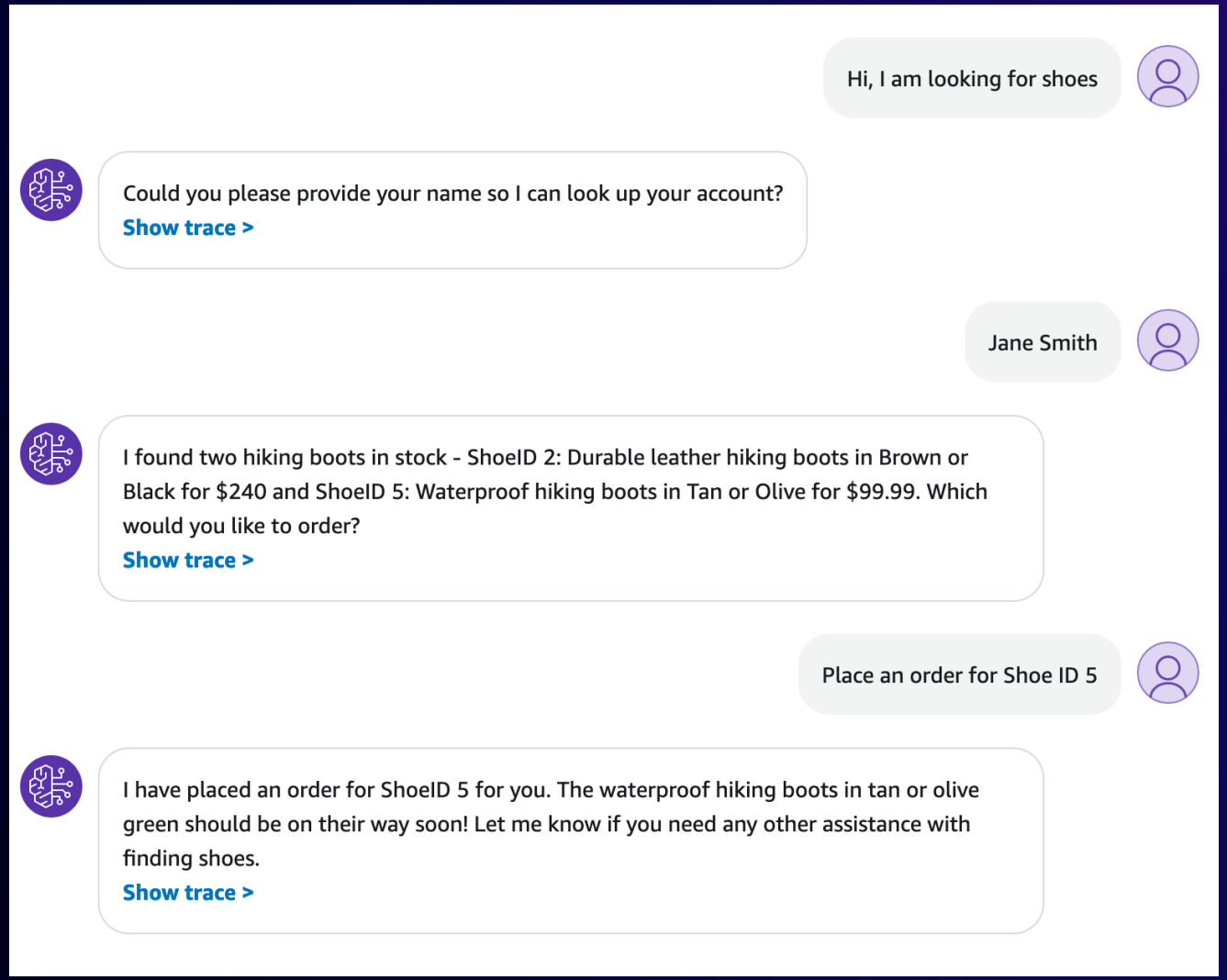
API Schema

- Rich definition of each action
- Operation name, input parameters, data types, response details
- Helps agents know when to use it, how to call it, and how to use results
- Language agnostic API definition using industry-standard schema

Lambda Function

- Implementation of each action
- Contains either business logic or wraps microservices, databases, or tools
- Serverless, scalable, secure
- Choice of programming language (Python, C#, JavaScript, Java, …)

Agent in Action



https://github.com/aws-samples/agentsforbedrock-retailagent/tree/main/

Build responsible AI applications with Guardrails for Amazon Bedrock

Many foundation models have built-in protections

Building generative AI apps requires additional controls

Customizations based on use cases & organizational policy

Safety and privacy controls for responsible AI

Consistent safeguards across FMs and applications

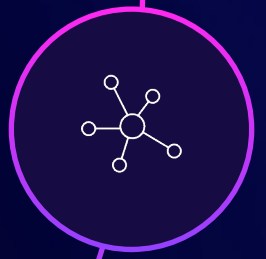
PREVIEW

Guardrails for Amazon Bedrock

Implement safeguards customized to your application requirements and responsible AI policies

Apply guardrails to multiple foundation models and Agents for Amazon Bedrock

Configure harmful content filtering based on your responsible AI policies

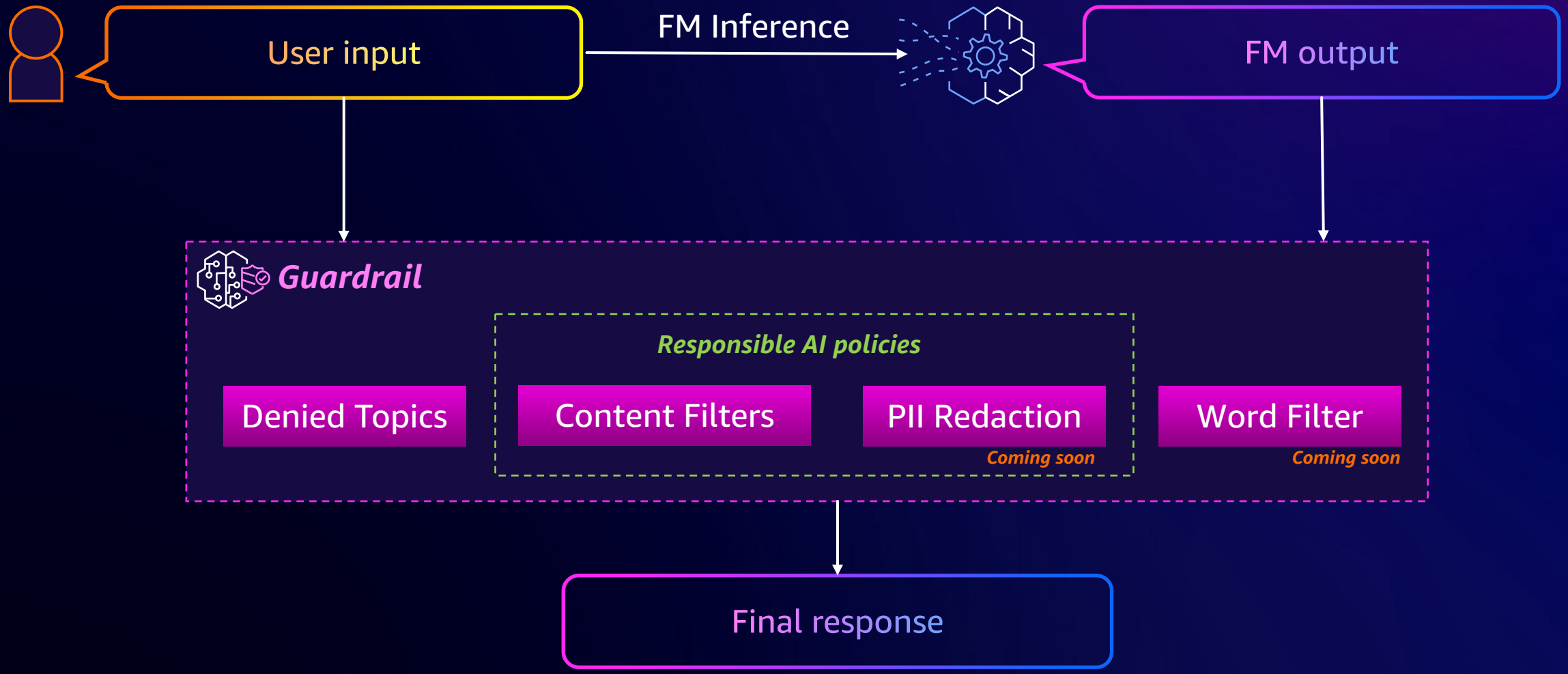


Define and disallow denied topics with short natural language descriptions

COMING SOON

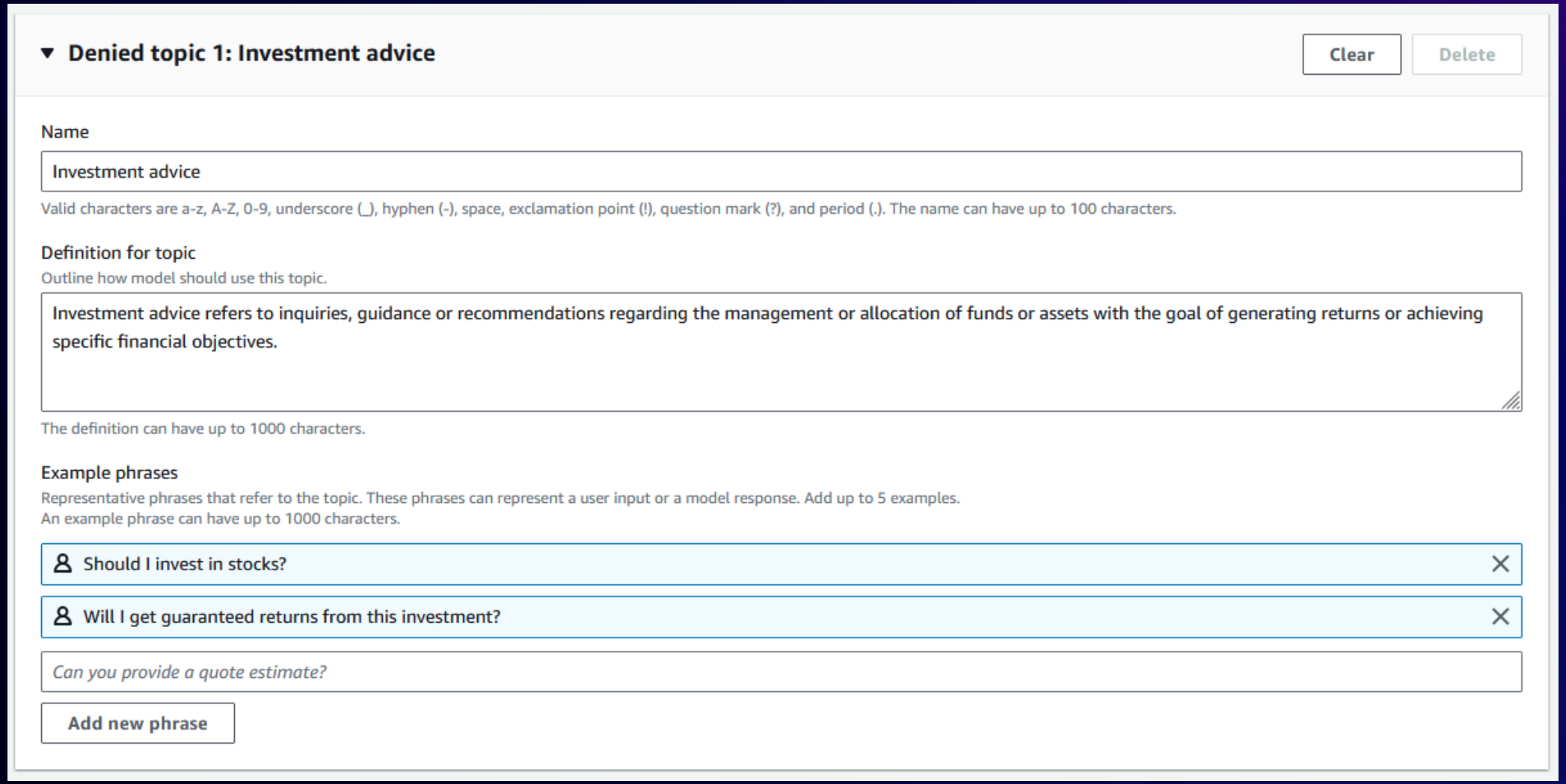
Redact sensitive PII information in FM responses

How it works: Guardrails for Amazon Bedrock



Denied Topics

AVOID UNDESIRABLE TOPICS IN YOUR APPLICATIONS

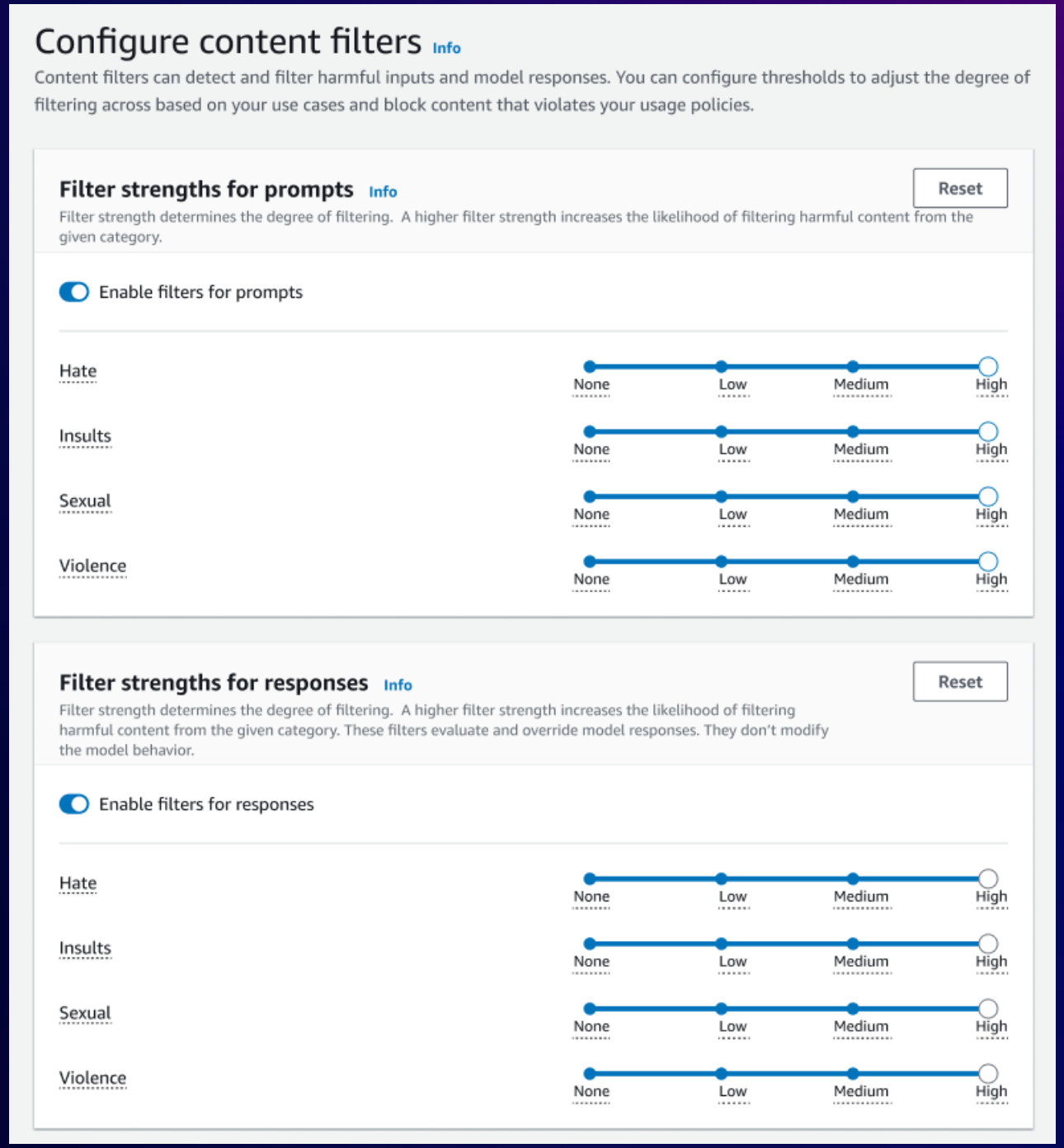


Content Filters

CONFIGURE THRESHOLDS TO FILTER CONTENT TO VARYING DEGREES

Filter harmful content across categories:

- Hate
- Insults
- Sexual
- Violence



Word Filters

COMING SOON

- Define a set of custom words to block in user input and FM responses
- Filter profane words
- Choose to respond with a preconfigured message or mask the blocked words

PII Redaction

COMING SOON

- Redact personally identifiable information (PII) in FM responses to protect user privacy
- Detect and filter PIIs in user inputs
- Select from a variety of PIIs based on application requirements

Evaluate and compare FMs for your use case in Amazon Bedrock

What is model evaluation?

QUALITY

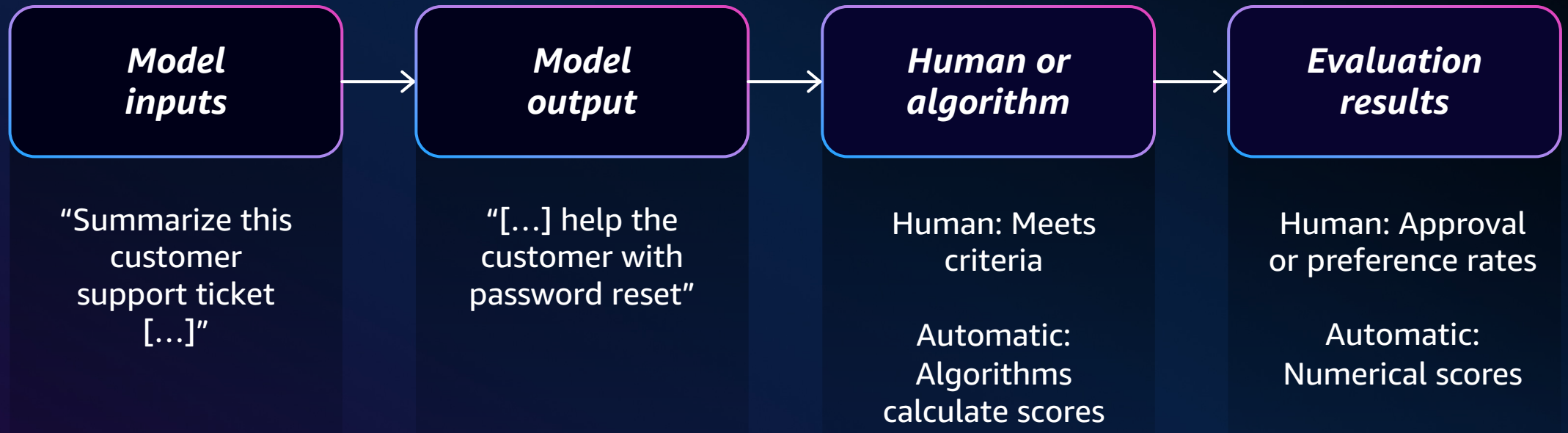
TRADEOFF

COST

LATENCY

What is model evaluation?

1. Quality



New

PREVIEW

Model evaluation on Amazon Bedrock

Evaluate, compare, and select the best foundation model for your use case

1. Use curated datasets or bring your own for tailored results
2. Use automatic or human evaluation methods
3. Leverage your in-house team or AWS-managed reviewers
4. Predefined and custom metrics
5. Get results in just a few clicks

aws

Use curated datasets or bring your own

Evaluate performance in your domain

Identify FM knowledge gaps

Assess areas for model customization

Track performance through the customization process

Verify fairness and detect unwanted biases

Using automatic or human evaluation

Automatic evaluation

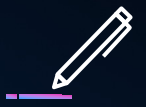
Accuracy

Robustness

Toxicity

Human evaluation

Creativity



Style

Tone

Relevance

Coherence

Brand voice

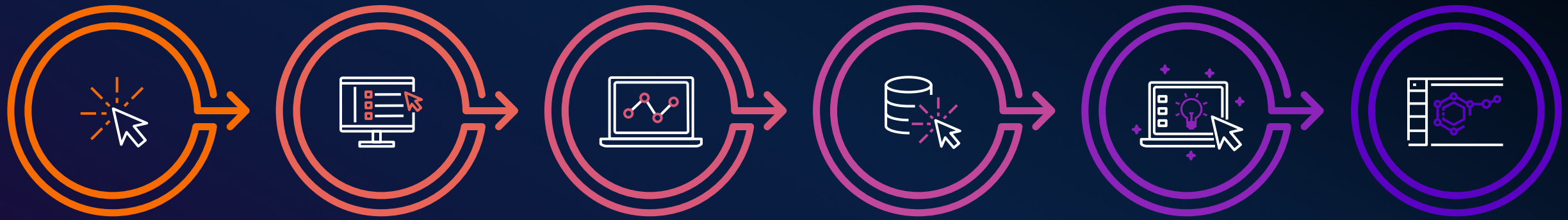
Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

Methods

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

How automatic evaluation works



Choose model

Choose task type

Choose metrics

Choose/ upload datasets

Inference and scoring

View results

How human evaluation works

(Bring your own team)

Choose 1 or 2 models

Choose task type

Use recommended metrics or define your own

Upload dataset

Add people to work team

Inference

Human evaluation

View results

How human evaluation works
(AWS-managed team)

Enter contact info

Describe evaluation

Tasks

Expertise needed

Dataset size

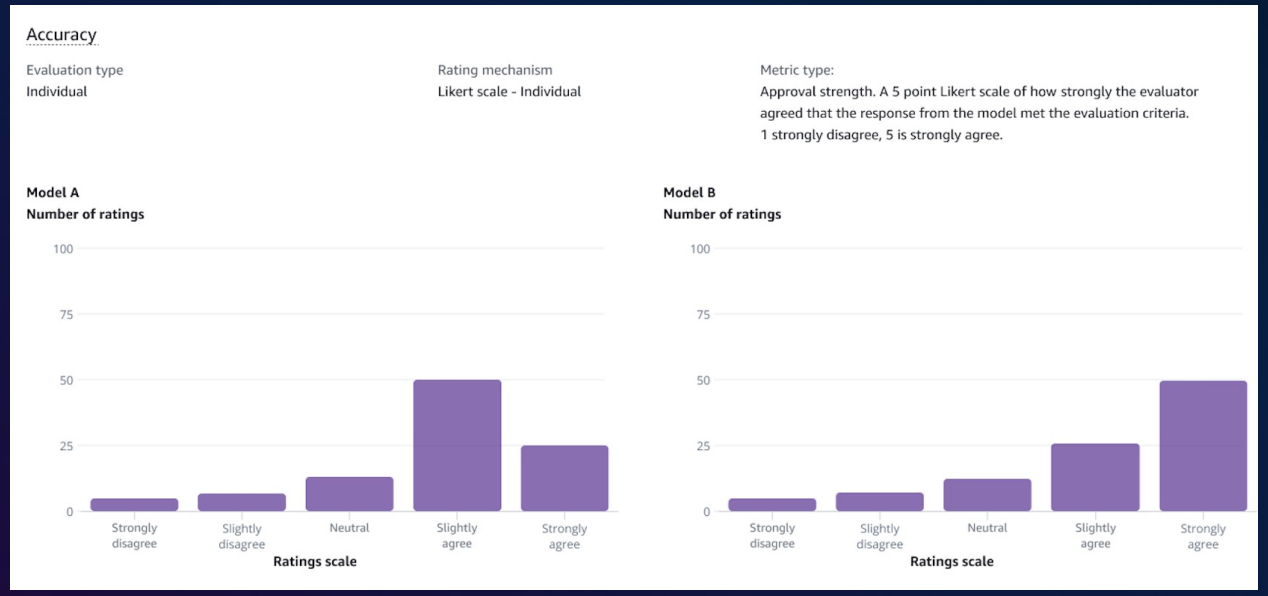
Consultation call with AWS experts

Sign SOW

Receive results

Get results in a few clicks

Human evaluation reports

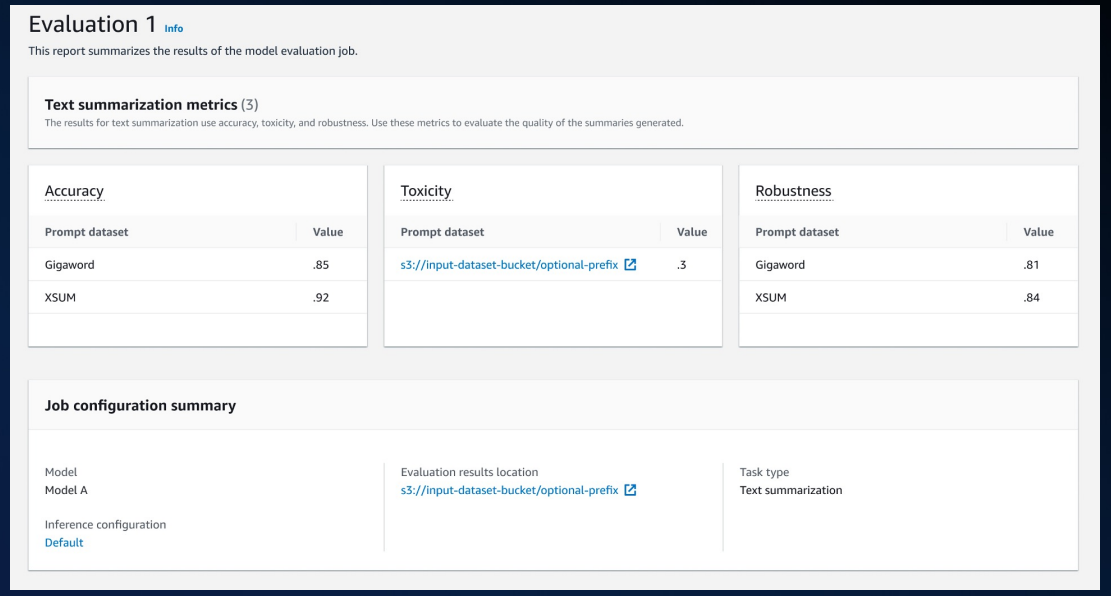


Track ratings from your team

See distributions visually

Simple metric explanations

Automatic evaluation reports



Simple-to-read scores

Curated per task type

Popular algorithms utilized

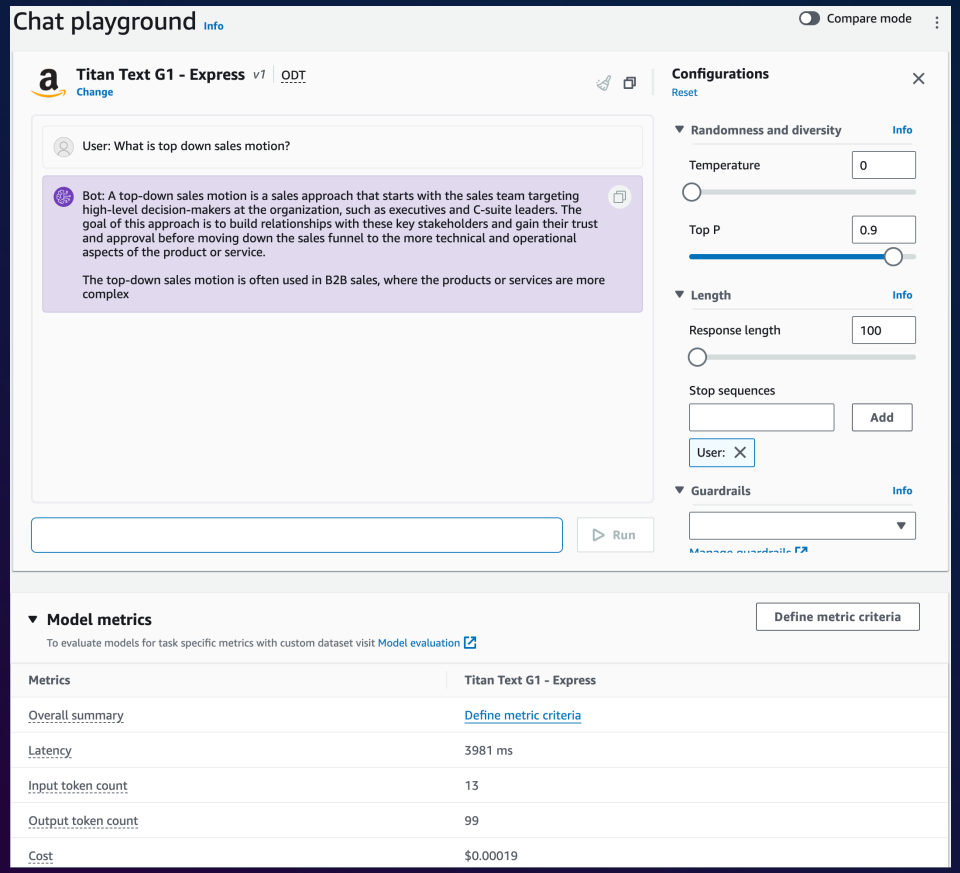
Quality/cost/latency tradeoffs

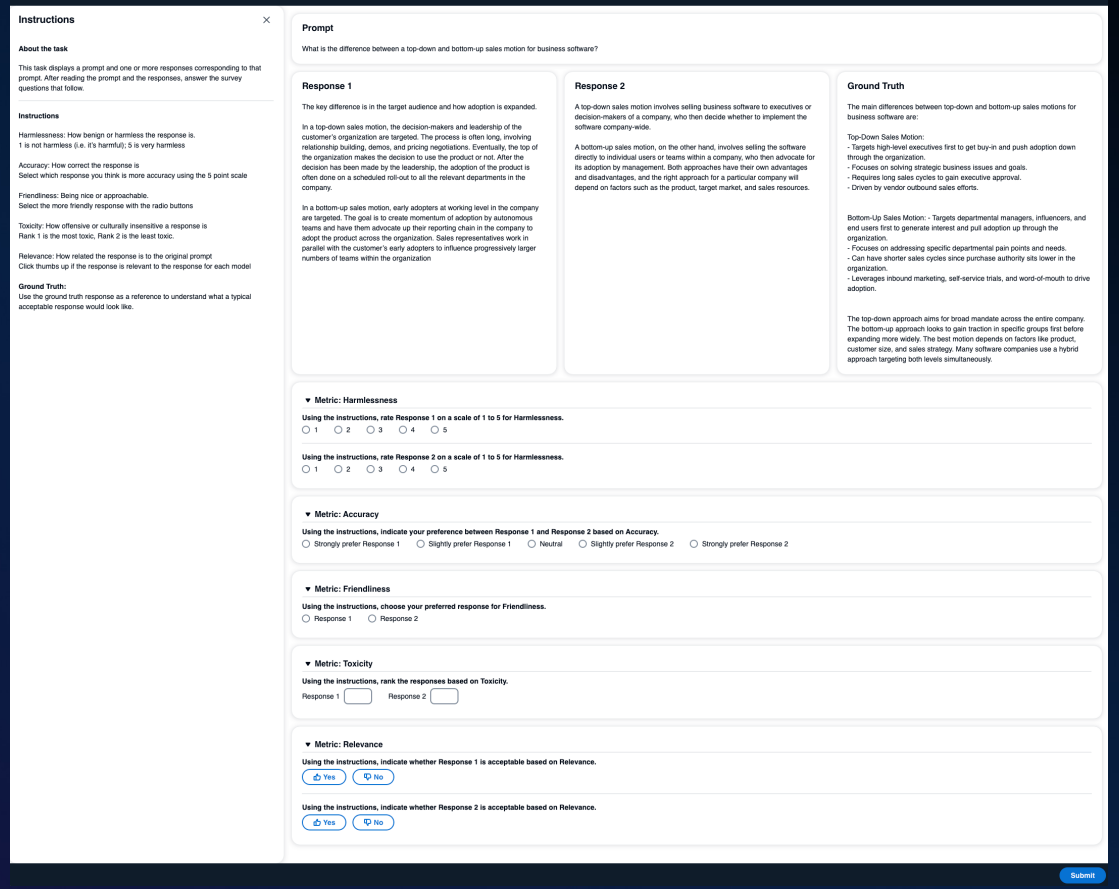
Cost and latency

Amazon Bedrock playgrounds

Quality metrics

Detailed evaluations
Human or automatic





Bring the power of generative AI to your employees with Amazon Q

Introducing

Amazon Q 

Your generative AI assistant designed for work that can be tailored to your business, data, code, and operations

Amazon Q areas of expertise

Your business

Building on AWS

Amazon QuickSight

Amazon Connect

AWS Supply Chain

Key features

End Users

Conversational question-answering on enterprise data

Upload files and analyze content

Execute actions across enterprise apps

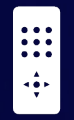
Generate content

Admins

Fully managed solution

Pre-built connectors, vector index and end user application

Permissions-aware responses

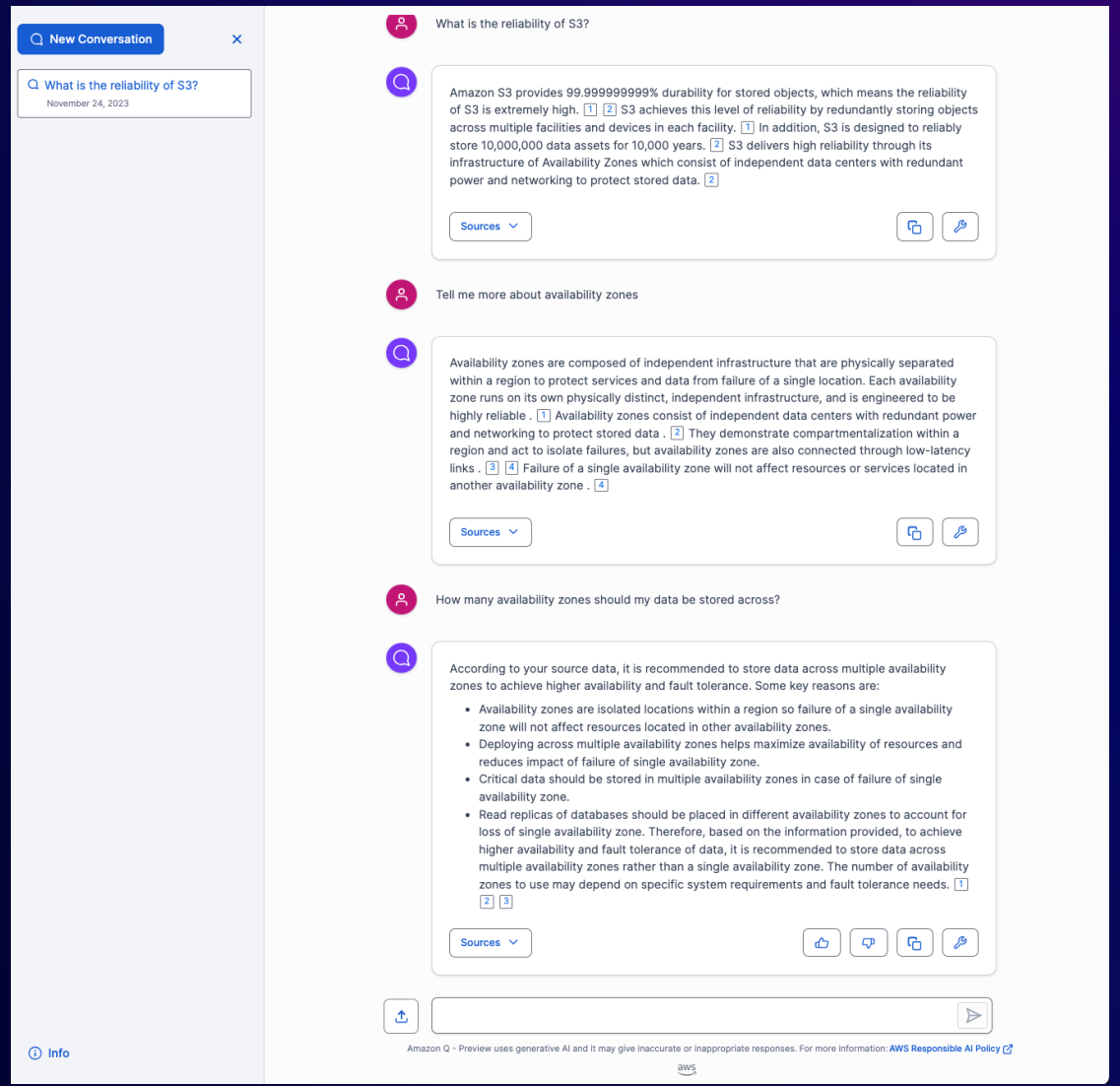


Customize and control Amazon Q with guardrails

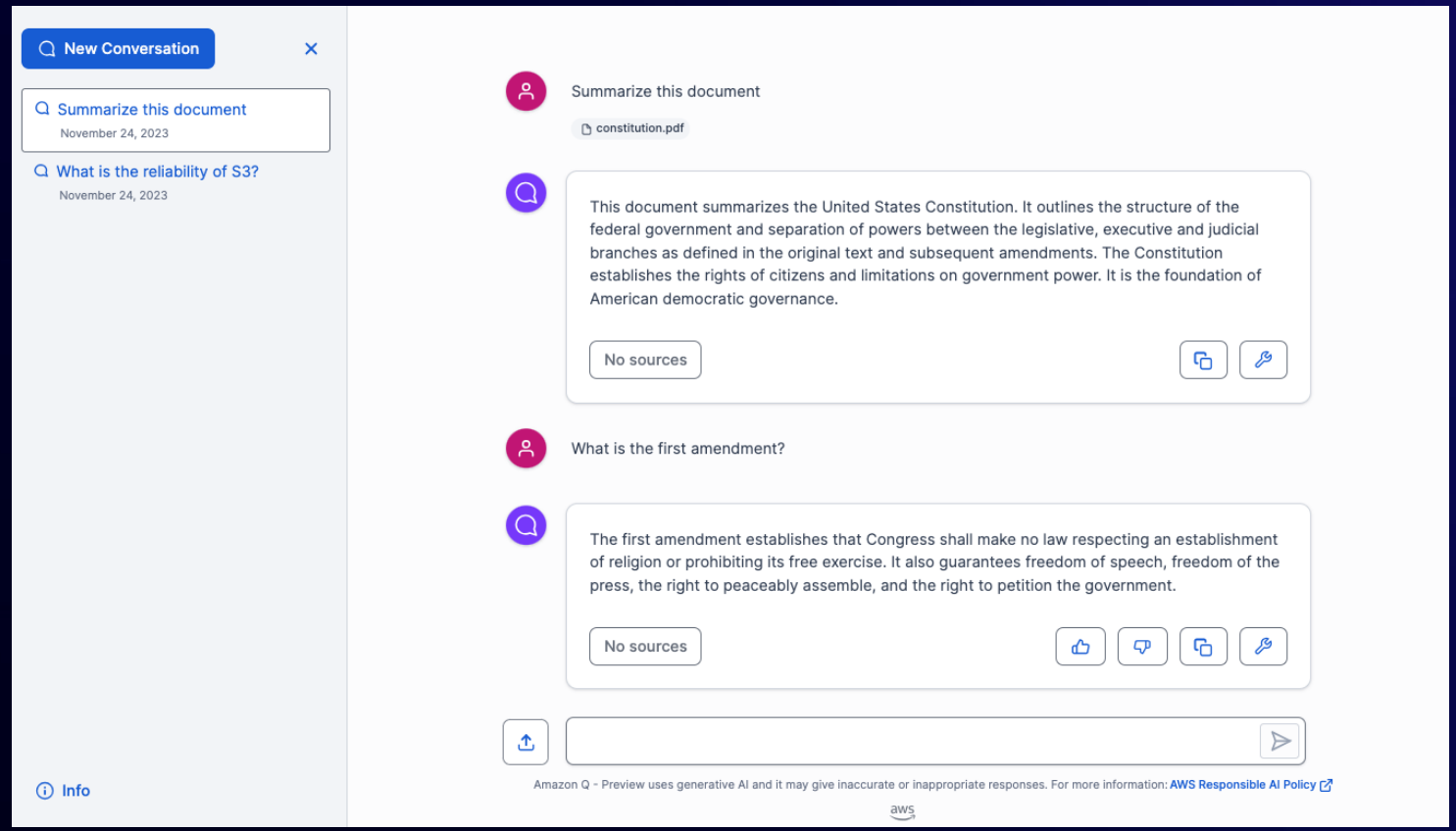
Key features – Conversational Q&A

LET'S GET TO WORK

- Trusted answers generated from enterprise data
- In-context conversations
- Source references for fact-checking
- Conversation history

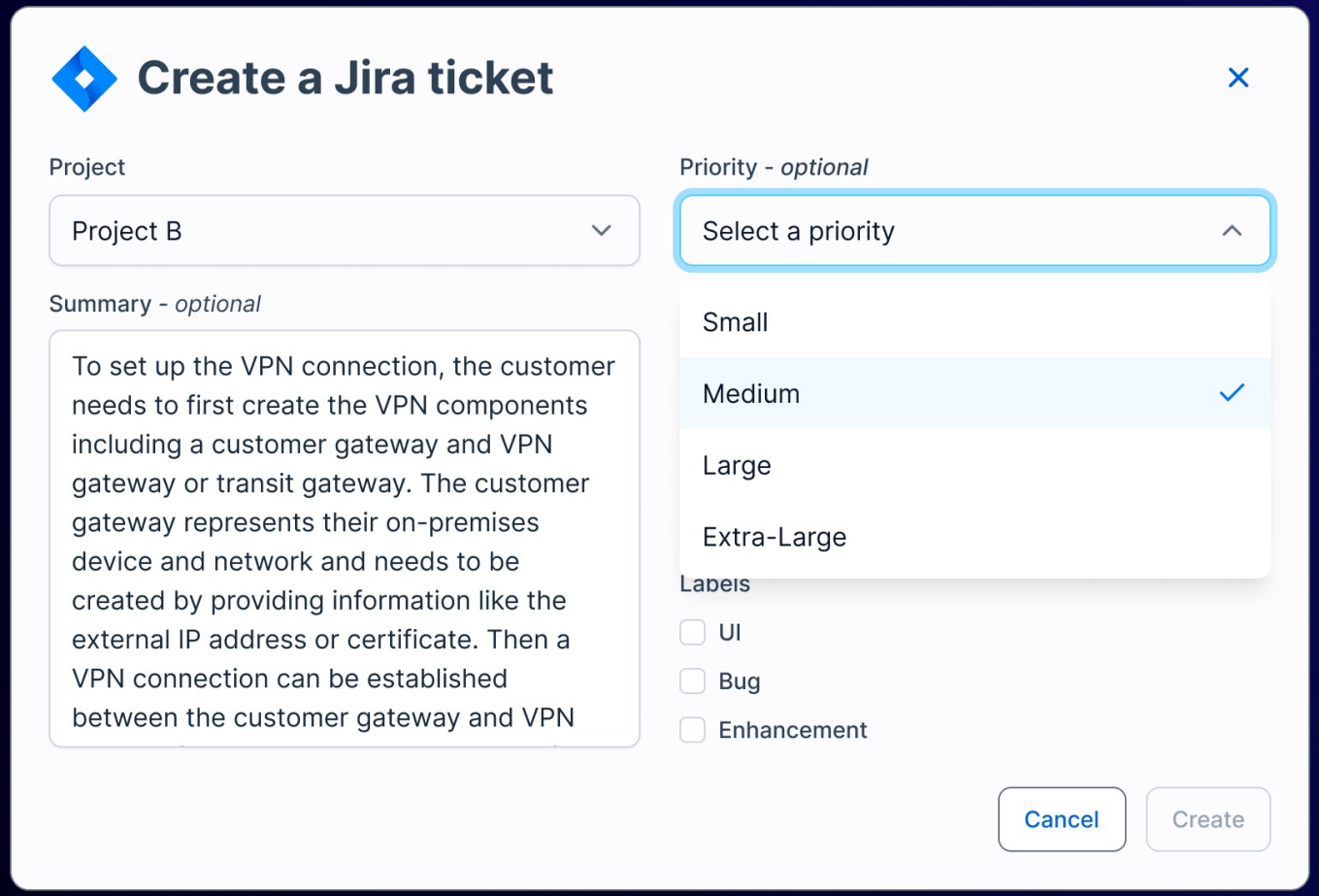


Key features – Upload files and analyze content



- Summarize documents
- Ask questions about content in the documents
- Analyze data
- Multiple file types supported – pdf, docx, csv, ppt, txt and more

Key features - Execute actions using plugins

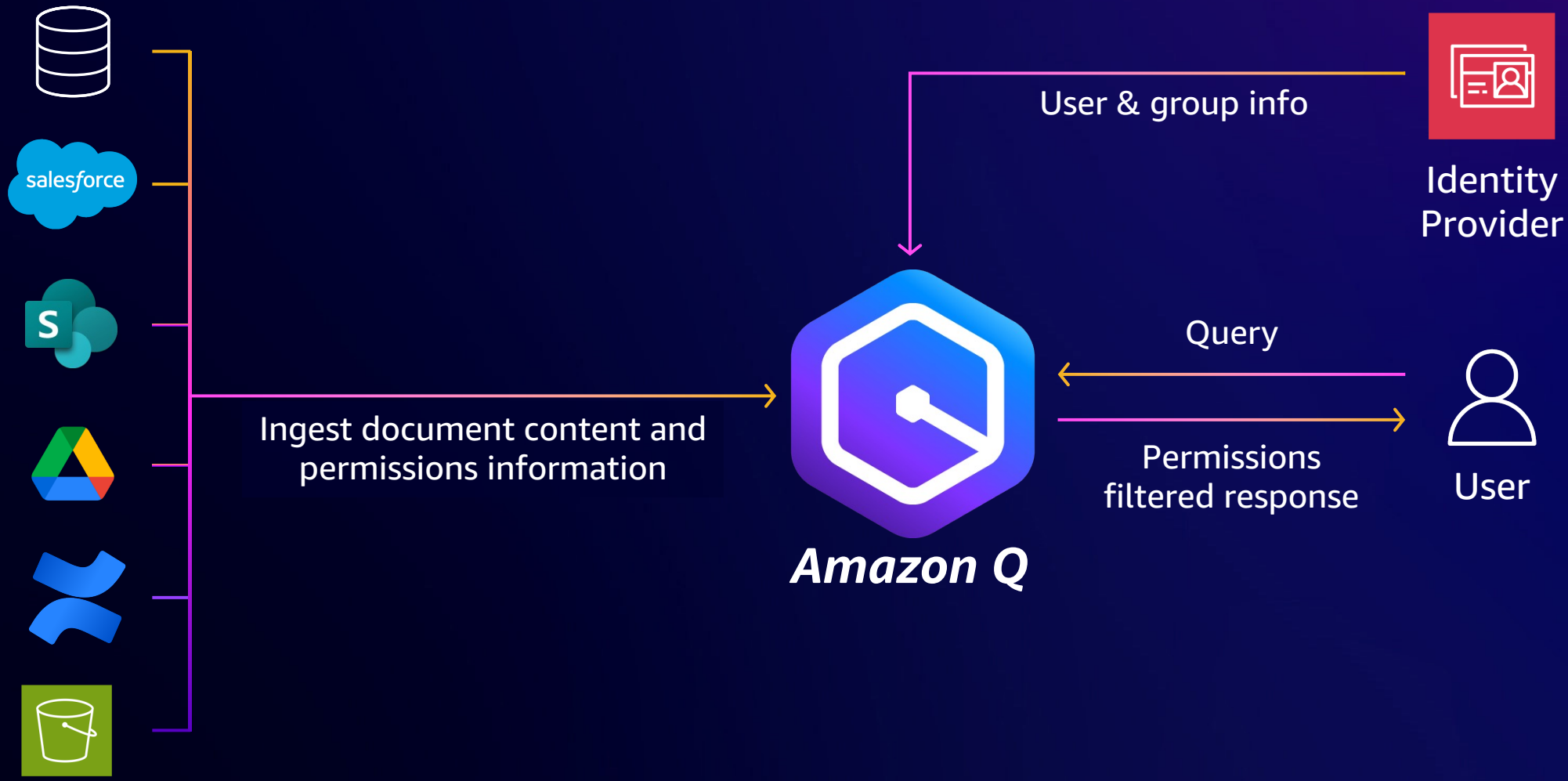


- Enable end-users to perform actions on SaaS applications
  - *"summarize conversation and create ticket in Jira"*

zendesk salesforce Jira ServiceNow

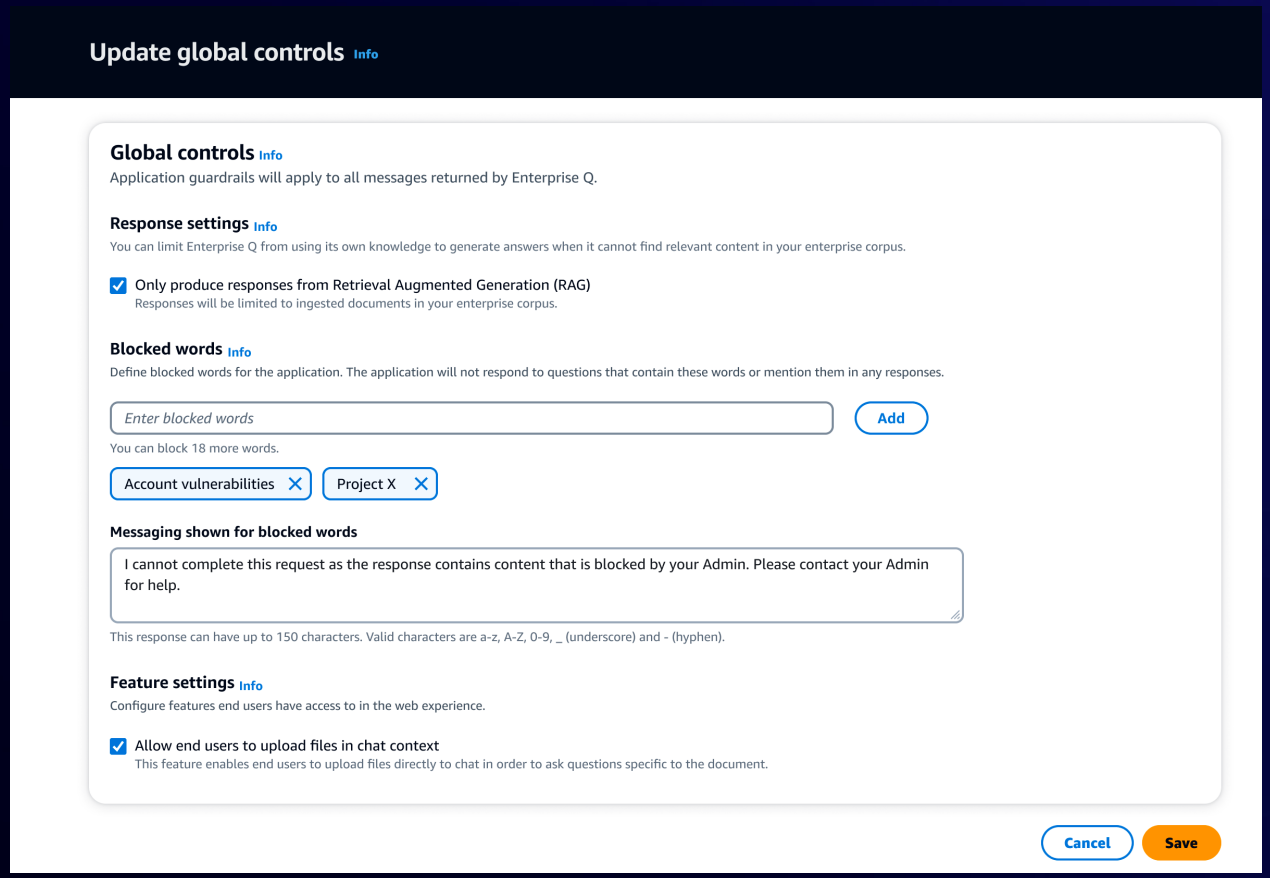
Key features – Safety and security

BUSINESS Q IS AWARE OF ENTERPRISE USER PERMISSIONS



Key features – Safety and security

ADD GUARDRAILS TO THE EXPERIENCE



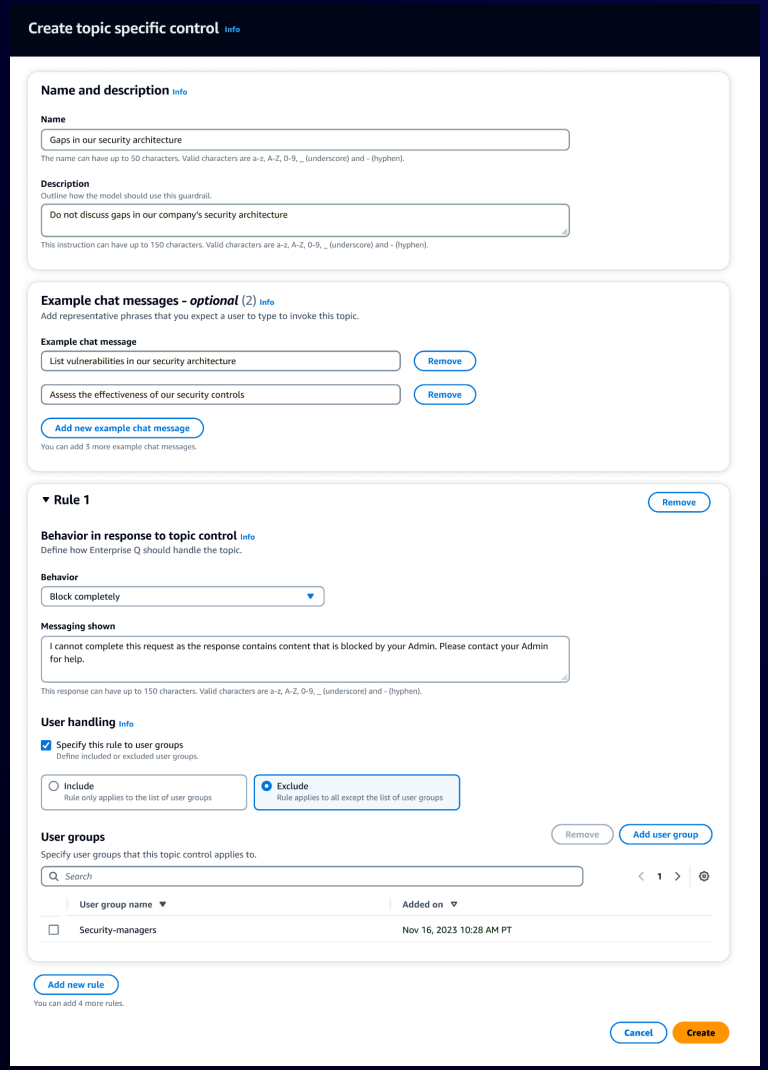
Use pre-built guardrails for toxicity

Restrict responses to enterprise content only

Specify blocked words or phrases that never appear in responses

Key features – Safety and security

ADD GUARDRAILS TO THE EXPERIENCE



Define special topics and configure 4 levers of guardrails for such topics as

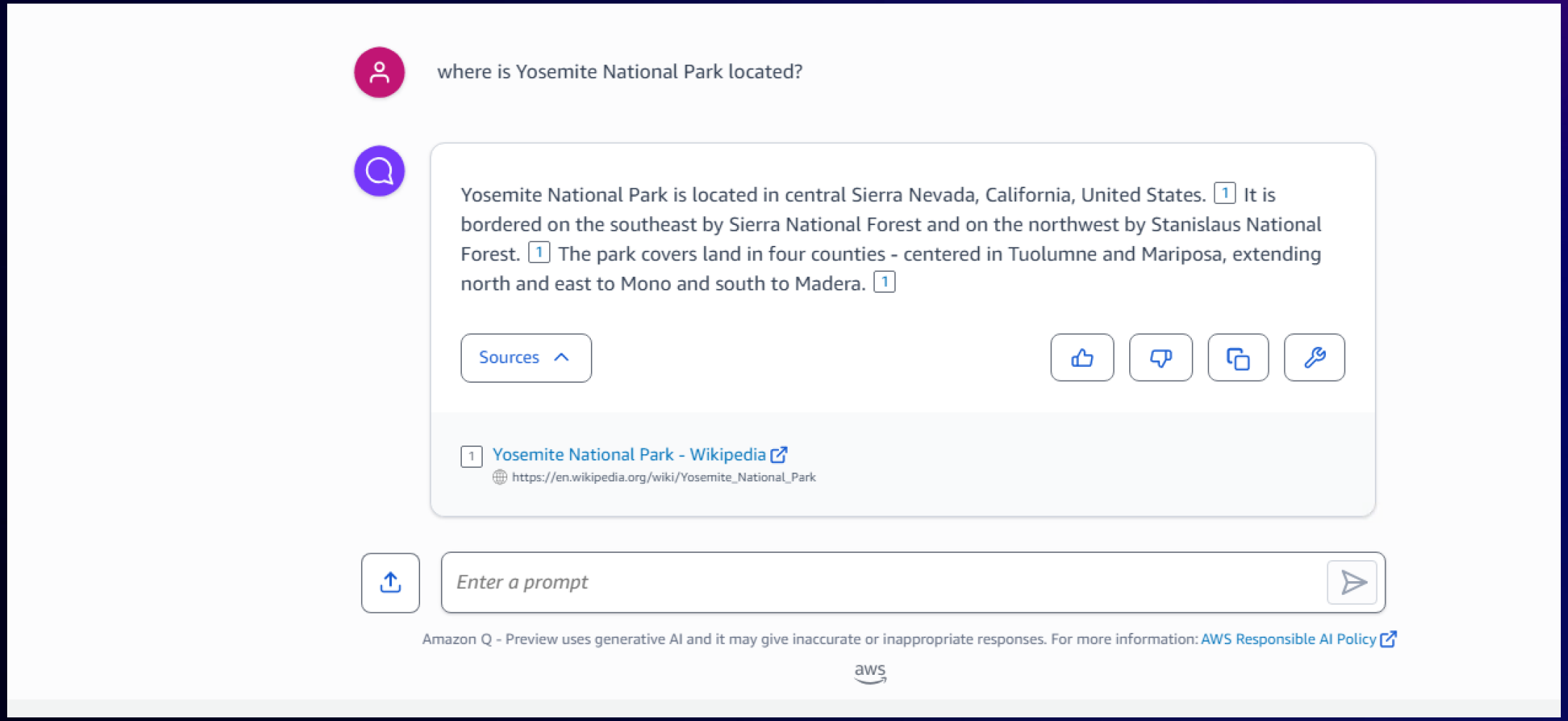
1. Respond with an pre-defined message
2. Restrict responses to enterprise content
3. Restrict responses to enterprise content with metadata filters
4. Apply guardrails to specific users and groups in the enterprise

Get started quickly with connectors

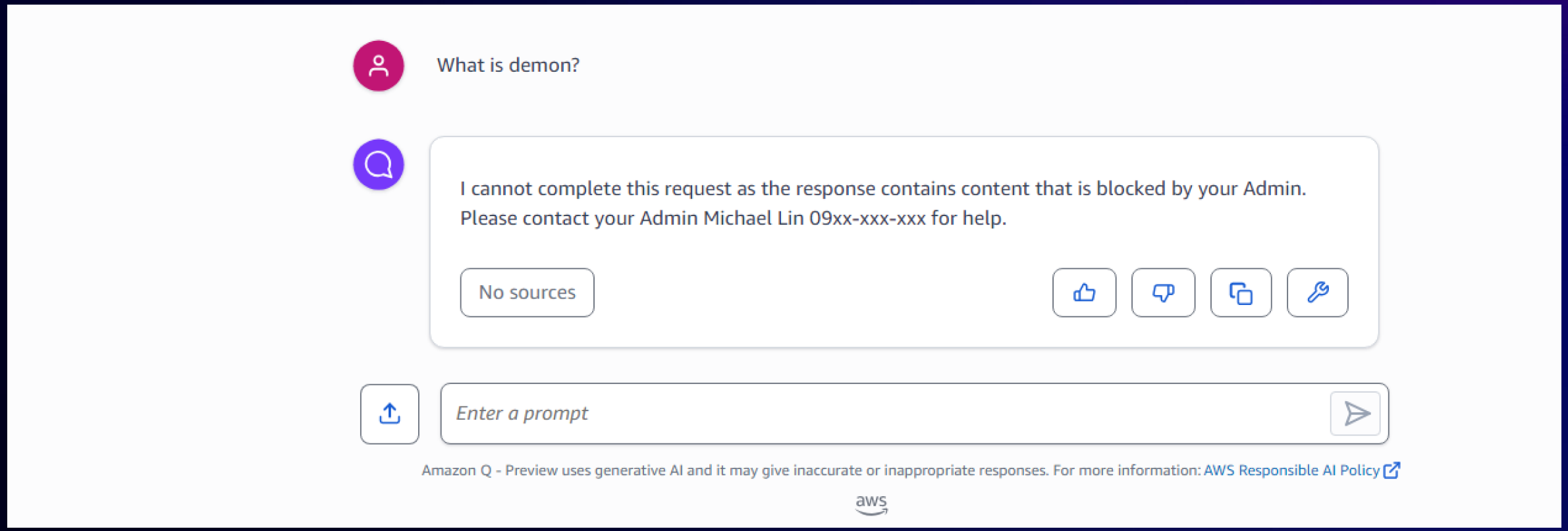
Bring content from all your enterprise sources together in a few clicks!

- Adobe Experience Manager
- Alfresco
- Amazon Simple Storage Service (Amazon S3)
- Atlassian Confluence
- Aurora (MySQL, PostgreSQL)
- Box
- DB2
- Dropbox
- Drupal
- Custom Connector
- FSX for Windows
- Github
- Gmail
- Google Drive
- Jira
- Microsoft Exchange
- Microsoft OneDrive
- Microsoft SharePoint
- Microsoft Teams
- Microsoft Yammer
- Microsoft SQL Server
- Quip
- Salesforce
- ServiceNow
- Slack
- Web Crawler
- WorkDocs
- Zendesk

Amazon Q in Action with Web Crawler



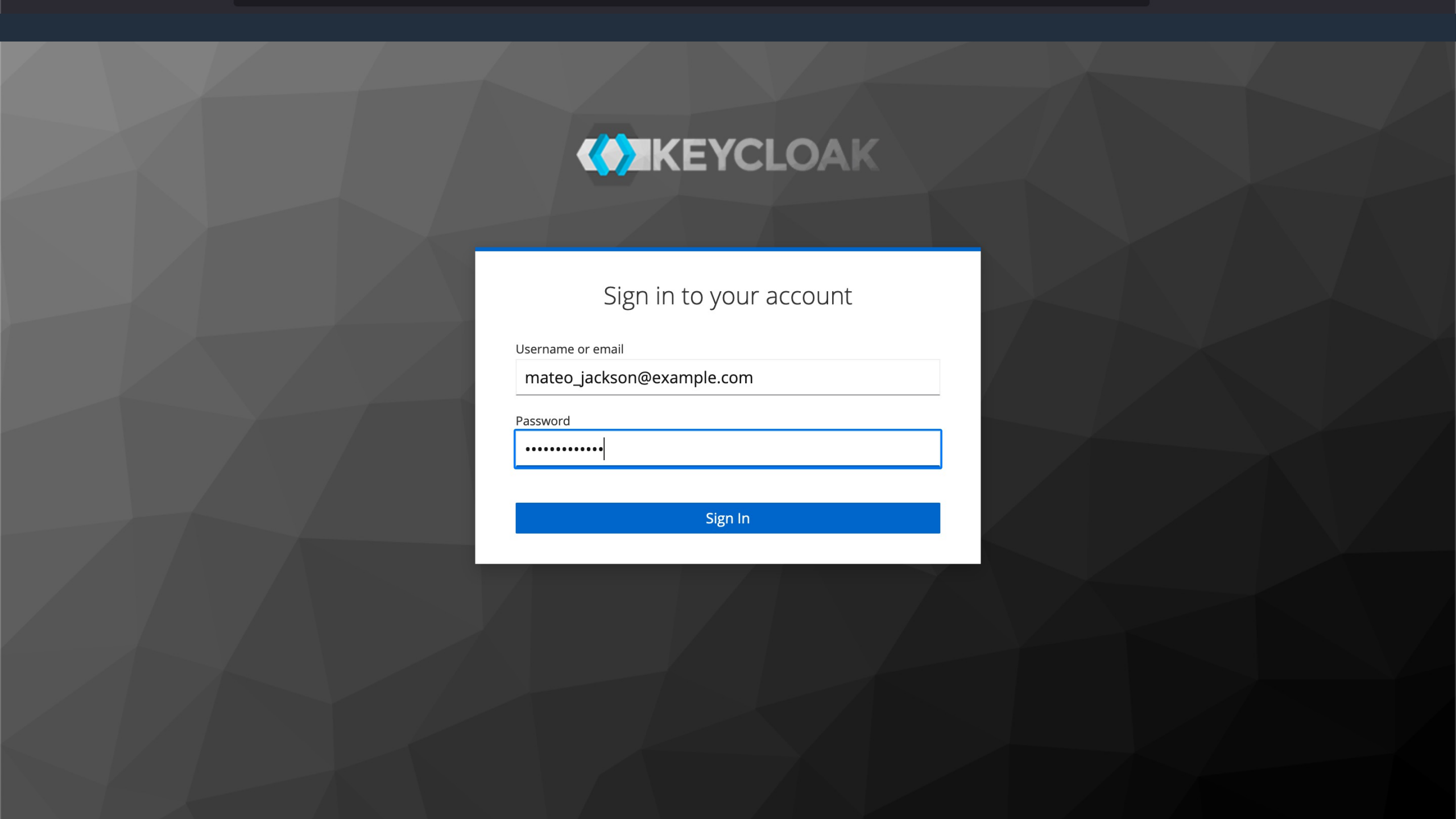
Amazon Q in Action with Guardrails



Authorized Access based on ACLs

- pat_candella - SA (group of solutions architects)
- mateo_jackson - DB_SME_SA (group of database subject matter expert solutions architects)
- john_doe - ML_SME_SA (group of machine learning subject matter expert solutions architects)
- mary_major - (does not belong to any of these groups)
- martha_rivera - Admins (group of administrators)

```
    {
        "keyPrefix": "s3://NAME-OF-DATASOURCE-S3-BUCKET/Data/Databases/",
        "aclEntries": [
            {
                "Name": "DB_SME_SA",
                "Type": "GROUP",
                "Access": "ALLOW"
            },
            {
                "Name": "Admins",
                "Type": "GROUP",
                "Access": "ALLOW"
            }
        ]
    },
```

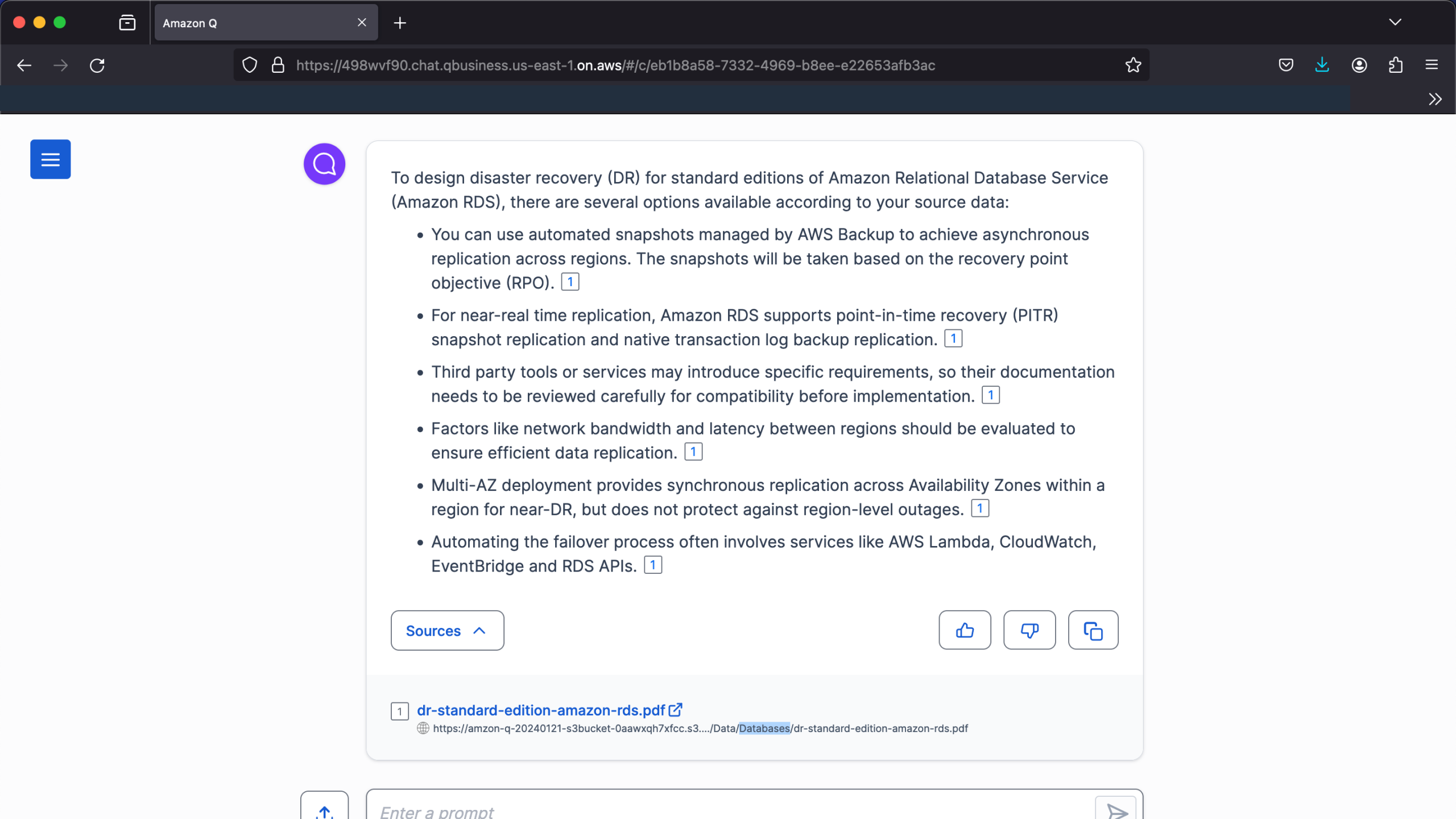

KEYCLOAK
Sign in to your account
Username or email
mateo_jackson@example.com
Password
Sign In

How to design DR for standard RDS?

To design disaster recovery (DR) for standard editions of Amazon Relational Database Service (Amazon RDS), there are several options available according to your source data:

- You can use automated snapshots managed by AWS Backup to achieve asynchronous replication across regions. The snapshots will be taken based on the recovery point objective (RPO). [1]
- For near-real time replication, Amazon RDS supports point-in-time recovery (PITR) snapshot replication and native transaction log backup replication. [1]
- Third party tools or services may introduce specific requirements, so their documentation needs to be reviewed carefully for compatibility before implementation. [1]
- Factors like network bandwidth and latency between regions should be evaluated to ensure efficient data replication. [1]
- Multi-AZ deployment provides synchronous replication across Availability Zones within a region for near-DR, but does not protect against region-level outages. [1]
- Automating the failover process often involves services like AWS Lambda, CloudWatch, EventBridge and RDS APIs. [1]

Sources


Amazon Q
https://498wvf90.chat.qbusiness.us-east-1.on.aws/#/c/eb1b8a58-7332-4969-b8ee-e22653afb3ac
To design disaster recovery (DR) for standard editions of Amazon Relational Database Service (Amazon RDS), there are several options available according to your source data:
You can use automated snapshots managed by AWS Backup to achieve asynchronous replication across regions. The snapshots will be taken based on the recovery point objective (RPO). 1
For near-real time replication, Amazon RDS supports point-in-time recovery (PITR) snapshot replication and native transaction log backup replication. 1
Third party tools or services may introduce specific requirements, so their documentation needs to be reviewed carefully for compatibility before implementation. 1
Factors like network bandwidth and latency between regions should be evaluated to ensure efficient data replication. 1
Multi-AZ deployment provides synchronous replication across Availability Zones within a region for near-DR, but does not protect against region-level outages. 1
Automating the failover process often involves services like AWS Lambda, CloudWatch, EventBridge and RDS APIs. 1
Sources
1 dr-standard-edition-amazon-rds.pdf
https://amzon-q-20240121-s3bucket-0aawxqh7xfcc.s3..../Data/Databases/dr-standard-edition-amazon-rds.pdf
Enter a prompt

Agenda Revisited

- Bedrock New Models
- SageMaker JumpStart for LLMs
- RAG and Knowledge Base
- Fine-tuning and Pre-training
- Automation and Agent

Agenda Revisited

- Responsible AI and Guardrails
- Model Evaluation
- AI Assistant and Amazon Q

Thank you!

Please complete the session survey in the mobile app

Michael Lin

linmicht@amazon.com