

AWS re:Invent 2023 *Generative AI Recap*

Michael Lin

Sr. Solutions Architect
Amazon Web Services



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

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

Agenda

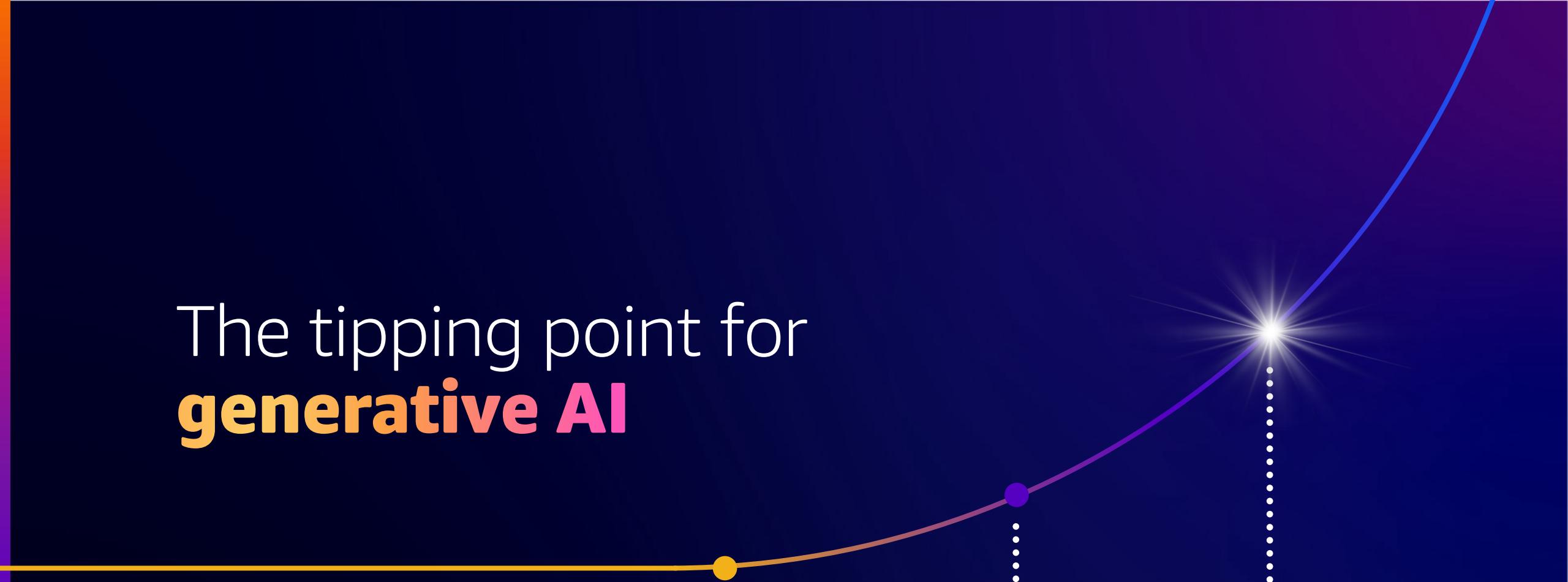
- Responsible AI and Guardrails
- Model Evaluation
- AI Assistant and Amazon Q

*Build your first generative AI
application with Amazon Bedrock*



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The tipping point for **generative AI**

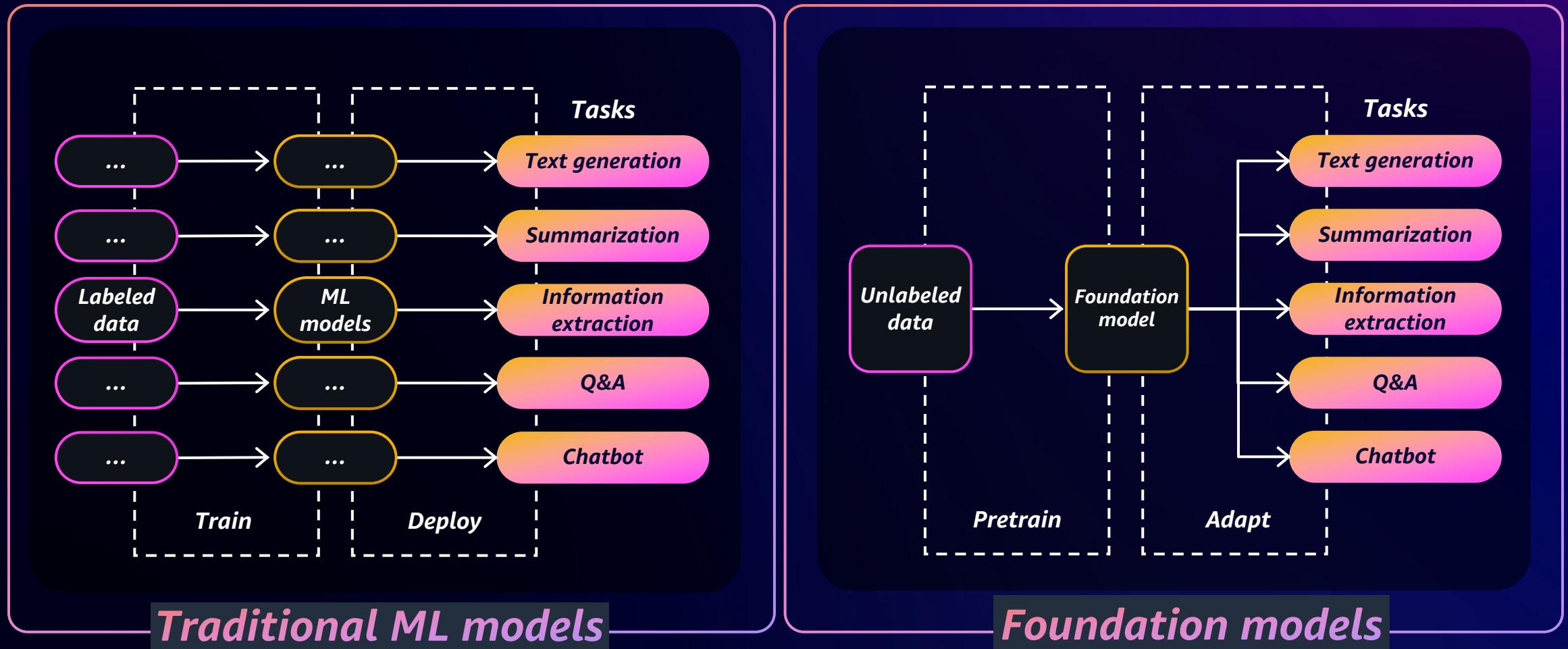


*Massive proliferation
of data*

*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

chegg

lonely planet

cimpress

PHILIPS

IBM | The Weather Company

nextiot

KONE

Sun Life

Neiman Marcus

RYANAIR

hellmann
WORLDWIDE LOGISTICS

WPS Office
Make It Simple

twilio

BRIDGEWATER

Showpad

coda

Booking.com



Amazon ***Bedrock***

Broad choice of foundation models

AI21labs

ANTHROPIC

cohere

Meta

stability.ai

amazon



Jurassic-2

Contextual answers,
summarization, paraphrasing

Claude

Summarization, complex
reasoning, writing, coding

Command & Embed

Text generation, search,
classification

Llama 2

Q&A and reading
comprehension

Stable Diffusion XL

High-quality images
and art

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/>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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/>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

**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/>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

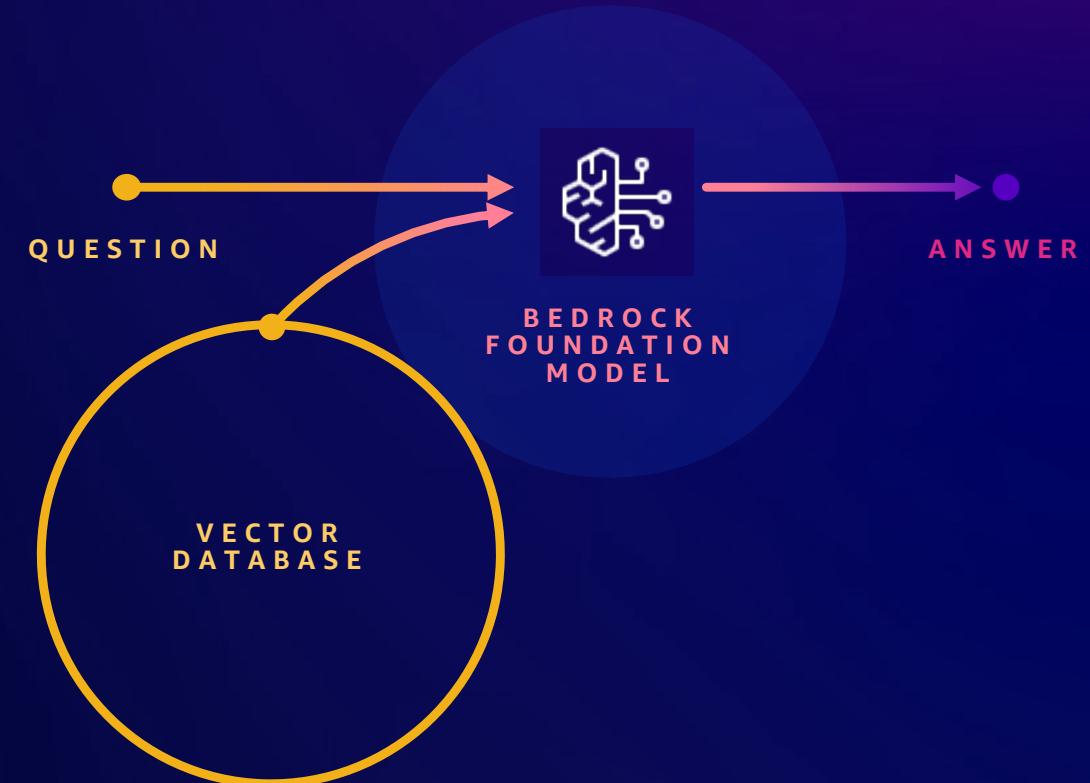
Knowledge base for Amazon Bedrock

USE RETRIEVAL AUGMENTED GENERATION (RAG)

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

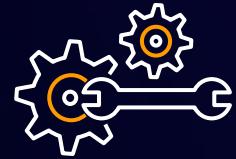
Enable automatic data source detection

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



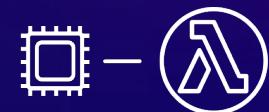
2

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 |

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



Amazon Titan Text

TEXT GENERATION MODELS



Titan Text Lite

Price-performance 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 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



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example

"Blue sneakers
without laces"

Result 1		10 results
1		text: N o Tie Shoelac es -
2		text: S hoe Laces No Tie
3		text: S hoe Laces No Tie
4		text: S hoe Laces No Tie
Result 2		10 results
1		text_description: C apelli New York Toddlers Unisex Slip On Sneaker
2		text_description: T esla Men's Ultra Lightweight Running Shoes L510 PR2
3		text_description: C rocs Men's Norlin Canvas Slip-on Upper: Canvas
4		text_description: B ed Stu Men's Bluegill Slip-On Loafer Slip on

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

Werner Vogels loves wearing white scarfs as he travels around India

Adult Female Person Woman Art Handicraft

Body Part Hand Person Adult Male Man Holding Hands

Person Shoving Adult Male Body Part Hand Holding Hands Machine Wheel

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



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Titan Image Generator

NOW IN PREVIEW



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

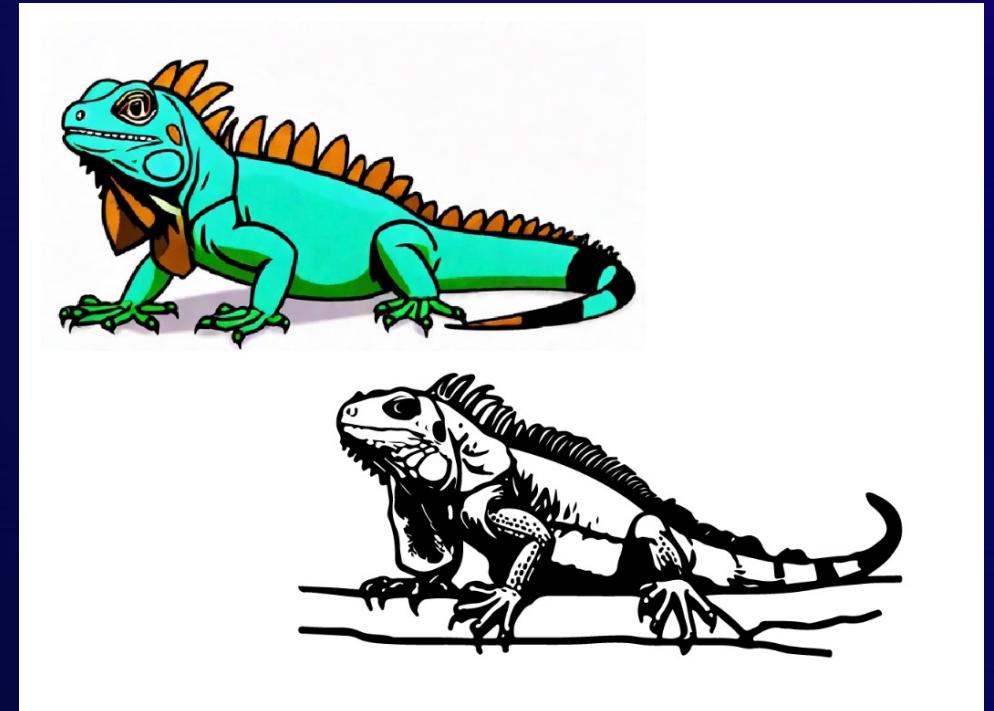


Image Variations

GENERATE VARIATIONS OF AN IMAGE

Amazon Bedrock > Image Playground

Image Playground

Compare mode

Titan Image Generator Change



orange iguana facing right in a rain forest

Run

Configurations

Reset

Mode: Generate

Negative prompt: Add system prompt

Reference image: 

Response image: Select style

Orientation: Landscape (checked)

Size: 512 x 512

Number of images: 3

Advanced configurations

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



Models

Jurassic-2
Ultra, Mid

Models

Llama 2 7B, 13B, 70B
Code Llama 7B, 13B,
34B
Open LlaMA

Models

Command
Cohere Light

Models

Falcon-7B, 40B,
180B,
Mistral 7B
RedPajama
MPT-7B
BloomZ 176B
Flan T-5
DistilGPT2
GPT NeoXT
Bloom

Models

Stable Diffusion
XL 1.0
2.1 base
Upscaling
Inpainting

Models

Lyra-Fr
10B, Mini

Models

Dolly

Models

AlexaTM 20B



Discover foundation models from multiple providers

The screenshot shows the SageMaker JumpStart interface. At the top, there's a navigation bar with 'Home', 'Quick actions' (including 'Open Launcher', 'Import & prepare data visually', and 'Open the'), and a search bar. Below the navigation is a sidebar with sections for 'Prebuilt ai' (Deploy built-in AI), 'Workflow' (Kick off a new workflow), 'Prepare data' (Connect to data, Transform, Store, Manage, Manage EMR), and 'SageMaker JumpStart'. The main content area is titled 'SageMaker JumpStart' and displays two sections: 'Foundation Models: Text Generation' and 'Foundation Models: Image Generation'.
Foundation Models: Text Generation
This section lists three models:

- Llama-2-70b-chat** (Meta AI) - Featured, Text Generation. Details: 70B fine-tuned model optimized for... Fine-tunable: Yes. Source: Meta. Buttons: View model >, View notebook >.
- Llama-2-7b** (Meta AI) - Featured, Text Generation. Details: 7B pretrained model. Fine-tunable: Yes. Source: Meta. Buttons: View model >, View notebook >.
- Jurassic-2 Ultra** (AI21 lab) - Featured, Text To Text. Details: Best-in-class instruction-following model. Fine-tunable: No. Provider: AI21. Buttons: View model >, View notebook >.

Foundation Models: Image Generation
This section lists three models:

- Stable Diffusion XL 1.0** (Stability AI) - Text To Image. Fine-tunable: No. Provider: Stability AI. Details: The leading generation model from... Buttons: View notebook >, View model >.
- Stable Diffusion XL Beta 0.8** (Stability AI) - Text To Image. Fine-tunable: No. Provider: Stability AI. Details: Beta version of SDXL, with native 512... Buttons: View notebook >, View model >.
- Stable Diffusion XL 1.0 (open)** (Stability AI) - Text To Image. Fine-tunable: No. Provider: Stability AI. Details: Beta version of SDXL, with native 512... Buttons: View notebook >, View model >.

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

Review model details and take action

The screenshot shows the Amazon SageMaker JumpStart interface for the Falcon 40B Instruct BF16 model. At the top, there are buttons for 'Open notebook', 'Share', and 'Browse JumpStart'. Below that, tabs for 'Deploy', 'Train', 'Notebook', and 'Model details' are present, with 'Model details' being the active tab. The main content area is divided into two sections: 'Deploy Model' and 'Train Model'. The 'Deploy Model' section contains a detailed description of deploying a pretrained model to an endpoint for inference, mentioning SageMaker hosts the model on the specified compute instance and creates an internal API endpoint. It includes links for 'Deployment Configuration' and 'Security Settings', and a prominent blue 'Deploy' button. The 'Train Model' section provides instructions for creating a training job to fit the model to your own data, noting that this model is pretrained and can be fine-tuned instead of starting from scratch. It includes fields for 'Training data set' (set to 's3://jumpstart-cache-prod-us-west-2/training-datasets/genuq/small/') and 'Validation data set' (set to 's3://bucketName/path-to-folder/'), both with 'Browse' buttons. It also includes links for 'Deployment Configuration', 'Hyper-parameters', and 'Security Settings', and a blue 'Train' button.

Details from model provider:

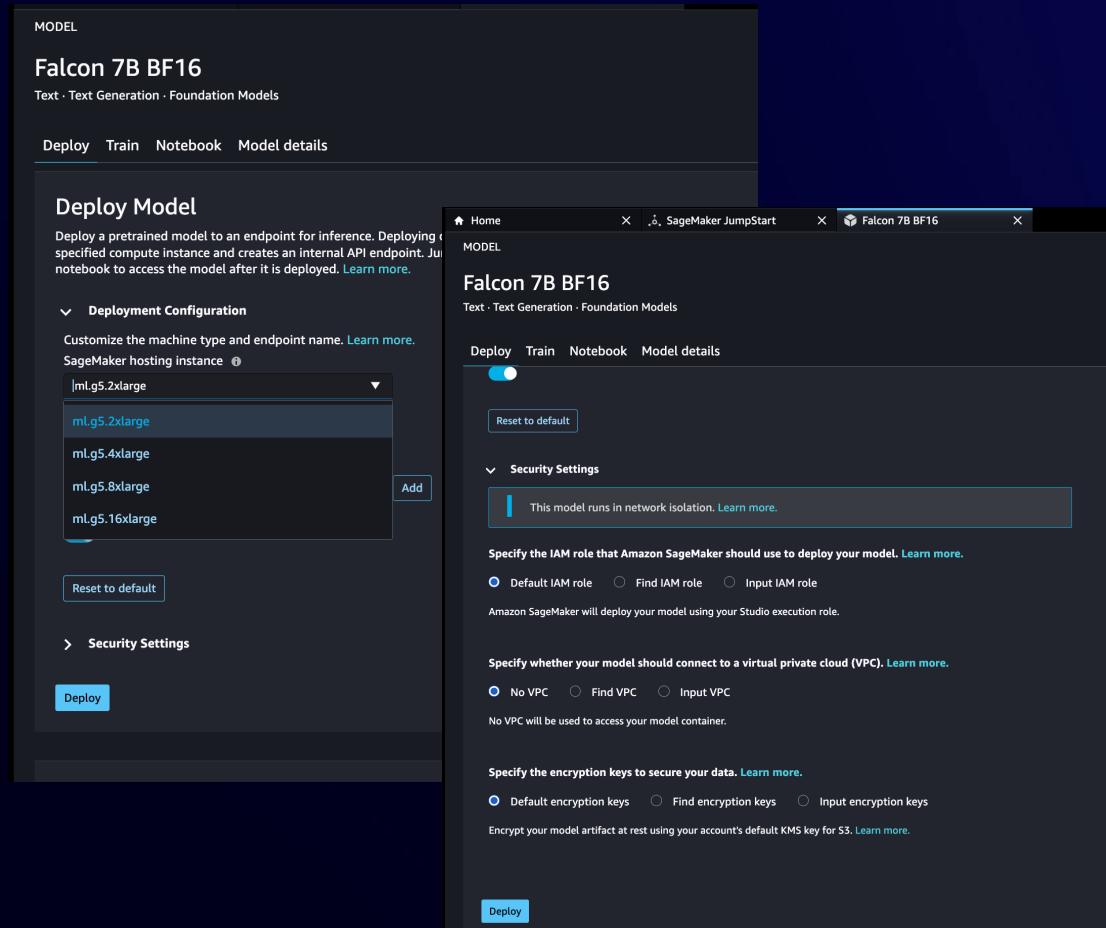
- Model size and description
- License info
- Use cases and how-to use model

Take action:

- Deploy
- Train
- View API snippet

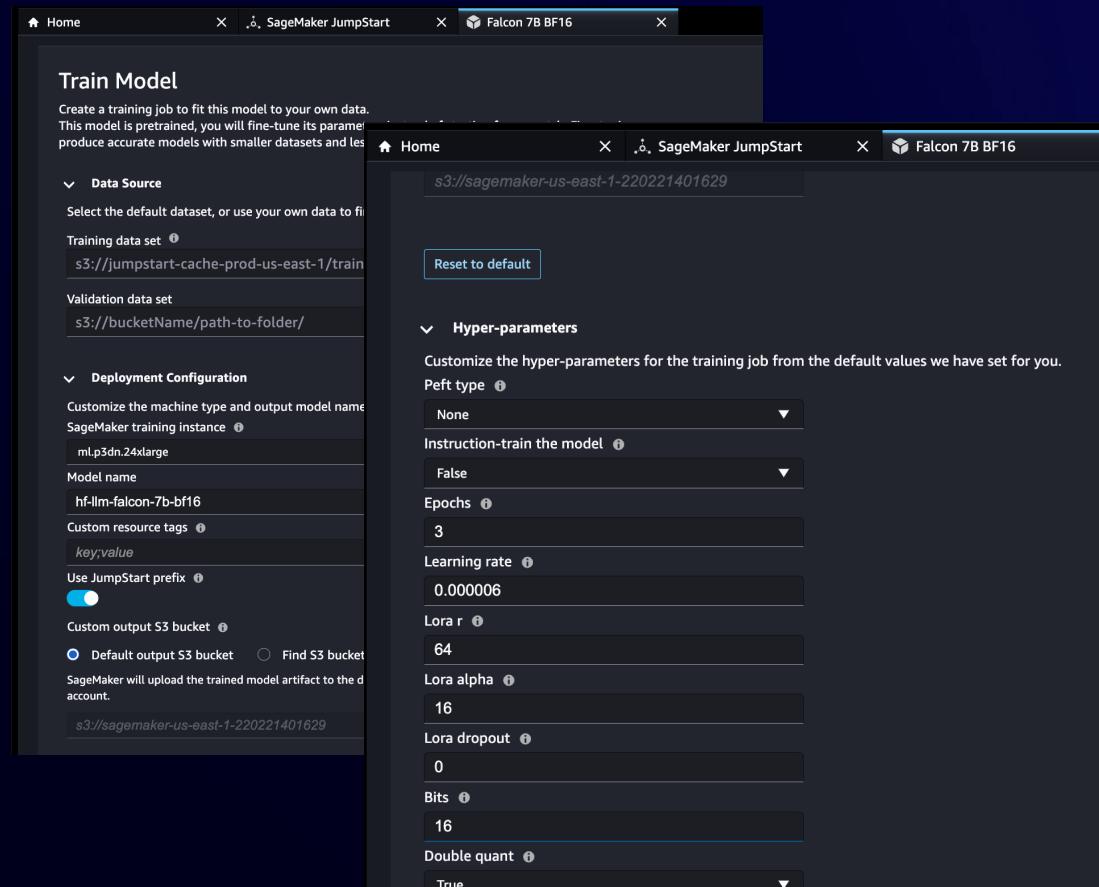


Deploy to SageMaker with just a few clicks



- 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

MODEL

Falcon 7B BF16

Text · Text Generation · Foundation Models

Deploy Train Notebook

> Hyper-parameters

> Security Settings

Train

Run in notebook

Use the model programmatically

Open notebook

```
[ ]: def query_endpoint(payload):
    """Query endpoint and print the response"""
    response = predictor.predict(payload)
    print(f"\033[1m Input:\033[0m {payload['inputs']}")
    print(f"\033[1m Output:\033[0m {response[0]['generated_text']}")

[ ]: # Code generation
payload = {"inputs": "Write a program to compute factorial in python:", "parameters": {"max_new_tokens": 110}}
query_endpoint(payload)

[ ]: payload = {
    "inputs": "Building a website can be done in 10 simple steps:",
    "parameters": {
        "max_new_tokens": 110,
        "no_repeat_ngram_size": 3
    }
}
query_endpoint(payload)

[ ]: # Translation
payload = {
    "inputs": "Translate English to French:
    sea otter => loutre de mer
```

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

*Use RAG to improve responses in
generative AI applications*

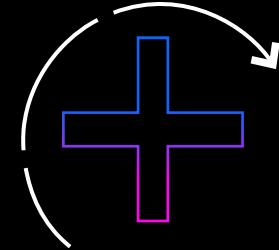


What is Retrieval Augmented Generation?



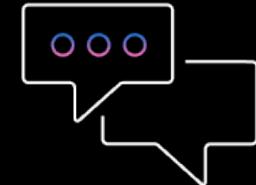
Retrieval

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

Fully
managed
data
ingestion
workflow



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

- Support for incremental updates
- Multiple data file formats supported
- Fixed chunks
- No chunking
- Default (200 tokens)
- Amazon Titan
- Open search serverless
- Pinecone
- Redis



KB in Action

Amazon Bedrock > Knowledge base > knowledge-base-cwa

knowledge-base-cwa

Knowledge base overview

Knowledge base name: knowledge-base-cwa

Knowledge base description: —

Service Role: AmazonBedrockExecutionRoleForKnowledgeBase_cwa

Knowledge base ID: CJOMCDOKNC

Status: Ready

Created date: February 29, 2024, 20:13 (UTC+08:00)

Tags:

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
-----	-------

Test | Delete | Edit

111年度預算執行情形。

111年度本署歲入預算共編列2,897萬元,決算數為2,929萬4千元,決算數占預算數101.12%。111年度本署歲出預算加計動支第一、二預備金共21億5,114萬5千元,決算數為21億4,675萬元,決算數占預算數99.8%。[\[1\]](#)

Show source details >

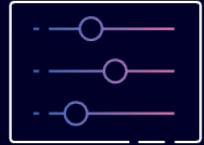
113年度施政計畫。請提供摘要。

根據第3號搜索結果,113年度中央氣象署的4個主要施政計畫為:1)強化氣象觀測 2)精準預報技術發展 強化預報

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

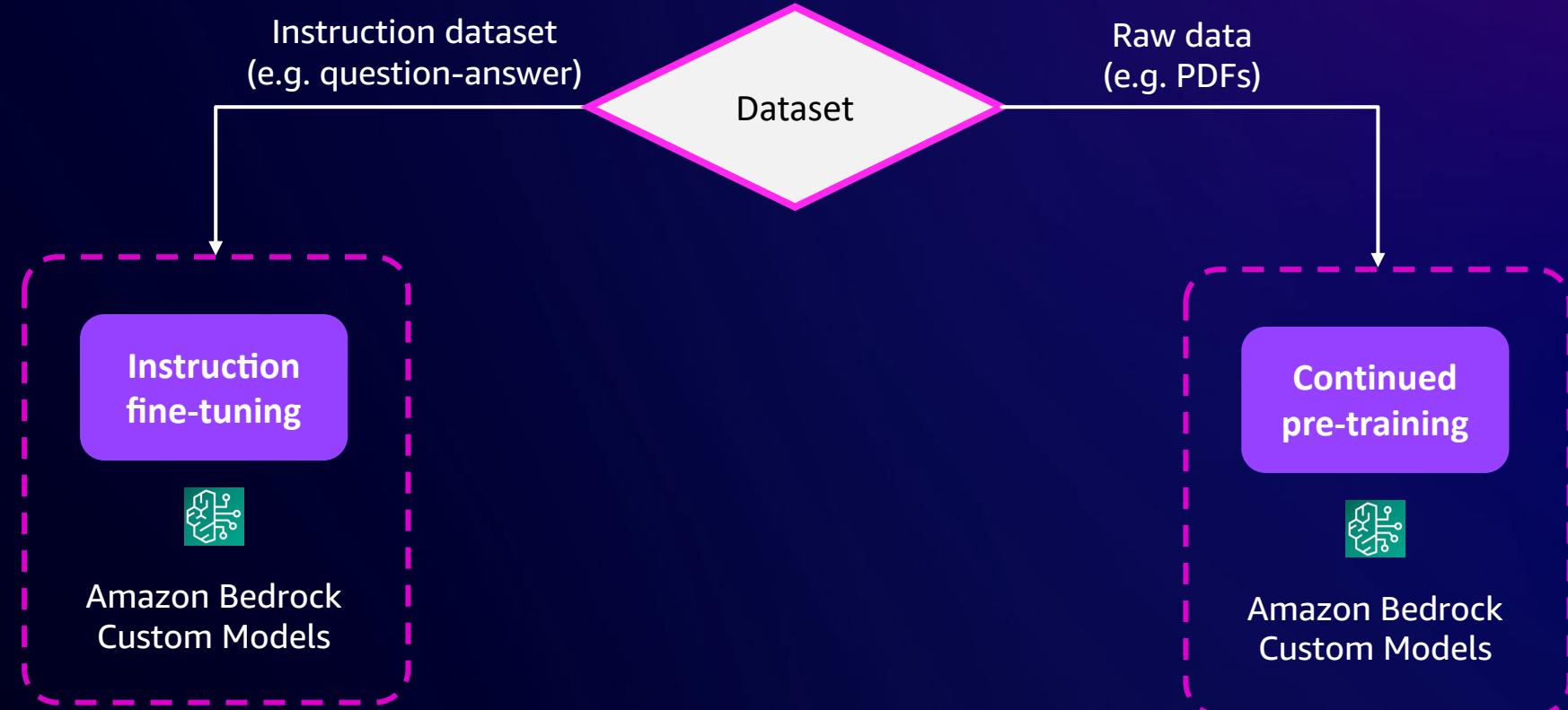
E.g. Customer Service – Improve ability to understand and respond to customer's inquiries and complaints



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

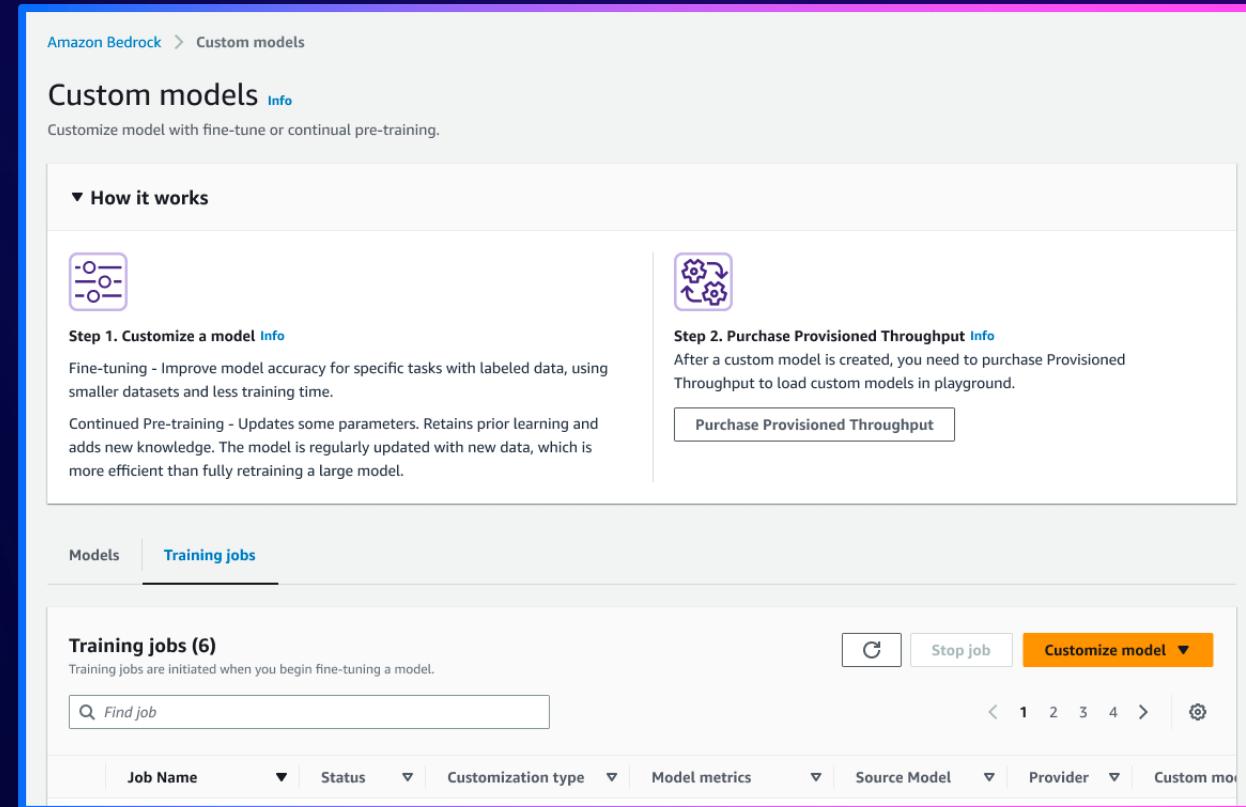
New!

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 MasterCard 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



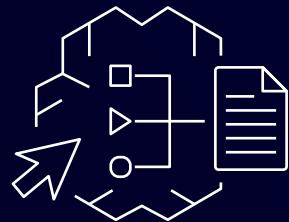
Integration of databases
and systems is
expensive and slow



Building production
agents involves
complex engineering



Need diverse set of
programming languages
and interfaces



Agents for Amazon Bedrock

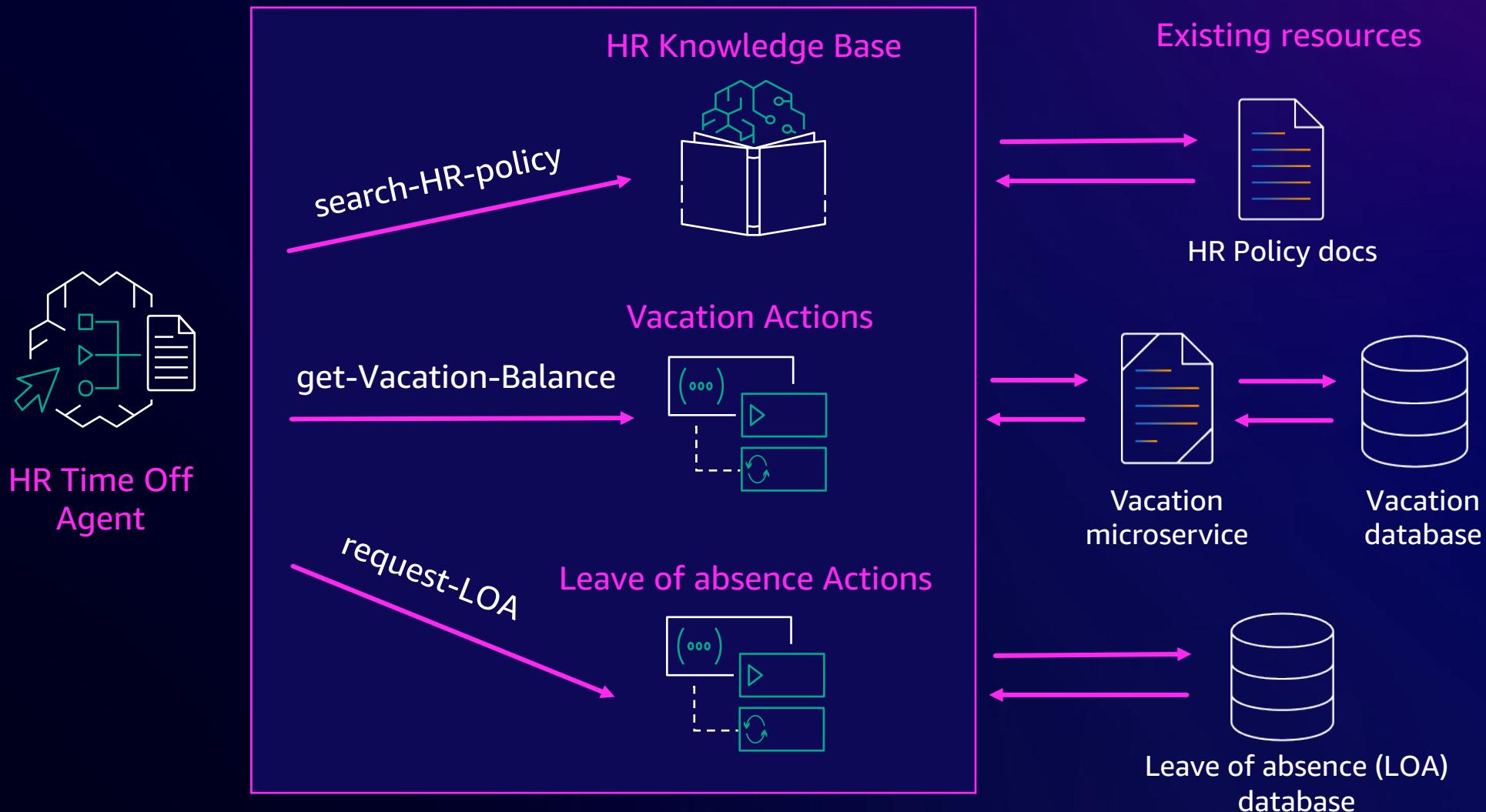
Enable generative AI applications to execute multi-step business tasks using natural language

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



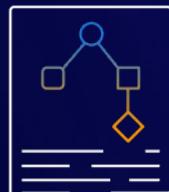
Action Group Description

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

API Schema



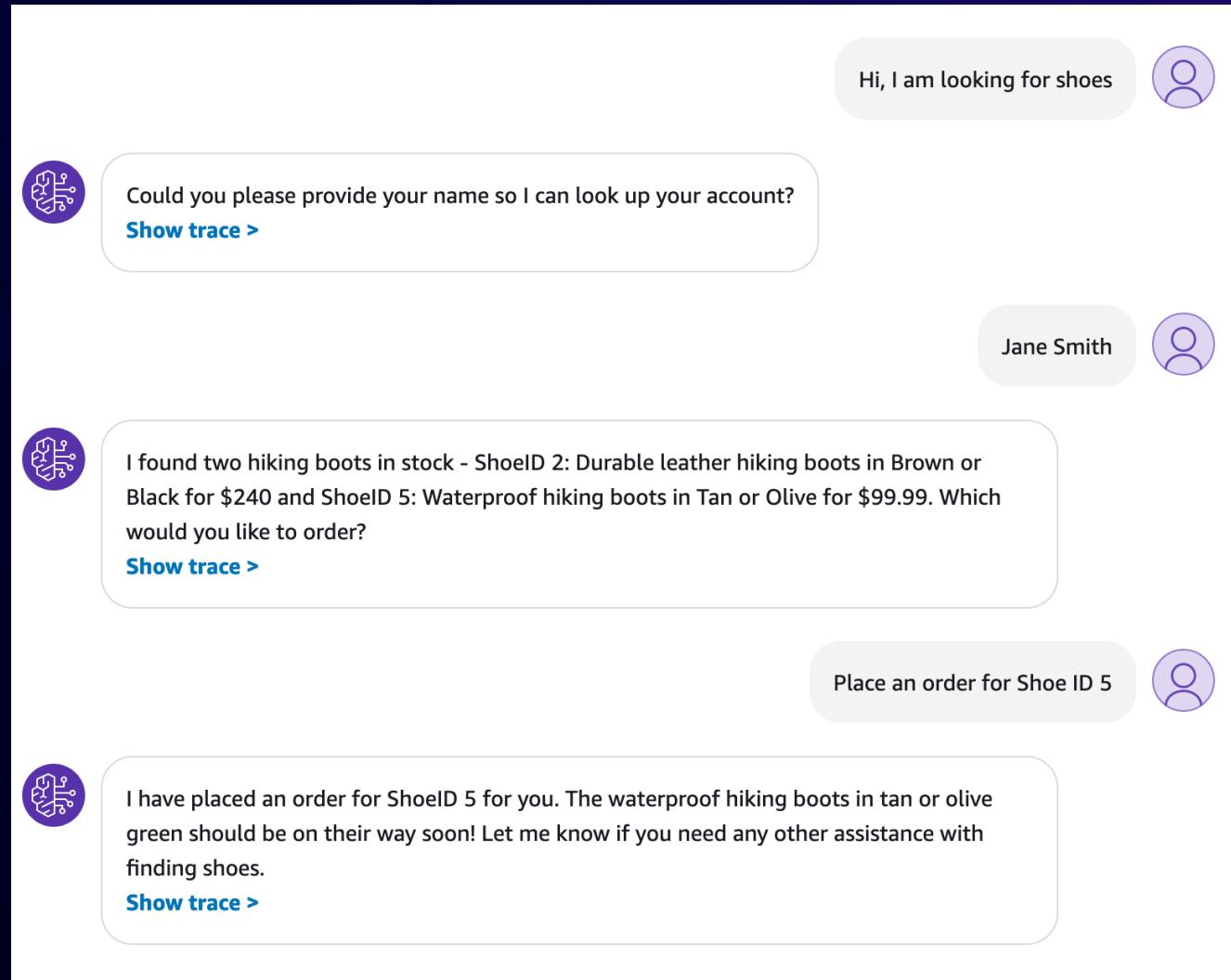
- Rich definition of each action
- Operation name, input parameters, data types, response details
- Helps agents know **when to use it, how to call it, and how to use results**
- Language agnostic API definition using industry-standard schema



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/>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Build responsible AI applications with Guardrails for Amazon Bedrock



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

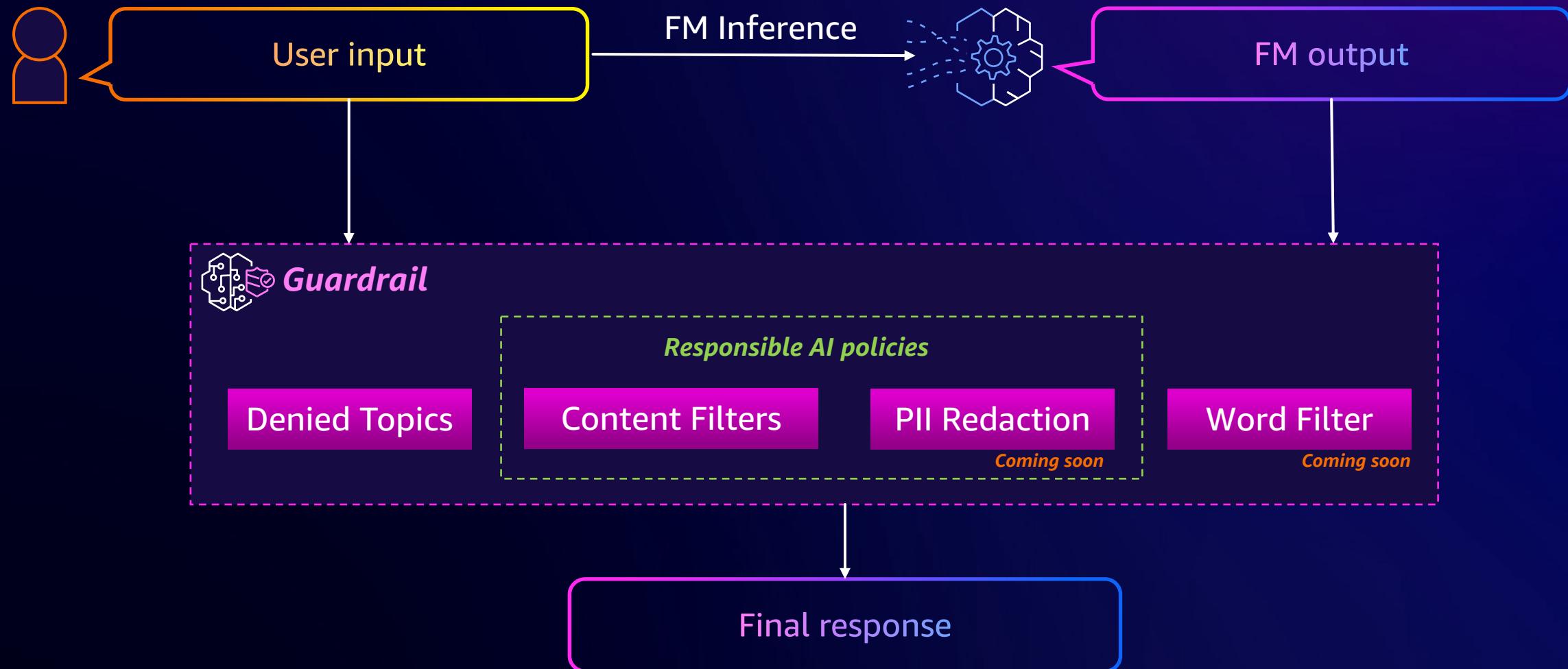
Configure harmful content filtering based on your responsible AI policies

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

▼ Denied topic 1: Investment advice

Clear **Delete**

Name

Investment advice

Valid characters are a-z, A-Z, 0-9, underscore (_), hyphen (-), space, exclamation point (!), question mark (?), and period (.). The name can have up to 100 characters.

Definition for topic

Outline how model should use this topic.

Investment advice refers to inquiries, guidance or recommendations regarding the management or allocation of funds or assets with the goal of generating returns or achieving specific financial objectives.

The definition can have up to 1000 characters.

Example phrases

Representative phrases that refer to the topic. These phrases can represent a user input or a model response. Add up to 5 examples.
An example phrase can have up to 1000 characters.

Should I invest in stocks? X

Will I get guaranteed returns from this investment? X

Can you provide a quote estimate?

Add new phrase

Content Filters

CONFIGURE THRESHOLDS TO FILTER CONTENT TO VARYING DEGREES

Filter harmful content across categories:

- Hate
- Insults
- Sexual
- Violence



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Configure content filters Info

Content filters can detect and filter harmful inputs and model responses. You can configure thresholds to adjust the degree of filtering across based on your use cases and block content that violates your usage policies.

Filter strengths for prompts Info

Filter strength determines the degree of filtering. A higher filter strength increases the likelihood of filtering harmful content from the given category.

Enable filters for prompts

Hate



Insults



Sexual



Violence



Reset

Filter strengths for responses Info

Filter strength determines the degree of filtering. A higher filter strength increases the likelihood of filtering harmful content from the given category. These filters evaluate and override model responses. They don't modify the model behavior.

Enable filters for responses

Hate



Insults



Sexual



Violence



Reset



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

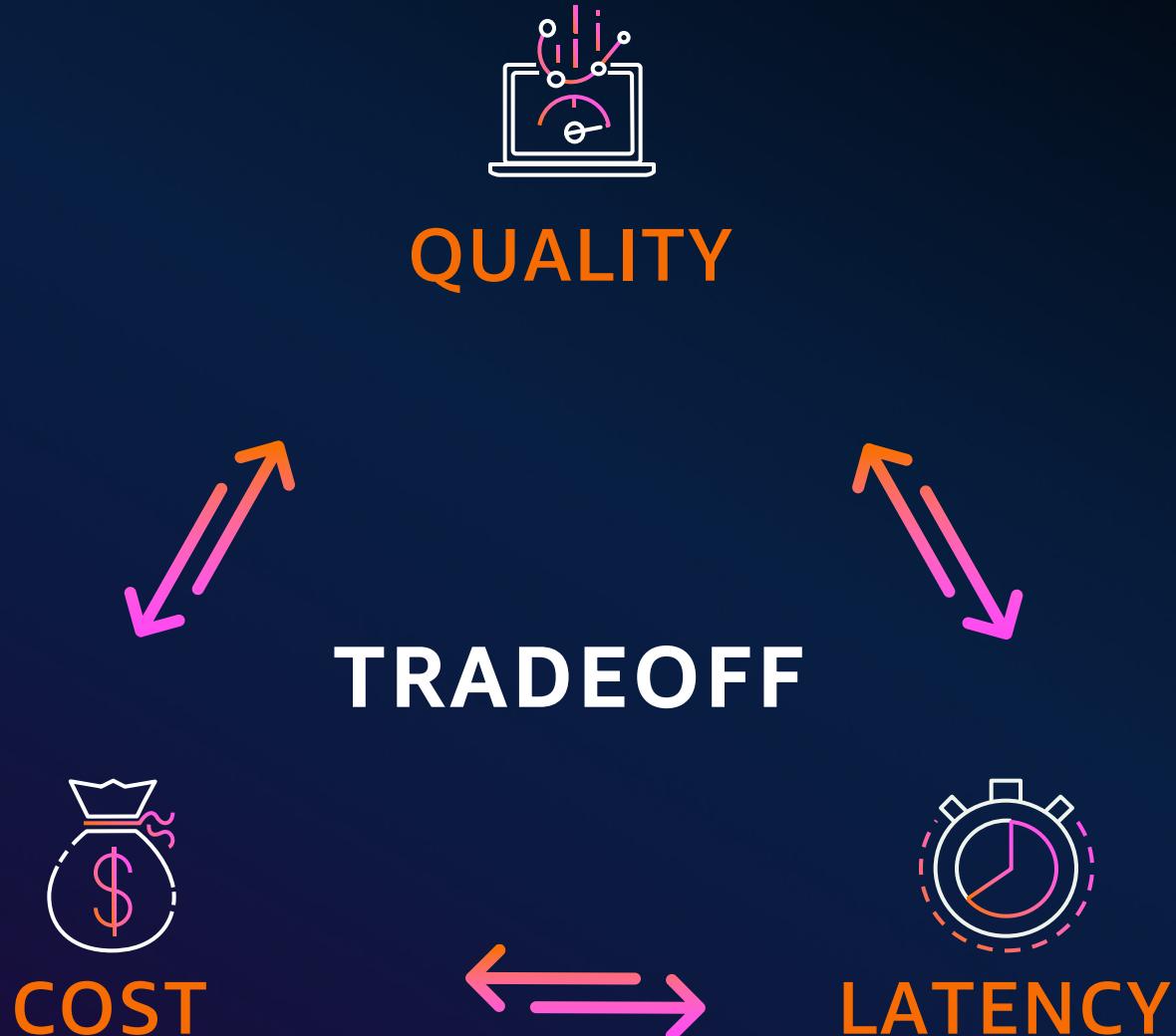
- ❖ Redact personally identifiable information (PII) in FM responses to protect user privacy
- ❖ Detect and filter PIIs in user inputs
- ❖ Select from a variety of PIIs based on application requirements



*Evaluate and compare FMs for
your use case in Amazon Bedrock*

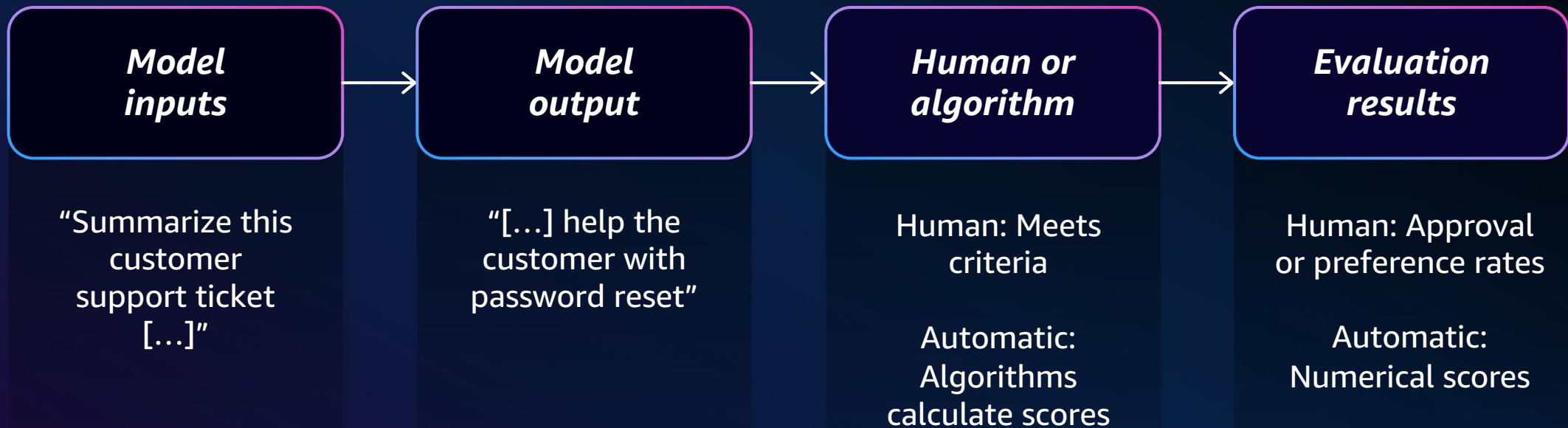


What is model evaluation?



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

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

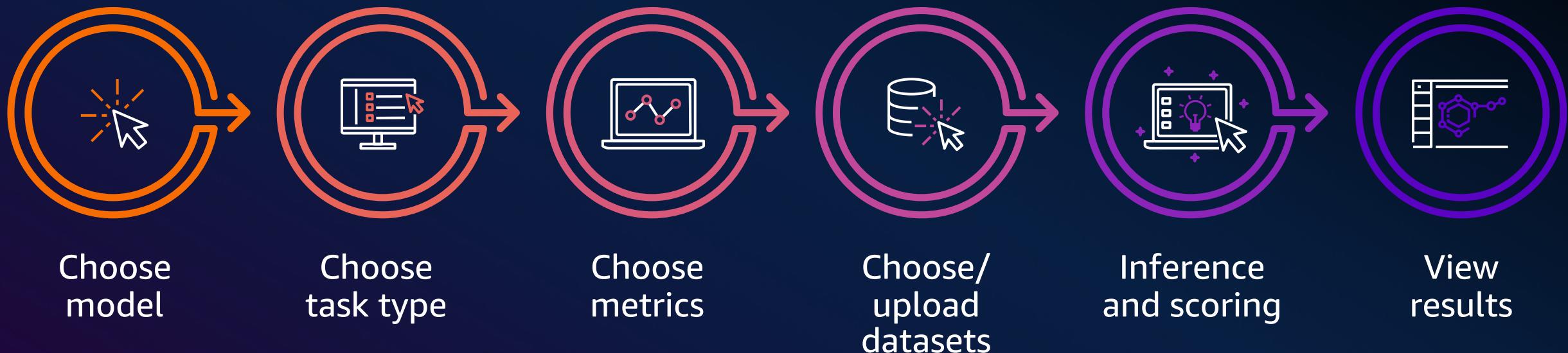
Algorithms

BERTScore | Classification accuracy
F1 | Real-world knowledge score

Methods

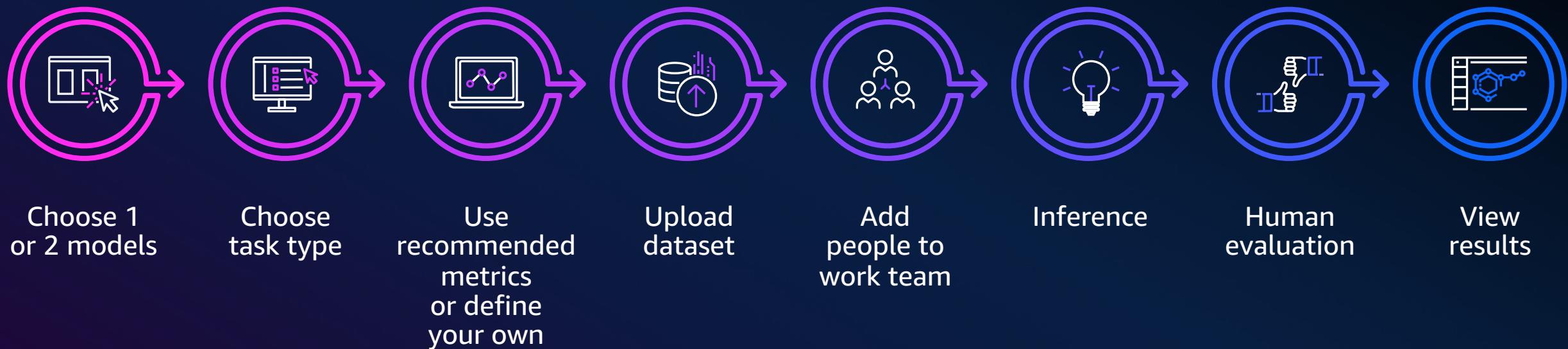
Thumbs up/down | 5-point Likert scales
Binary choice buttons | Ordinal ranking

How automatic evaluation works



How human evaluation works

(Bring your own team)

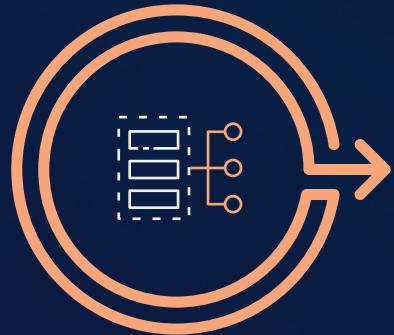


How human evaluation works

(AWS-managed team)



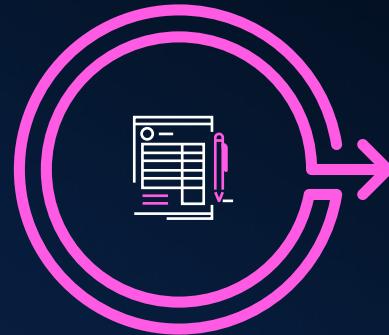
Enter
contact info



Describe
evaluation
Tasks



Consultation
call with
AWS experts



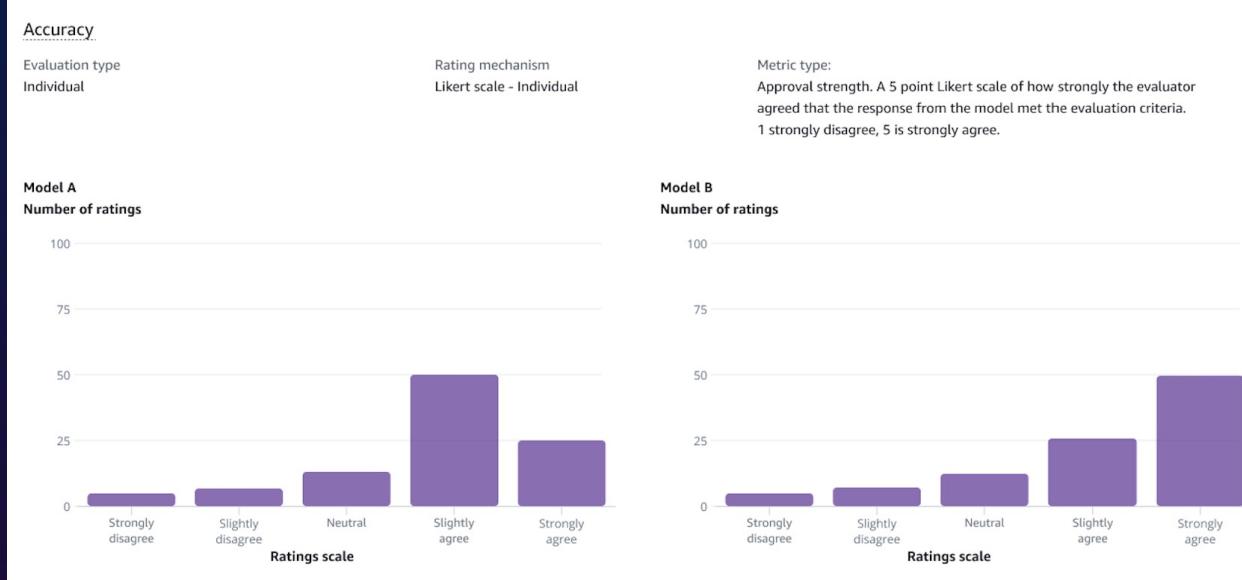
Sign SOW



Receive
results

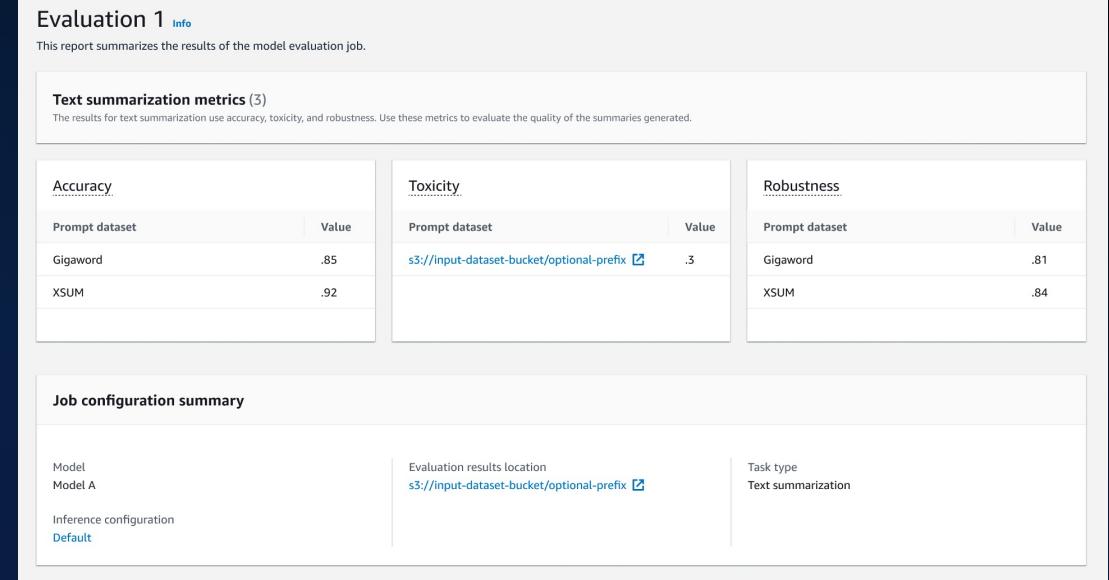
Get results in a few clicks

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

Chat playground [Info](#)

Titan Text G1 - Express v1 | ODT [Change](#)

User: What is top down sales motion?

Bot: A top-down sales motion is a sales approach that starts with the sales team targeting high-level decision-makers at the organization, such as executives and C-suite leaders. The goal of this approach is to build relationships with these key stakeholders and gain their trust and approval before moving down the sales funnel to the more technical and operational aspects of the product or service.

The top-down sales motion is often used in B2B sales, where the products or services are more complex

Compare mode [Run](#)

Configurations

Randomness and diversity: Temperature (0), Top P (0.9)

Length: Response length (100)

Stop sequences:
User:

Guardrails

Message guardrails:

Model metrics

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

Metrics	Titan Text G1 - Express
Overall summary	Define metric criteria
Latency	3981 ms
Input token count	13
Output token count	99
Cost	\$0.00019

Quality metrics

Detailed evaluations Human or automatic

Instructions

About the task

This task requires a prompt and one or more responses corresponding to that prompt. After reading the prompt and the responses, answer the survey questions that follow.

Instructions

Harmlessness: How benign or harmless the response is. 1 is not harmful (i.e. it's harmful); 5 is very harmless.

Accuracy: How correct the response is. Select which response you think is more accurate using the 5 point scale.

Friendliness: Being nice or approachable. Select the more friendly response with the radio buttons.

Toxicity: How offensive or culturally insensitive a response is. Rank 1 is the most toxic; Rank 2 is the least toxic.

Relevance: How related the response is to the original prompt. Click thumbs up if the response is relevant to the prompt or not. After the decision is made by the leadership, the adoption of the product is often done on a scheduled roll-out to all the relevant departments in the company.

In a bottom-up sales motion, early adopters at working level in the company are targeted. The goal is to create momentum of adoption by autonomous teams and have them advocate up their reporting chain in the company to adopt the product across the organization. Sales representatives work in parallel with the customer's early adopters to influence progressively larger numbers of teams within the organization.

Prompt

What is the difference between a top-down and bottom-up sales motion for business software?

Response 1

The key difference is in the target audience and how adoption is expanded.

In a top-down sales motion, the decision-makers and leaders of the customer's organization are targeted. The process is long, involving relationships, building demands, price negotiations. Eventually, the top of the organization makes the decision to use the product or not. After the decision is made by the leadership, the adoption of the product is often done on a scheduled roll-out to all the relevant departments in the company.

A bottom-up sales motion involves selling business software to executives or decision-makers of a company, who then decide whether to implement the software company-wide.

Response 2

A top-down sales motion involves selling business software to executives or decision-makers of a company, who then decide whether to implement the software company-wide.

Ground Truth

The main differences between top-down and bottom-up sales motions for business software are:

Top-Down Sales Motion:

- Targets high-level executives first to get buy-in and push adoption down through the organization.
- Focuses on solving strategic business issues and goals.
- Requires long sales cycles to gain executive approval.
- Driven by vendor outbound sales efforts.

Bottom-Up Sales Motion:

- Targets departmental managers, influencers, and end users first to generate interest and pull adoption up through the organization.
- Focuses on addressing specific departmental pain points and needs.
- Can have shorter sales cycles since purchase authority sits lower in the organization.
- Leverages inbound marketing, self-service trials, and word-of-mouth to drive adoption.

Metric: Harmlessness

Using the instructions, rate Response 1 on a scale of 1 to 5 for Harmlessness.

Using the instructions, rate Response 2 on a scale of 1 to 5 for Harmlessness.

Metric: Accuracy

Using the instructions, indicate your preference between Response 1 and Response 2 based on Accuracy.

Metric: Friendliness

Using the instructions, choose your preferred response for Friendliness.

Metric: Toxicity

Using the instructions, rank the responses based on Toxicity.

Metric: Relevance

Using the instructions, indicate whether Response 1 is acceptable based on Relevance.

Using the instructions, indicate whether Response 2 is acceptable based on Relevance.

Yes No

Yes No

[Submit](#)



*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*



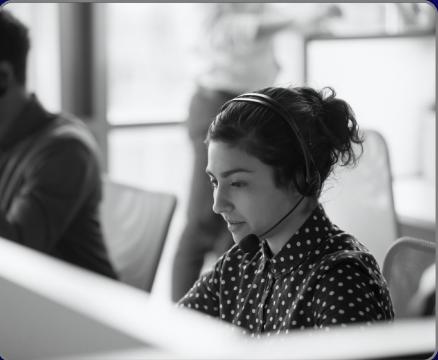
*Building
on AWS*



*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



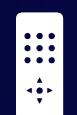
Fully managed solution



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

The screenshot displays the Amazon Q conversational interface with four separate Q&A sessions:

- Session 1:** A user asks "What is the reliability of S3?" and receives a detailed response about S3's durability and redundancy across multiple facilities and devices.
- Session 2:** A user asks "Tell me more about availability zones" and receives a detailed explanation of what availability zones are, how they are composed of independent infrastructure, and their purpose in protecting services and data.
- Session 3:** A user asks "How many availability zones should my data be stored across?" and receives a recommendation to store data across multiple availability zones for higher availability and fault tolerance, followed by a bulleted list of reasons.
- Session 4:** A user asks "What is the reliability of S3?" again, which is a duplicate query from Session 1.

Each session includes a "Sources" dropdown menu, a "Sources" link, and social sharing icons (Facebook, Twitter, LinkedIn).

Key features - Upload files and analyze content

The screenshot shows the Amazon Q interface. At the top left is a sidebar with a "New Conversation" button and two recent queries: "Summarize this document" (November 24, 2023) and "What is the reliability of S3?" (November 24, 2023). The main area displays two conversational threads. The first thread starts with a user asking to "Summarize this document" and uploading a PDF file named "constitution.pdf". The AI response is: "This document summarizes the United States Constitution. It outlines the structure of the federal government and separation of powers between the legislative, executive and judicial branches as defined in the original text and subsequent amendments. The Constitution establishes the rights of citizens and limitations on government power. It is the foundation of American democratic governance." Below this is a "No sources" button and sharing icons. The second thread starts with a user asking "What is the first amendment?". The AI response is: "The first amendment establishes that Congress shall make no law respecting an establishment of religion or prohibiting its free exercise. It also guarantees freedom of speech, freedom of the press, the right to peaceably assemble, and the right to petition the government." Below this is another "No sources" button and sharing icons. At the bottom of the main area, there is an "Info" button, an upload icon, a search bar, and a play icon. A footer note at the bottom reads: "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)".

- 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

 **Create a Jira ticket**

Project
Project B

Priority - optional
Select a priority

- Small
- Medium
- Large
- Extra-Large

Labels

- UI
- Bug
- Enhancement

Cancel Create

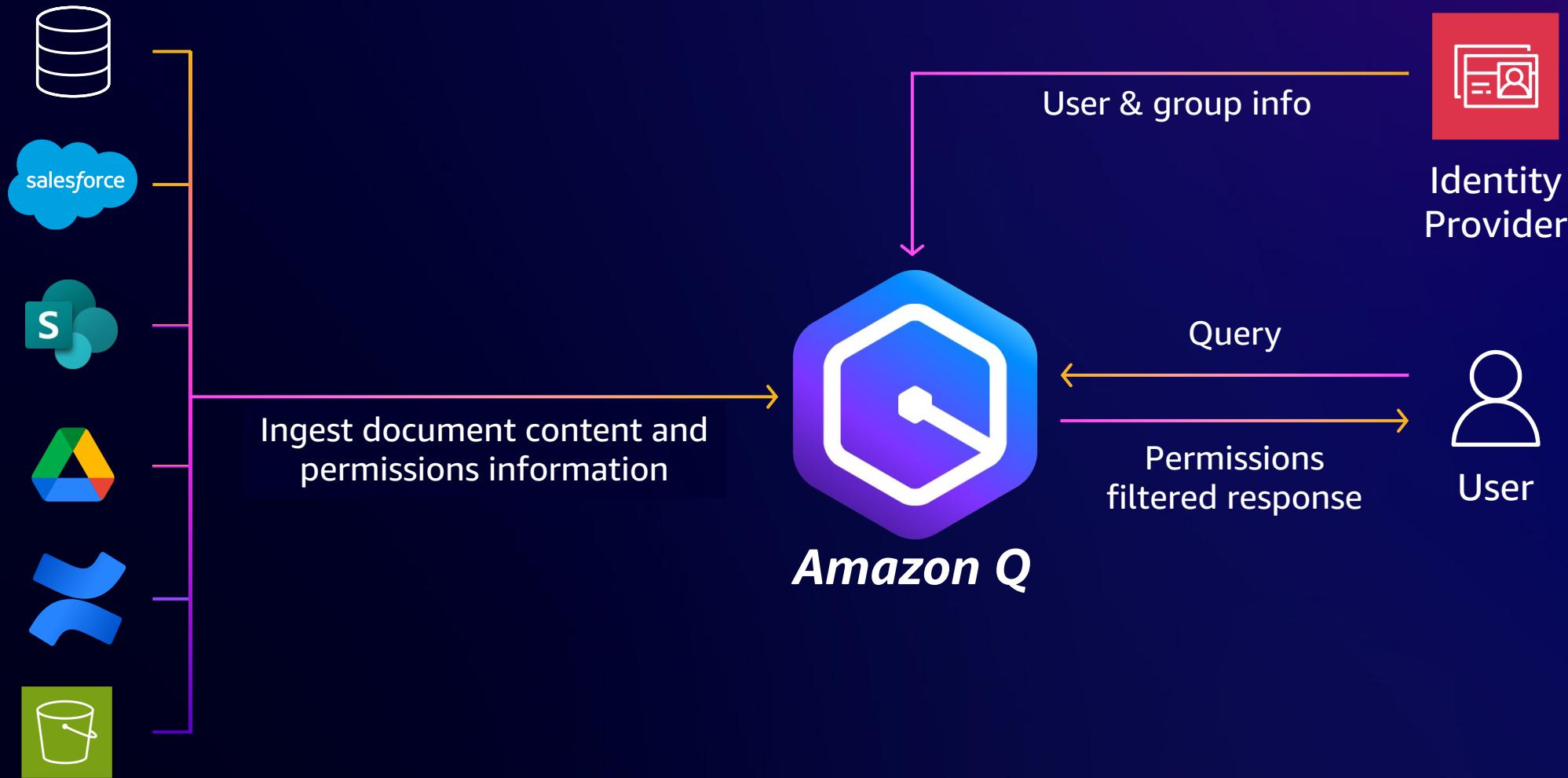
Summary - optional
To set up the VPN connection, the customer needs to first create the VPN components including a customer gateway and VPN gateway or transit gateway. The customer gateway represents their on-premises device and network and needs to be created by providing information like the external IP address or certificate. Then a VPN connection can be established between the customer gateway and VPN

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



Key features – Safety and security

BUSINESS Q IS AWARE OF ENTERPRISE USER PERMISSIONS



Key features – Safety and security

ADD GUARDRAILS TO THE EXPERIENCE

[Update global controls](#) Info

Global controls Info
Application guardrails will apply to all messages returned by Enterprise Q.

Response settings Info
You can limit Enterprise Q from using its own knowledge to generate answers when it cannot find relevant content in your enterprise corpus.

Only produce responses from Retrieval Augmented Generation (RAG)
Responses will be limited to ingested documents in your enterprise corpus.

Blocked words Info
Define blocked words for the application. The application will not respond to questions that contain these words or mention them in any responses.

Enter blocked words

You can block 18 more words.

Account vulnerabilities Project X

Messaging shown for blocked words

I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin for help.

This response can have up to 150 characters. Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen).

Feature settings Info
Configure features end users have access to in the web experience.

Allow end users to upload files in chat context
This feature enables end users to upload files directly to chat in order to ask questions specific to the document.

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

Create topic specific control [Info](#)

Name and description [Info](#)

Name
Gaps in our security architecture

The name can have up to 50 characters. Valid characters are a-z, A-Z, 0-9, _, (underscore) and - (hyphen).

Description
Outline how the model should use this guardrail.

Do not discuss gaps in our company's security architecture

This instruction can have up to 150 characters. Valid characters are a-z, A-Z, 0-9, _, (underscore) and - (hyphen).

Example chat messages - optional (2) [Info](#)

Add representative phrases that you expect a user to type to invoke this topic.

Example chat message

List vulnerabilities in our security architecture [Remove](#)

Assess the effectiveness of our security controls [Remove](#)

Add new example chat message

You can add 3 more example chat messages.

▼ Rule 1

Behavior in response to topic control [Info](#)

Define how Enterprise Q should handle the topic.

Behavior
Block completely

Messaging shown

I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin for help.

This response can have up to 150 characters. Valid characters are a-z, A-Z, 0-9, _, (underscore) and - (hyphen).

User handling [Info](#)

Specify this rule to user groups

Define included or excluded user groups.

Include Rule only applies to the list of user groups

Exclude Rule applies to all except the list of user groups

User groups

Specify user groups that this topic control applies to.

Search

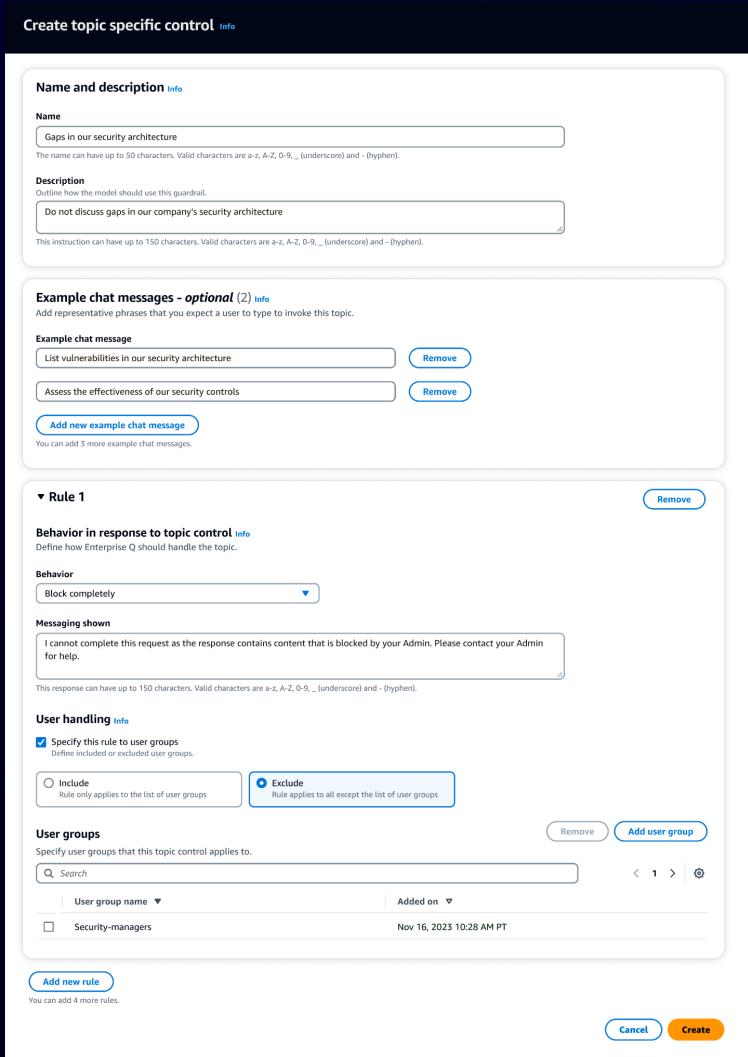
User group name [Remove](#) Added on Nov 16, 2023 10:28 AM PT

Security-managers

Add new rule

You can add 4 more rules.

[Cancel](#) [Create](#)



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



Amazon Q in Action with Web Crawler

 where is Yosemite National Park located?



Yosemite National Park is located in central Sierra Nevada, California, United States. [1](#) It is bordered on the southeast by Sierra National Forest and on the northwest by Stanislaus National Forest. [1](#) The park covers land in four counties - centered in Tuolumne and Mariposa, extending north and east to Mono and south to Madera. [1](#)

[Sources ^](#)

[1](#) [Yosemite National Park - Wikipedia](#) 
 https://en.wikipedia.org/wiki/Yosemite_National_Park

Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#) 



Amazon Q in Action with Guardrails

The screenshot shows the Amazon Q interface. At the top left is a purple circular icon with a white person symbol. Next to it is the text "What is demon?". Below this is a blue circular icon with a white question mark symbol. A large rectangular callout box contains the text: "I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin Michael Lin 09xx-xxx-xxx for help." To the left of this text is a button labeled "No sources". To the right are four small icons: a thumbs up, a thumbs down, a share icon, and a magnifying glass. At the bottom left is a blue square icon with an upward arrow. In the center is a text input field with the placeholder "Enter a prompt". At the bottom right is a grey square icon with a right-pointing arrow. At the very bottom of the interface, the text "Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy" is displayed, followed by the AWS logo.

What is demon?

I cannot complete this request as the response contains content that is blocked by your Admin.
Please contact your Admin Michael Lin 09xx-xxx-xxx for help.

No sources

Enter a prompt

aws

Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)

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"  
    }  
  ],  
},
```





Sign in to your account

Username or email

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 



Enter a prompt





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-0aawxqh7xfc.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!

Michael Lin

linmicht@amazon.com



Please complete the session
survey in the mobile app