aws INNOVATE

GENERATIVE AI + DATA

aws INNOVATE

GENERATIVE AI + DATA

FRA01

Build responsible AI applications with Guardrails for Amazon Bedrock

Christian Kamwangala

(he/him)
Solutions Architect
Amazon Web Services

Today's agenda

- Challenges for using generative AI responsibly
- NEW! Safeguard applications with Guardrails for Amazon Bedrock
- Deploy guardrails with Agents for Amazon Bedrock
- Demos and walkthroughs

Generative AI is powering multiple use cases

Boost employee productivity

Design & Content Creation

Search, Summarization & Analysis

BI and Report Generation

Code Generation

Enhance customer experiences

Chatbots & Virtual Assistants

Customer Care Agent Assist

Reviews & Conversation Analytics

Personalize User Experiences

Optimize business processes

Document Processing

Data Augmentation

Intelligent Process Automation

Market & Supply Chain Insights

Building generative apps brings new challenges

Undesirable and Irrelevant Topics

Controversial queries and responses

Toxicity & Safety (incl. brand risk)

Harmful or offensive responses

Privacy Protection

Protect user information or sensitive data

Bias/Stereotype Propagation

Biased results or unfair user outcomes

Many foundation models have built-in protections

AI21 labs	ANTHROPIC	co:here	Meta	stability.ai	amazon
JURASSIC	CLAUDE	COMMAND + EMBED	LLAMA 2	SDXL 1.0	AMAZON TITAN

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

NEW

Safeguard generative AI applications with Guardrails for Amazon Bedrock

aws

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

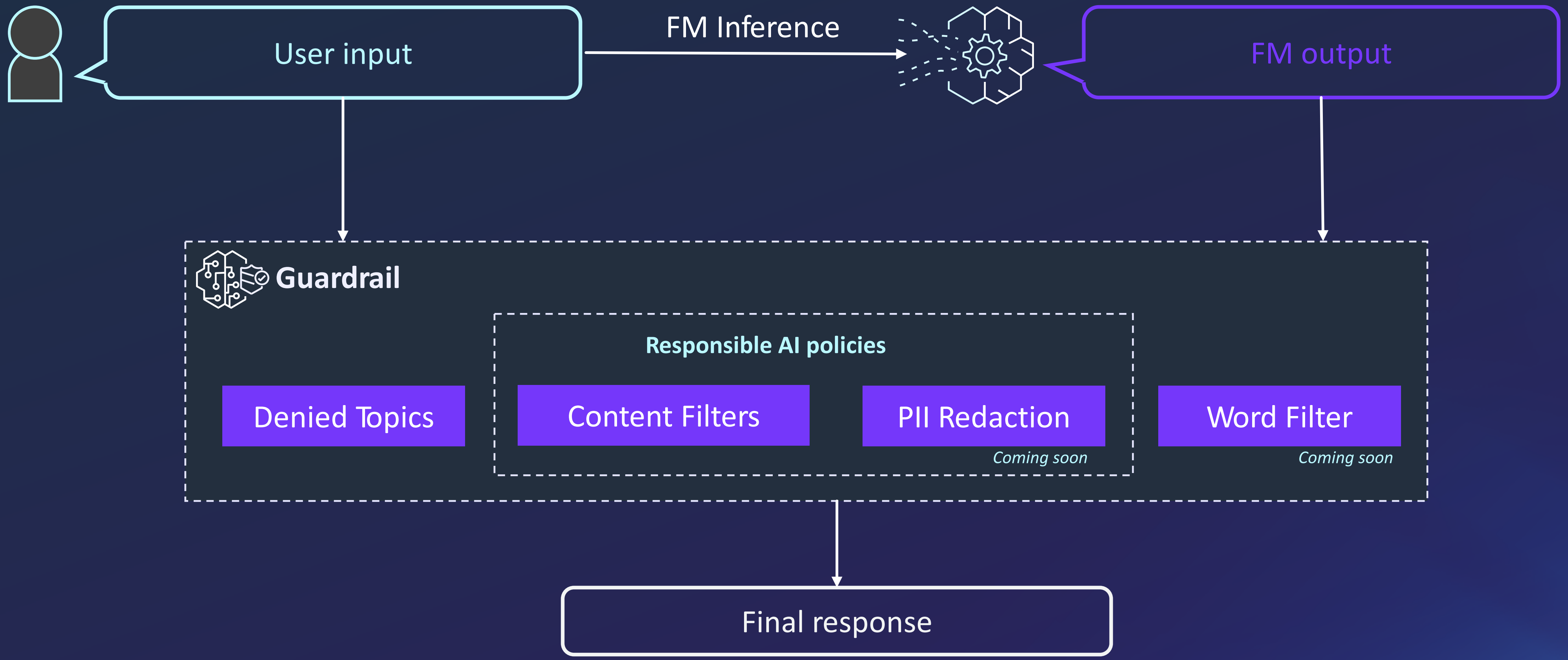
Configure harmful content filtering based on your responsible AI policies

Define and disallow denied topics with short natural language descriptions

COMING SOON

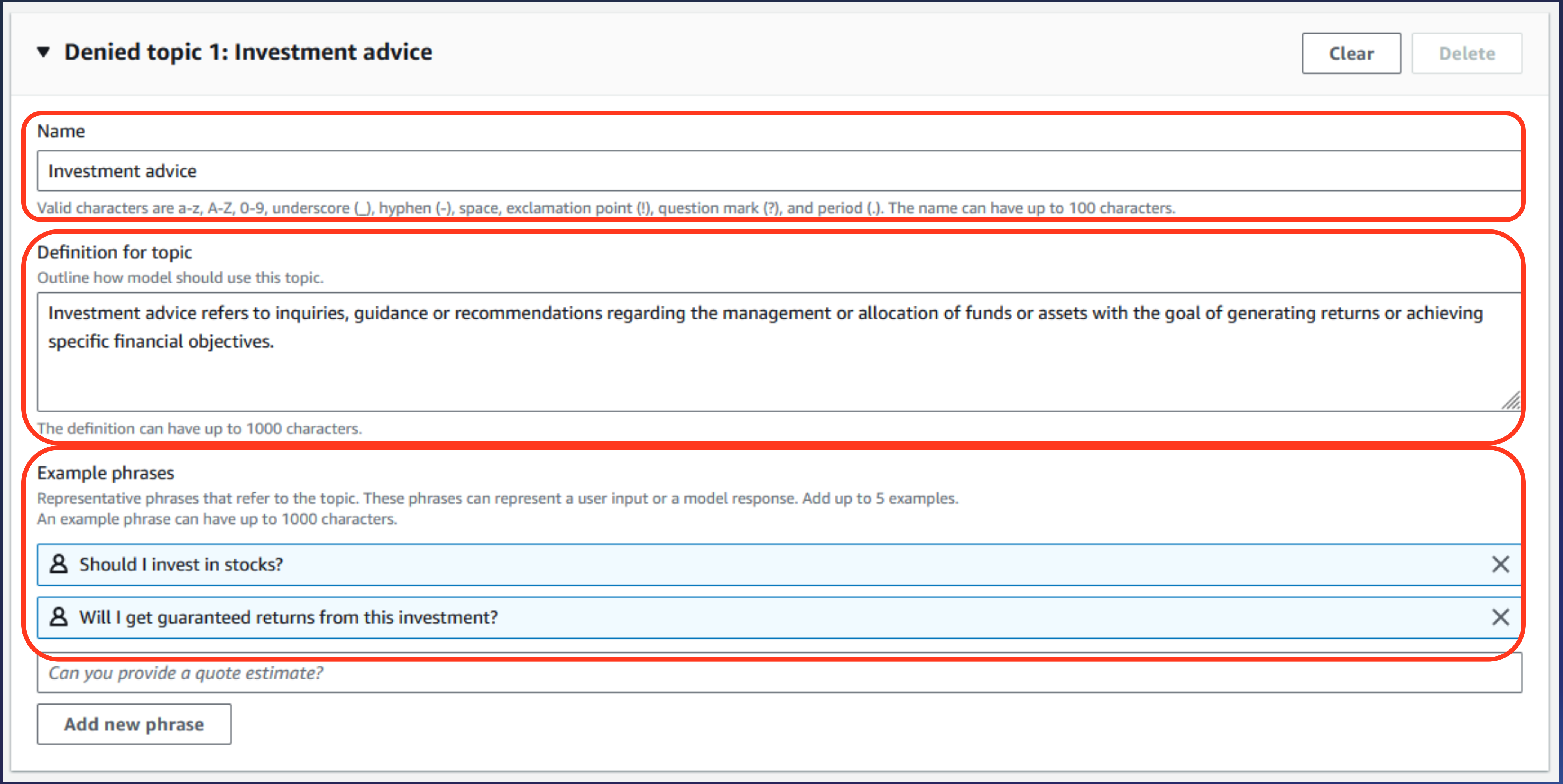
Redact sensitive PII information in FM responses

How it works: Guardrails for Amazon Bedrock



Denied Topics

AVOID Undesirable TOPICS in your applications



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?
- Will I get guaranteed returns from this investment?

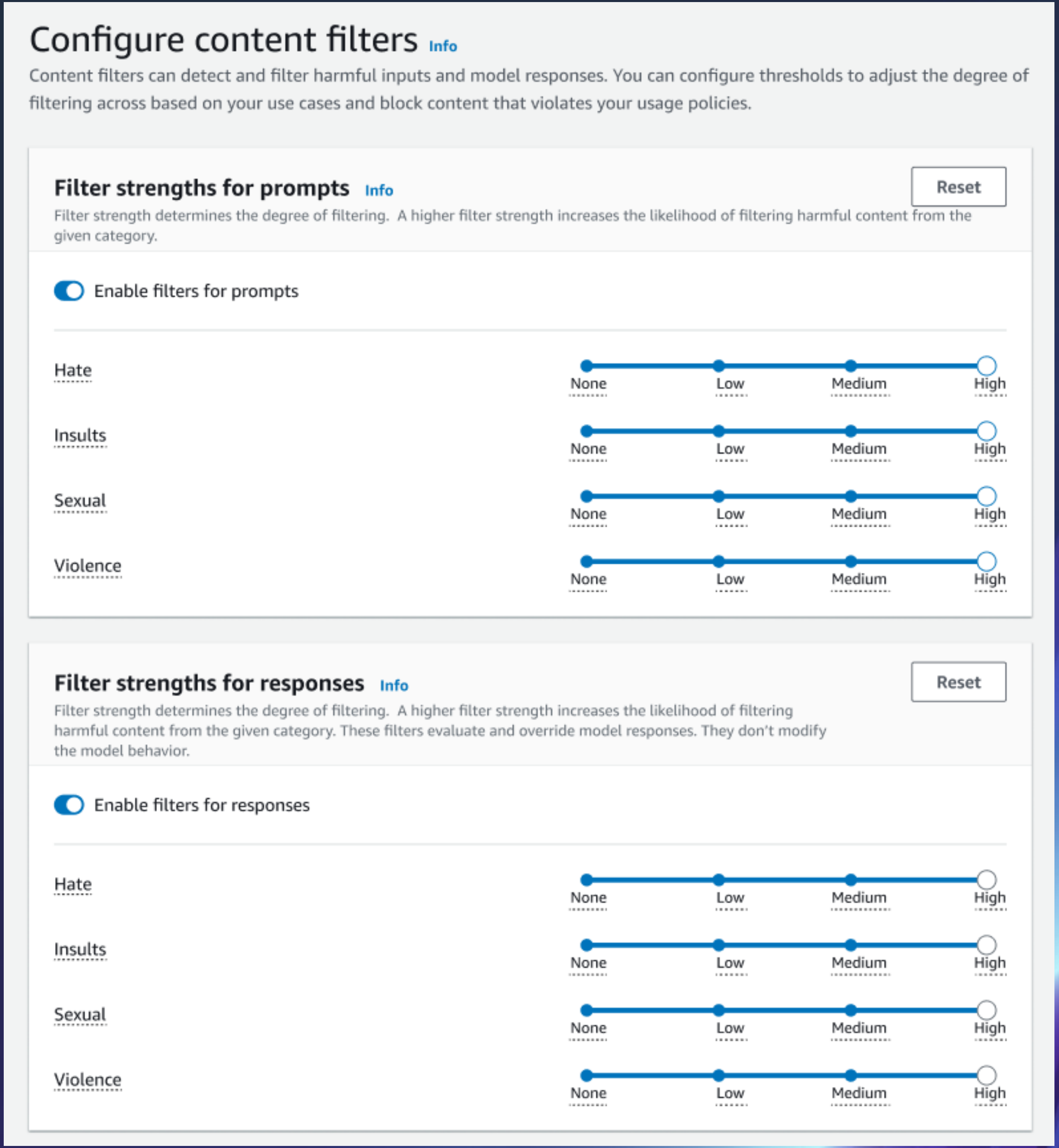
Can you provide a quote estimate?

Content Filters

CONFIGURE THRESHOLDS TO FILTER CONTENT TO VARYING DEGREES

Filter harmful content across categories:

- Hate
- Insults
- Sexual
- Violence



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

Reset

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: None Low Medium [✓ High]

Insults: None Low Medium [✓ High]

Sexual: None Low Medium [✓ High]

Violence: None Low Medium [✓ High]

Filter strengths for responses Info

Reset

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: None Low Medium [✓ High]

Insults: None Low Medium [✓ High]

Sexual: None Low Medium [✓ High]

Violence: None Low Medium [✓ High]

Word Filters

COMING SOON

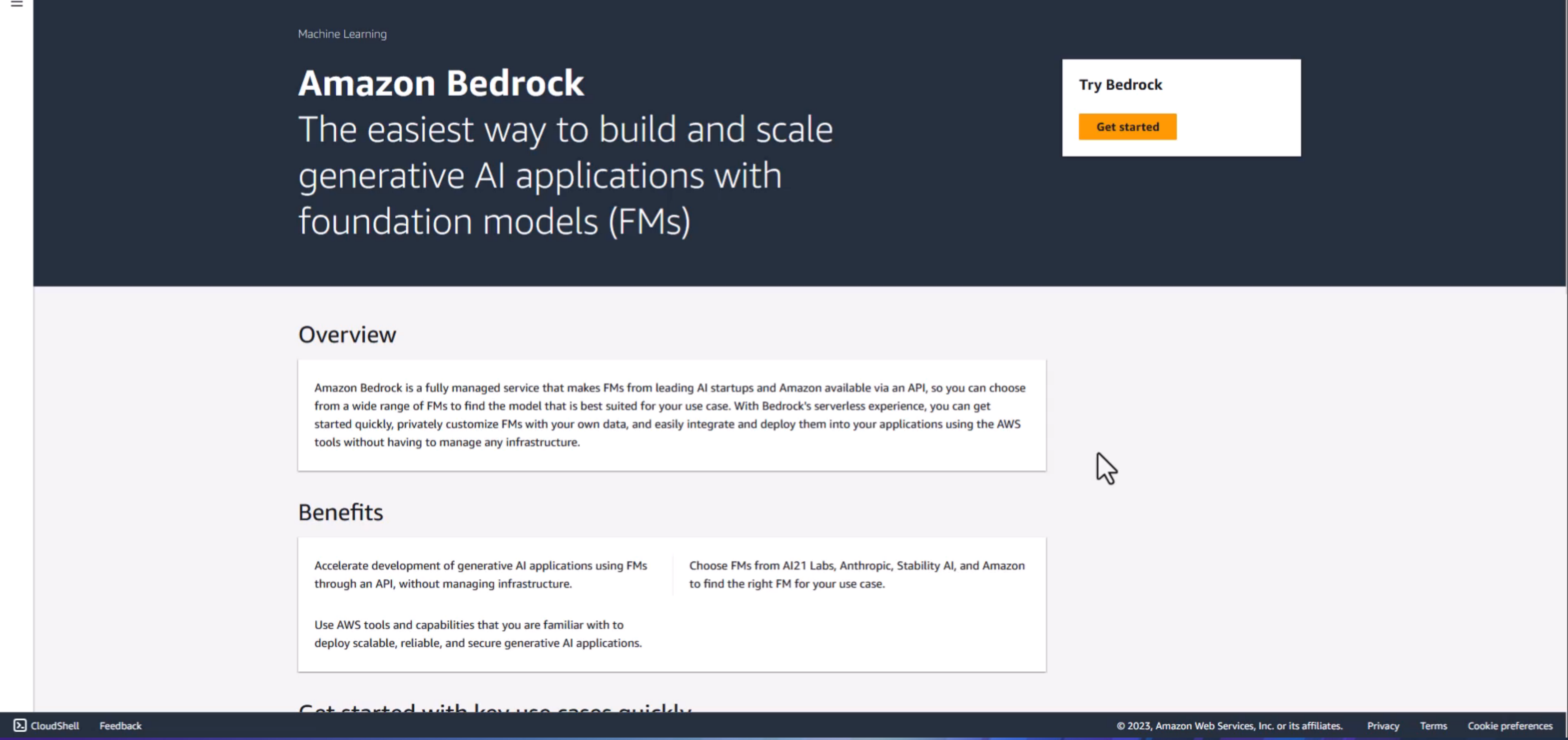
- ❖ Define a set of custom words to block in user input and FM responses
- ❖ Filter profane words
- ❖ Choose to respond with a preconfigured message or mask the blocked words

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

Demo



Amazon Bedrock

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

Overview

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

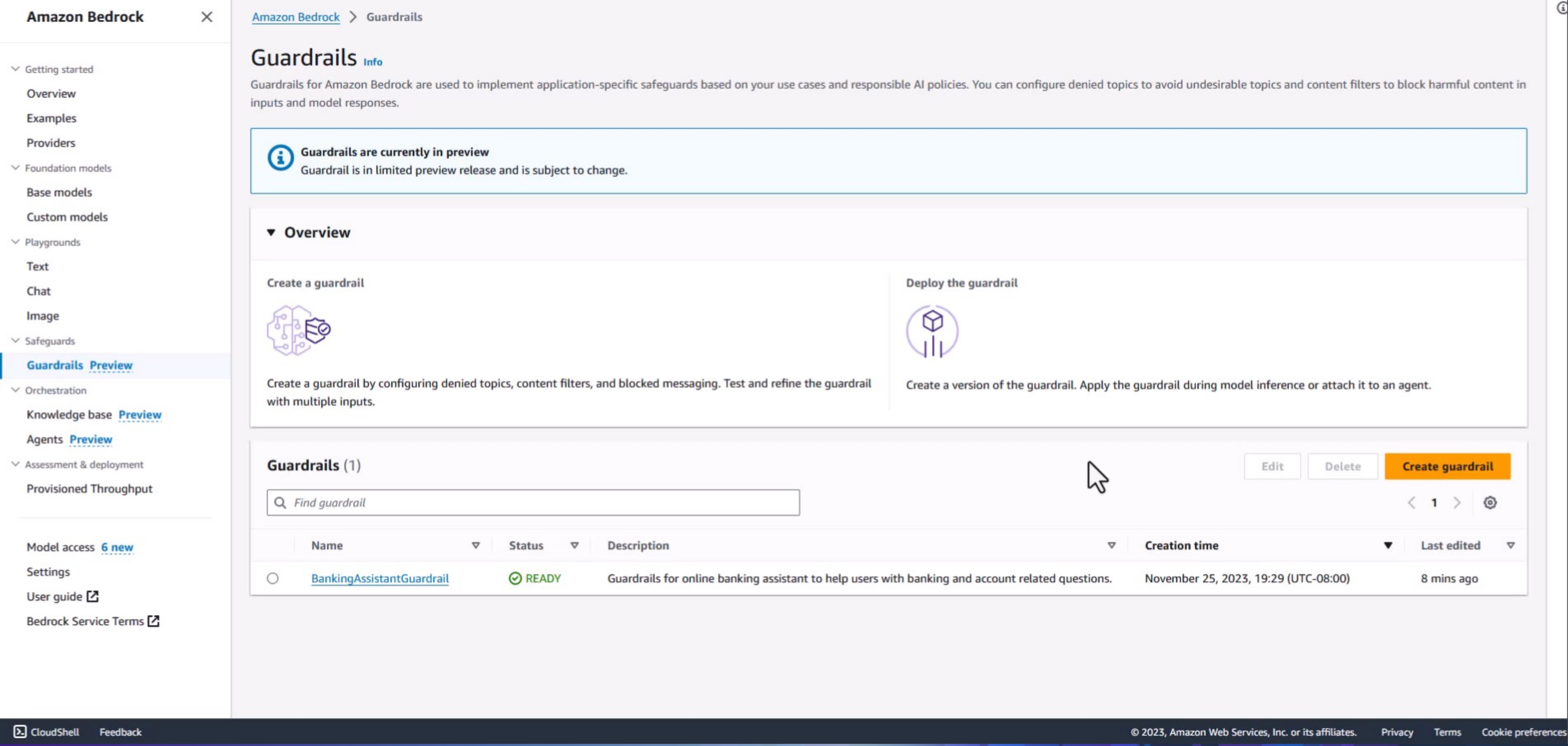
Benefits

Accelerate development of generative AI applications using FMs through an API, without managing infrastructure.

Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case.

Use AWS tools and capabilities that you are familiar with to deploy scalable, reliable, and secure generative AI applications.

Get started with key use cases quickly



Guardrails Info

Guardrails for Amazon Bedrock are used to implement application-specific safeguards based on your use cases and responsible AI policies. You can configure denied topics to avoid undesirable topics and content filters to block harmful content in inputs and model responses.

Guardrails are currently in preview
Guardrail is in limited preview release and is subject to change.

Overview

Create a guardrail

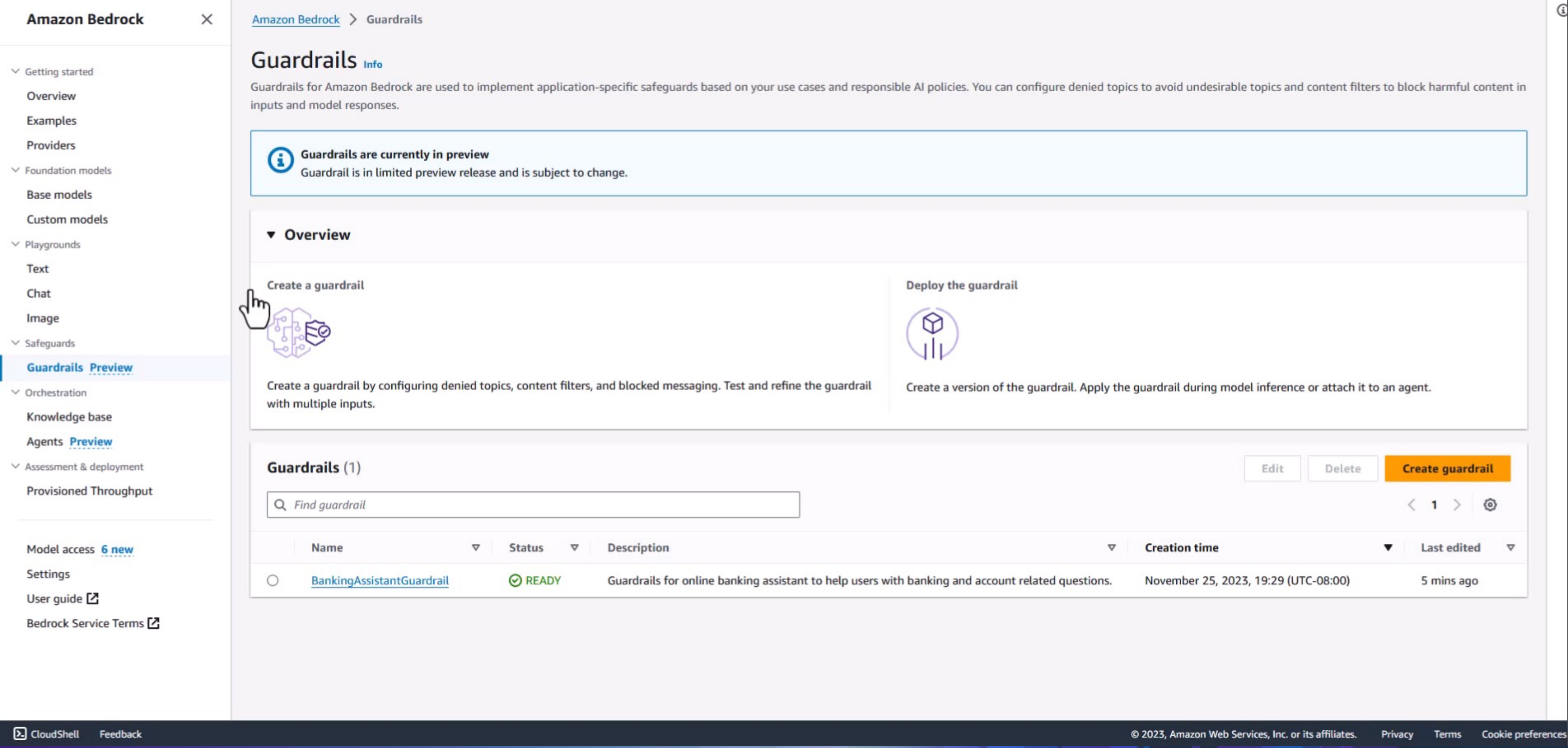
Create a guardrail by configuring denied topics, content filters, and blocked messaging. Test and refine the guardrail with multiple inputs.

Deploy the guardrail

Create a version of the guardrail. Apply the guardrail during model inference or attach it to an agent.

Guardrails (1)

	Name	Status	Description	Creation time	Last edited
○	BankingAssistantGuardrail	READY	Guardrails for online banking assistant to help users with banking and account related questions.	November 25, 2023, 19:29 (UTC-08:00)	8 mins ago



Guardrails Info

Guardrails for Amazon Bedrock are used to implement application-specific safeguards based on your use cases and responsible AI policies. You can configure denied topics to avoid undesirable topics and content filters to block harmful content in inputs and model responses.

Guardrails are currently in preview
Guardrail is in limited preview release and is subject to change.

▼ Overview

Create a guardrail

Create a guardrail by configuring denied topics, content filters, and blocked messaging. Test and refine the guardrail with multiple inputs.

Deploy the guardrail

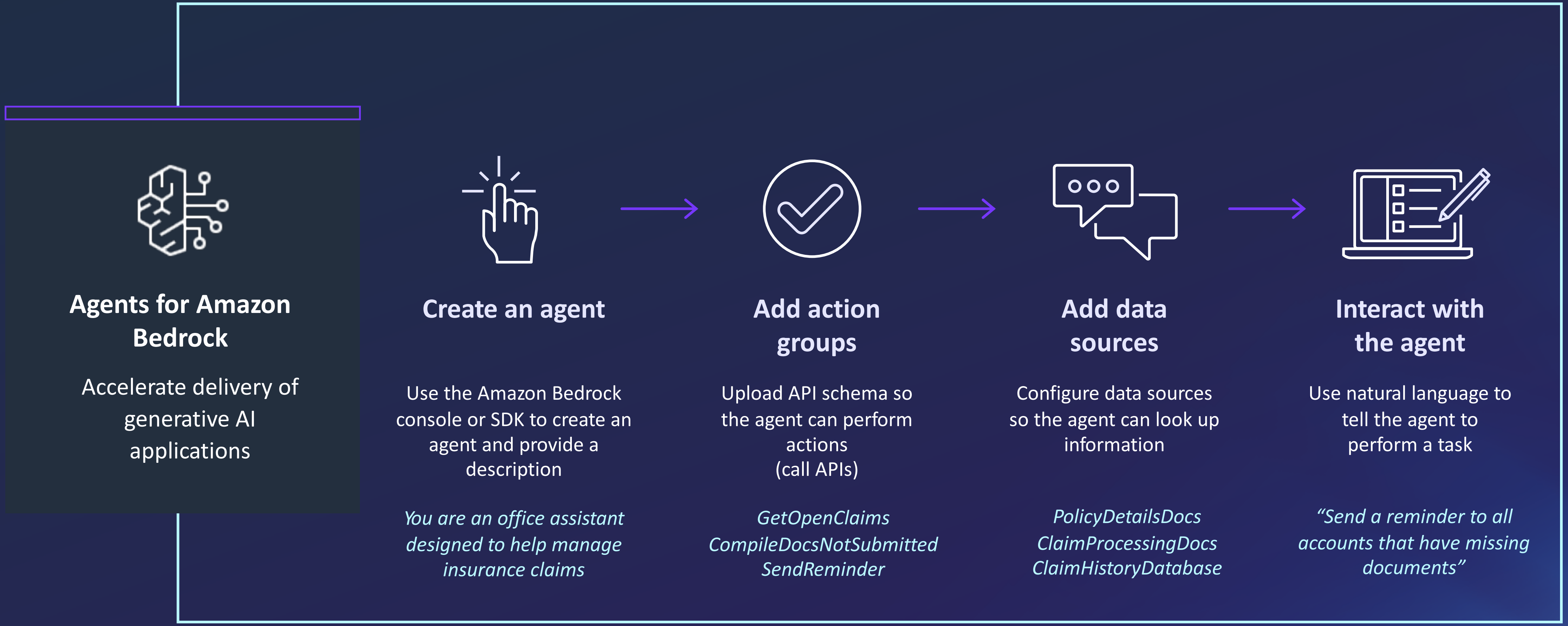
Create a version of the guardrail. Apply the guardrail during model inference or attach it to an agent.

Guardrails (1)

	Name	Status	Description	Creation time	Last edited
○	BankingAssistantGuardrail	✓ READY	Guardrails for online banking assistant to help users with banking and account related questions.	November 25, 2023, 19:29 (UTC-08:00)	5 mins ago

Agents for Amazon Bedrock

How to create Agents in Amazon Bedrock



Agents for Amazon Bedrock

Accelerate delivery of generative AI applications

Create an agent

Use the Amazon Bedrock console or SDK to create an agent and provide a description

You are an office assistant designed to help manage insurance claims

Add action groups

Upload API schema so the agent can perform actions (call APIs)

GetOpenClaims
CompileDocsNotSubmitted
SendReminder

Add data sources

Configure data sources so the agent can look up information

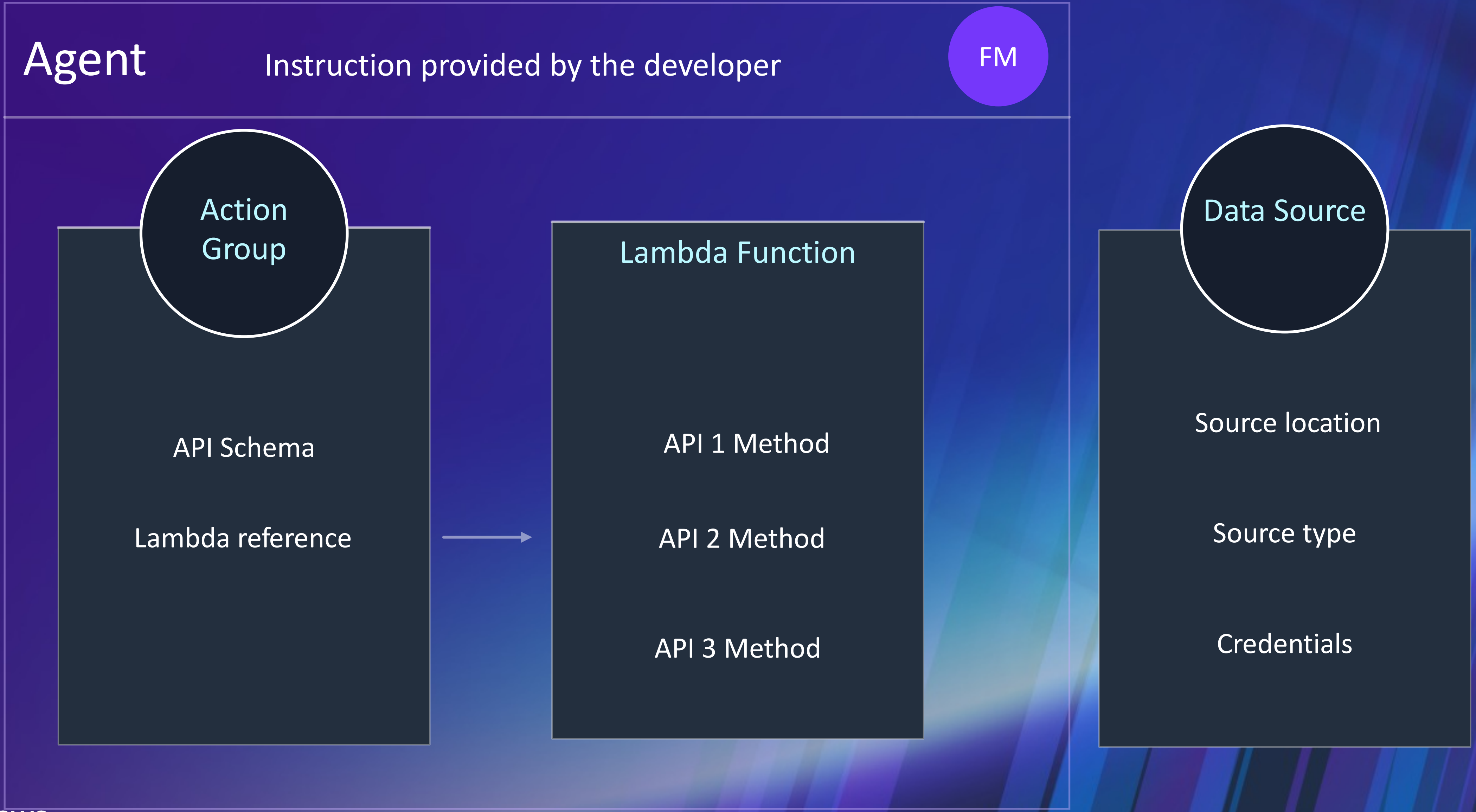
PolicyDetailsDocs
ClaimProcessingDocs
ClaimHistoryDatabase

Interact with the agent

Use natural language to tell the agent to perform a task

"Send a reminder to all accounts that have missing documents"

Agent structure



Agent

Instruction provided by the developer

FM

Action Group

API Schema

Lambda reference

Lambda Function

API 1 Method

API 2 Method

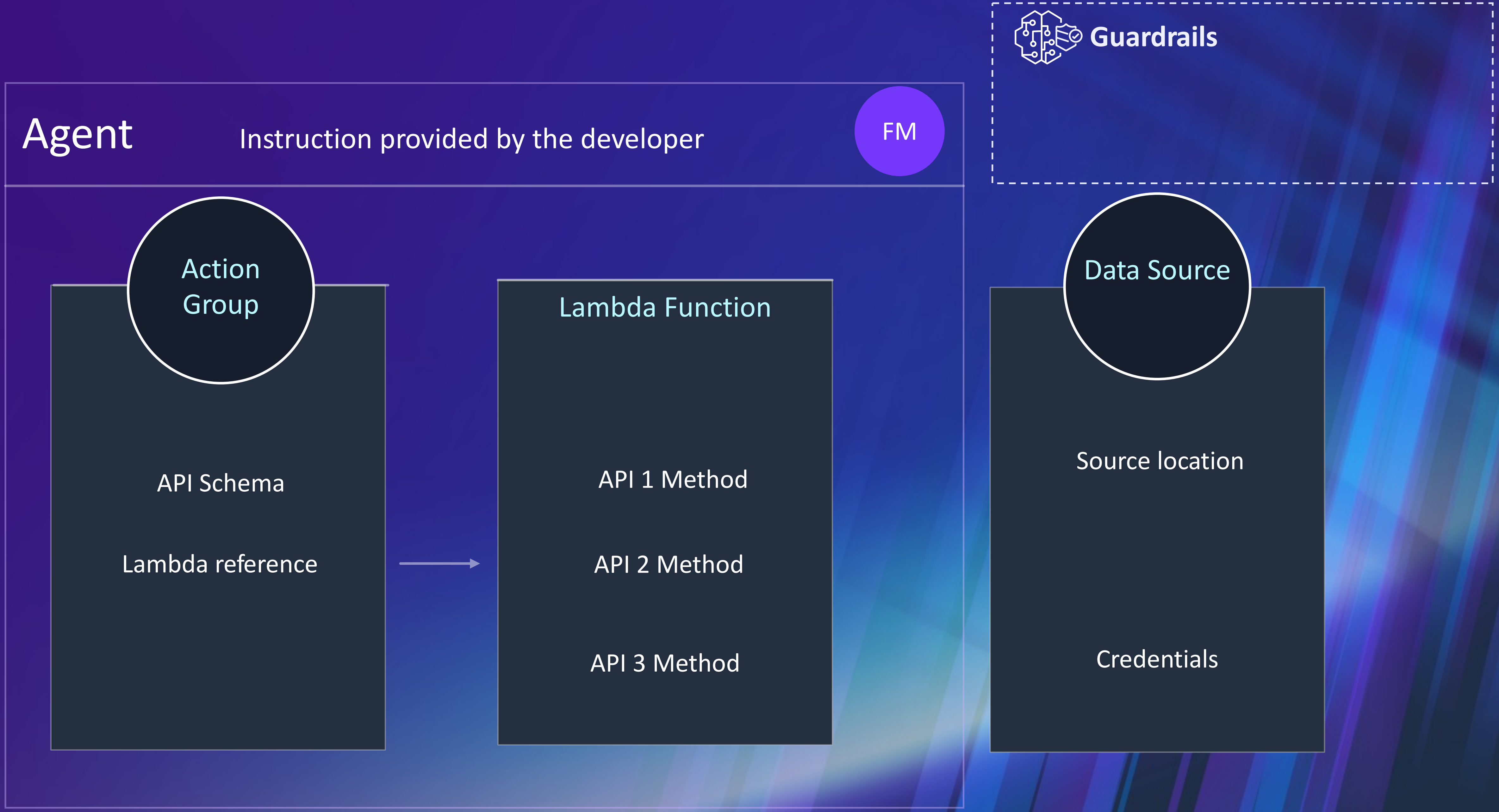
API 3 Method

Data Source

Source location

Source type

Credentials



Guardrails

Agent

Instruction provided by the developer

FM

Action Group

API Schema

Lambda reference

Lambda Function

API 1 Method

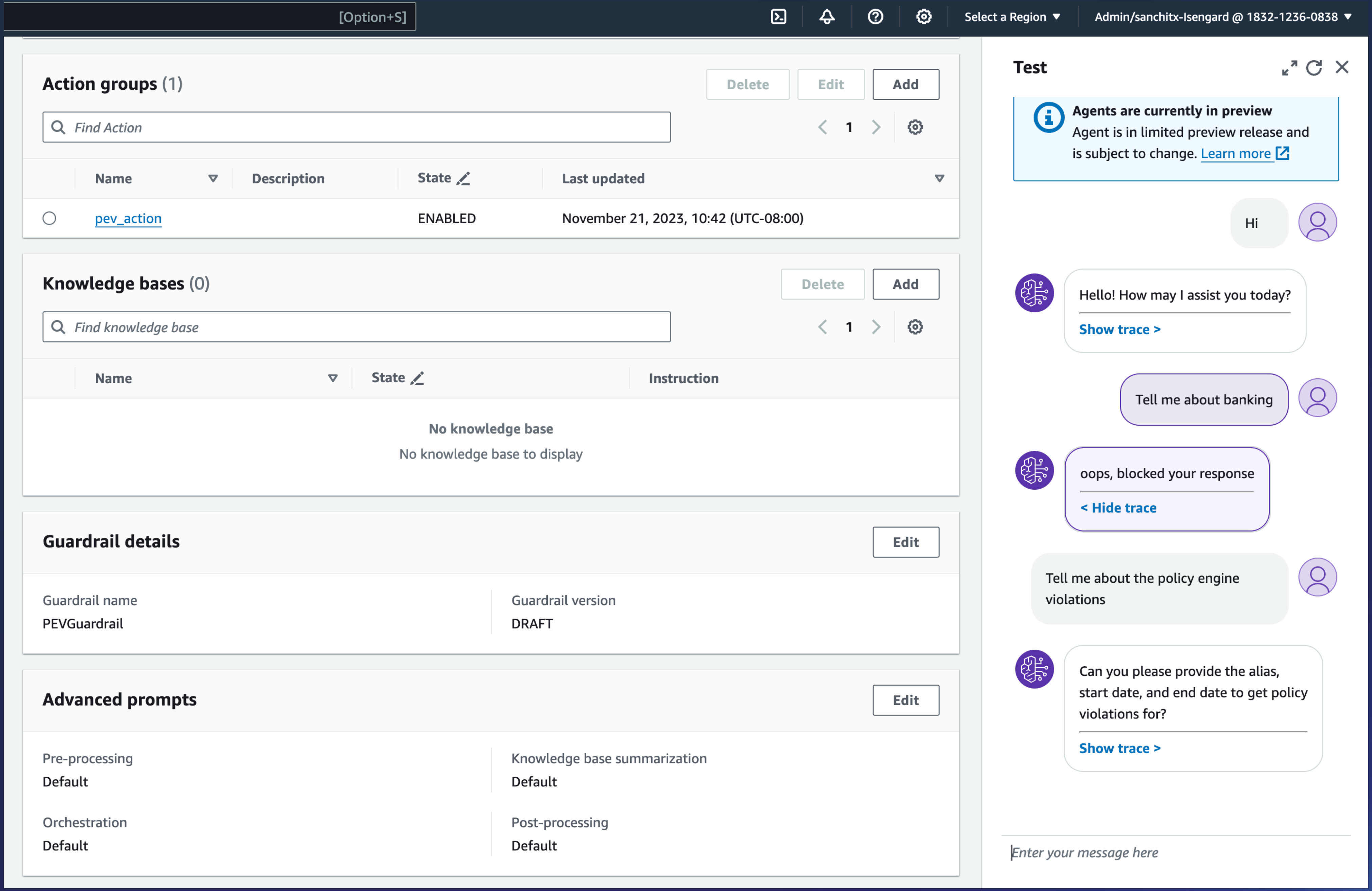
API 2 Method

API 3 Method

Data Source

Source location

Credentials



Action groups (1)

	Name	Description	State	Last updated
○	pev_action		ENABLED	November 21, 2023, 10:42 (UTC-08:00)

Knowledge bases (0)

	Name	State	Instruction

No knowledge base

No knowledge base to display

Guardrail details

Guardrail name
PEVGuardrail

Guardrail version
DRAFT

Advanced prompts

Pre-processing
Default

Knowledge base summarization
Default

Orchestration
Default

Post-processing
Default

Test

Agents are currently in preview
Agent is in limited preview release and is subject to change. Learn more

Hi

Hello! How may I assist you today?
Show trace >

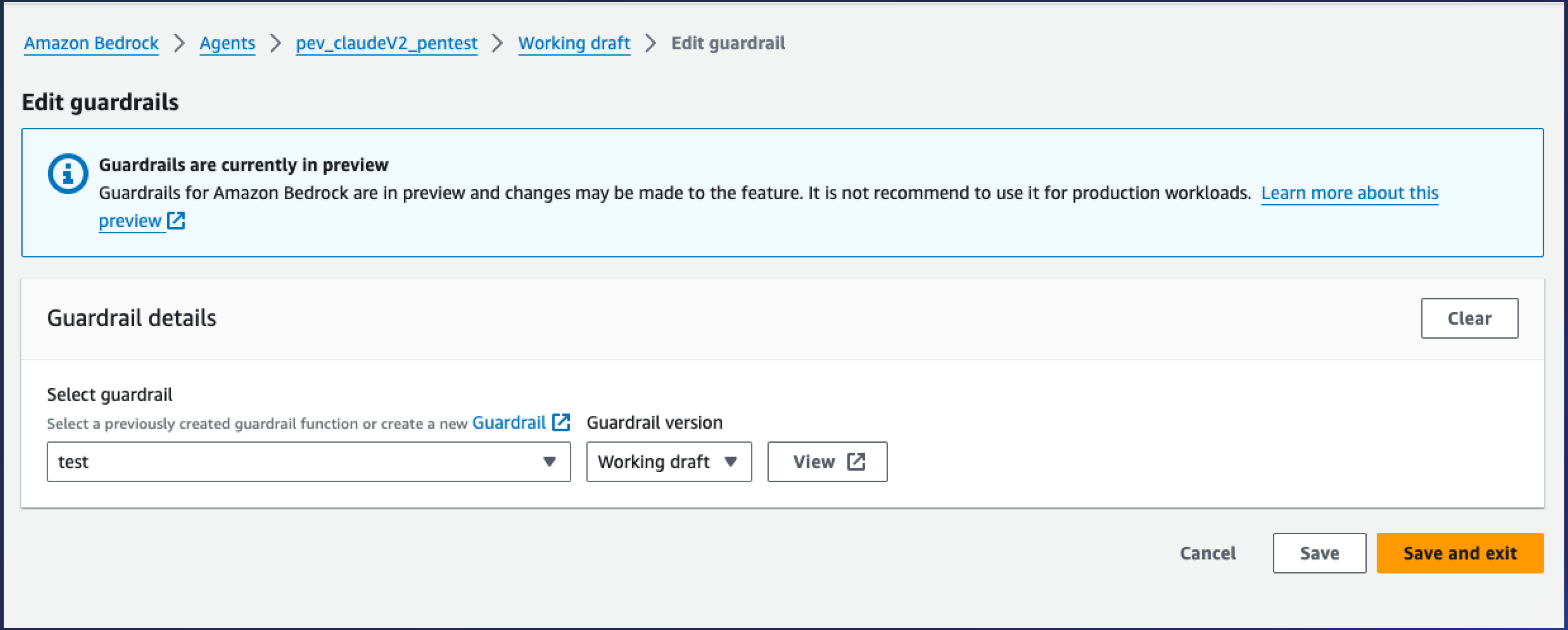
Tell me about banking

oops, blocked your response
< Hide trace

Tell me about the policy engine violations

Can you please provide the alias, start date, and end date to get policy violations for?
Show trace >

Enter your message here



Amazon Bedrock > Agents > pev_claudeV2_pentest > Working draft > Edit guardrail

Edit guardrails

Guardrails are currently in preview

Guardrails for Amazon Bedrock are in preview and changes may be made to the feature. It is not recommend to use it for production workloads. Learn more about this preview

Guardrail details

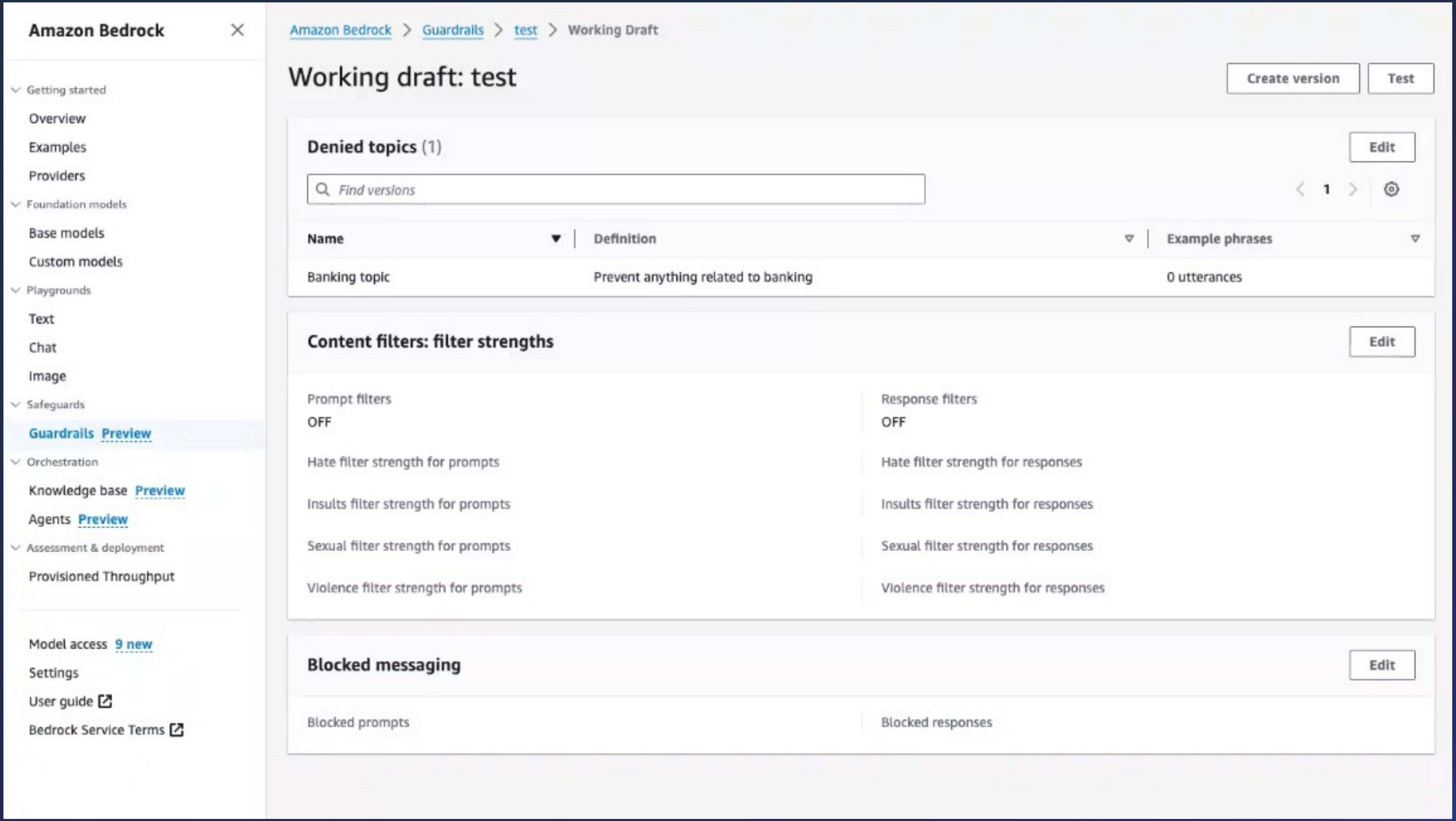
Select guardrail

Select a previously created guardrail function or create a new Guardrail

test

Guardrail version

Working draft



Working draft: test

Denied topics (1)

Name	Definition	Example phrases
Banking topic	Prevent anything related to banking	0 utterances

Content filters: filter strengths

Prompt filters
OFF

Response filters
OFF

Hate filter strength for prompts

Hate filter strength for responses

Insults filter strength for prompts

Insults filter strength for responses

Sexual filter strength for prompts

Sexual filter strength for responses

Violence filter strength for prompts

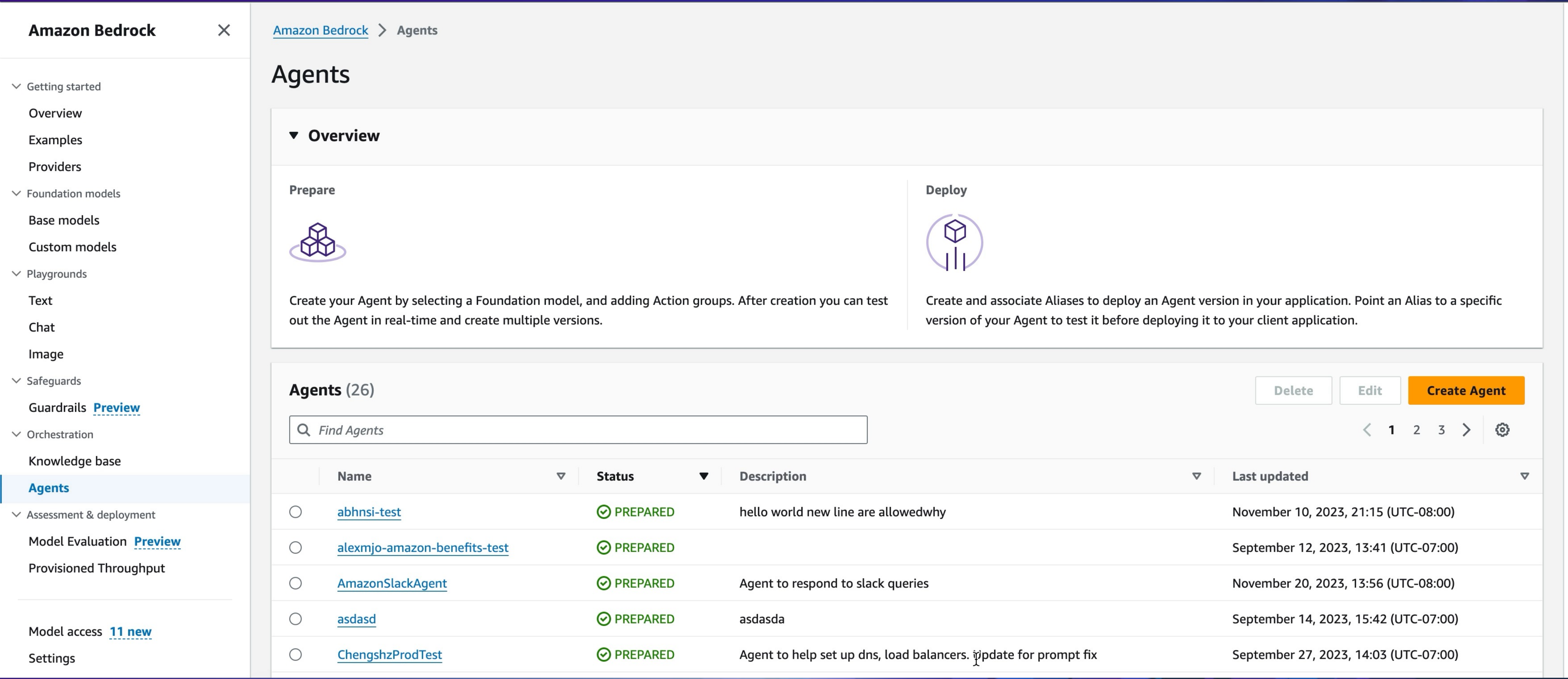
Violence filter strength for responses

Blocked messaging

Blocked prompts

Blocked responses

Demo: Agents + Guardrails



Agents

Overview

Prepare

Create your Agent by selecting a Foundation model, and adding Action groups. After creation you can test out the Agent in real-time and create multiple versions.

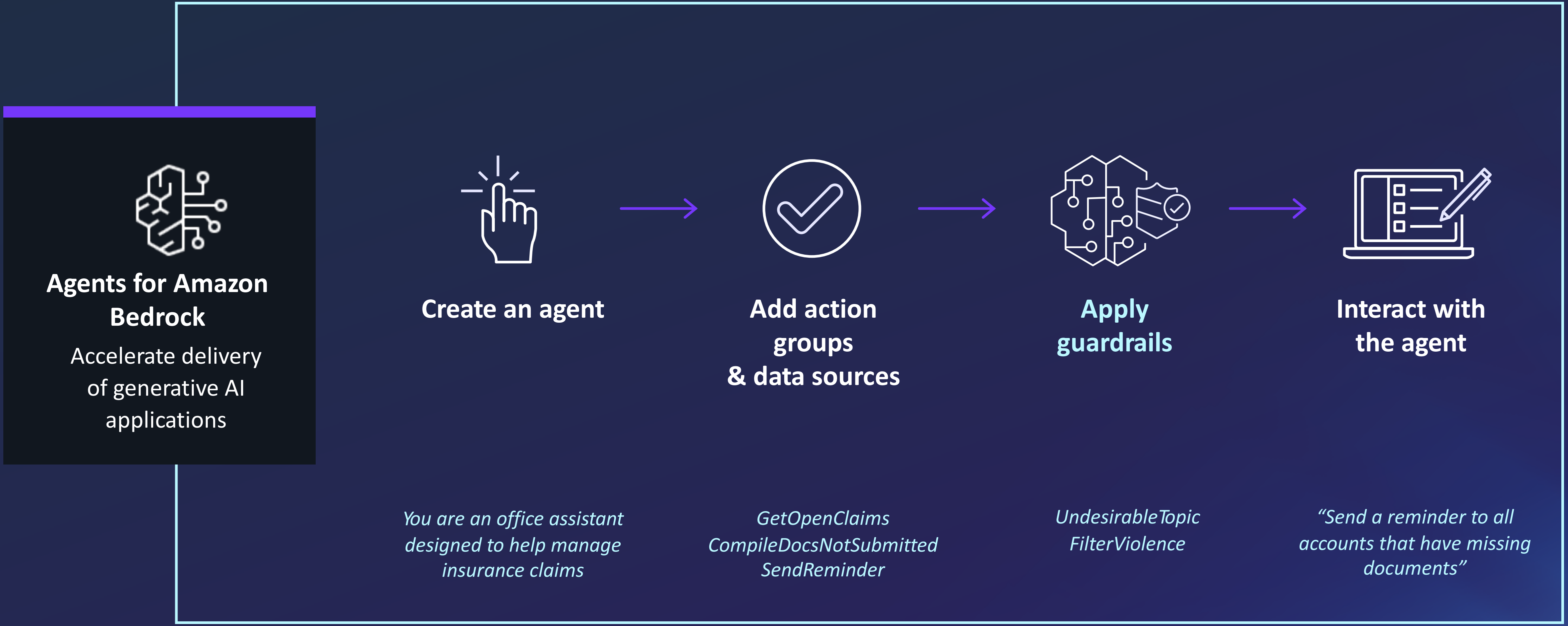
Deploy

Create and associate Aliases to deploy an Agent version in your application. Point an Alias to a specific version of your Agent to test it before deploying it to your client application.

Agents (26)

	Name	Status	Description	Last updated
○	abhnsi-test	✓ PREPARED	hello world new line are allowedwhy	November 10, 2023, 21:15 (UTC-08:00)
○	alexmjo-amazon-benefits-test	✓ PREPARED		September 12, 2023, 13:41 (UTC-07:00)
○	AmazonSlackAgent	✓ PREPARED	Agent to respond to slack queries	November 20, 2023, 13:56 (UTC-08:00)
○	asdasd	✓ PREPARED	asdasda	September 14, 2023, 15:42 (UTC-07:00)
○	ChengshzProdTest	✓ PREPARED	Agent to help set up dns, load balancers. update for prompt fix	September 27, 2023, 14:03 (UTC-07:00)

How to create Agents in Amazon Bedrock



Agents for Amazon Bedrock

Accelerate delivery of generative AI applications

Create an agent

You are an office assistant designed to help manage insurance claims

Add action groups & data sources

GetOpenClaims
CompileDocsNotSubmitted
SendReminder

Apply guardrails

UndesirableTopic
FilterViolence

Interact with the agent

“Send a reminder to all accounts that have missing documents”

Thank you!

Christian Kamwangala

linkedin.com/in/christian-kamwangala