AWS re:Inforce

JUNE 10 – 12, 2024 | PHILADELPHIA, PA

GRC325

Build responsible AI applications with Guardrails for Amazon Bedrock

Anubhav Mishra

(he/him)
Principal Product Manager
AWS

Building generative apps brings new challenges

Undesirable and irrelevant topics

Controversial queries and responses

Toxicity and safety (including 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

AI21labs	amazon	ANTHROP\C	cohere	Meta	MISTRAL AI_	stability.ai
Jurassic-2	Titan	Claude	Command + Embed	Llama	Mistral + Mixtral	Stable Diffusion

Building generative AI apps requires additional controls

Customizations based on use cases and organizational policy

Safety and privacy controls for responsible AI

Consistent safeguards across FMs and applications

Guardrails for Amazon Bedrock

Implement safeguards customized to your application requirements and responsible AI policies

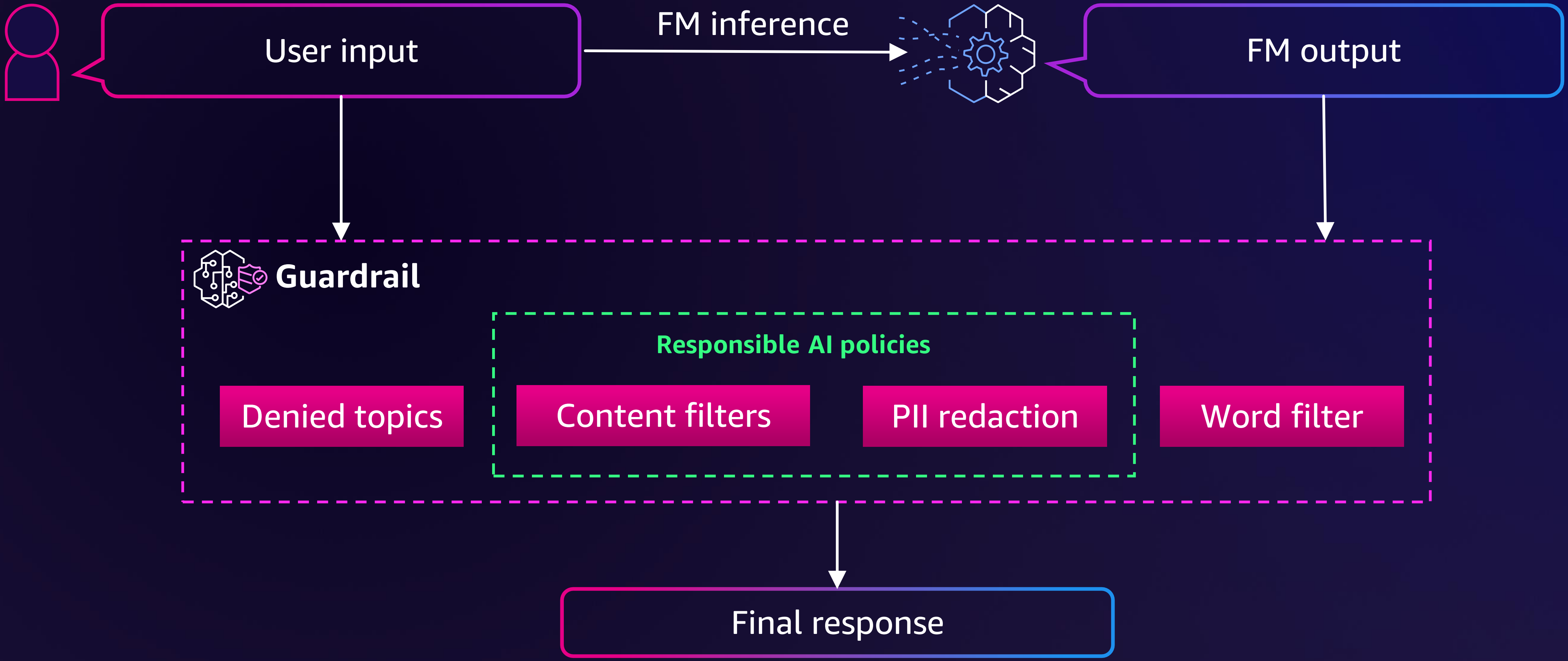
Apply guardrails to multiple foundation models, knowledge bases, and agents for Amazon Bedrock

Filter harmful content and safeguard against prompt injection and jailbreaks

Define and disallow denied topics with short natural language descriptions

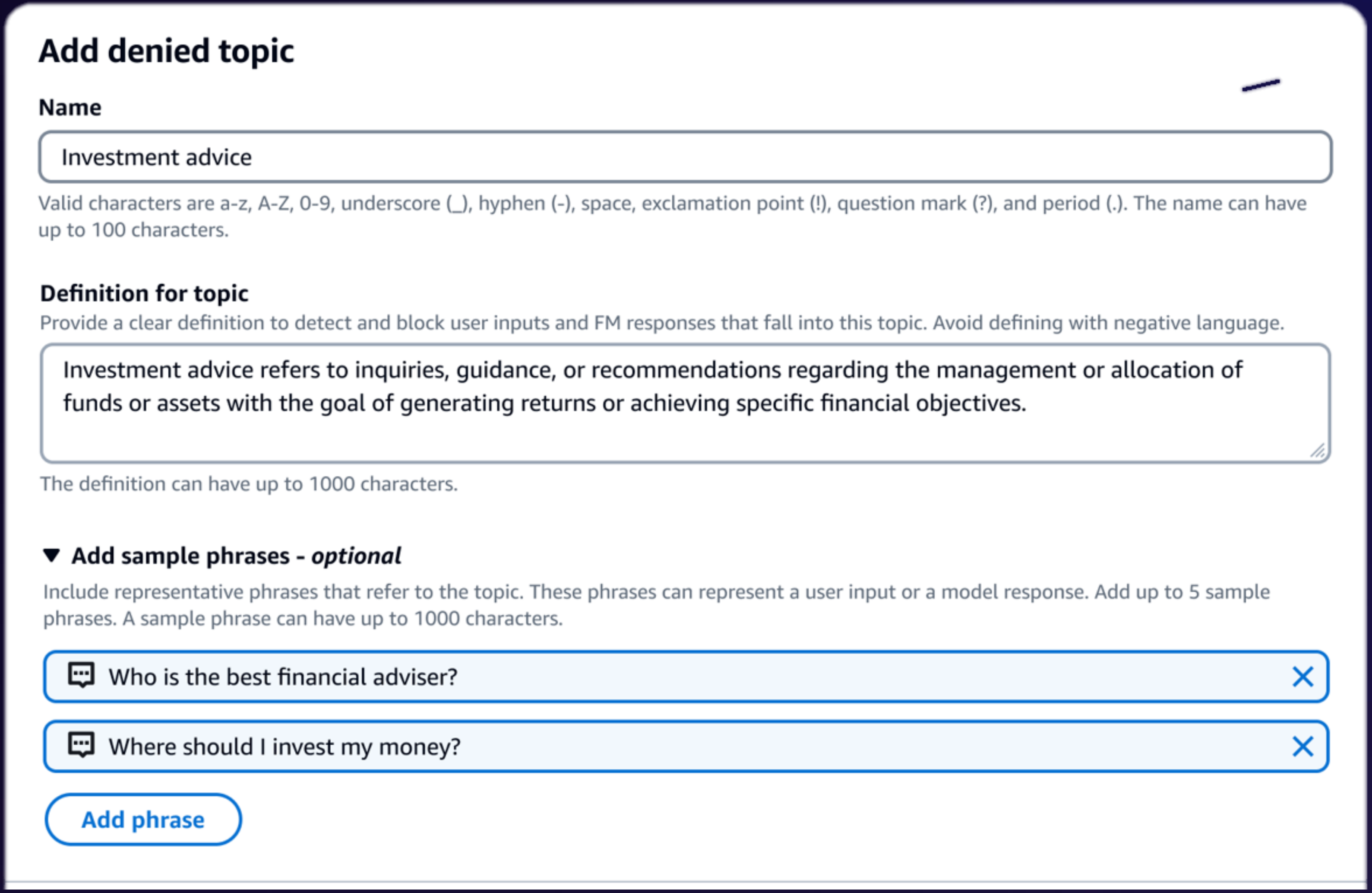
Redact or block sensitive information, such as PIIs, and custom regex (regular expressions)

How it works: Guardrails for Amazon Bedrock



Denied topics

AVOID UNDESIRABLE TOPICS IN YOUR APPLICATIONS



Add denied topic

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

Provide a clear definition to detect and block user inputs and FM responses that fall into this topic. Avoid defining with negative language.

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.

▼ **Add sample phrases - *optional***

Include representative phrases that refer to the topic. These phrases can represent a user input or a model response. Add up to 5 sample phrases. A sample phrase can have up to 1000 characters.

Who is the best financial adviser? ✕

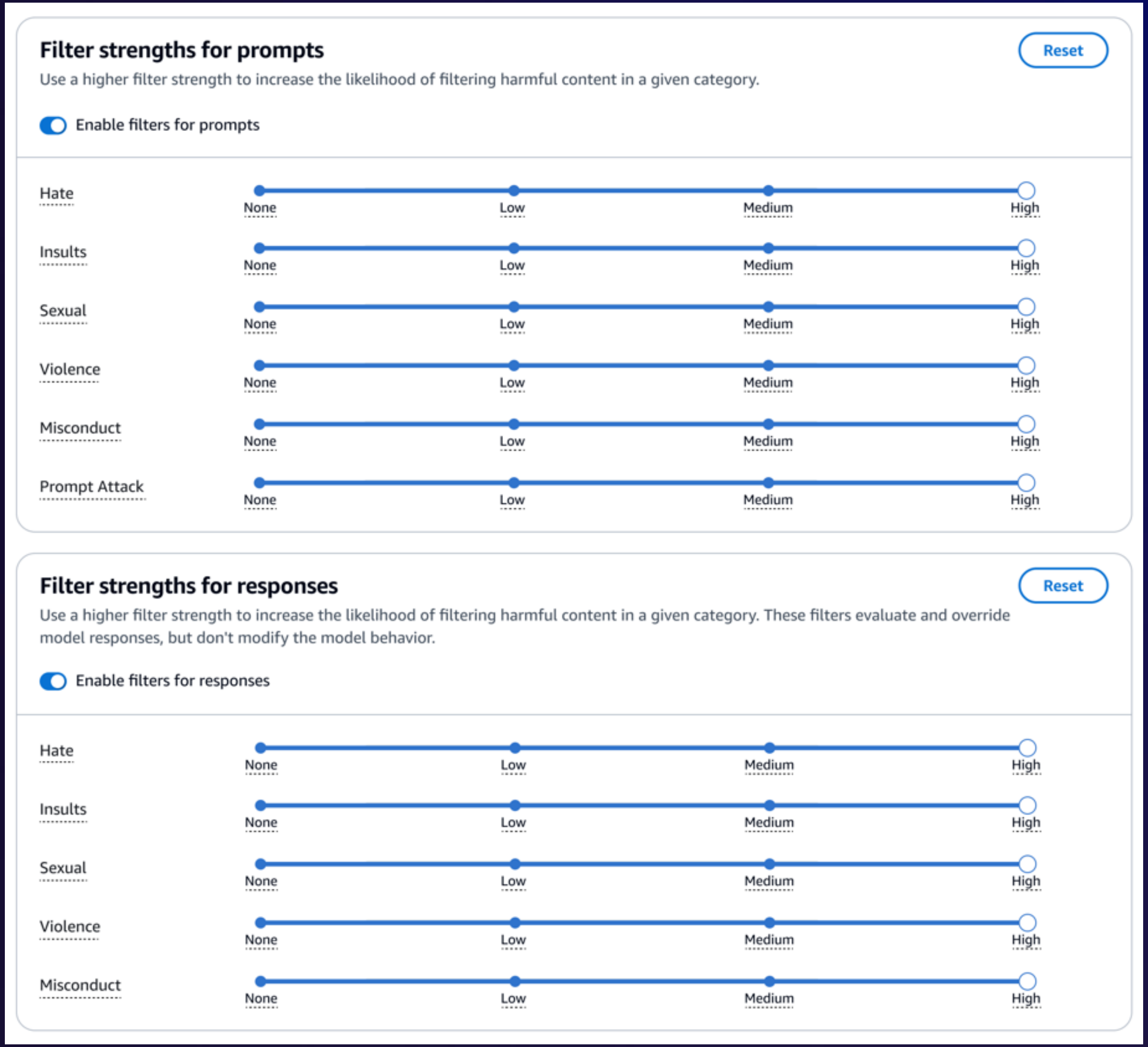
Where should I invest my money? ✕

Content filters

CONFIGURE THRESHOLDS TO FILTER CONTENT TO VARYING DEGREES

Filter harmful content across categories:

- Hate
- Insults
- Sexual
- Violence
- Misconduct (criminal activity)
- Prompt Attack (jailbreak and prompt injection)



Filter strengths for prompts

Reset

Use a higher filter strength to increase the likelihood of filtering harmful content in a 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]
Misconduct	None	Low	Medium	[✓ High]
Prompt Attack	None	Low	Medium	[✓ High]

Filter strengths for responses

Reset

Use a higher filter strength to increase the likelihood of filtering harmful content in a given category. These filters evaluate and override model responses, but don't modify the model behavior.

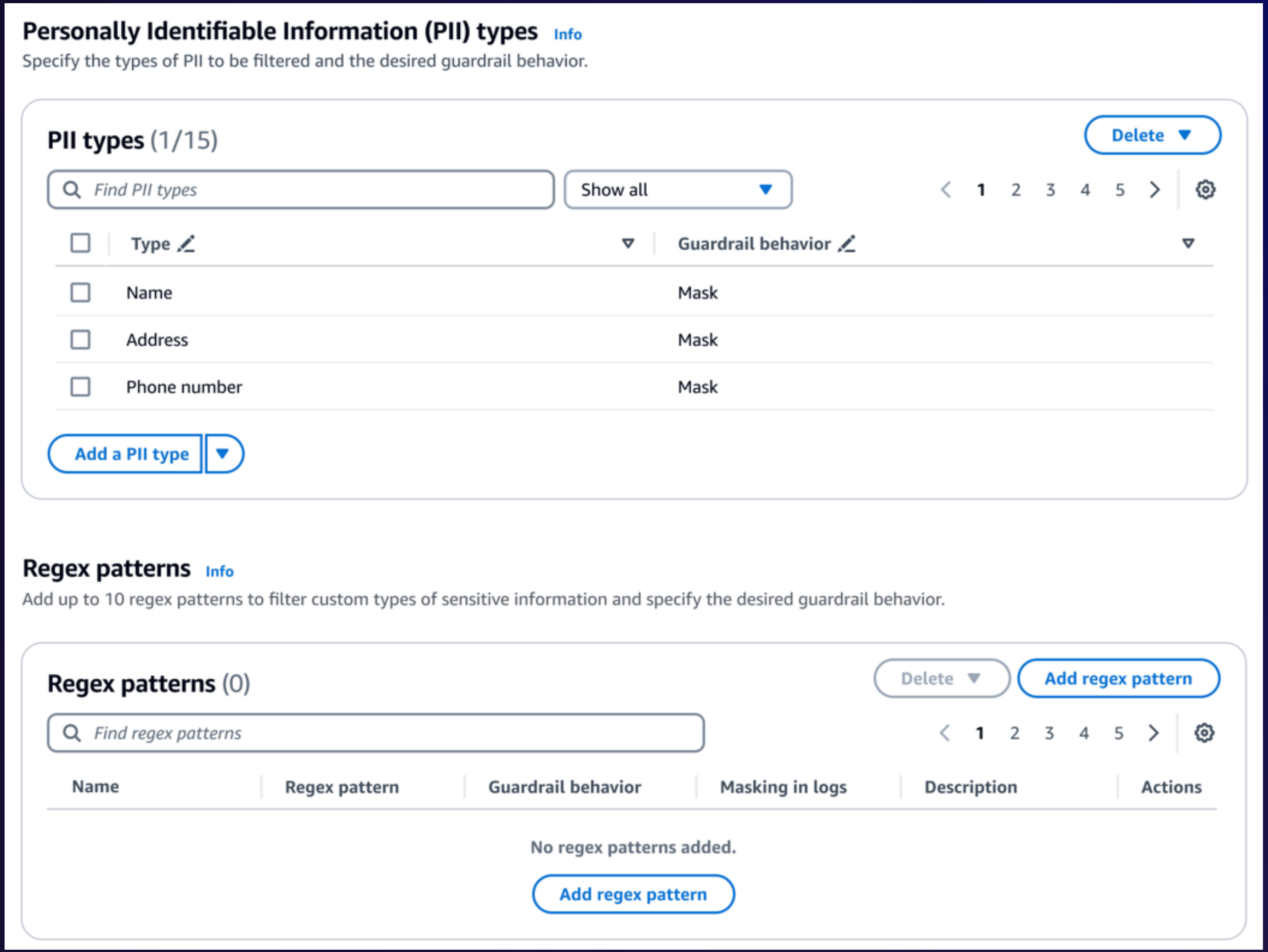
☑ Enable filters for responses

Hate	None	Low	Medium	[✓ High]
Insults	None	Low	Medium	[✓ High]
Sexual	None	Low	Medium	[✓ High]
Violence	None	Low	Medium	[✓ High]
Misconduct	None	Low	Medium	[✓ High]

Sensitive information filter

PROTECT SENSITIVE INFORMATION AND PRESERVE USER PRIVACY

- ➢ 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
- ➢ Define your own sensitive information using regular expressions (regex)



Personally Identifiable Information (PII) types

Specify the types of PII to be filtered and the desired guardrail behavior.

PII types (1/15)

□	Type	Guardrail behavior
□	Name	Mask
□	Address	Mask
□	Phone number	Mask

Regex patterns

Add up to 10 regex patterns to filter custom types of sensitive information and specify the desired guardrail behavior.

Regex patterns (0)

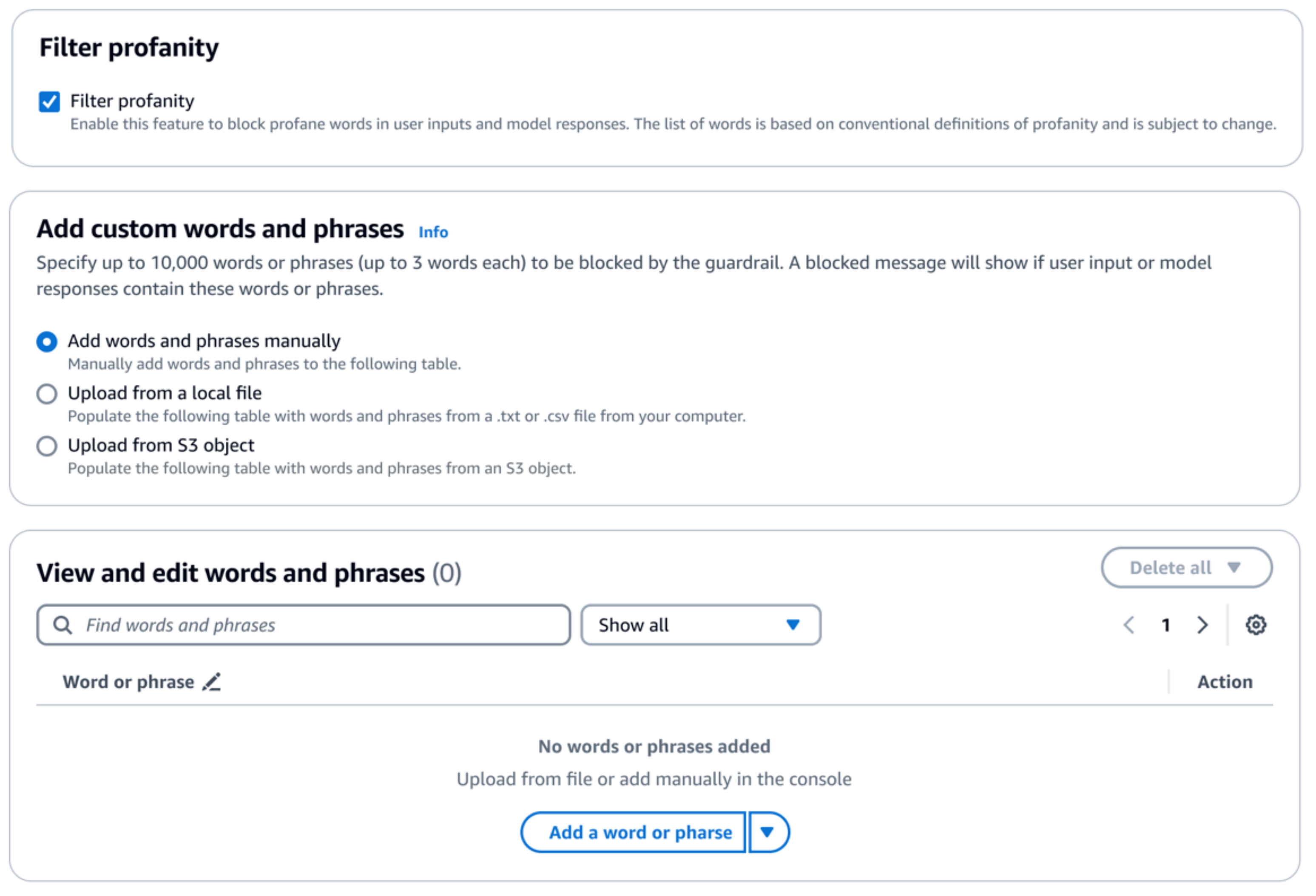
Name	Regex pattern	Guardrail behavior	Masking in logs	Description	Actions

No regex patterns added.

Word filters

BLOCK CUSTOM WORDS AND PROFANITY

- Filter profane words
- Define a set of custom words to block user input and FM responses



Filter profanity

☑ Filter profanity
Enable this feature to block profane words in user inputs and model responses. The list of words is based on conventional definitions of profanity and is subject to change.

Add custom words and phrases Info

Specify up to 10,000 words or phrases (up to 3 words each) to be blocked by the guardrail. A blocked message will show if user input or model responses contain these words or phrases.

● Add words and phrases manually
Manually add words and phrases to the following table.

○ Upload from a local file
Populate the following table with words and phrases from a .txt or .csv file from your computer.

○ Upload from S3 object
Populate the following table with words and phrases from an S3 object.

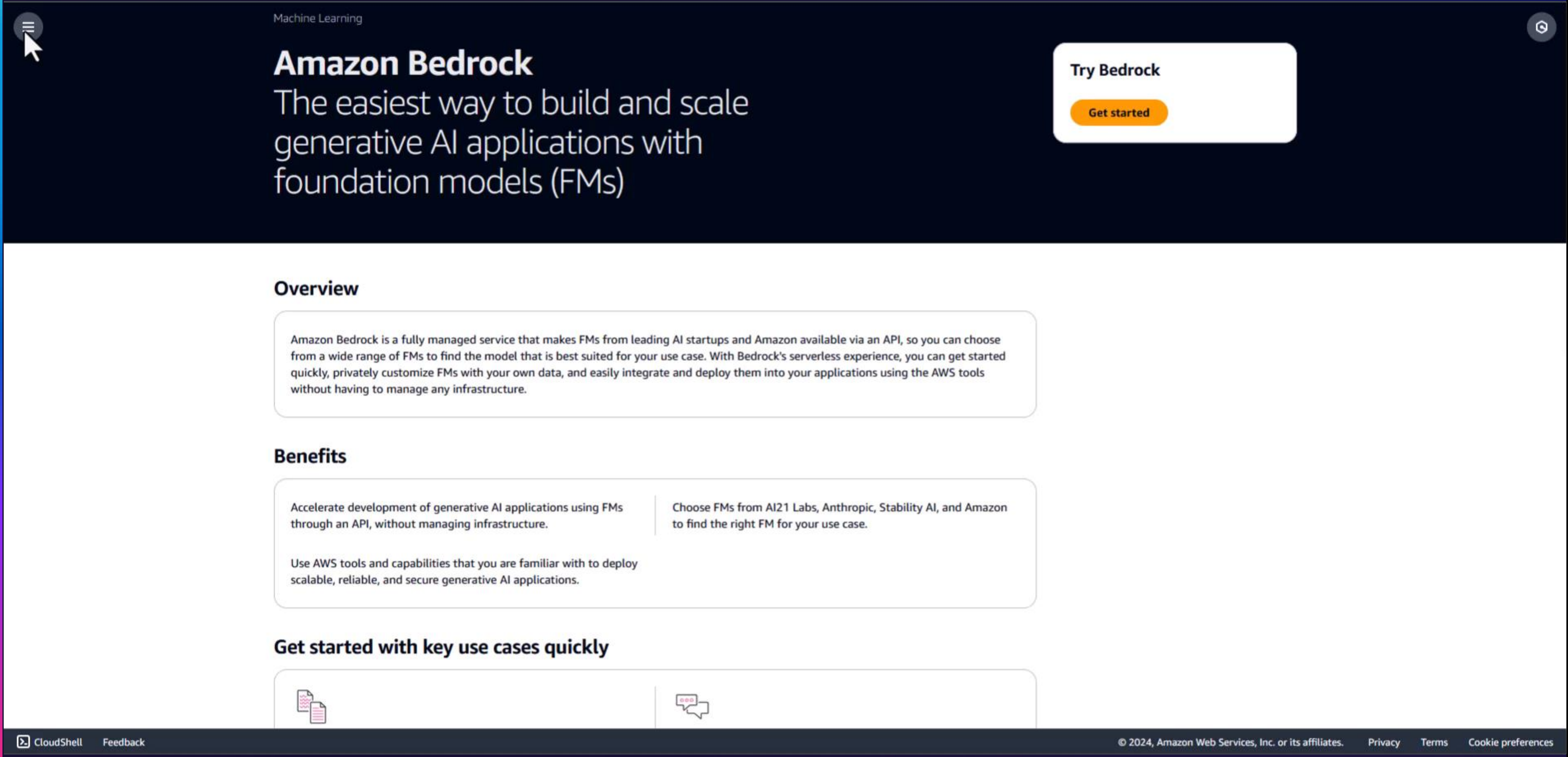
View and edit words and phrases (0)

No words or phrases added

Upload from file or add manually in the console

Guardrails demo

Guardrails demo



Machine Learning

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.

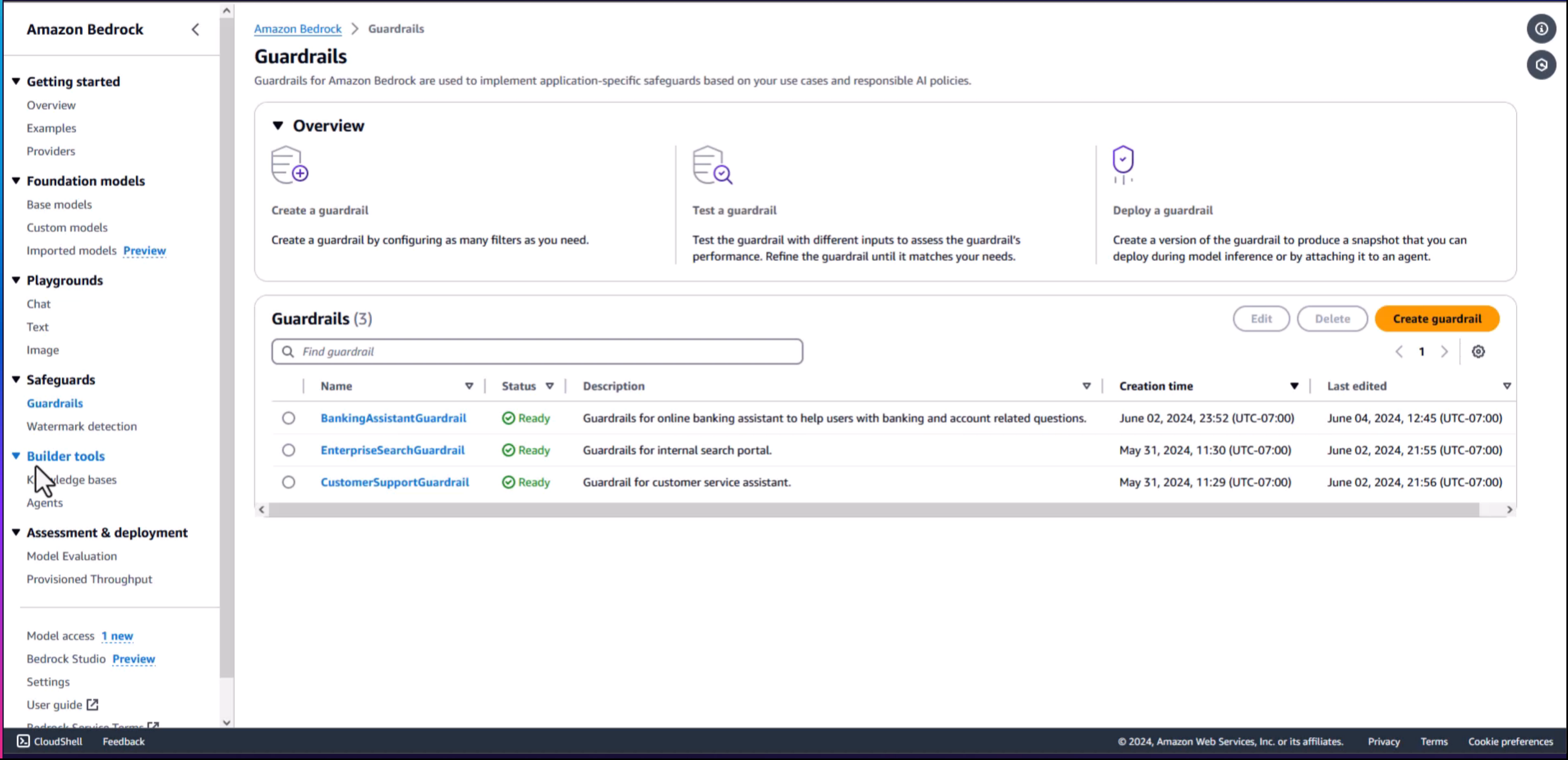
Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case.

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

Agents with guardrail demo

Agents with guardrail demo



Amazon Bedrock > Guardrails

Guardrails

Guardrails for Amazon Bedrock are used to implement application-specific safeguards based on your use cases and responsible AI policies.

Overview

Create a guardrail

Create a guardrail by configuring as many filters as you need.

Test a guardrail

Test the guardrail with different inputs to assess the guardrail's performance. Refine the guardrail until it matches your needs.

Deploy a guardrail

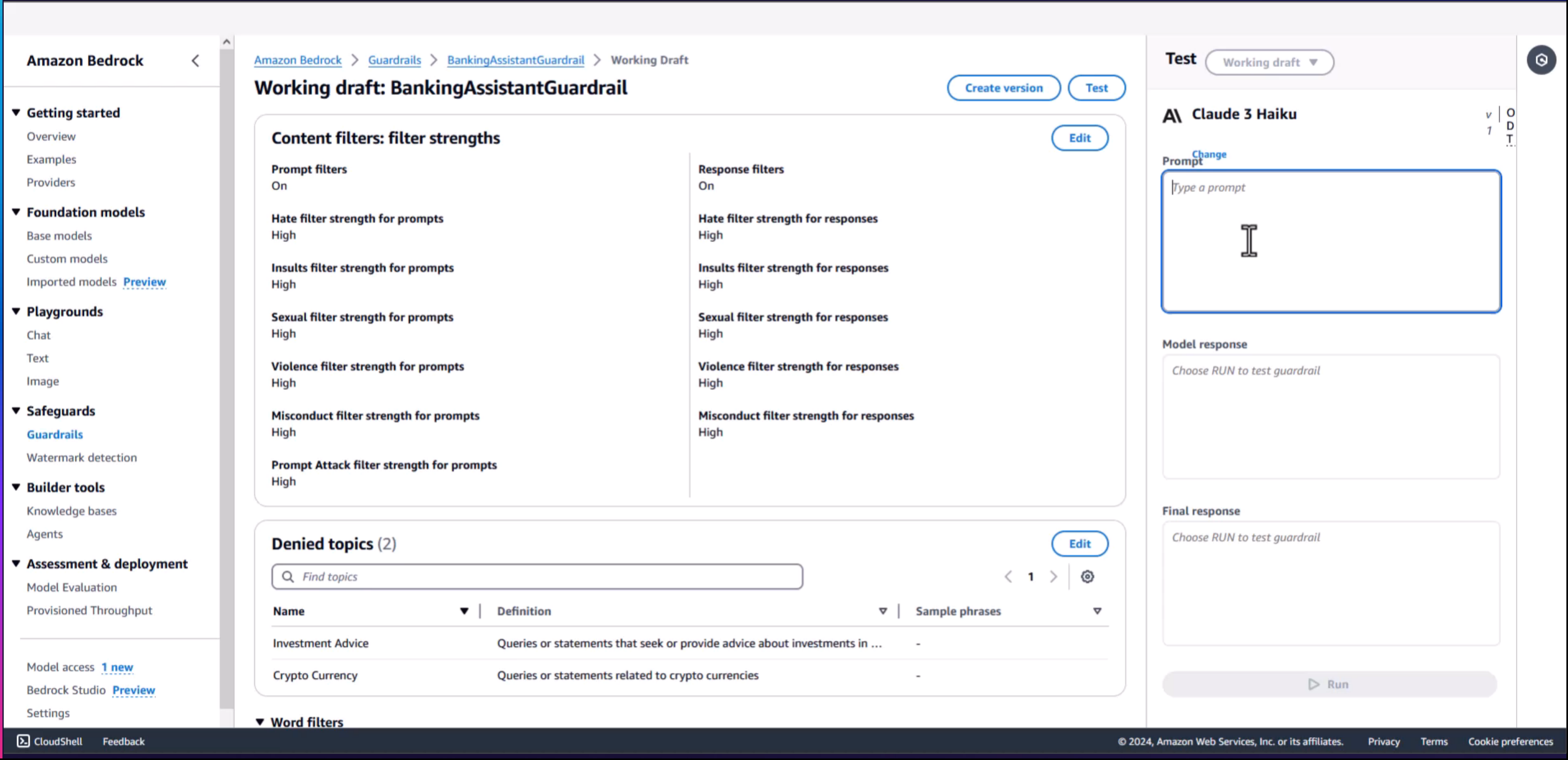
Create a version of the guardrail to produce a snapshot that you can deploy during model inference or by attaching it to an agent.

Guardrails (3)

	Name	Status	Description	Creation time	Last edited
○	BankingAssistantGuardrail	✓ Ready	Guardrails for online banking assistant to help users with banking and account related questions.	June 02, 2024, 23:52 (UTC-07:00)	June 04, 2024, 12:45 (UTC-07:00)
○	EnterpriseSearchGuardrail	✓ Ready	Guardrails for internal search portal.	May 31, 2024, 11:30 (UTC-07:00)	June 02, 2024, 21:55 (UTC-07:00)
○	CustomerSupportGuardrail	✓ Ready	Guardrail for customer service assistant.	May 31, 2024, 11:29 (UTC-07:00)	June 02, 2024, 21:56 (UTC-07:00)

Sensitive information filter demo

Sensitive information filter demo



Working draft: BankingAssistantGuardrail

Content filters: filter strengths

Prompt filters	Response filters
Prompt filters On	**Response filters** On
Hate filter strength for prompts High	**Hate filter strength for responses** High
Insults filter strength for prompts High	**Insults filter strength for responses** High
Sexual filter strength for prompts High	**Sexual filter strength for responses** High
Violence filter strength for prompts High	**Violence filter strength for responses** High
Misconduct filter strength for prompts High	**Misconduct filter strength for responses** High
Prompt Attack filter strength for prompts High	

Denied topics (2)

Find topics

Name	Definition	Sample phrases
Investment Advice	Queries or statements that seek or provide advice about investments in ...	-
Crypto Currency	Queries or statements related to crypto currencies	-

▼ Word filters

Prompt

Type a prompt

Model response

Choose RUN to test guardrail

Final response

Choose RUN to test guardrail

© 2024, Amazon Web Services, Inc. or its affiliates.