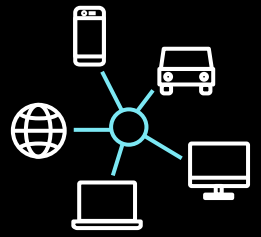
Amazon Bedrock

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

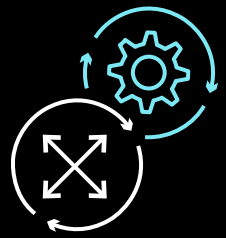
Agenda

- Customer needs for generative AI
- Service overview
- Key features
- Customers
- Getting started

What generative AI customers are asking for



Which model should I use?



How can I move quickly?

How can I keep my data secure and private?

Amazon Bedrock

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

- Choice of leading FMs through a single API
- Model customization
- Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Security, privacy, and safety

Amazon Bedrock simplifies

Choice

Customization

Integration

Security and governance

Amazon Bedrock

Broad choice of models

AI21 labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI_	stability.ai
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, Q&A, text classification, text completion, code generation	High-quality images and art
Jurassic-2 Ultra	Amazon Titan Text Premier	Claude 3 Opus	Command	Llama 3 8B	Mistral Large	Stable Diffusion XL1.0
Jurassic-2 Mid	Amazon Titan Text Lite	Claude 3 Sonnet	Command Light	Llama 3 70B	Mistral 7B	Stable Diffusion XL 0.8
	Amazon Titan Text Express	Claude 3 Haiku	Embed English	Llama 2 13B	Mixtral 8x7B	
	Amazon Titan Text Embeddings	Claude 2.1	Embed Multilingual	Llama 2 70B		
	Amazon Titan Text Embeddings V2	Claude 2	Command R+			
	Amazon Titan Multimodal Embeddings	Claude Instant	Command R			
	Amazon Titan Image Generator					

aws

Meta Llama 3 (8B and 70B)

Llama 3 is designed for developers, researchers, and businesses to build, experiment, and responsibly scale generative AI ideas

Meta

1. Generation over generation, Llama 3 demonstrates state-of-the-art performance on a wide range of industry benchmarks and offers new capabilities, including improved reasoning

2. **Llama 3 8B:**
   ideal for limited computational power and resources, faster training times, and edge devices

3. **Llama 3 70B:**
   ideal for content creation, conversational AI, language understanding, research development, and enterprise applications

Cohere Command R+ and Command R *(Coming Soon)*

Build enterprise generative AI and advanced multilingual applications with Cohere

1. **Command R+:**
   Cohere's most powerful generative language model optimized for long-context tasks, such as RAG and multistep tool use

2. **Command R:**
   a generative language model optimized for long-context tasks, such as RAG and tools, and large-scale production workloads

3. These models balance strong accuracy and efficiency to empower businesses to move beyond proof of concept and start using AI in day-to-day operations

Claude 3 family in Amazon Bedrock

CHOOSE THE EXACT COMBINATION OF INTELLIGENCE, SPEED, AND COST TO SUIT YOUR NEEDS

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku
Use case	Most intelligence and highest performance	Balance between intelligence, speed, and cost	Fastest performance at the lowest cost
Context	200K	200K	200K
Vision	✓	✓	✓
Cost* **Input:** **Output:**	$0.015 $0.075	$0.003 $0.015	$0.00025 $0.00125

*Per 1K tokens

Amazon Titan

FMs

AMAZON TITAN TEXT EMBEDDINGS

Numerical representations of text

AMAZON TITAN TEXT LITE

Summarization, copywriting, fine-tuning

AMAZON TITAN TEXT EXPRESS

Open-ended text generation, conversational chat, RAG support

AMAZON TITAN TEXT PREMIER

Enterprise-grade text generation, optimized for RAG and Agents

AMAZON TITAN MULTIMODAL EMBEDDINGS

Search, recommendation, personalization

AMAZON TITAN IMAGE GENERATOR

Realistic, studio-quality images

Model evaluation on Amazon Bedrock

Evaluate, compare, and select the best FM for your use case

1. Use curated datasets or bring your own for tailored results
2. Apply automatic or human evaluation methods
3. Use your in-house team or reviewers managed by AWS
4. Provides predefined and custom metrics
5. Get results in just a few quick steps

Model evaluation in Amazon Bedrock

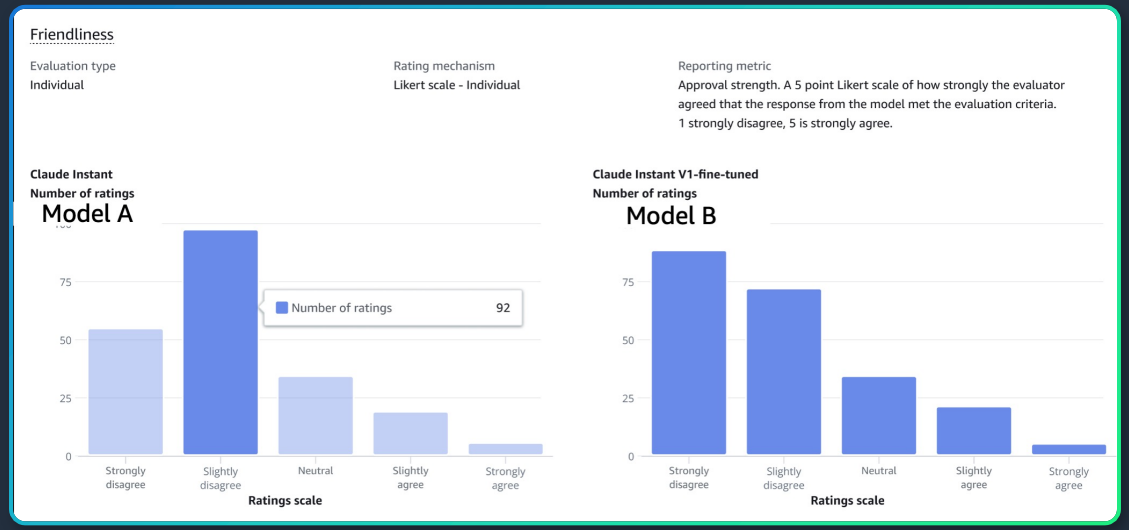
EVALUATE FMs TO SELECT THE BEST ONE FOR YOUR USE CASE

Automatic or human evaluation method

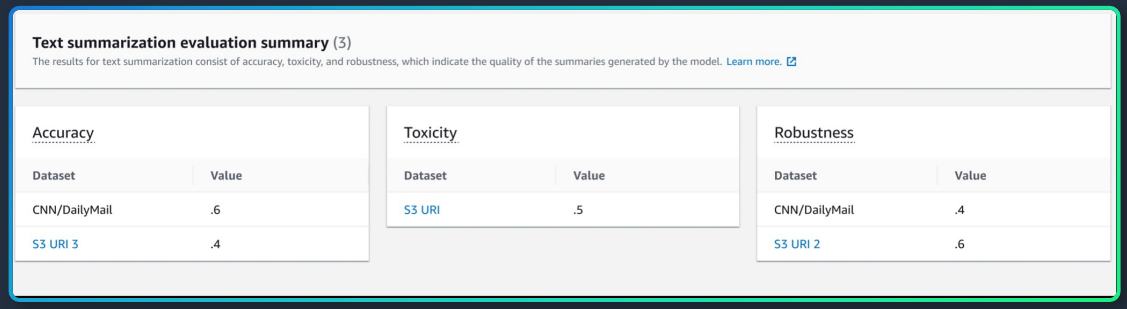
Curated datasets or bring your own

Predefined and custom metrics

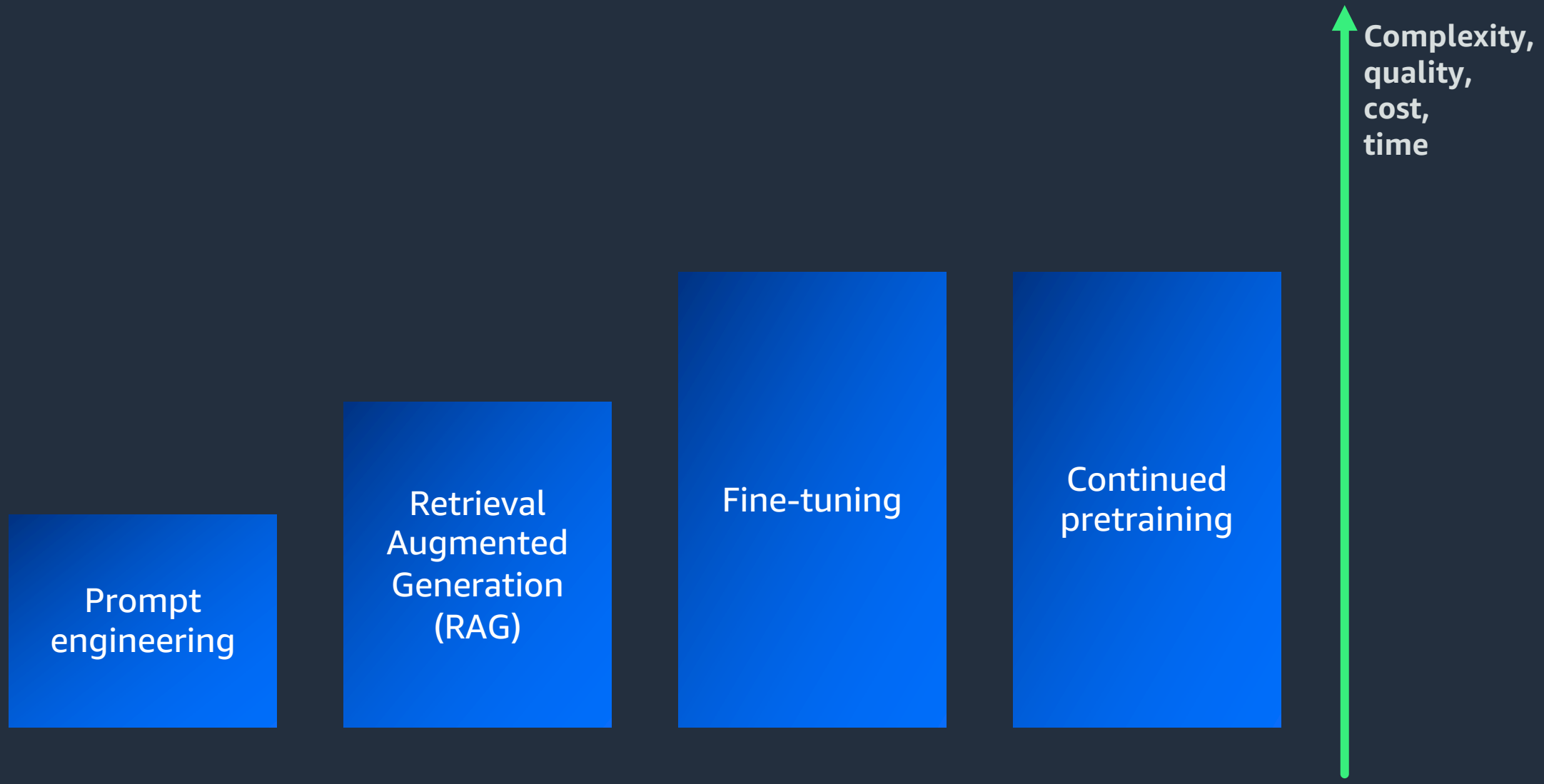
Human evaluation report



Automatic evaluation report



Common approaches for customizing FMs



Knowledge Bases now simplifies asking questions on a single document

Ask questions and summarize data from a document, without setting up a vector database

1. Ask questions, summarize content, and more without needing to ingest data into a vector data base.
2. Documents are retained only for the session. Low-cost method to use your single document for content retrieval and generation related tasks.
3. No data preparation required.

Knowledge Bases for Amazon Bedrock

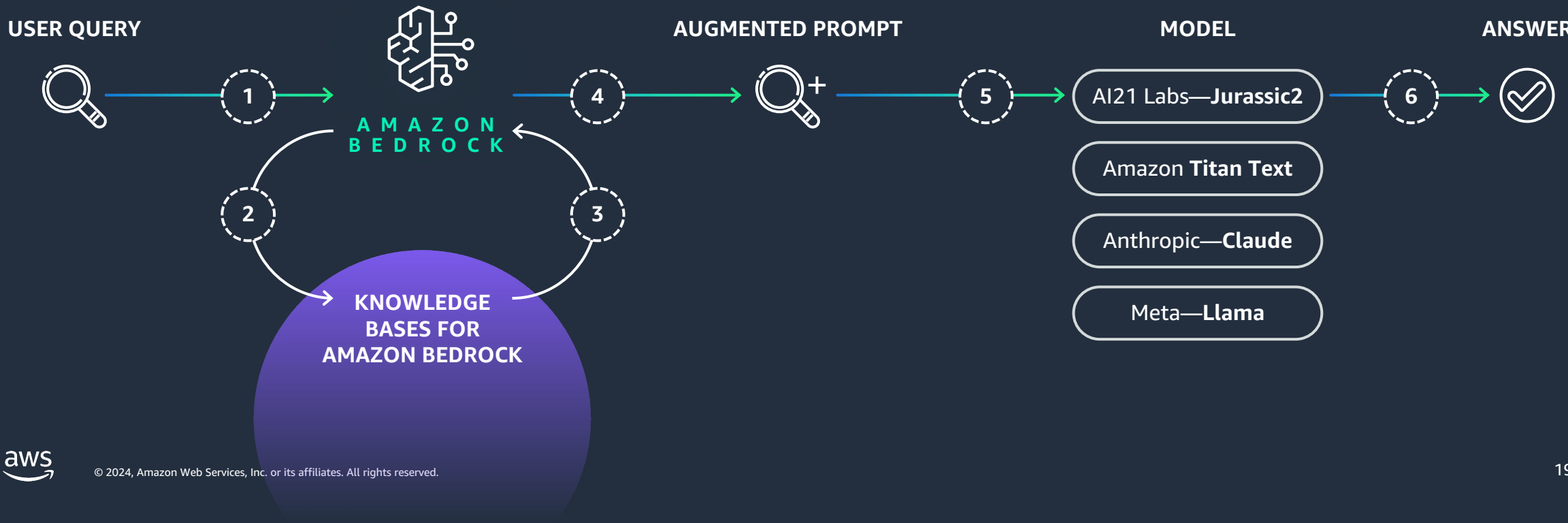
NATIVE SUPPORT FOR RAG

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

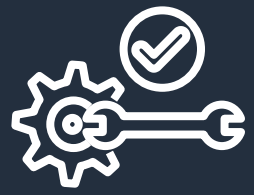
Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multiturn conversations

Automatic citations with retrievals to improve transparency



Customizing model responses for your business



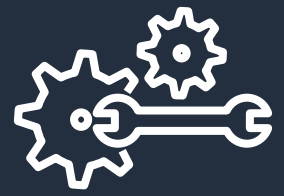
Fine-tuning

PURPOSE

Maximizing accuracy for **specific tasks**

DATA NEED

Small number of labeled examples



Continued pretraining

PURPOSE

Maintaining model accuracy for **your domain**

DATA NEED

Large number of unlabeled datasets

Enabling semantic (vector) search across our services

Amazon OpenSearch Service

Amazon OpenSearch Serverless

Amazon Aurora PostgreSQL

Amazon RDS for PostgreSQL

Amazon DocumentDB

Amazon DynamoDB via zero-ETL

Amazon MemoryDB for Redis

Amazon Neptune

Storing vectors and data together

Use familiar tools that meet your requirements

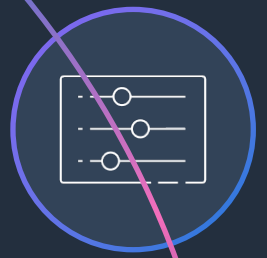
Avoid additional licensing and management

Provide a faster experience to end users

Reduce the need for data sync and movement

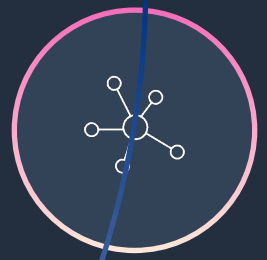
Guardrails for Amazon Bedrock

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

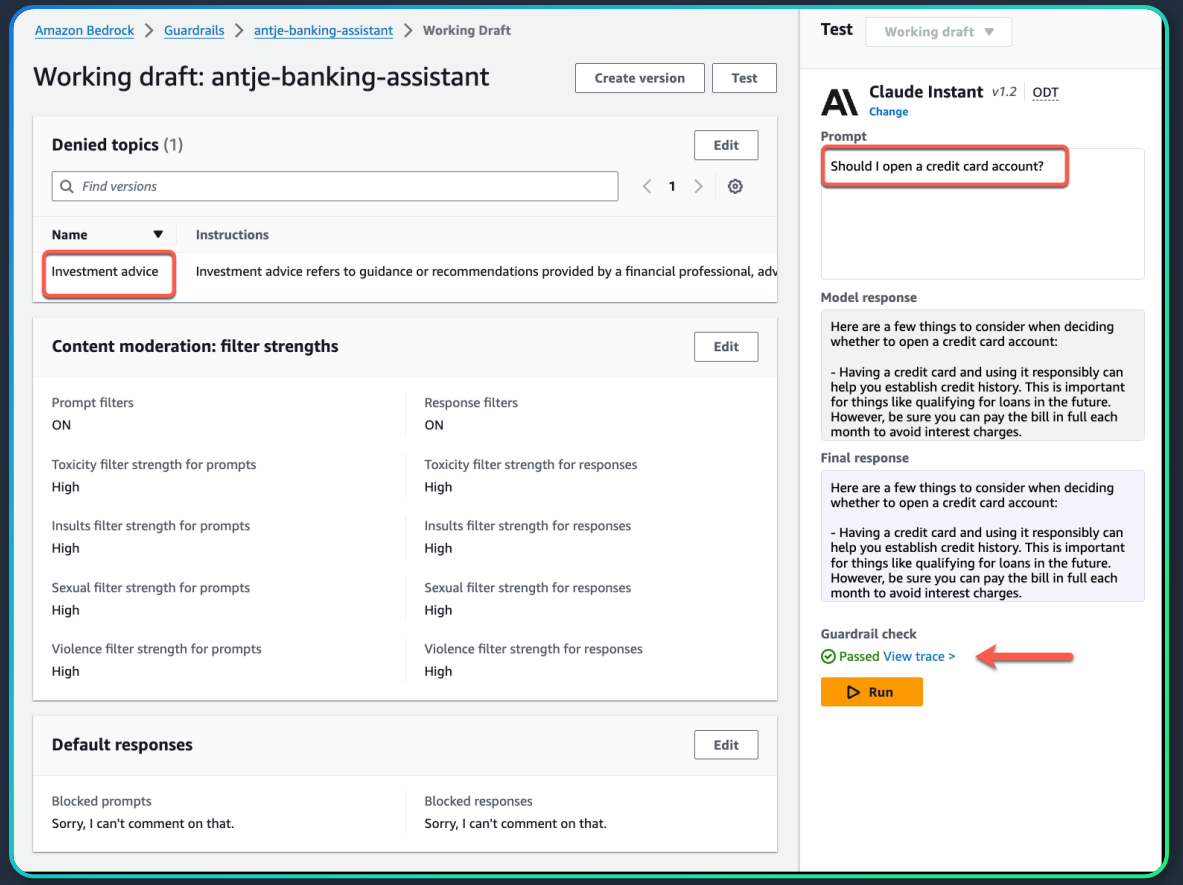
Redact or block sensitive information such as PIIs, and custom Regex.

aws

Guardrails for Amazon Bedrock

Guardrails for Amazon Bedrock is the only solution offered by a major cloud provider that enables customers to build and customize safety and privacy protections for their generative AI applications in a single solution.

It helps customers block as much as 85% more harmful content than protection natively provided by FM.



Amazon Bedrock

Helps keep your data secure and private

- None of the customer's data is used to train the underlying models
- All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC
- Data remains in the Region where the API is processed
- Support for GDPR, SOC, ISO, CSA compliance, and HIPAA eligibility

Amazon Bedrock inference consumption options

On demand

Pay-as-you-go, no commitment

- Pricing based on input and output token count for LLMs
- Great for prototyping, POCs, and small workloads with more relaxed requirements for throughput and latency
- Requests per minute (RPM) and tokens per minute (TPM) limits enforced

Provisioned throughput

Provision sufficient throughput to meet your application's performance requirements

- Reserve throughput (input/output tokens per minute) at a fixed cost
- Flexible commitment term of 1 month or 6 months
- Hourly PT (2MU), no commit, across different models
- Pay hourly rate, discounted for extended commit
- Great for production workloads or inference on custom models

Batch mode (preview)

Efficiently run model inference on large volumes of data

Avoids throttling when running large jobs

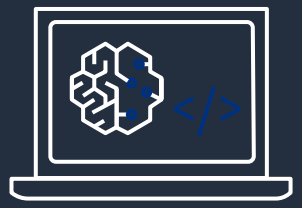
Fully managed model invocation jobs

No need to write code to handle failures and restarts

Works with base and custom models

Bedrock Studio

BUILD GEN AI APPLICATIONS FASTER AND MORE SECURELY



Easy to use playground

Projects based collaboration

Easy access with corporate SSO

Getting started

Get started with Amazon Bedrock

Discover features with a step-by-step tutorial

Dive deep with a hands-on workshop

Thank you!