Amazon Bedrock

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

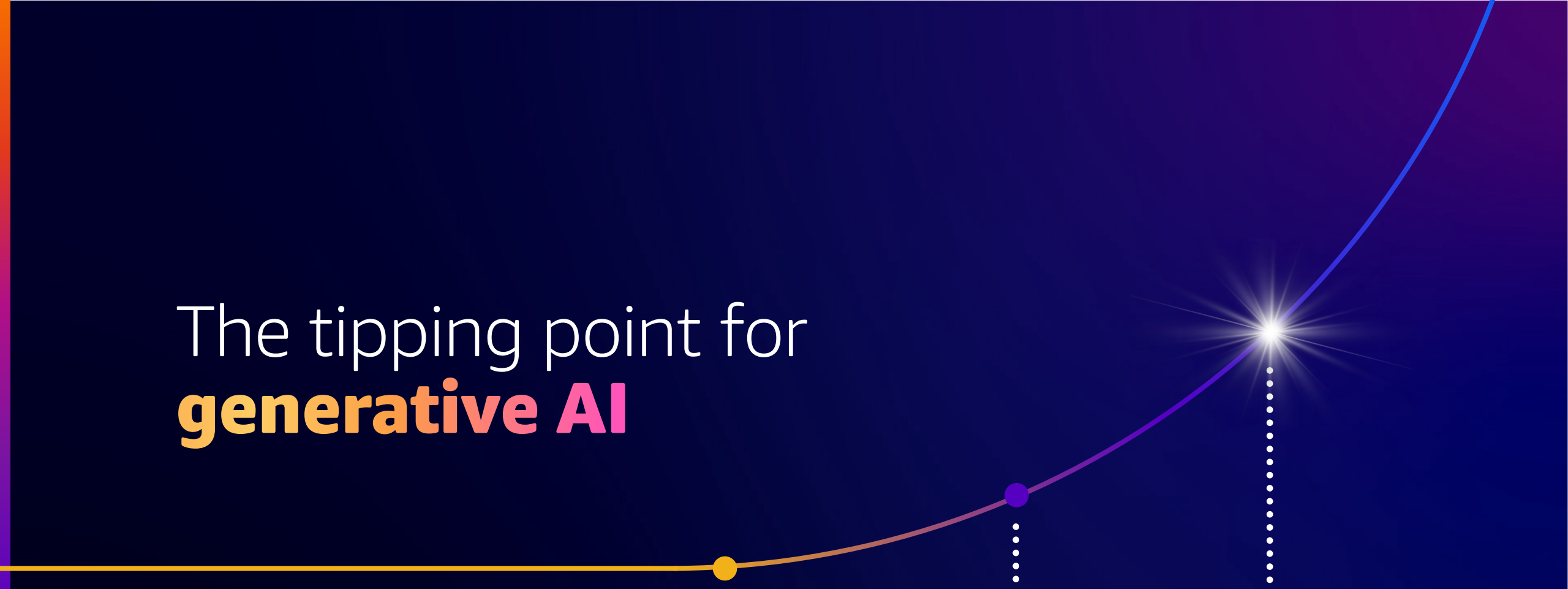
Michael Lin

Sr. Solutions Architect
Amazon Web Services

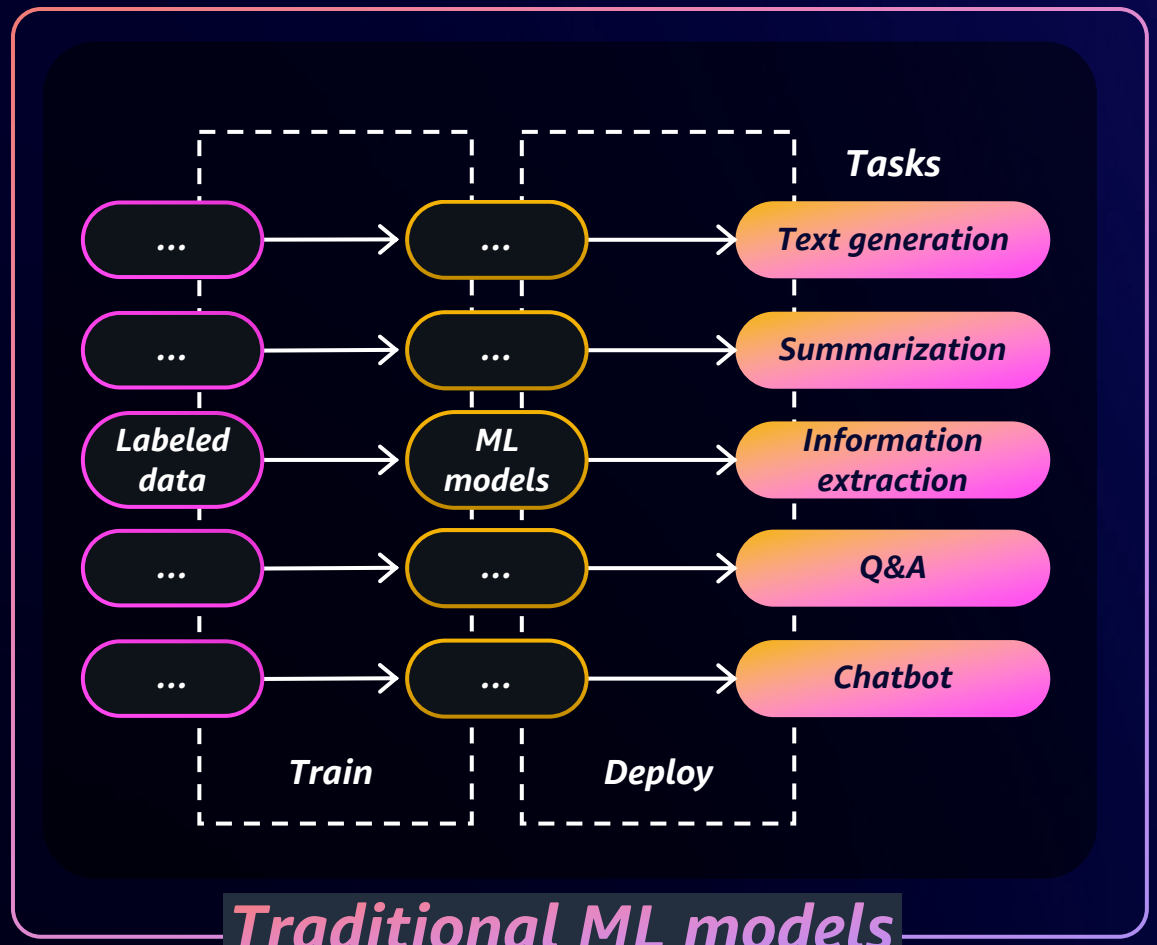
Agenda

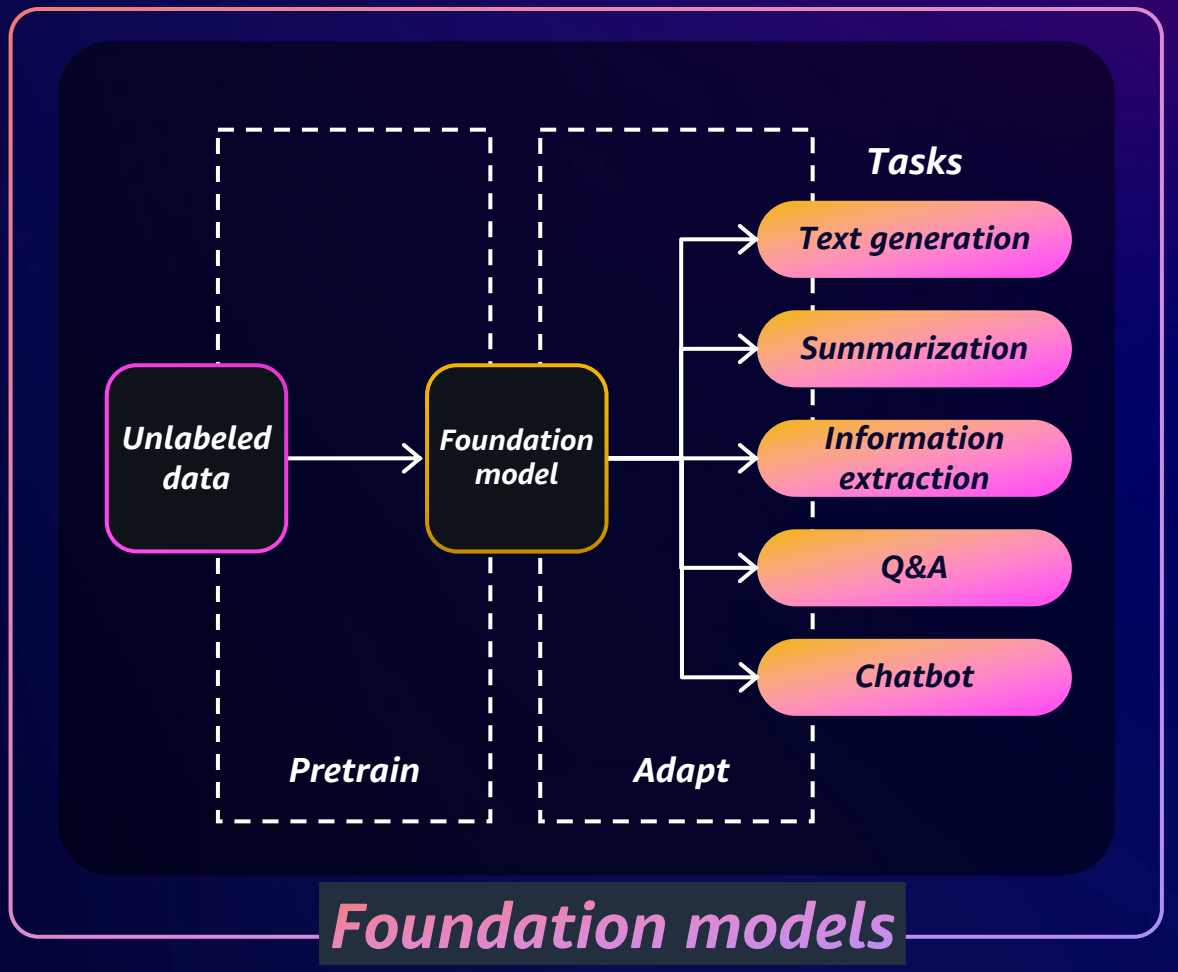
- Customer needs for generative AI
- Service overview
- Key features
- Customers
- Getting started

The tipping point for **generative AI**



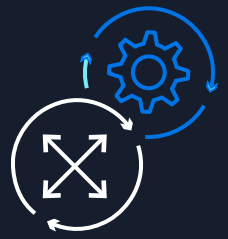
Generative AI is powered by foundation models (FMs)





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

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, text classification, text completion, code generation, Q&A*	*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					

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:	$0.015	$0.003	$0.00025
	Output:	$0.075	$0.015	$0.00125

*Per 1K tokens

Claude 3 can only be used via the Messages API

Text Completions API

```
Today is December 19, 2023.

Human: What are 3 ways to cook apples? Output your answer in numbered <method> XML tags.

Assistant: <method 1>
```

*The full prompt above includes the words after "Assistant". This is a technique called **prefilling Claude's response** - we'll talk about it in later slides*

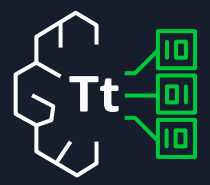
Messages API

```
  "system": "Today is December 19, 2023.",
  "messages": [
    { "role": "user", "content": "What are 3 ways to cook apples? Output your answer in numbered <method> XML tags." },
    { "role": "assistant", "content": "<method 1>" }
  ]
```

https://docs.anthropic.com/claude/reference/migrating-from-text-completions-to-messages

Amazon Titan

FMS

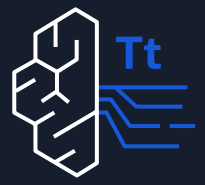


Amazon Titan Text Embeddings

Numerical representations of text

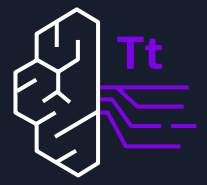
Amazon Titan Text Lite

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

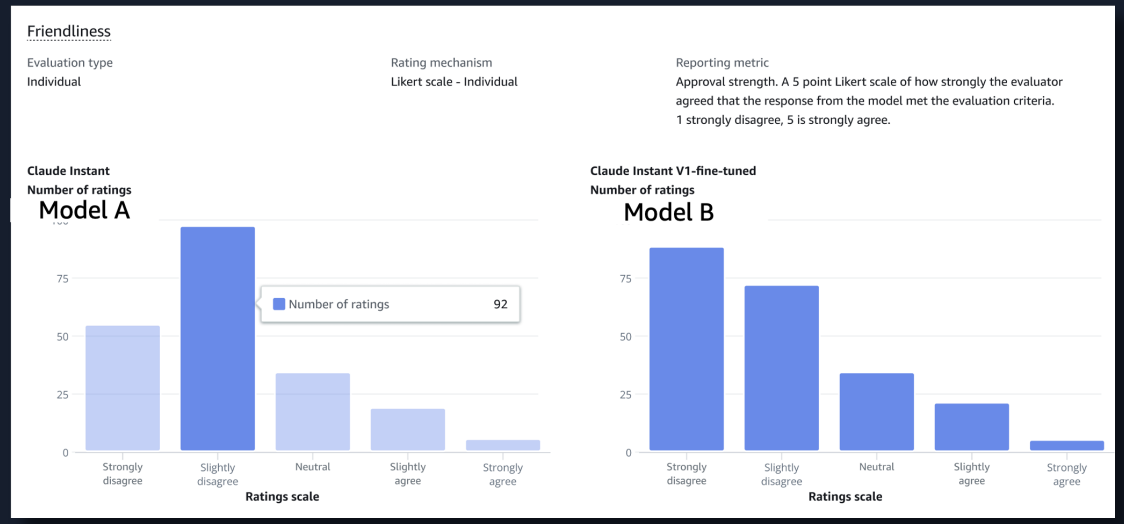
EVALUATE FMS TO SELECT THE BEST ONE FOR YOUR USE CASE

Automatic or human evaluation method

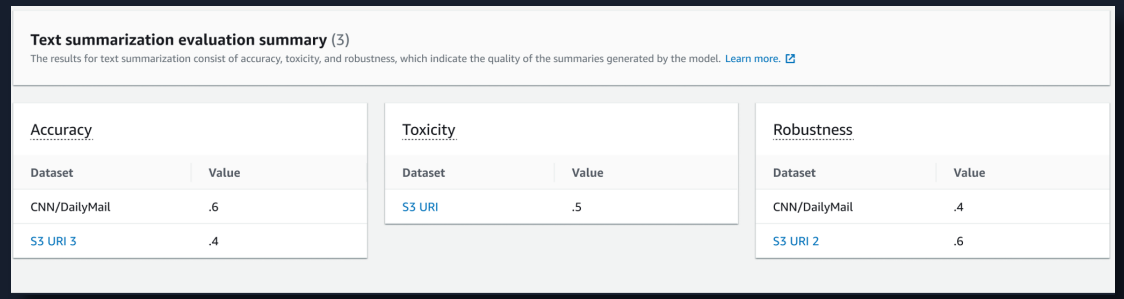
Curated datasets or bring your own

Predefined and custom metrics

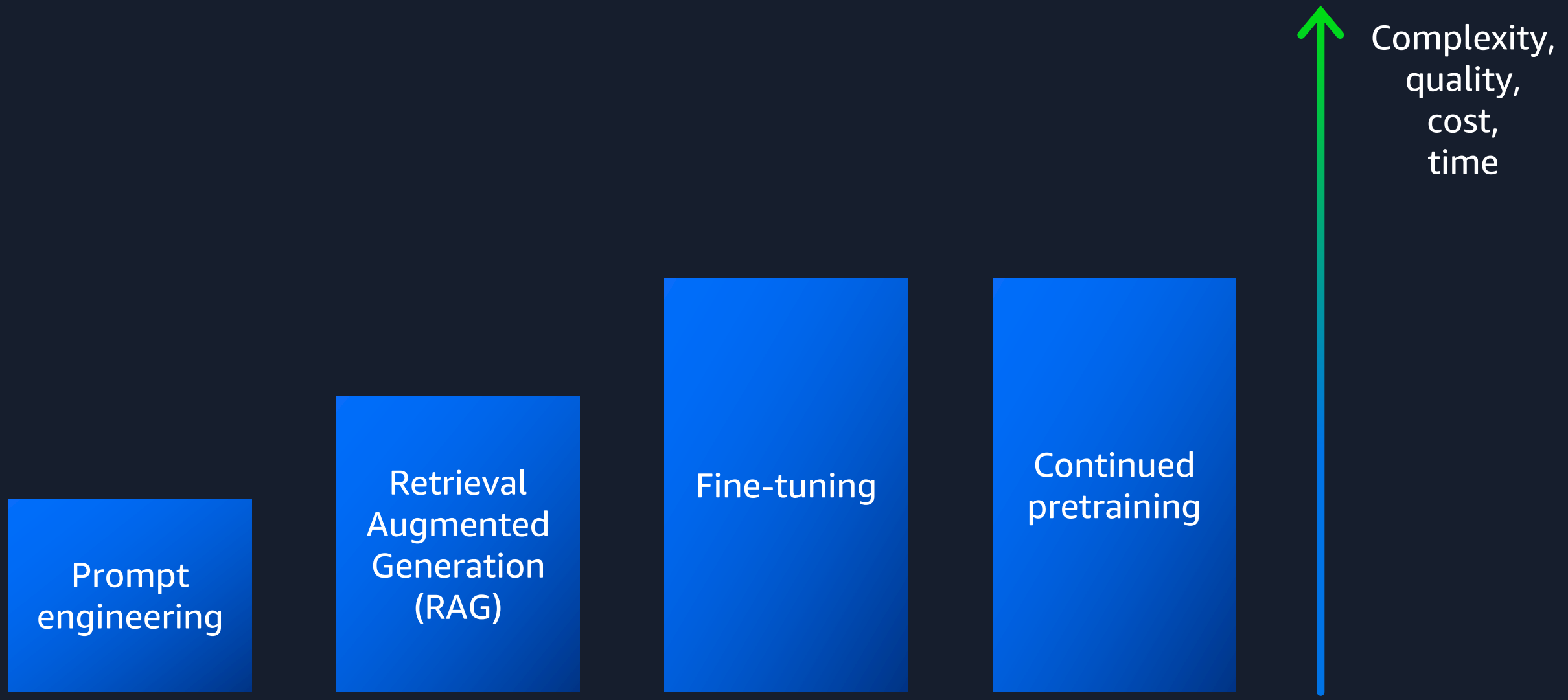
Human evaluation report



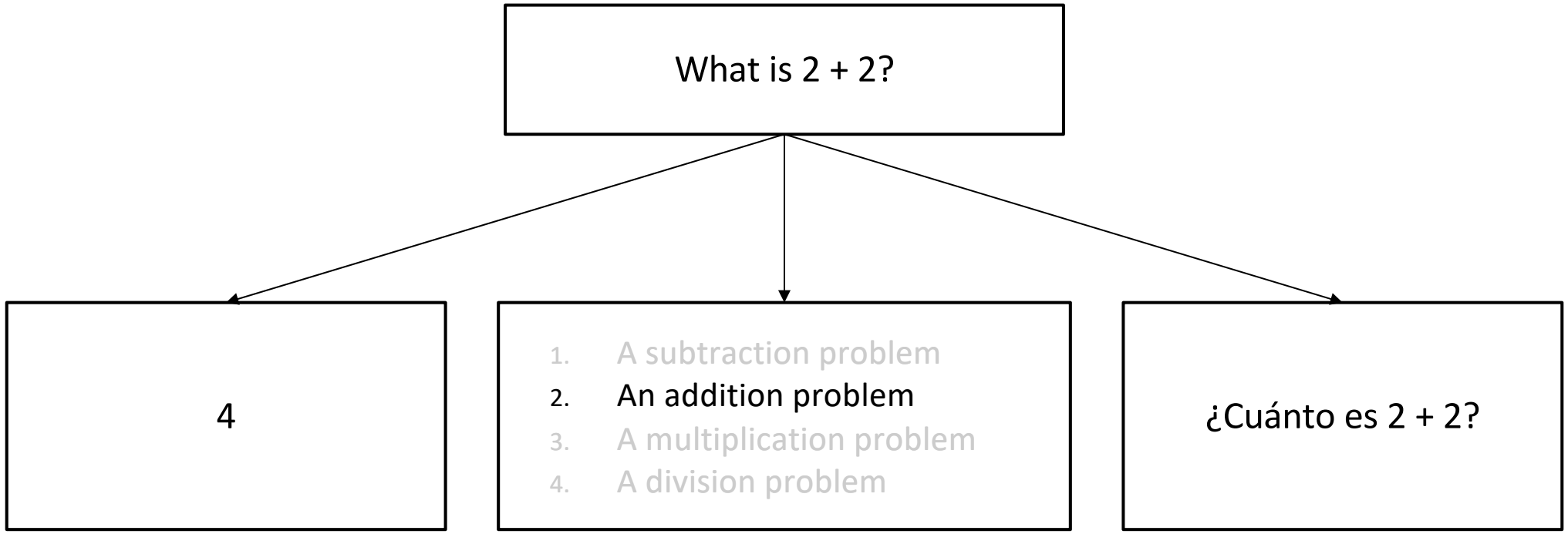
Automatic evaluation report



Common approaches for customizing FMs



What is prompt engineering?



Prompt engineering is the process of **controlling model behavior** by **optimizing your prompt to elicit high performing LLM responses** (as assessed by rigorous evaluations tailored to your use case).

Parts of a prompt

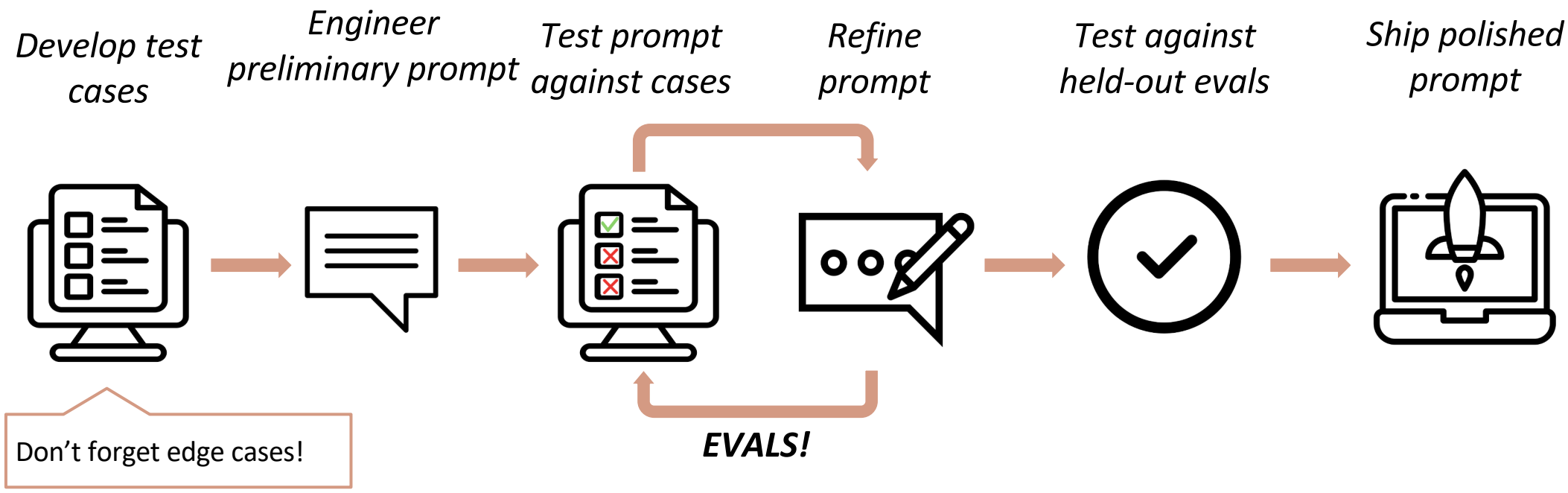
1. Task context
2. Tone context
3. Background data, documents, and images
4. Detailed task description & rules
5. Examples
6. Conversation history
7. Immediate task description or request
8. Thinking step by step / take a deep breath
9. Output formatting
10. Prefilled response (if any)

Example:

User	You will be acting as an AI career coach named Joe created by the company AdAstra Careers. Your goal is to give career advice to users. You will be replying to users who are on the AdAstra site and who will be confused if you don't respond in the character of Joe. You should maintain a friendly customer service tone. Here is the career guidance document you should reference when answering the user: <guide>{{DOCUMENT}}</guide> Here are some important rules for the interaction: - Always stay in character, as Joe, an AI from AdAstra careers - If you are unsure how to respond, say “Sorry, I didn’t understand that. Could you repeat the question?” - If someone asks something irrelevant, say, “Sorry, I am Joe and I give career advice. Do you have a career question today I can help you with?” Here is an example of how to respond in a standard interaction: <example> User: Hi, how were you created and what do you do? Joe: Hello! My name is Joe, and I was created by AdAstra Careers to give career advice. What can I help you with today? </example> Here is the conversation history (between the user and you) prior to the question. It could be empty if there is no history: <history> {{HISTORY}} </history> Here is the user’s question: <question> {{QUESTION}} </question> How do you respond to the user’s question? Think about your answer first before you respond. Put your response in <response></response> tags.
Assistant (prefill)	<response>

How to engineer a good prompt

Empirical science: always test your prompts & iterate often!



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 database.
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

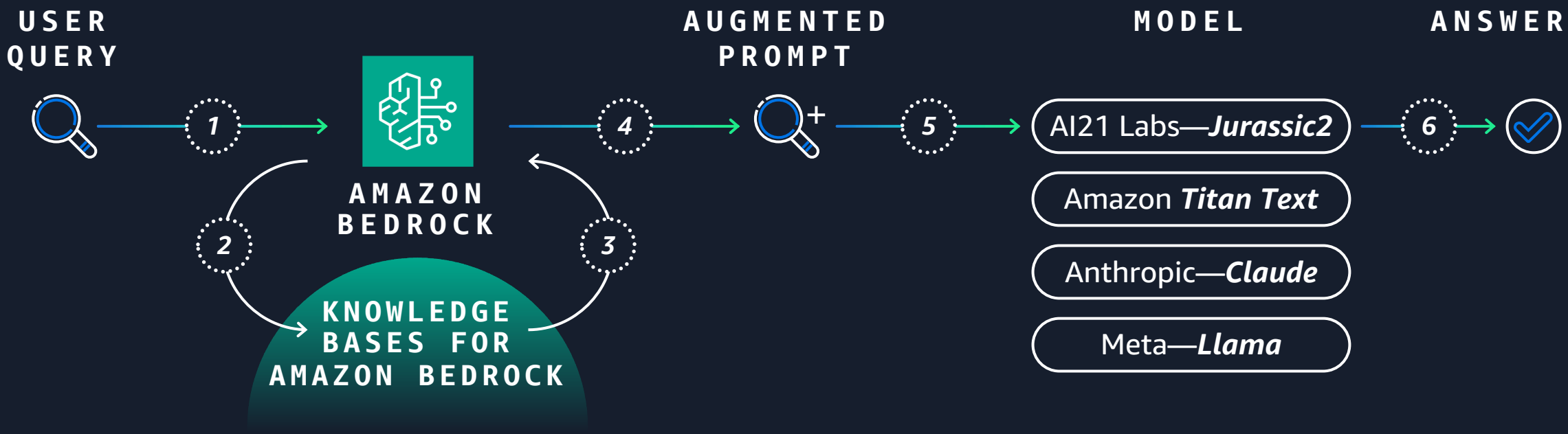
NATIVE SUPPORT FOR RAG

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



Amazon OpenSearch Service

Amazon OpenSearch Serverless

Amazon Aurora PostgreSQL

Amazon RDS for PostgreSQL

Enabling semantic (vector) search across our services

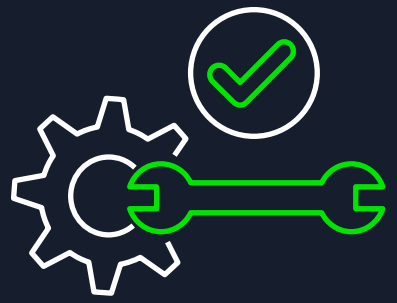
Amazon DocumentDB

Amazon DynamoDB via zero-ETL

Amazon MemoryDB for Redis

Amazon Neptune

Customizing model responses for your business



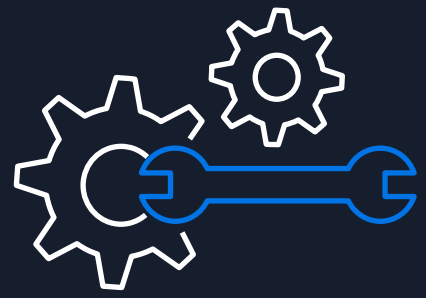
Fine-tuning

PURPOSE

Maximizing accuracy for ***specific tasks***

DATA NEED

Small number of labeled examples



Continued pretraining

PURPOSE

Maintaining model accuracy for ***your domain***

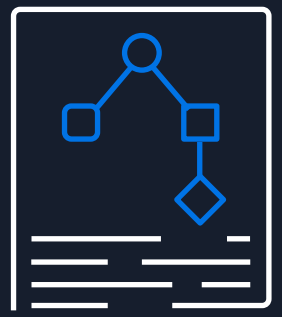
DATA NEED

Large number of unlabeled datasets

Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO EXECUTE MULTISTEP TASKS USING COMPANY SYSTEMS AND DATA SOURCES

Decompose into steps using available actions and Knowledge Bases for Amazon Bedrock



Execute action or search knowledge base

Observe results

Think about next step

Until final answer

Agents for Amazon Bedrock now faster and easier to use

Quickly create Agents for Amazon Bedrock with a simplified console experience and attain better performance with Sonnet and Haiku

1. Gain better performance and greater accuracy, now also including Claude 3 Sonnet and Haiku
2. Fewer steps to create agents using the new Agents for Amazon Bedrock console experience
3. Simplified schemas enable faster schema creation and remove the need to conform to the OpenAPI Specification
4. Return of control allows business logic implementation for actions using your backend services

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

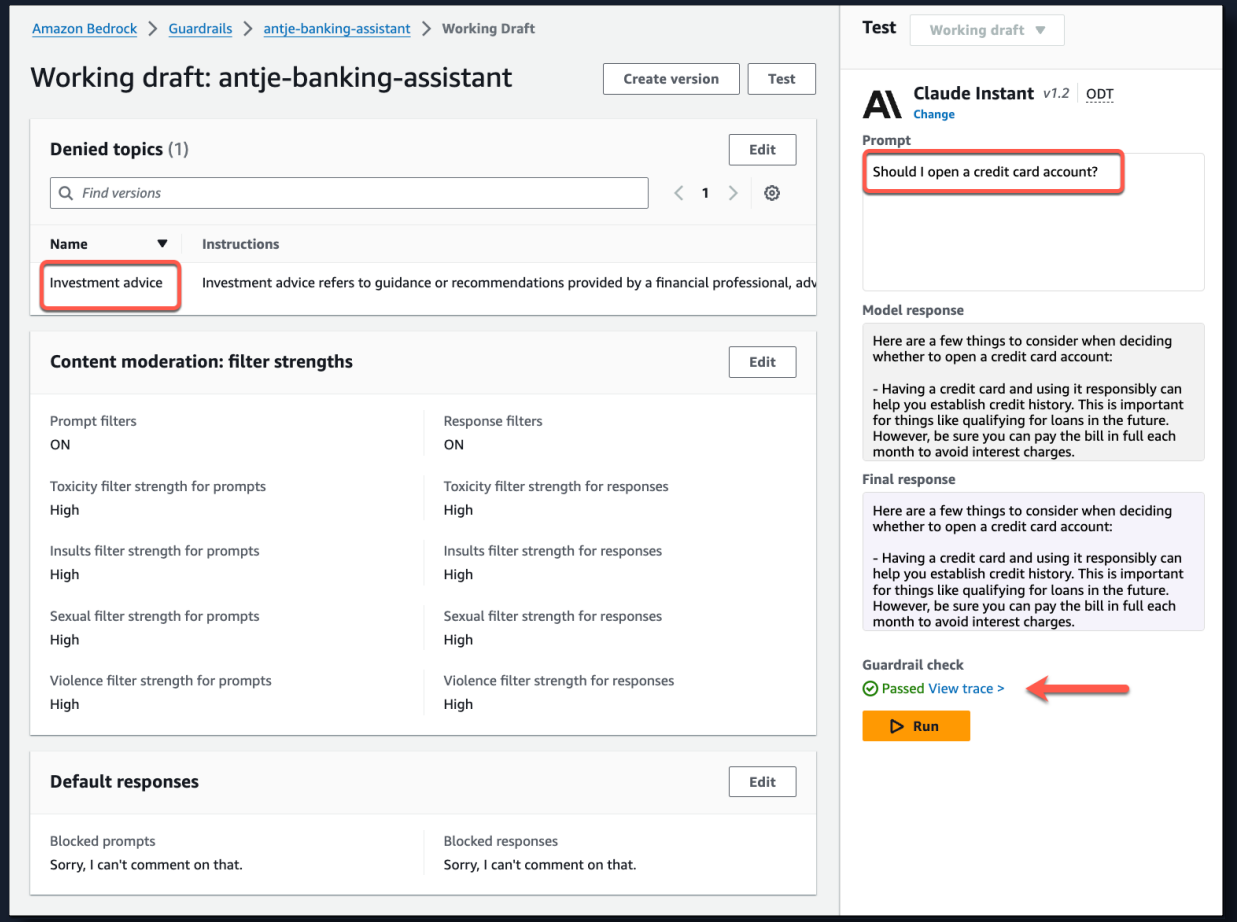
Define and disallow denied topics with short natural language descriptions

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

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

Adidas accelerated development of a conversational AI assistant, providing their engineers unified access to knowledge and support

"We were excited to be part of the Amazon Bedrock preview and get our hands on the service. Bedrock quickly became a highly valued addition to our generative AI toolkit, allowing us to focus on the core aspects of our LLM projects, while letting it handle the heavy lifting of managing infrastructure. Using Bedrock, we have developed a generative AI solution that gives the community of Adidas engineers the ability to find information and answers from our knowledge base through a single conversational interface, covering everything from getting started to highly technical questions."

Daniel Eichten

VP Enterprise Architecture at Adidas

Thank you!

Michael Lin

linmicht@amazon.com