AWS State, Local, and Education Learning Days

New York

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws Learning Days
State, Local, and Education

AI/ML Track

10:15am – 11:15am

200 level

Practical AI for Public Sector

Explore Generative AI's potential in public sector and higher education

11:30am – 12:30pm

200 level

Increase productivity and satisfaction with an intelligent contact center

Adding intelligence to your call center

1:30pm – 3:00pm

300 level

Workshop: Building Agentic Workflows

Advanced AWS workshop on building autonomous AI workflows with Amazon Bedrock for developers

3:15pm – 4:15pm

300 level

Generative AI Masterclass

Learn how AI can modernize customer service operations and improve satisfaction

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

GenAI Master Class

Vinod Kisanagaram

Solutions Architect
vinodaws@amazon.com

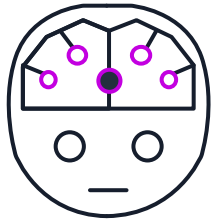
Mark Bronnimann

Sr Solutions Architect
mbronni@amazon.com

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws Learning Days
State, Local, and Education

AIML/GenAI hierarchy

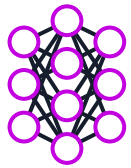


Artificial Intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically



Deep learning (DL)

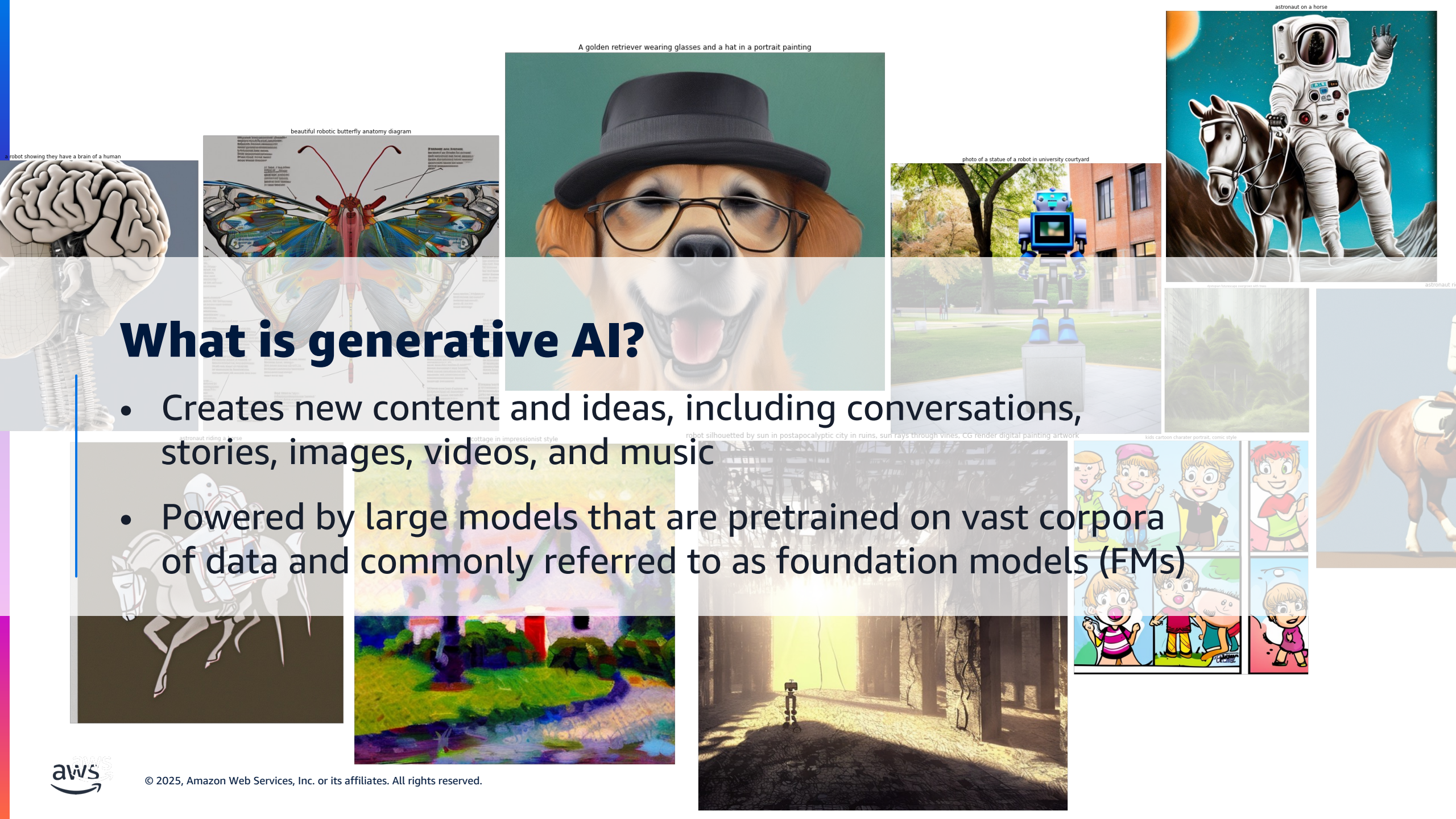
A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

Generative AI

Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



What is generative AI?

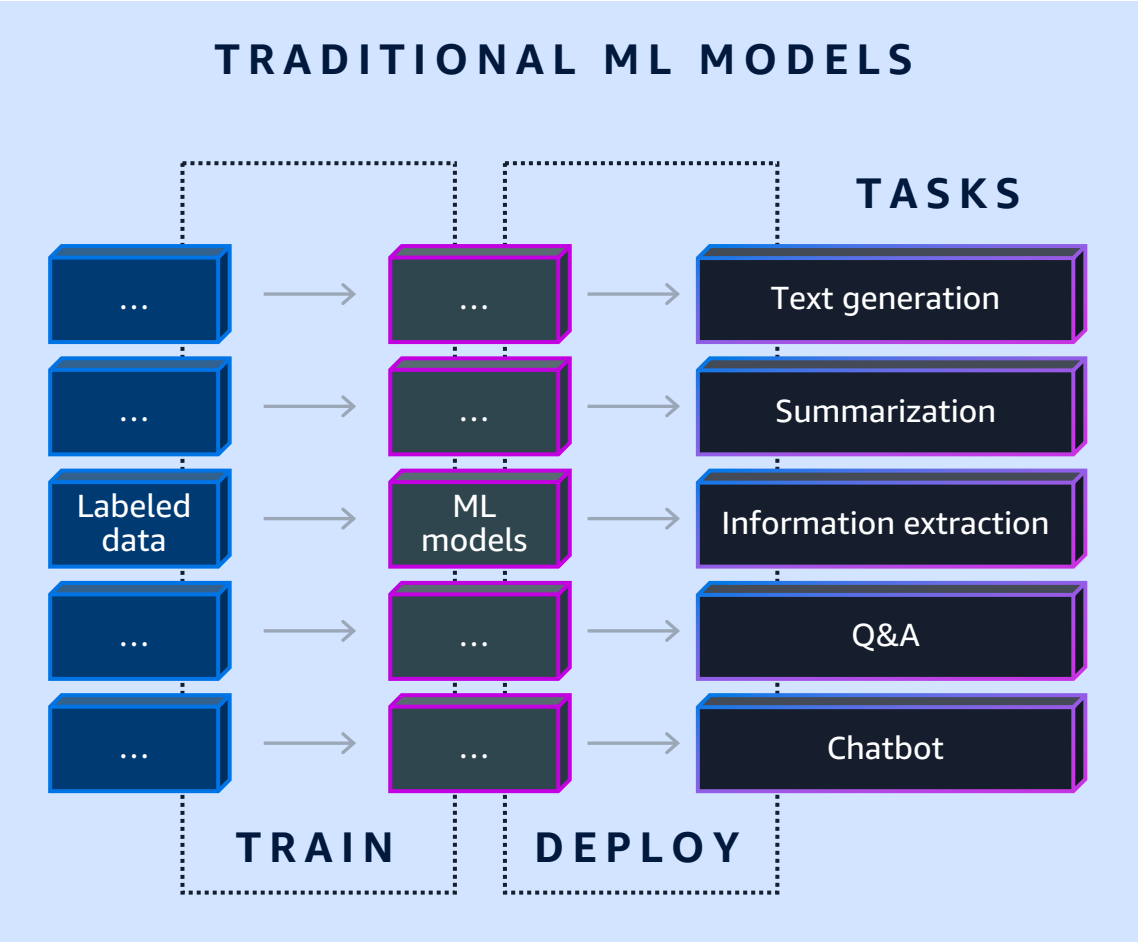
- Creates new content and ideas, including conversations, stories, images, videos, and music
- Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

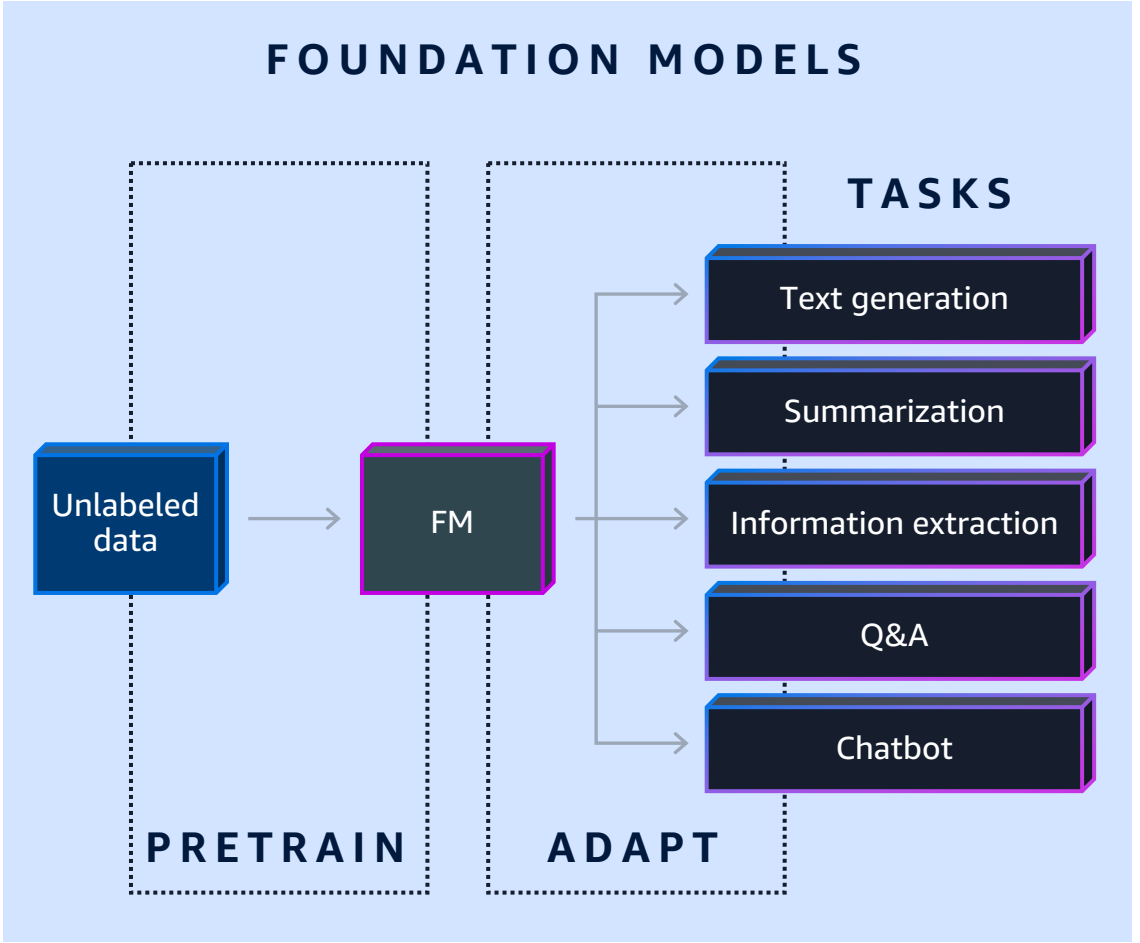
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

GenAI in Action

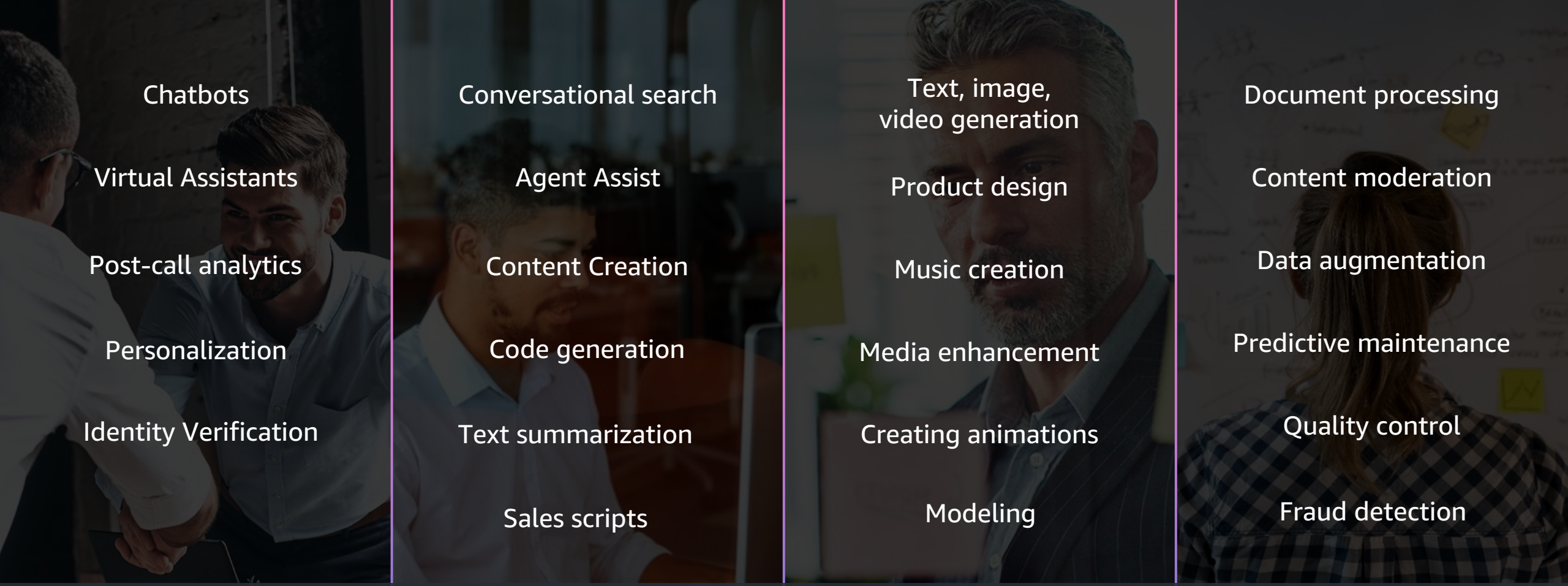
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why foundation models?





© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Chatbots

Virtual Assistants

Post-call analytics

Personalization

Identity Verification

Enhance customer experience

Conversational search

Agent Assist

Content Creation

Code generation

Text summarization

Sales scripts

Boost employee productivity

Text, image, video generation

Product design

Music creation

Media enhancement

Creating animations

Modeling

Creativity & Content Creation

Document processing

Content moderation

Data augmentation

Predictive maintenance

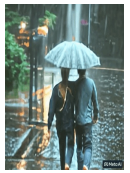
Quality control

Fraud detection

Improve business operations

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Unimodal Models

Text input	→	Foundation model	→	Output
"Summarize this article ……."		Text generation model		[Text] "This article talks about …."
"a photo of an astronaut riding a horse on mars"		Image generation model		
"A young couple walking in rain." "Children singing nature songs" "Write Python code to sort array …"		Video Audio Code → generation model		[Video]  [Audio]  [Code] 
"This is the insurance policy for life, clause number 1 is …."		Embeddings model		[Vectors] 0.11,-2.33,4.75

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multimodal Models

Any input → Foundation model → Any Output

"Summarize this article ……."

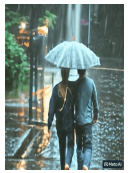
"a photo of an astronaut riding a horse on mars"

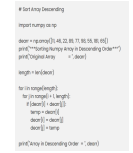
"A young couple walking in rain."
"Children singing songs about Nature"
"Write Python code to sort array …"

Multimodal model

[Text] "This article talks about …."

[Video] [Audio] [Code]





"a photo of an astronaut riding a horse on mars"

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What generative AI does well (for now)

- Summarization
- Content generation (Text, image, audio, video)
- Language Translation
- Correction/paraphrasing
- Classification

Use case examples

- Extract insights from your documents
- Create a product description
- Generate draft marketing copy
- Create a job posting
- Summarize your meetings
- Create an ad from a product description
- Explain complex data in plain English
- Easily draft documents, emails, or design

aws

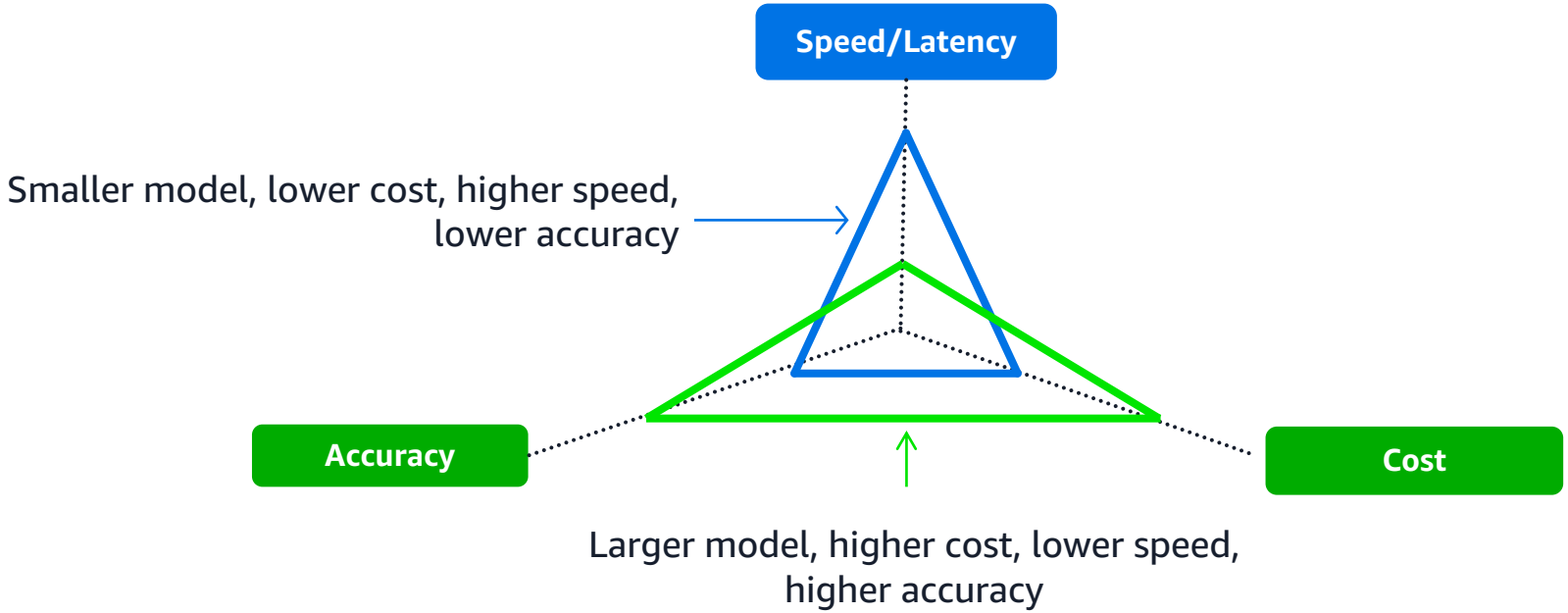
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Current limitations

- Explainability of the model and results
- Hallucinations and Biases
- Data Staleness/update and IP infringements
- Not good at complex math and reasoning (yet)
- Not good at large scale code translation (yet)

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Considerations when selecting a generative AI model



Model Provider – Ethical AI, Data Privacy, Security, Risk, and Sustainability Commitments

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Strategies for Implementation

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demystifying common GenAI terms (and why they are important to outcomes)

Vector

Numerical representations of text that saves the meaning and the semantic relationship with other words

Prompt Engineering

The process where you guide Foundation Models to generate desired outputs by providing text instructions.

Pre-Training

When you create an Foundation Model from scratch.

Embeddings

Collection of Vectors that belong to a piece of text. Embeddings models transform text into Vectors, that are then stored in Vector DBs.

RAG

Retrieval Augmented Generation, use internal company documentation to find answers to questions

Re-Training

When you use a FM model and unlabeled company data to teach the model about that data so that it can then respond questions about it. Increases accuracy in the answers.

Parameters

Memory of the models, the large the number of parameters the better it tends to recall. Also it's slower and more expensive to run.

i.e. Llama 2 - 7B, Llama 3.1 405B

Agentic System

Using the reasoning capabilities of an LLM to automate proceses/task

Fine Tuning

When you use a FM model and labelled company data to teach the model about that data so that it can then respond questions about it. Increases accuracy in the answers.

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Strategies for implementation and their trade-offs

1

Prompt engineering

Using context instructions, examples and output indicators

2

Retrieval Augmented Generation (RAG)

Text generation based on specified corpus of data, to generate accurate responses with no hallucination

3

Agentic Systems

Using LLMs to understand tasks, gather information, decide best path and execute actions

4

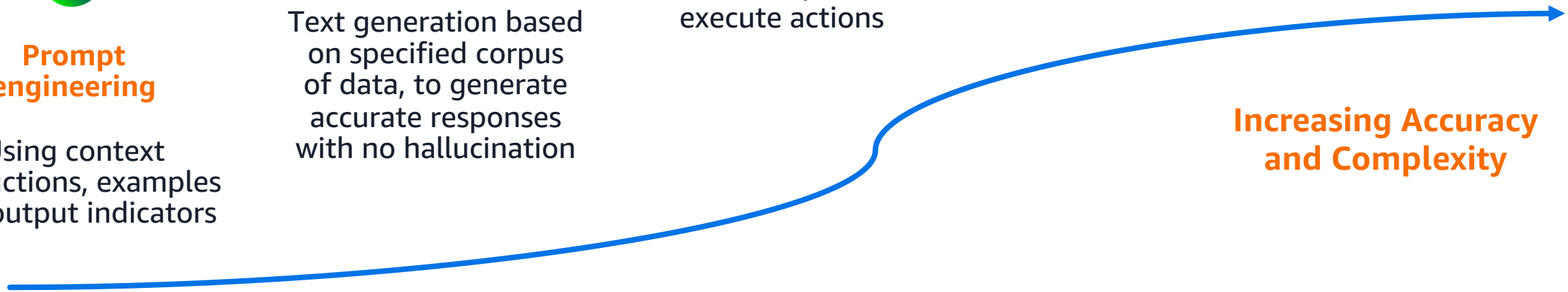
Fine-tuning

Give a model company or domain specific data, or instructions examples

5

Re-training

Give a model a different dataset, or building a model from scratch



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt Engineering

Zero shot learning

Prompt

Review: “Earnings per share have beaten analyst expectations"

What is the sentiment?

Input

The text explains that earnings have been expectations, that is generally a good signal in financial reporting, therefore the review is positive.

Output

Few shot learning

Prompt

Review: " Earnings per share have beaten analyst expectations "
Sentiment: positive

Review: “sales remained constant over the past quarter but EBIDTA has decreased"
Sentiment: negative

Review: “S&P500 Tops 5,600 for first time as tech rallies"
Sentiment:

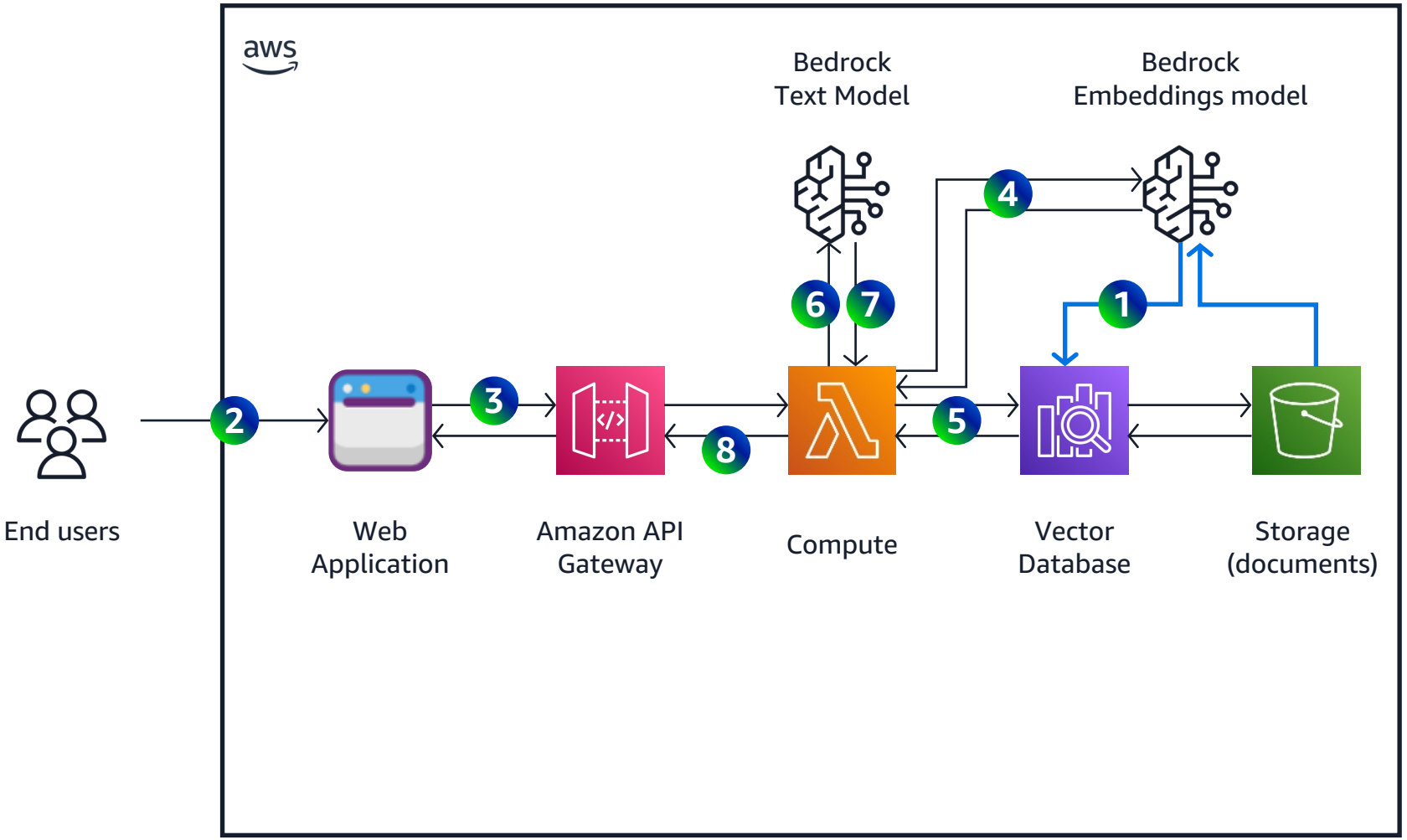
Input

positive

Output

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval augmented generation (RAG) – “Chat with my Docs”



1. Documents are converted into vectors using the embeddings model on Bedrock and then they are stored in OpenSearch
2. User opens the application UX
3. The web application sends a request to Amazon API Gateway that calls a Lambda function
4. The Lambda function sends the prompt to Bedrock which converts the prompt in embeddings
5. The vectors are sent to Lambda and then to OpenSearch to find the paragraphs that contain the answer to the question that sit on S3
6. Those paragraphs together with the prompt is sent to the model
7. The model generates the answer
8. The answer is sent it back to the UX

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agentic Systems

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agents can combine Actions and Knowledge Bases

how much vacation do I get per year?

as a full-timer with 3 years tenure, you get 15 days

cool. I'd like to take off December 8 to 15

approved, enjoy. you have 8 more days available

HR Policy Assistant

Instructions: "you are an HR agent, helping employees understand HR policies and manage vacation time"

HR Knowledge Base

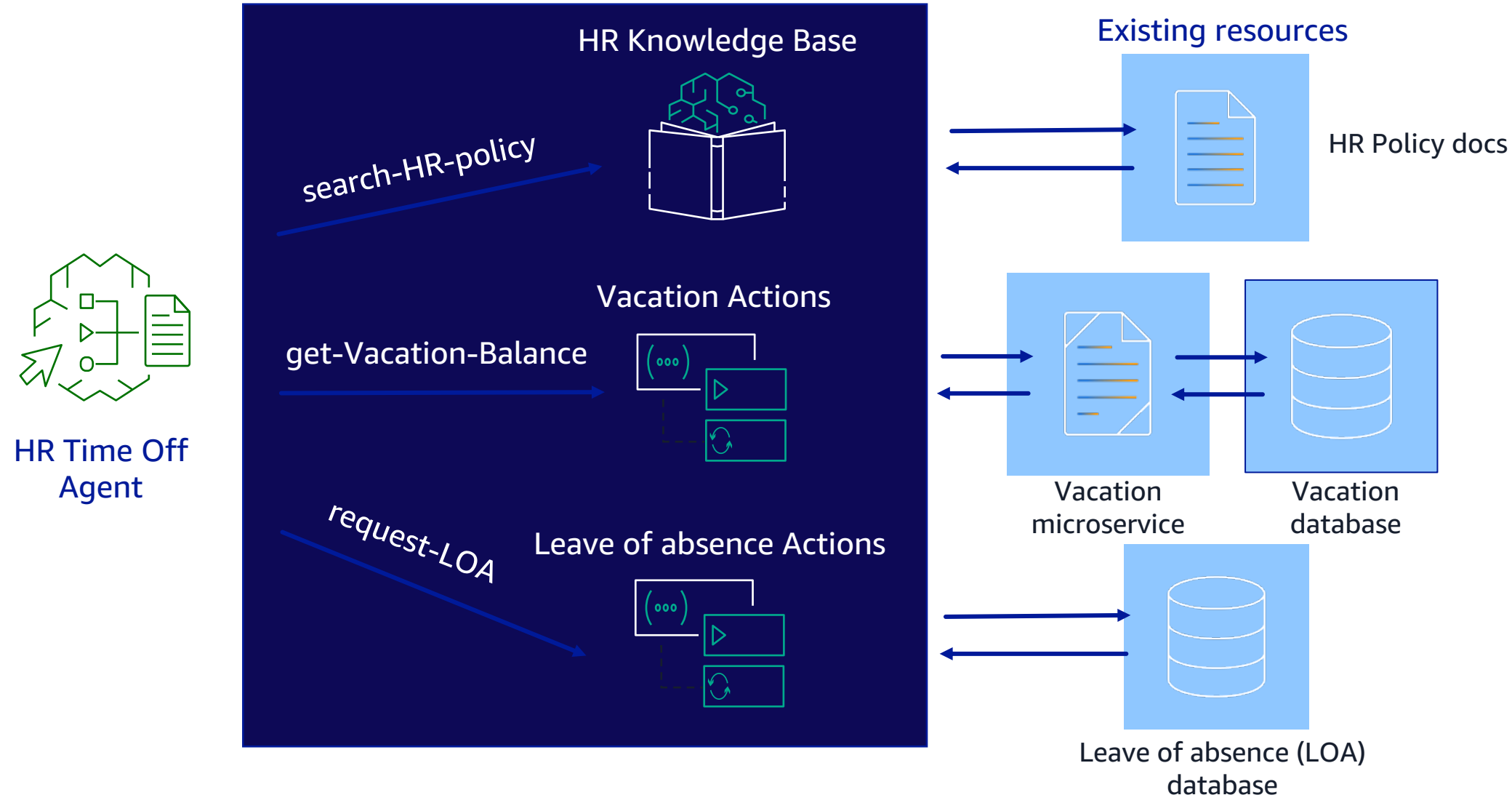
Vacation Policy
Contains the entire vacation policy for the company

HR Actions

Request Vacation
In: start date, end date
Out: approval status, remaining balance

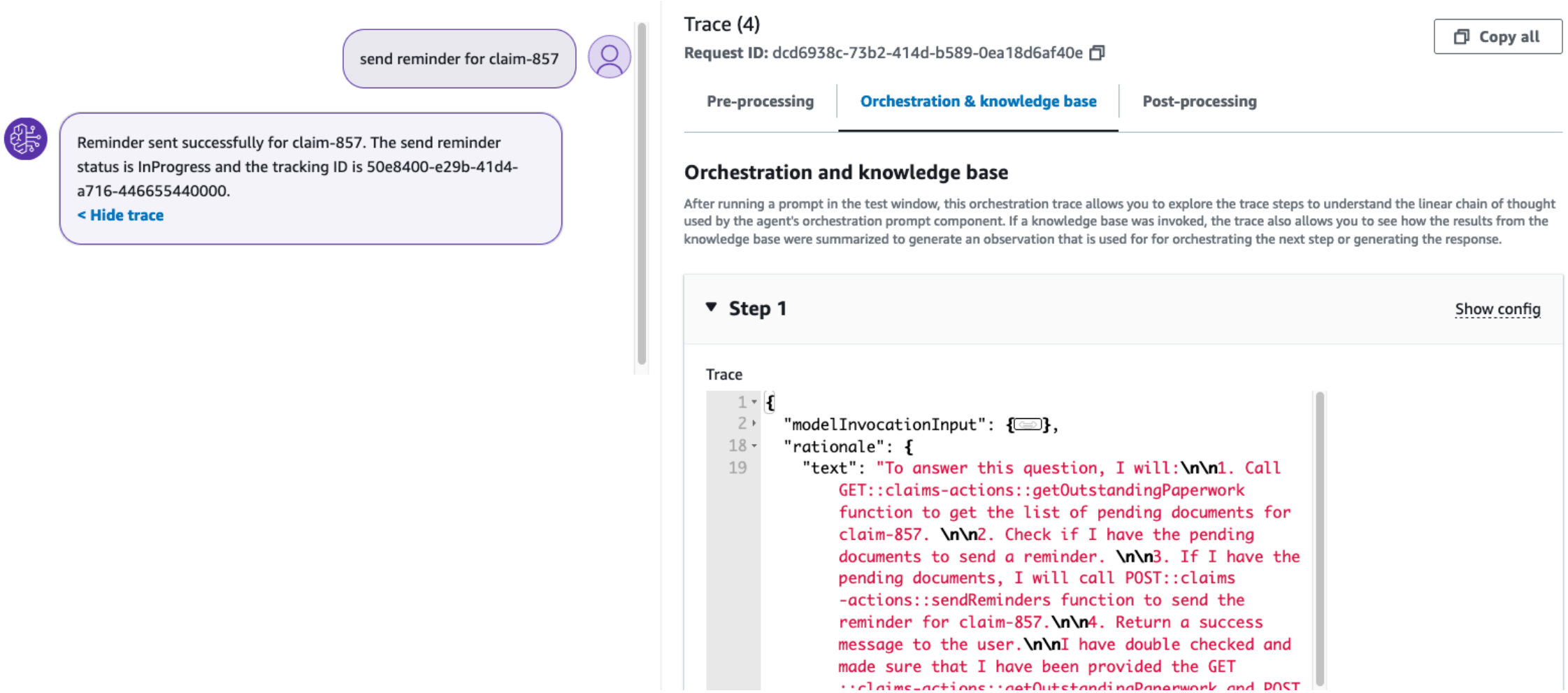
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agents build on existing enterprise resources



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

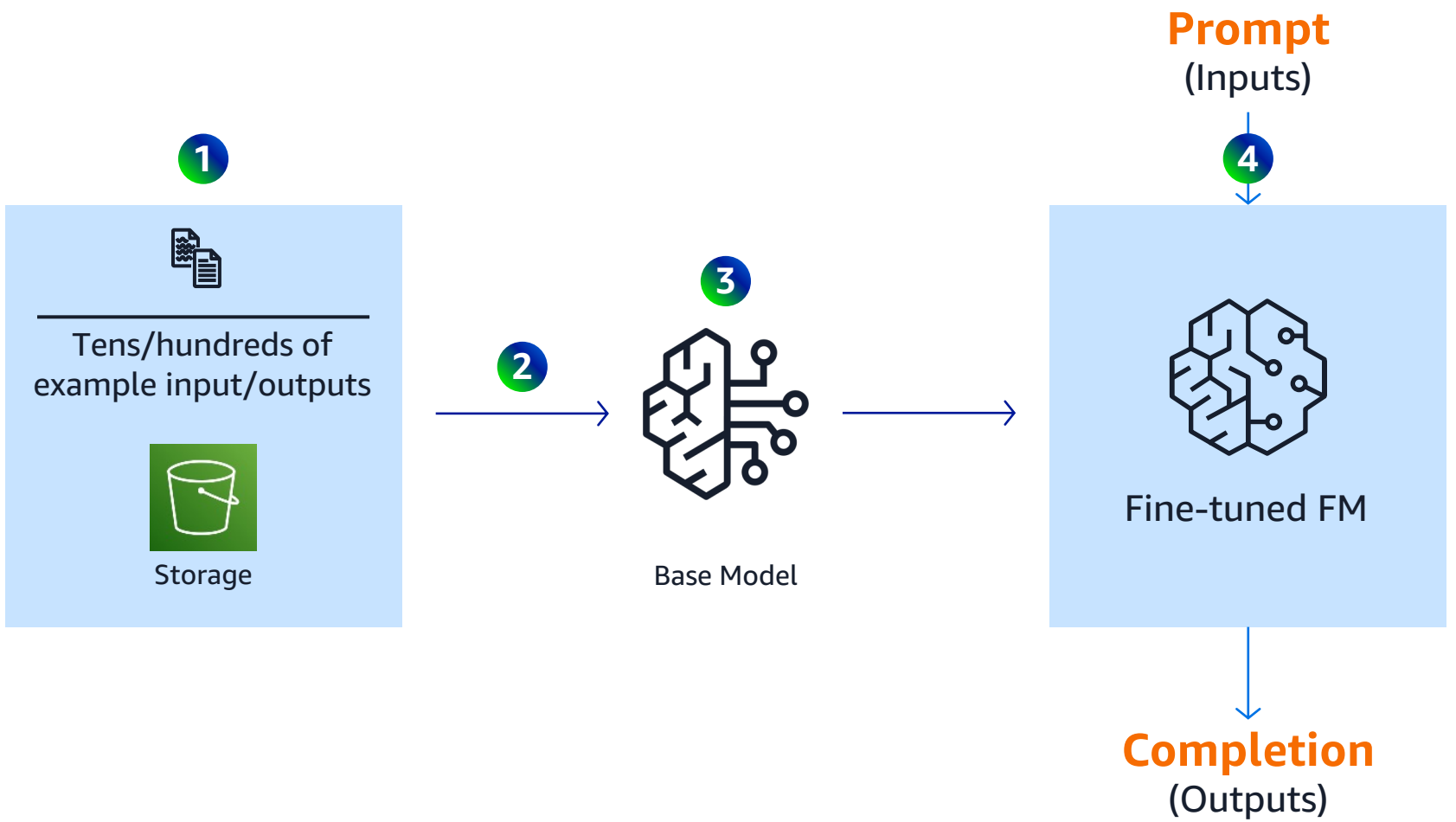
Agent orchestration is transparent – Trace



Detailed orchestration trace in the console and from the SDK

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

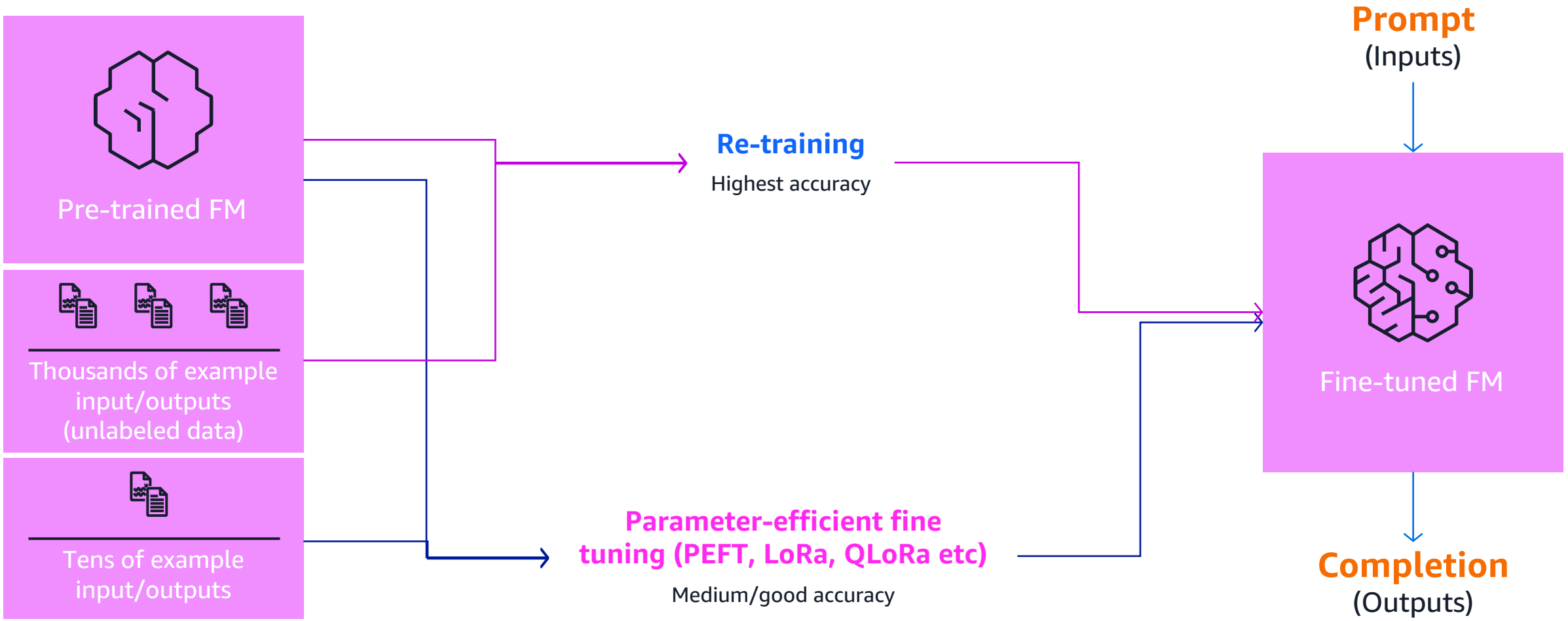
Fine-tuning (Task specific)



1. Tens/hundreds of instructions or examples are stored
2. Instructions are sent to the base model via APIs
3. A base model is fine-tuned with these instructions (minutes, hours, days)
4. The new fine-tuned model is available for the customer to star inferencing

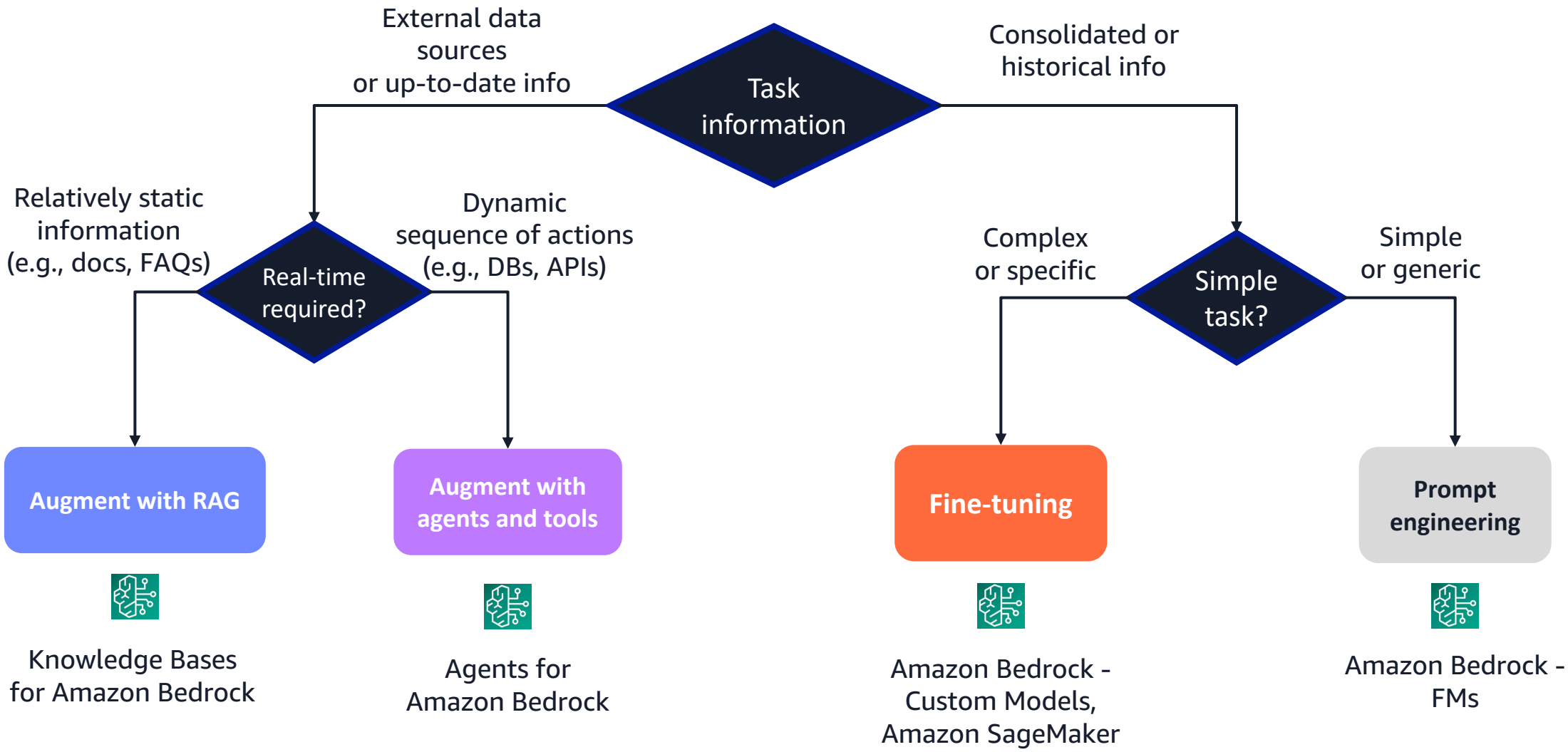
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Advanced fine-tuning & re-training (domain adaptation)



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Decision Tree



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock

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

Choice of leading FMs through a single API

Optimization for cost, latency and accuracy

Customization with your data

Safety and responsible AI checks

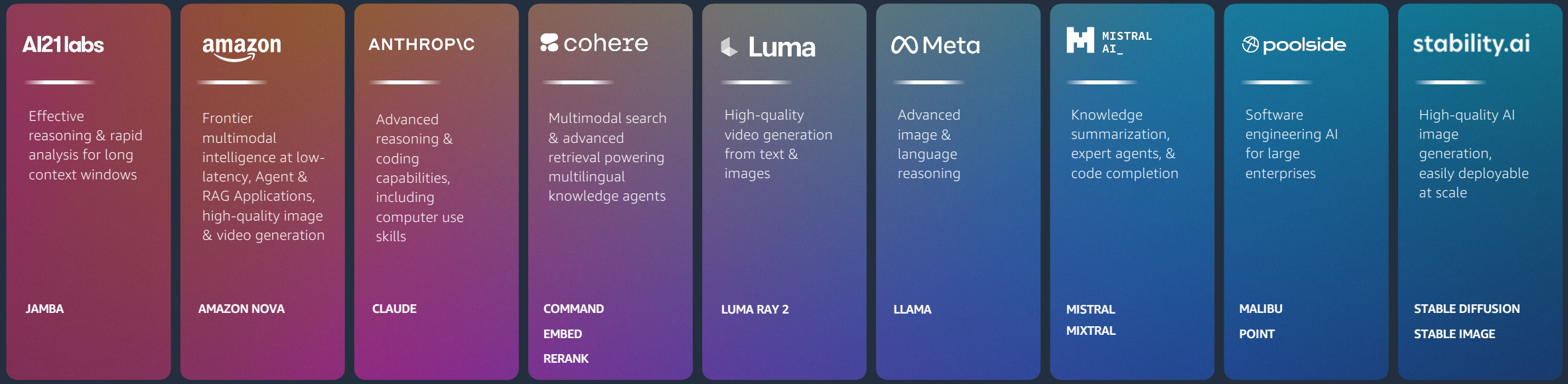
Agents that execute complex tasks

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock

BROAD CHOICE OF MODELS



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

Amazon Nova Foundation Models

State-of-the-art foundation models that deliver frontier intelligence and industry leading price performance.

Understanding models

Creative content generation models

Amazon Nova **Micro**

Our text only model that delivers the lowest latency responses at very low cost

GENERALLY AVAILABLE

Amazon Nova **Lite**

Our lowest cost multimodal model that is lightning fast for lightweight tasks

GENERALLY AVAILABLE

Amazon Nova **Pro**

Our highly capable multimodal model with best combination of accuracy, speed, and cost for a wide range of tasks

GENERALLY AVAILABLE

Amazon Nova **Premier**

Our most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

COMING SOON

Amazon Nova **Canvas**

State-of-the-art image generation model

GENERALLY AVAILABLE

Amazon Nova **Reel**

State-of-the-art video generation model

GENERALLY AVAILABLE

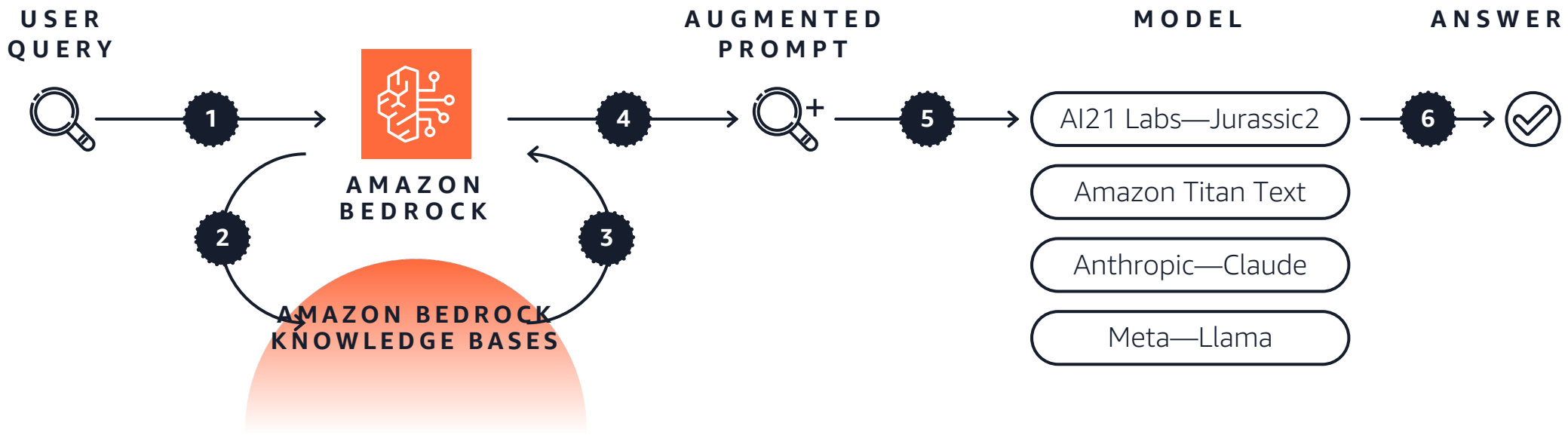
Lower Cost & Latency → **Increasing Intelligence**

aws

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

Amazon Bedrock Knowledge Bases

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

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building RAG-based Apps on Amazon Bedrock

CHOOSE YOUR FAVORITE KNOWLEDGE BASE OR VECTOR SEARCH ENABLED SERVICE

NATIVELY AVAILABLE BEDROCK KNOWLEDGE BASES:

Vector engine
for Amazon OpenSearch
Serverless

Amazon Aurora
PostgreSQL
with pgvector

Pinecone

Redis Enterprise
Cloud

MongoDB

AMAZON DATABASES ENABLED WITH VECTOR SEARCH:

Amazon
OpenSearch
Serverless

Amazon
OpenSearch
Service

Amazon
Aurora
PostgreSQL

Amazon RDS
PostgreSQL

Amazon
DynamoDB
via zero-ETL

Amazon
MemoryDB

Amazon
Neptune

Amazon
DocumentD
B

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

New data sources for Amazon Bedrock Knowledge Bases

Extract structured data, metadata, and other information from documents

Inclusion/exclusion content filters

Incremental content syncs for added, updated, deleted content

Source attribution

Web Crawler
Single or multiple URLs

Atlassian Confluence
Confluence Cloud

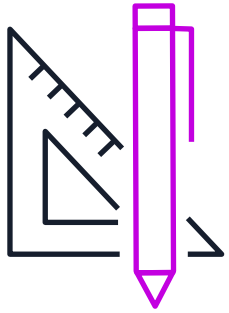
Microsoft SharePoint
SharePoint Online

Salesforce
Salesforce Standard and Custom objects

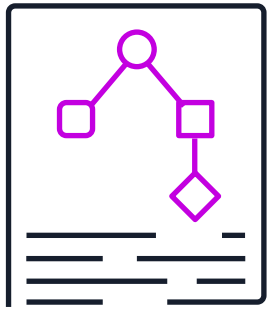
aws
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Agents

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



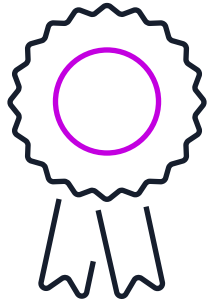
Decompose into steps using available actions and Amazon Bedrock Knowledge Bases



Execute action or search knowledge base

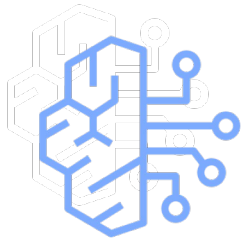
Observe results

Think about next step



Until final answer

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Amazon Bedrock Security

Helps keep your data secure and private

None of the customer's data is used to train the underlying model

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

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Guardrails

Implement safeguards customized to your application requirements and aligned to your responsible AI policies

Block as much as 85% more harmful content than protection natively provided by some FMs on Amazon Bedrock today, and filters over 75% hallucinated responses for RAG and summarization workloads

Evaluate prompts and model responses for agents, knowledge bases, FMs in Amazon Bedrock, and custom or third-party FMs

Configure thresholds to filter harmful content, jailbreaks and prompt injection attacks

Define and disallow denied topics with short natural language descriptions

Remove personally identifiable information (PII) and sensitive information in gen AI apps

Filter hallucinations by detecting groundedness and relevance of model responses based on context

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Agent Demo: Filing Noise Complaint

This demo showcases how an Amazon Bedrock Agent can collect user information and file a noise complaint using natural language, streamlining the process for both citizens and authorities.

Noise Complaint Demo Details

Workflow

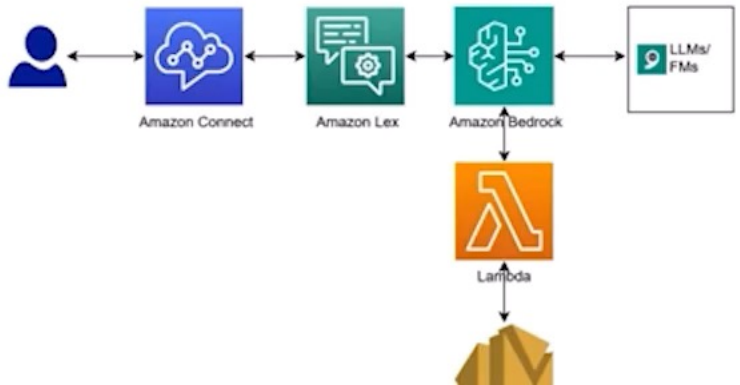
- User provides details like name, address, and noise type
- Bedrock Agent collects and verifies the information
- Agent processes the complaint in real time
- Notifies authorities via email or CAD system
- Confirms submission to the user

Key Benefits

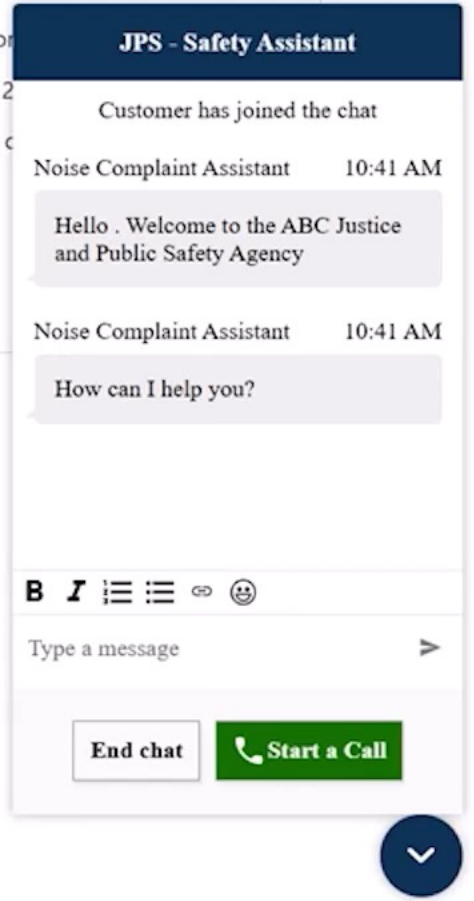
- Natural language input simplifies complaint filing
- Fast and scalable with emailing option
- Integrates with Bedrock, Lambda, and SES
- Reduces manual workload for authorities

Prompts to Ask

- "I want to file a noise co
- "Report loud music at 12
- "File a complaint about



Architecture



Security considerations for generative AI

COMPLIANCE & GOVERNANCE

The policies, procedures, and reporting needed to empower the business while minimizing risk

---

Create generative AI usage guidelines

Establish process for output validation

Develop monitoring & reporting processes

LEGAL & PRIVACY

The specific regulatory, legal, and privacy requirements for using or creating generative AI solutions.

Retain control of your data

Encrypt data in transit and at rest

Support regulatory standards

CONTROLS

The implementation of security controls that are used to mitigate risk.

Human-in-the-loop

Explainability & auditability

Testing strategy

Identity and access management

RISK MANAGEMENT

Identification of potential threats to generative AI solutions and recommended mitigations.

Threat modeling

Third-party risk assessments

Ownership of data, including prompts and responses

RESILIENCE

How to architect generative AI solutions to maintain availability and meet business SLAs.

Data management strategy

Availability

High Availability and Disaster Recovery strategy

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Responsible AI Dimensions

FAIRNESS

Considering impacts on different groups of stakeholders

EXPLAINABILITY

Understanding and evaluating system outputs

CONTROLLABILITY

Having mechanisms to monitor and steer AI system behavior

SAFETY

Preventing harmful system output and misuse

PRIVACY & SECURITY

Appropriately obtaining, using and protecting data and models

GOVERNANCE

Incorporating best practices into the AI supply chain, including providers and deployers

TRANSPARENCY

Enabling stakeholders to make informed choices about their engagement with an AI system

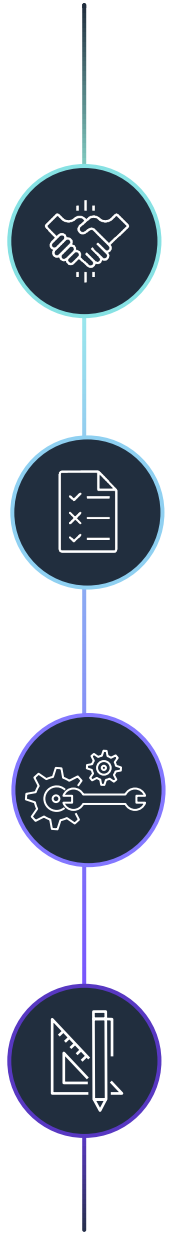
VERACITY & ROBUSTNESS

Achieving correct system outputs, even with unexpected or adversarial inputs

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Responsible AI:
Best practices



Put your people first

Assess risk on a (use) case-by-case basis

Iterate across the AI lifecycle

Test, test again, and then test again

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Vinod Kisanagaram

Solutions Architect
vinodaws@amazon.com

Mark Bronnimann

Sr Solutions Architect
mbronni@amazon.com

Please complete the survey for this session

Track : AI/ML

Session: (Workshop)
Building Agentic Workflows

aws Learning Days
State, Local, and Education

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.