

AWS State, Local, and Education Learning Days

Boston, MA

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

300
level

Generative AI Masterclass

A comprehensive masterclass on AI, covering technology evolution, implementation strategies, responsible practices.

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

200
level

Intelligent contact centers

Learn how AI can modernize customer service operations and improve satisfaction



GenAI Master Class

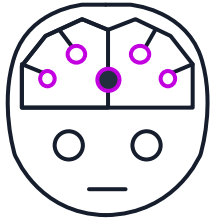
Vlad Fatu (he/him)

Solutions Architecture Manager
vladfatu@amazon.com

Maria Fassinger (she/her)

Solutions Architect
fascmari@amazon.com

AIML/GenAI hierarchy



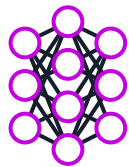
Artificial Intelligence (AI)

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



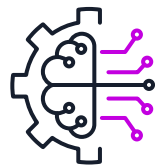
Machine learning (ML)

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



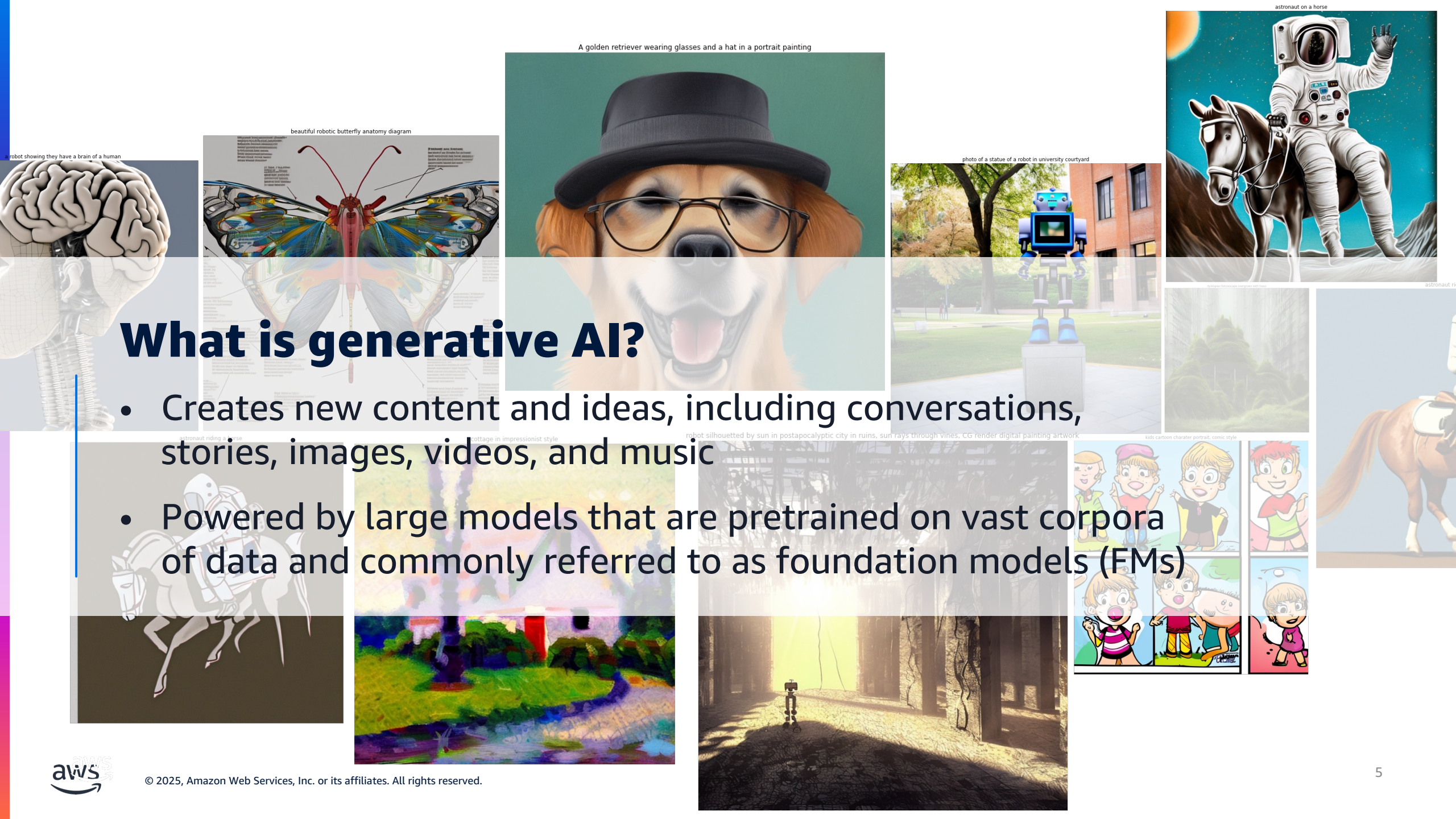
Deep learning (DL)

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)



A golden retriever wearing glasses and a hat in a portrait painting

beautiful robotic butterfly anatomy diagram

photo of a statue of a robot in university courtyard

astronaut on a horse

robot showing they have a brain of a human

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)

astronaut riding a horse

cottage in impressionist style

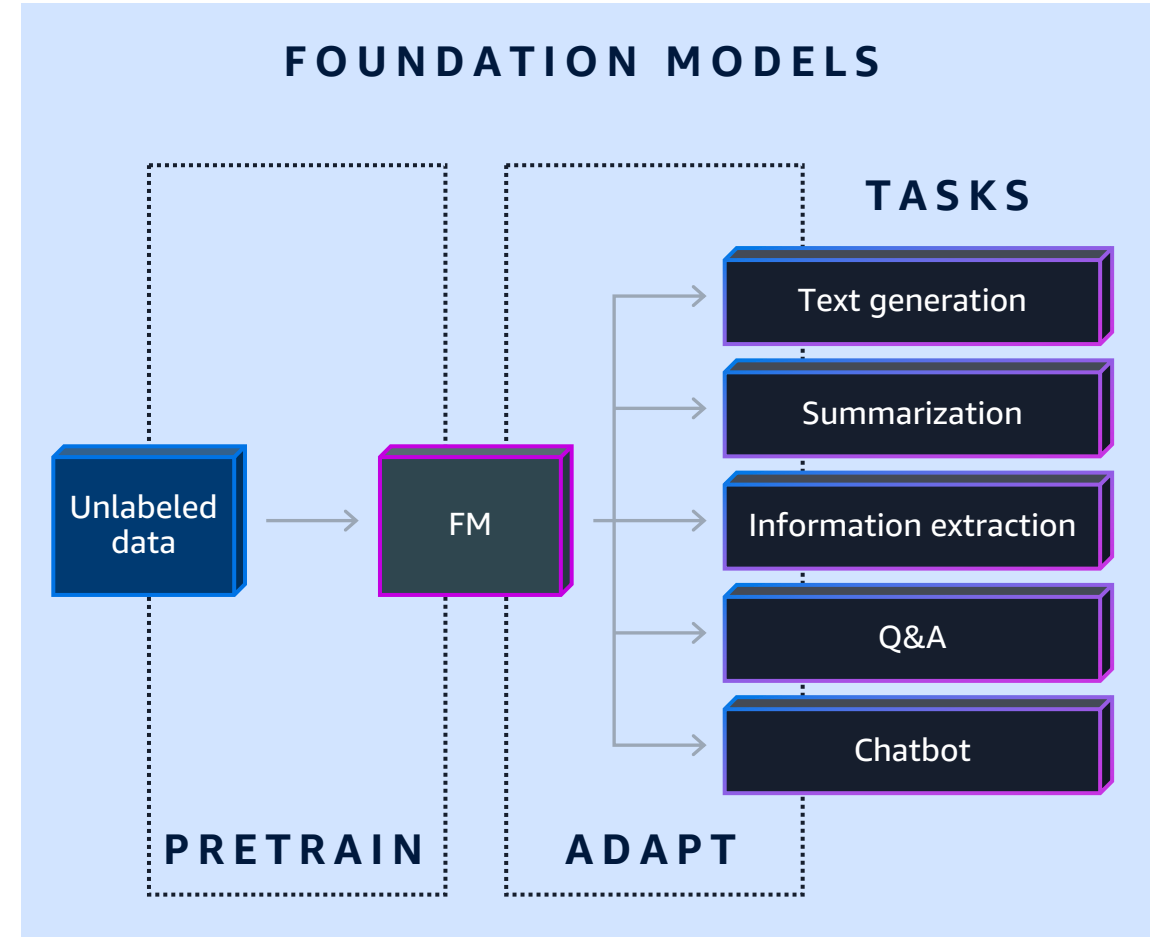
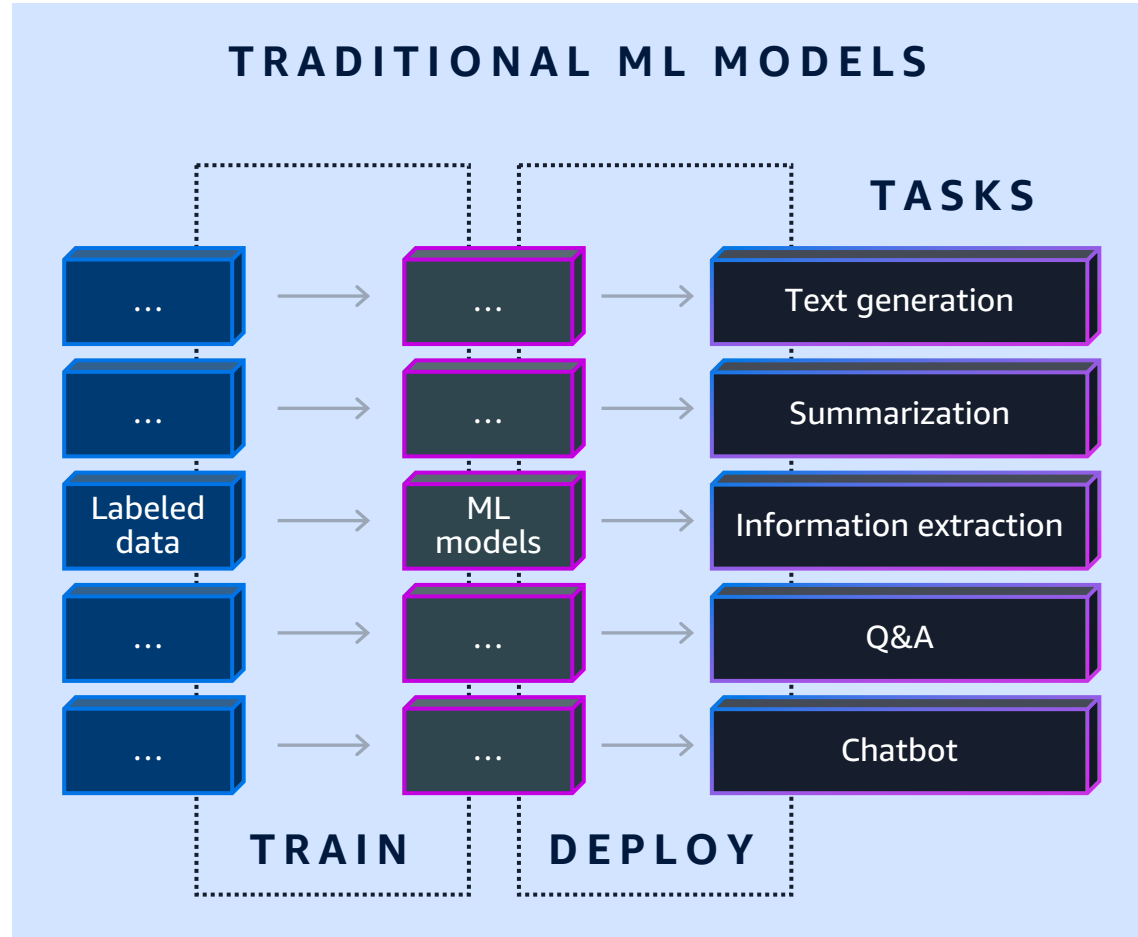
robot silhouetted by sun in postapocalyptic city in ruins, sun rays through vines, CG render digital painting artwork

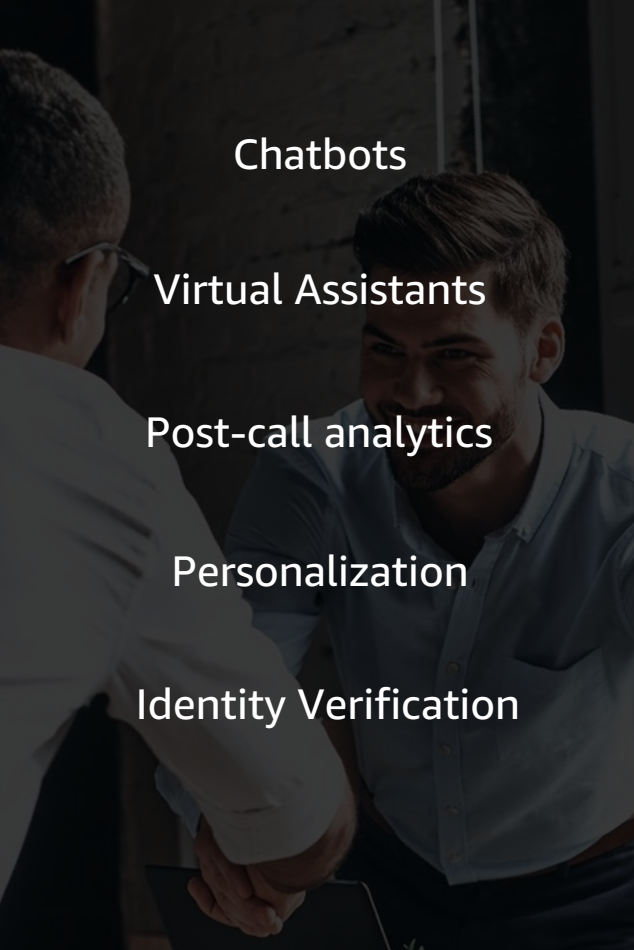
kids cartoon character portrait, comic style

GenAI in Action



Why foundation models?





Chatbots

Virtual Assistants

Post-call analytics

Personalization

Identity Verification



Conversational search

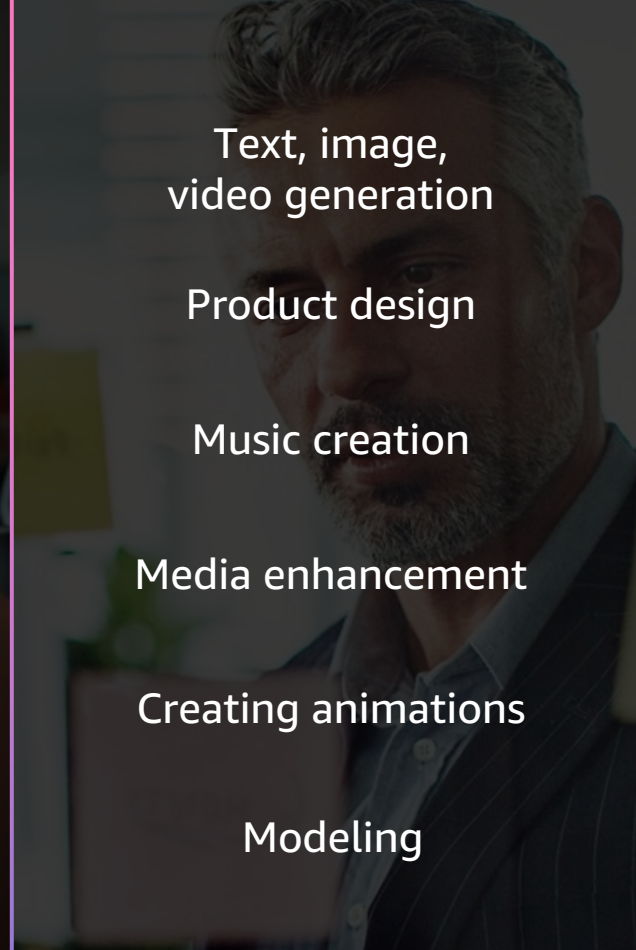
Agent Assist

Content Creation

Code generation

Text summarization

Sales scripts



Text, image,
video generation

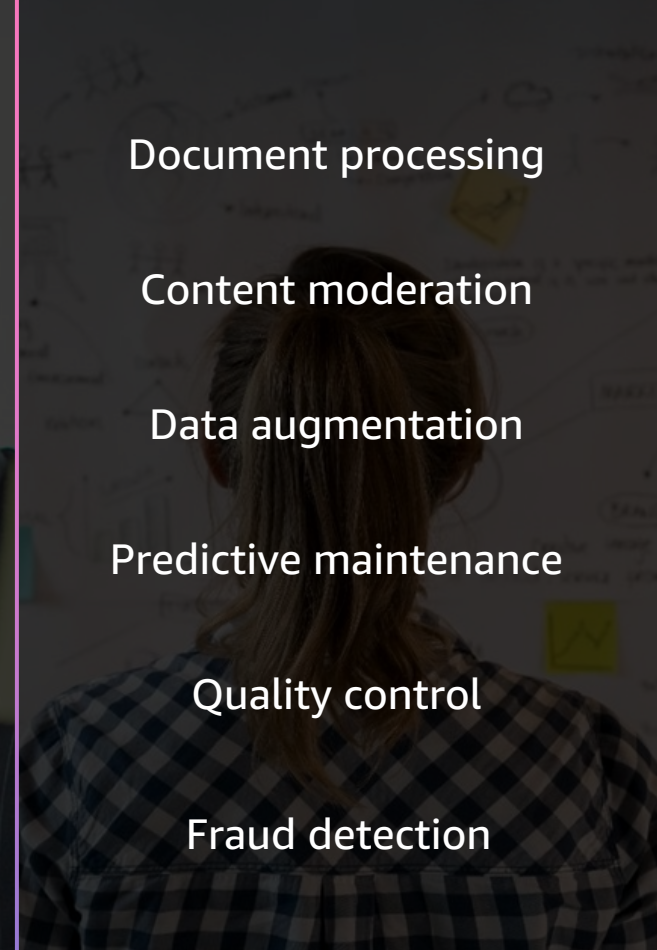
Product design

Music creation

Media enhancement

Creating animations

Modeling



Document processing

Content moderation

Data augmentation

Predictive maintenance

Quality control

Fraud detection

**Enhance
customer
experience**

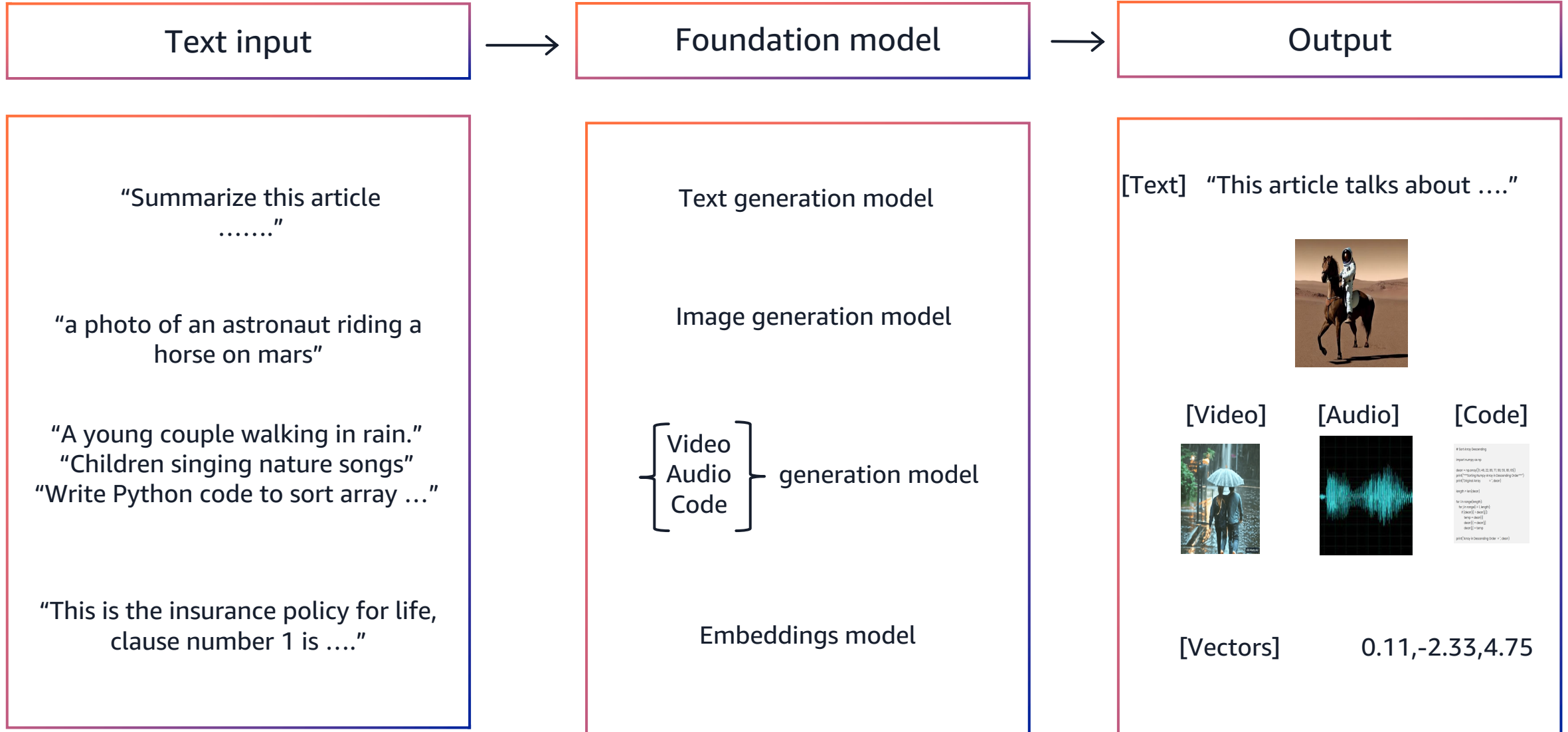
**Boost
employee
productivity**

**Creativity &
Content
Creation**

**Improve
business
operations**



Unimodal Models



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]

```
# Sort array descending
import numpy as np

array = np.array([45, 32, 11, 98, 55, 91, 67])
print("Original Array is descending order")
print(array)

length = len(array)

for i in range(length):
    for j in range(i + 1, length):
        if array[i] < array[j]:
            array[i], array[j] = array[j], array[i]
```

“a photo of an astronaut riding a
horse on mars”

What generative AI does well (for now)



Summarization



Content generation (Text, image, audio, video)



Language Translation



Correction/paraphrasing



Classification

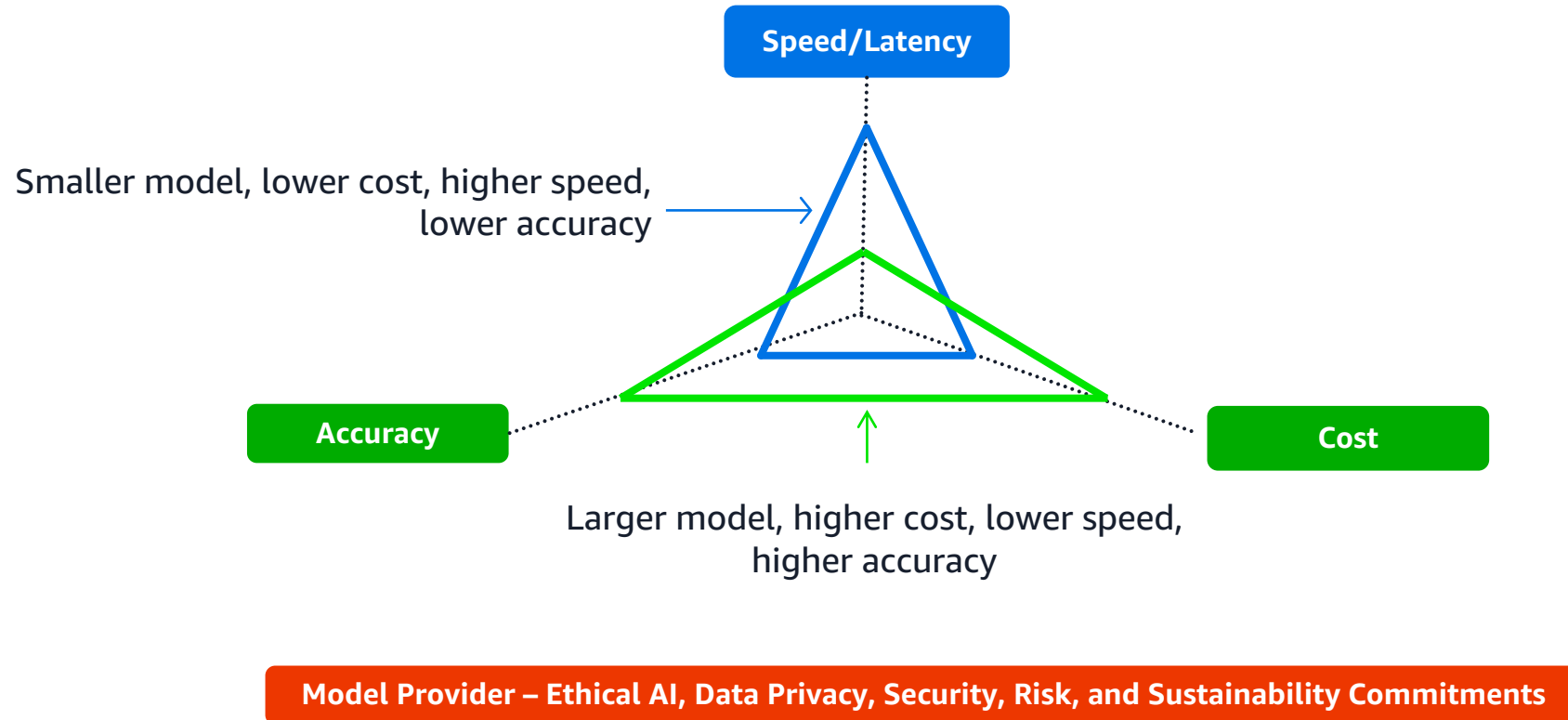
Use case examples

- 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

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)

Considerations when selecting a generative AI model



Strategies for Implementation

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

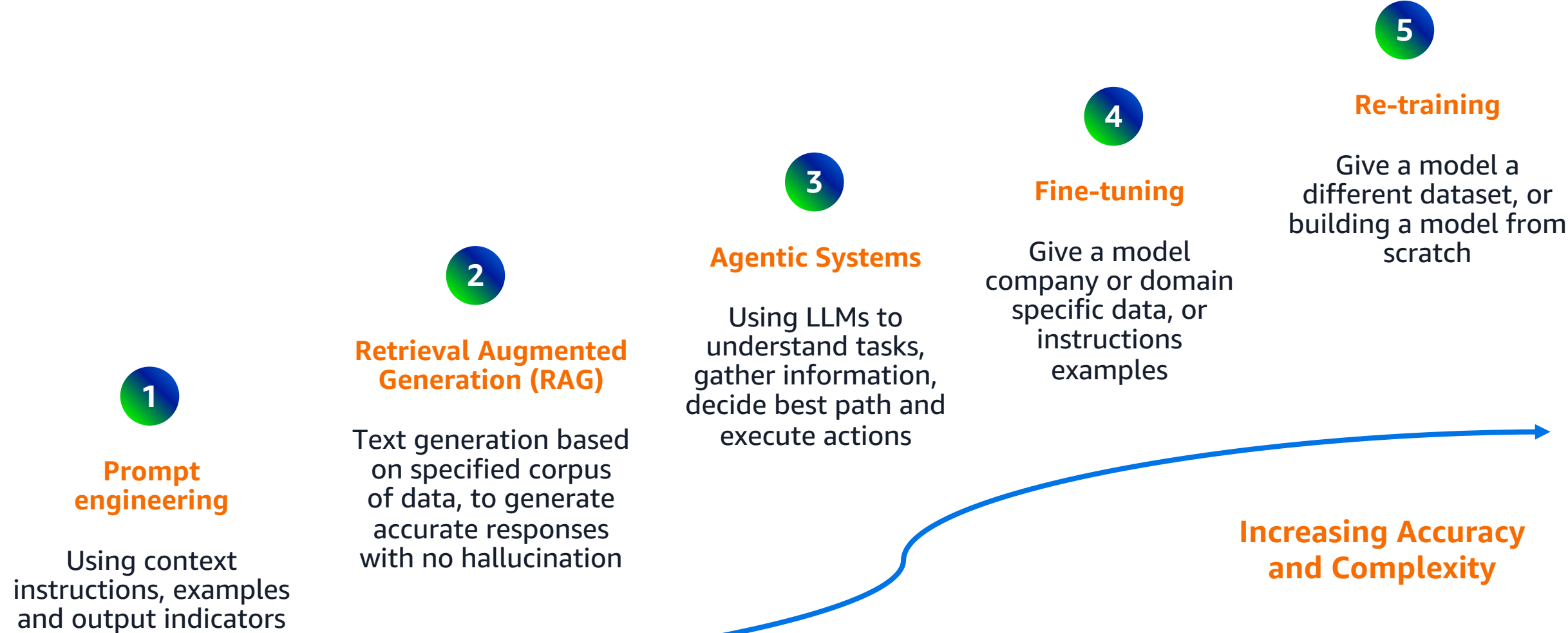
Agentic System

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.

Strategies for implementation and their trade-offs



Prompt Engineering

Zero shot learning

Prompt

Review: "Earnings per share have beaten analyst expectations"

What is the sentiment?

Input



Output



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

Few shot learning

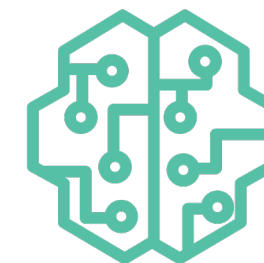
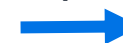
Prompt

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

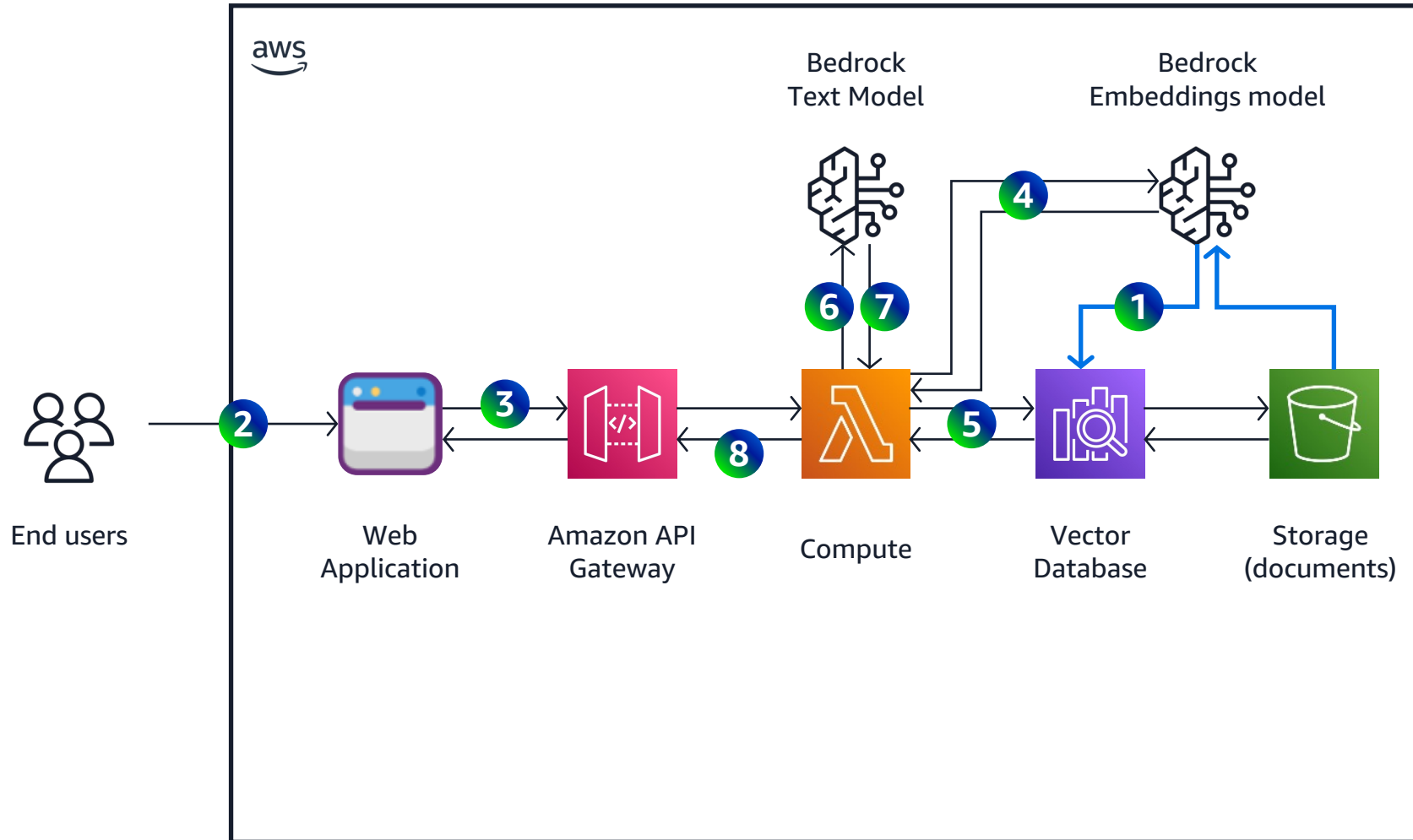


Output



positive

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



Agentic Systems

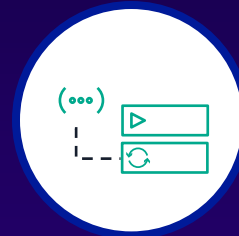


do this
for me...

Done. Here's
the result...

Agent

Instructions: "you are an agent that ..."



Actions



Knowledge Bases

Agents can combine **Actions** and **Knowledge Bases**

HR Policy Assistant



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

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

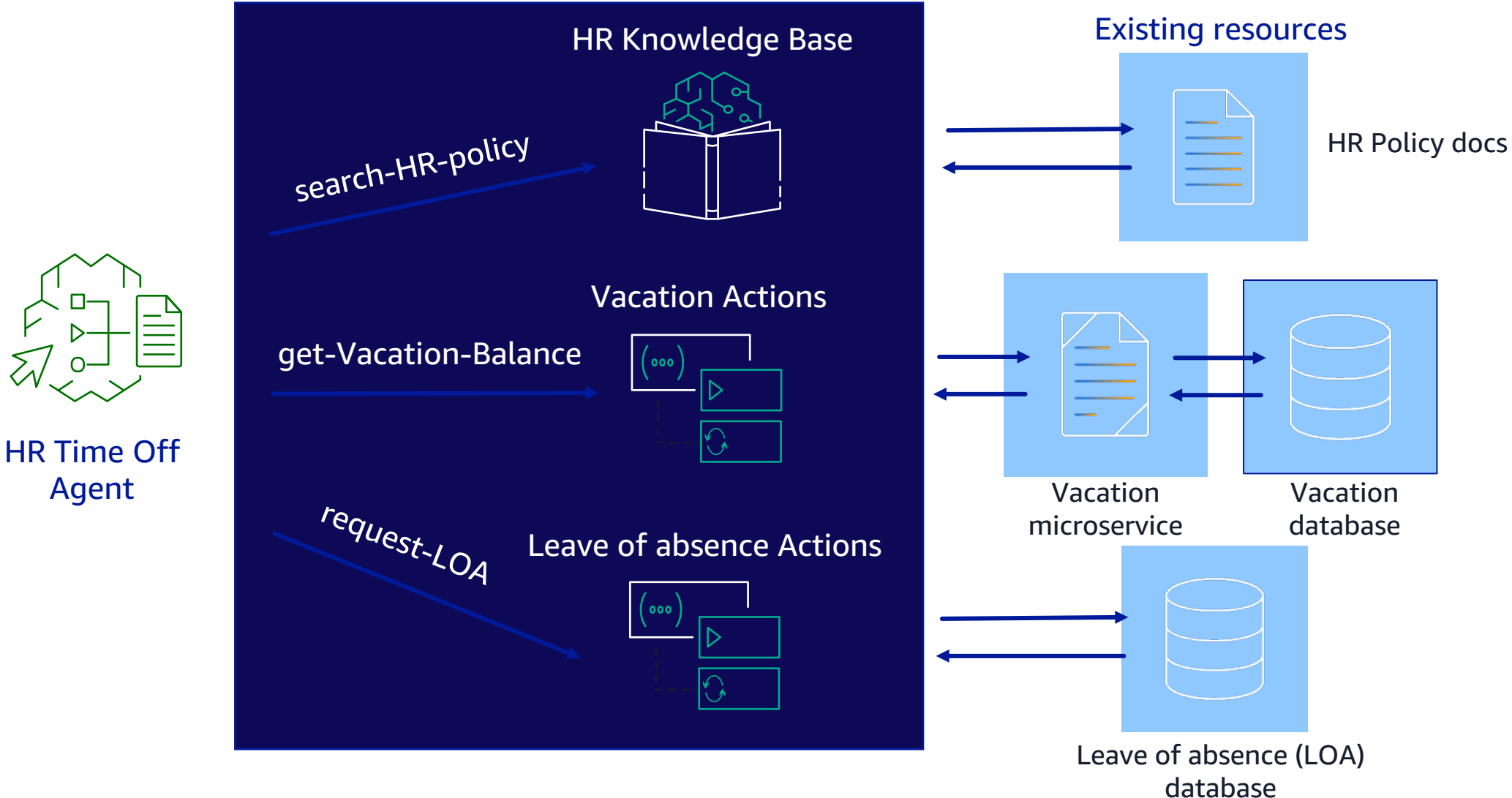
HR Knowledge Base

Vacation Policy
Contains the entire vacation policy for the company

HR Actions

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

Agents build on existing enterprise resources



Agent orchestration is transparent – Trace



send reminder for claim-857



Reminder sent successfully for claim-857. The send reminder status is InProgress and the tracking ID is 50e8400-e29b-41d4-a716-446655440000.
[< Hide trace](#)

Trace (4)

Request ID: dcd6938c-73b2-414d-b589-0ea18d6af40e

Copy all

Pre-processing

Orchestration & knowledge base

Post-processing

Orchestration and knowledge base

After running a prompt in the test window, this orchestration trace allows you to explore the trace steps to understand the linear chain of thought used by the agent's orchestration prompt component. If a knowledge base was invoked, the trace also allows you to see how the results from the knowledge base were summarized to generate an observation that is used for for orchestrating the next step or generating the response.

▼ Step 1

[Show config](#)

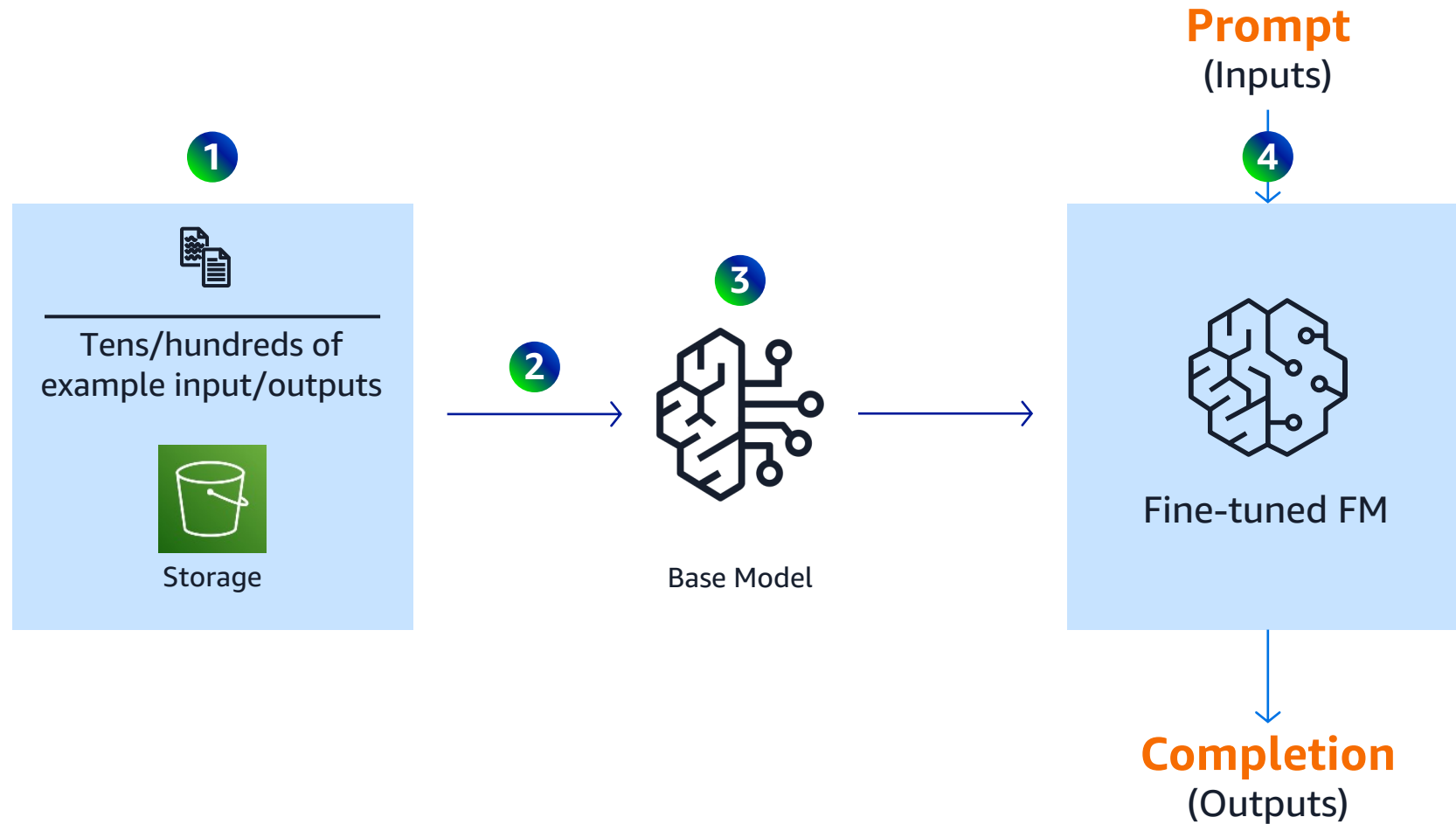
Trace

```
1 {
2   "modelInvocationInput": {},
18  "rationale": {
19    "text": "To answer this question, I will:\n\n1. Call
           GET::claims-actions::getOutstandingPaperwork
           function to get the list of pending documents for
           claim-857. \n\n2. Check if I have the pending
           documents to send a reminder. \n\n3. If I have the
           pending documents, I will call POST::claims-
           actions::sendReminders function to send the
           reminder for claim-857.\n\n4. Return a success
           message to the user.\n\nI have double checked and
           made sure that I have been provided the GET
           ::claims-actions::getOutstandingPaperwork and POST
```

Detailed orchestration **trace** in the console and from the SDK

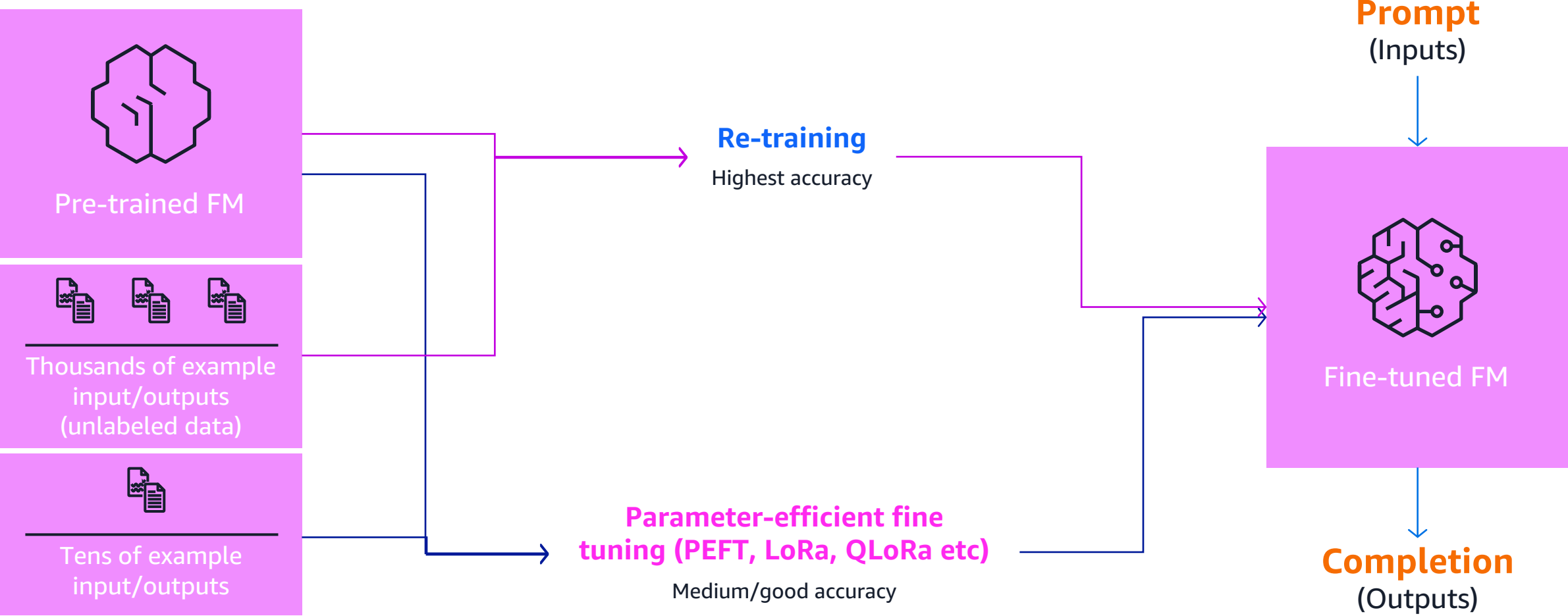


Fine-tuning (Task specific)

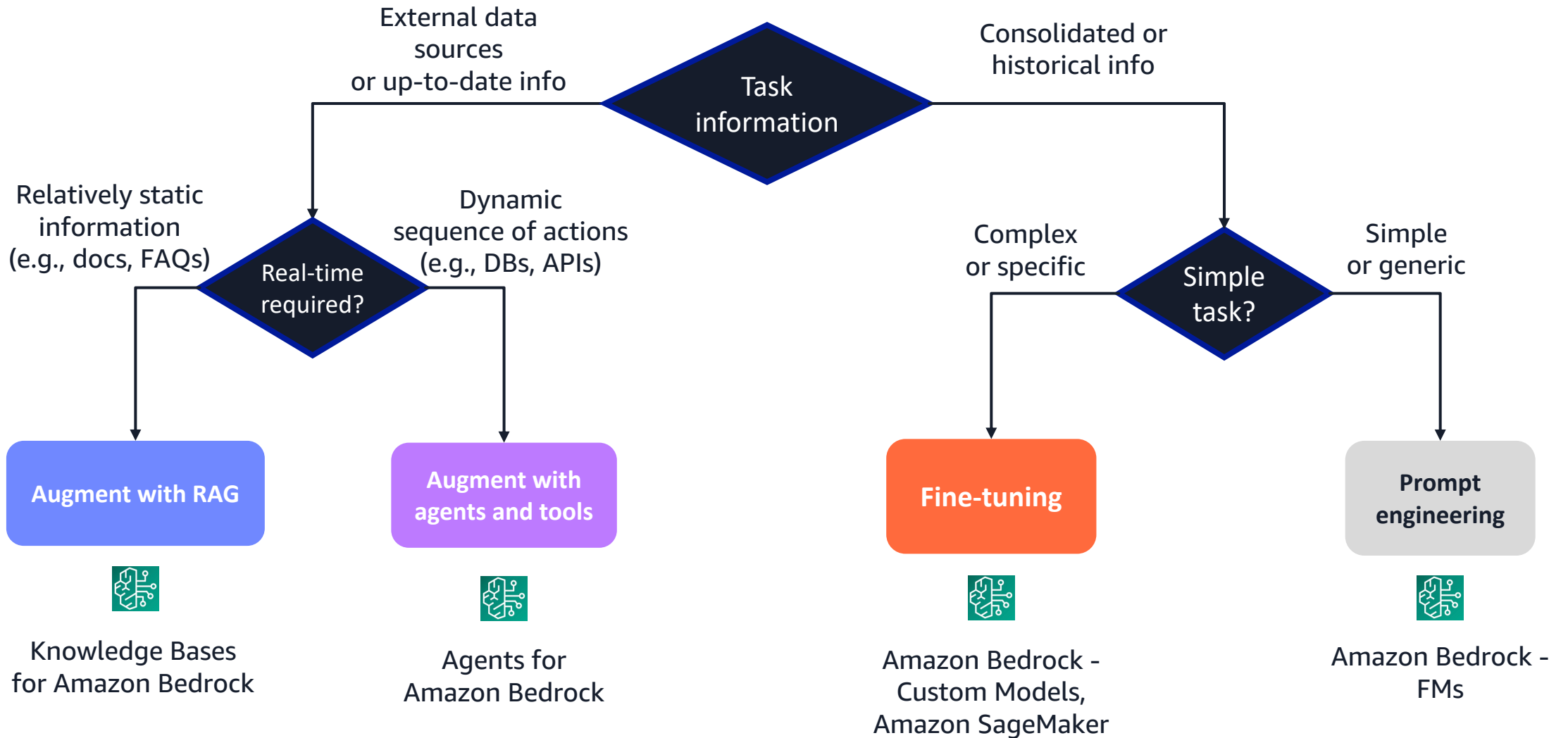


- 1** Tens/hundreds of instructions or examples are stored
- 2** Instructions are sent to the base model via APIs
- 3** A base model is fine-tuned with these instructions (minutes, hours, days)
- 4** The new fine-tuned model is available for the customer to start inferencing

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



Decision Tree





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

Amazon Bedrock

BROAD CHOICE OF MODELS

AI21labs

Effective reasoning & rapid analysis for long context windows

JAMBA

amazon

Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation

AMAZON NOVA

ANTHROPIC

Advanced reasoning & coding capabilities, including computer use skills

CLAUDE

cohere

Multimodal search & advanced retrieval powering multilingual knowledge agents

COMMAND
EMBED
RERANK

Luma

High-quality video generation from text & images

LUMA RAY 2

Coming soon

Meta

Advanced image & language reasoning

LLAMA

MISTRAL AI

Knowledge summarization, expert agents, & code completion

MISTRAL
MIXTRAL

poolside

Software engineering AI for large enterprises

MALIBU
POINT

Coming soon

stability.ai

High-quality AI image generation, easily deployable at scale

STABLE DIFFUSION
STABLE IMAGE



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

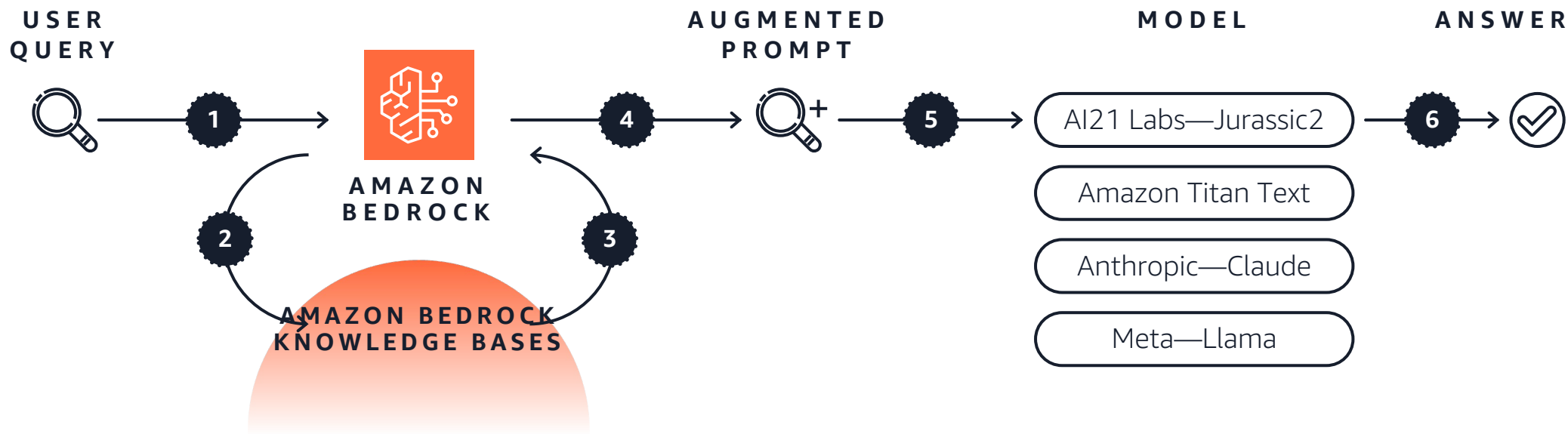
← Lower Cost & Latency

→ Increasing Intelligence



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

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

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



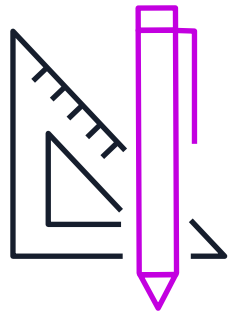
Microsoft SharePoint
SharePoint Online



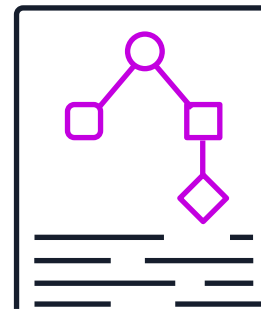
Salesforce
Salesforce Standard and Custom objects

Amazon Bedrock Agents

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



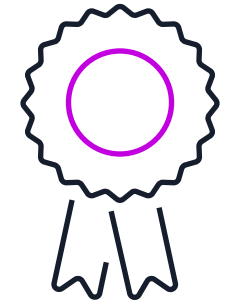
Decompose into steps using available actions and Amazon Bedrock Knowledge Bases



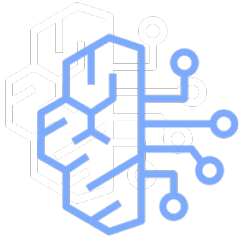
Execute action or search knowledge base

Observe results

Think about next step



Until final answer



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

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

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 complaint"
- "Report loud music at 123 Main St"
- "File a complaint about construction noise"

Architecture



JPS - Safety Assistant

Customer has joined the chat

Noise Complaint Assistant 10:41 AM

Hello . Welcome to the ABC Justice and Public Safety Agency

Noise Complaint Assistant 10:41 AM

How can I help you?

B **I** **≡** **≡** **🔗** **😊**

Type a message

End chat **Start a Call**



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

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



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



Thank you!

Vlad Fatu (he/him)

Solutions Architecture Manager
vladfatu@amazon.com

Maria Fassinger (she/her)

Solutions Architect
fascmari@amazon.com

1:30pm – 3:00pm

300
level

**Workshop:
Building Agentic
Workflows**

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

Please complete the survey
for this session



Track : AI/ML

**Session: (Workshop)
Building Agentic Workflows**