AWS State, Local, and Education Learning Days

Salt Lake City, Utah

11:30am – 12:30pm

300 level

Generative AI Masterclass

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

aws Learning Days
State, Local, and Education

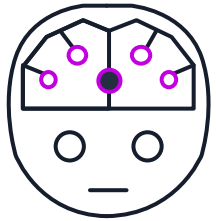
GenAI Master Class

Kai Dickman

AIML Specialist Solutions Architect
kaimatt@amazon.com

aws Learning Days
State, Local, and Education

AIML/GenAI hierarchy

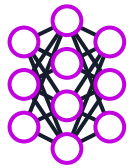


Artificial Intelligence (AI)

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)

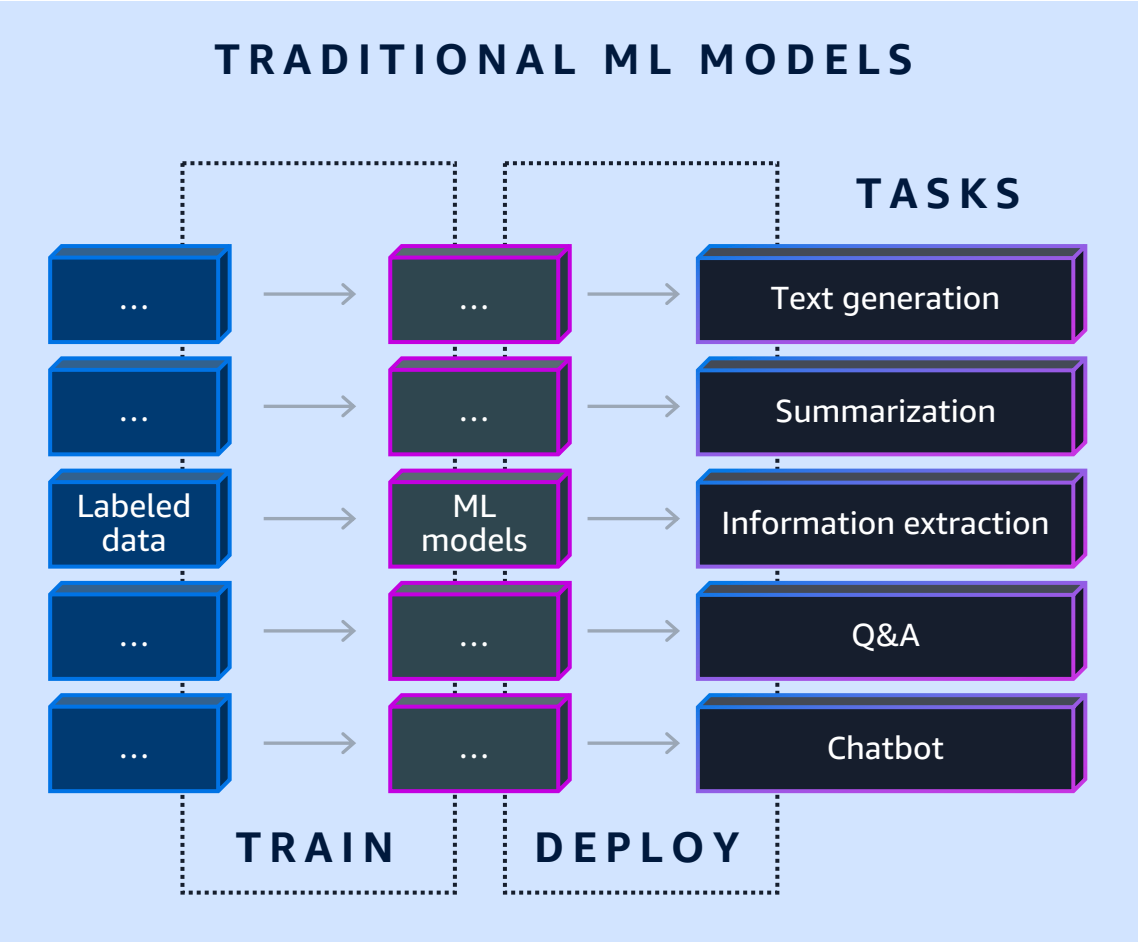
A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

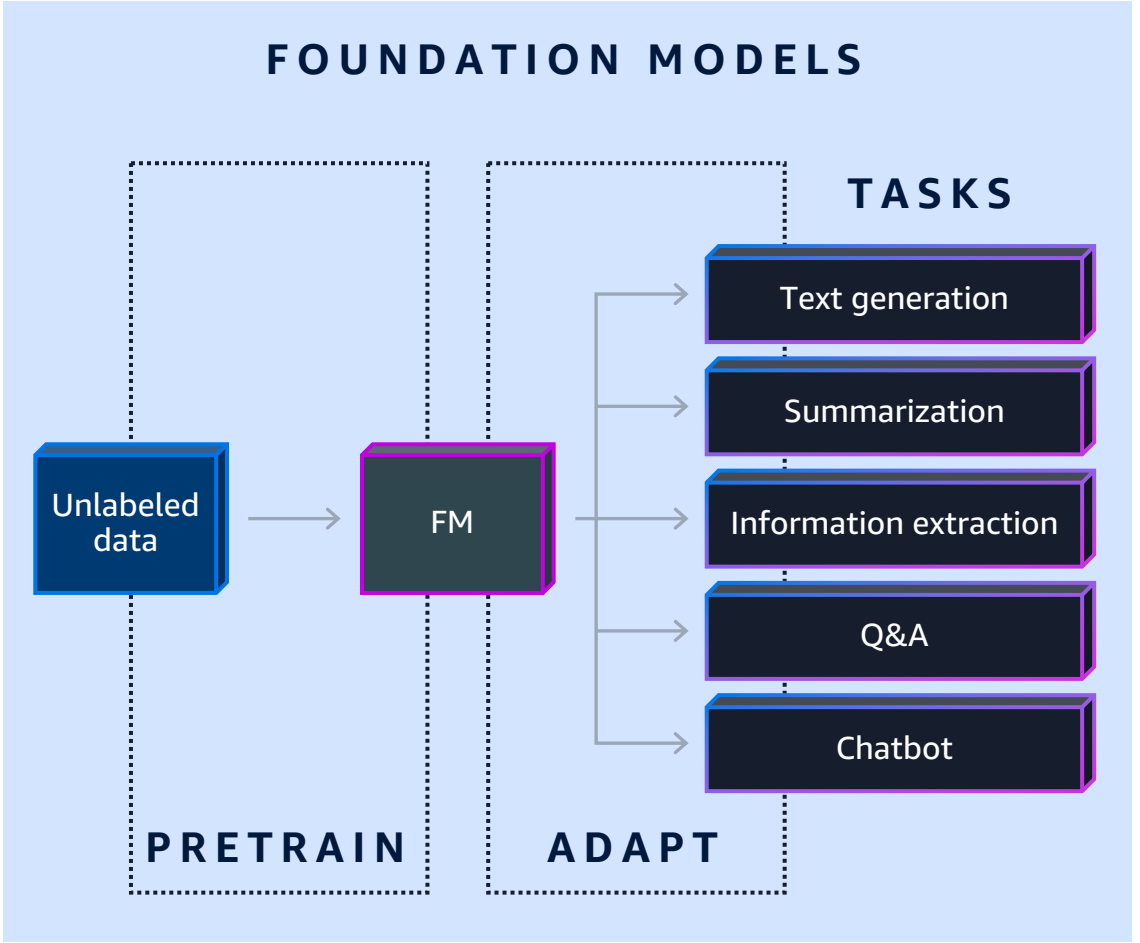
Generative AI

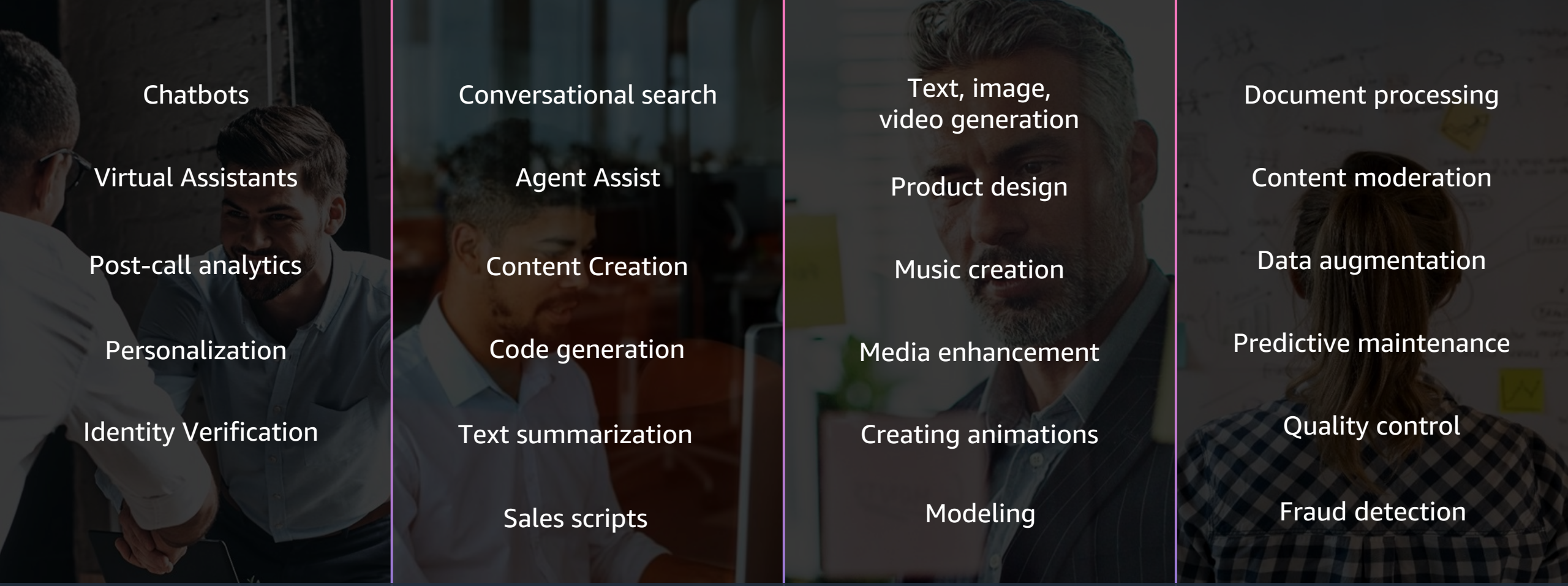
Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

GenAI in Action

Why foundation models?







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

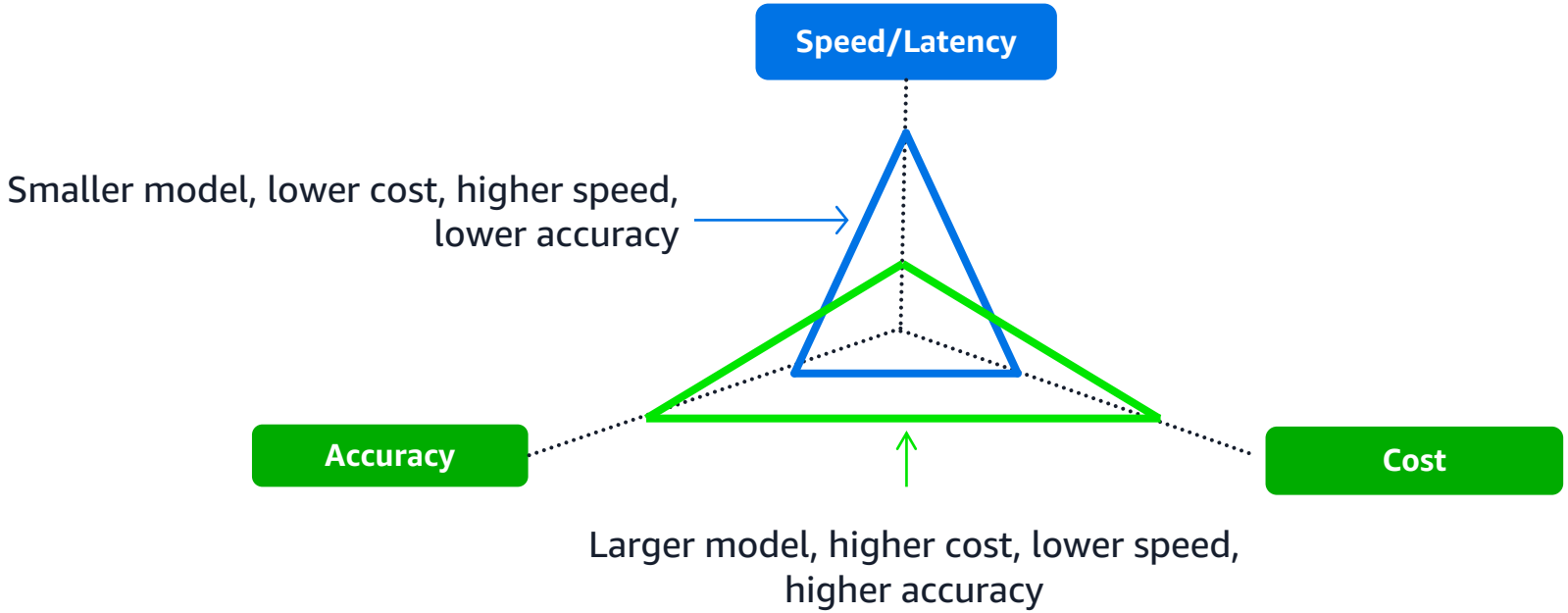
What generative AI does well (for now)

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

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



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

Strategies for Implementation

Data is your foundation


Generative AI
Application


Generative AI
Application
Data
Foundation
STORAGE
GOVERNANCE
& COMPLIANCE
DATABASES,
ANALYTICS,
& DATA LAKES
DATA
INTEGRATION

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

Vector

Numerical representations of data such as text, images, audio, and files

Embeddings

Collection of vectors that captures the meaning and semantic relationship between data points

Parameters

The millions or billions of adjustable values that determine how the AI processes and responds to information

Prompt Engineering

The process of providing text instructions to models so as to generate the desired outputs

RAG

Leveraging data to augment the context being passed to a model

Agentic System

Using the reasoning capabilities of an LLM to automate processes and tasks

Pre-training

The process of creating a foundation model from scratch

Continuous pre-training

Expanding a model's understanding of a specific domain and improve the model's overall competency for your needs

Fine-tuning

Using a small set of labelled data to increase accuracy and knowledge of a model.

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

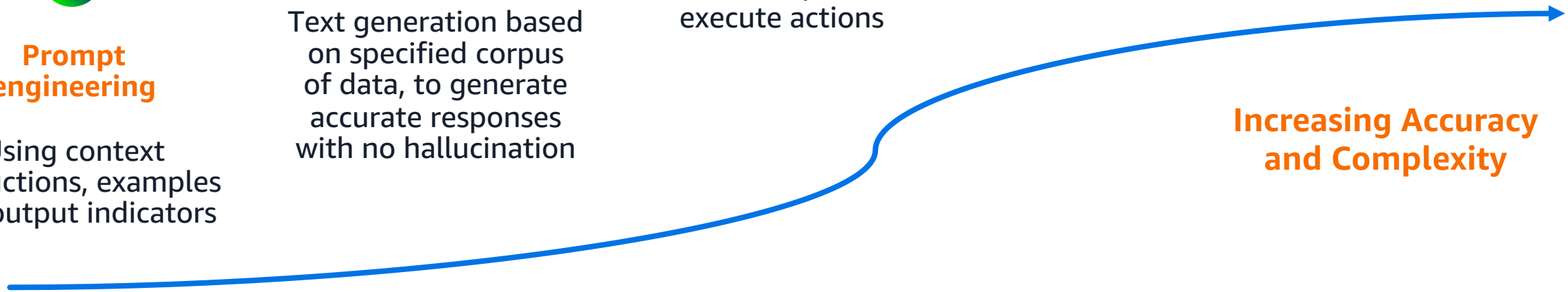
Fine-tuning

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



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

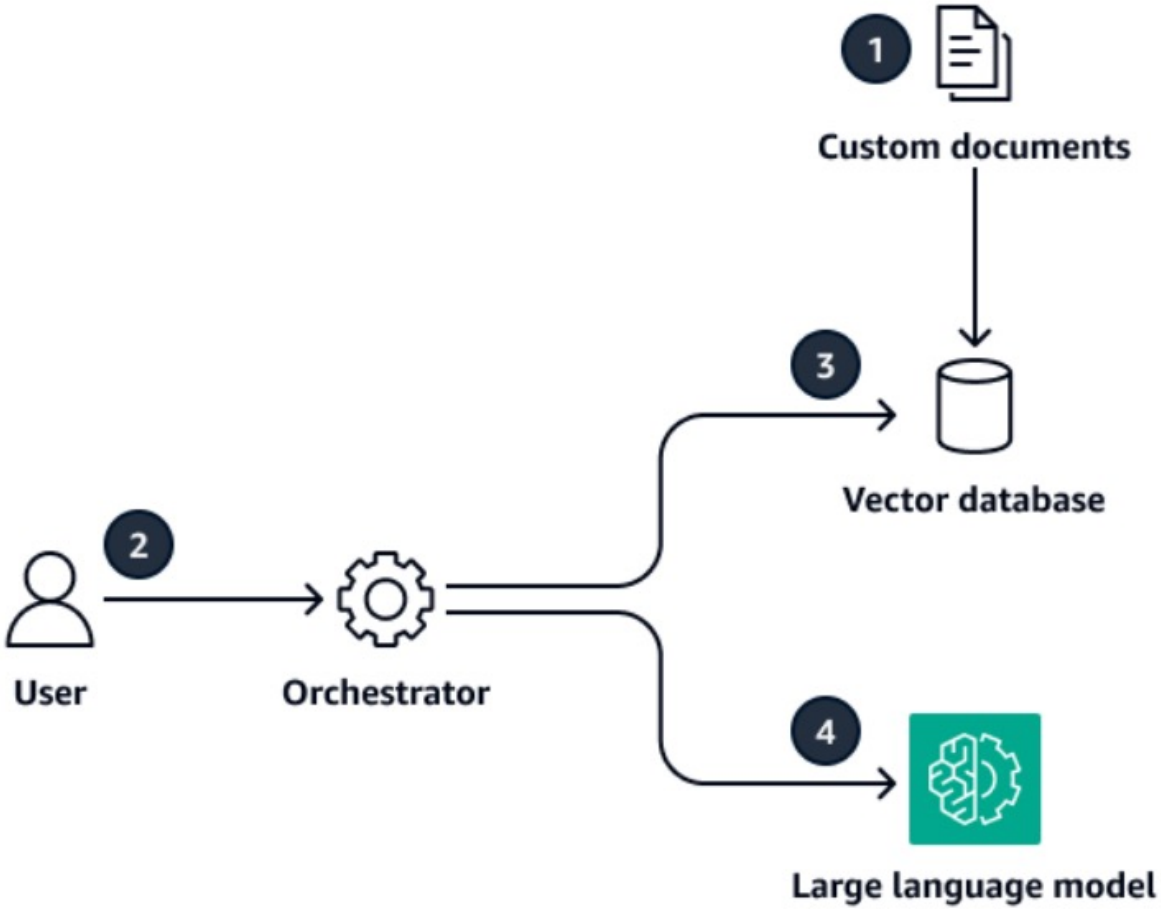
Review: “S&P500 Tops 5,600 for first time as tech rallies"
Sentiment:

Input

positive

Output

Retrieval augmented generation (RAG)



Agentic Systems

do this for me…

Done. Here’s the result…

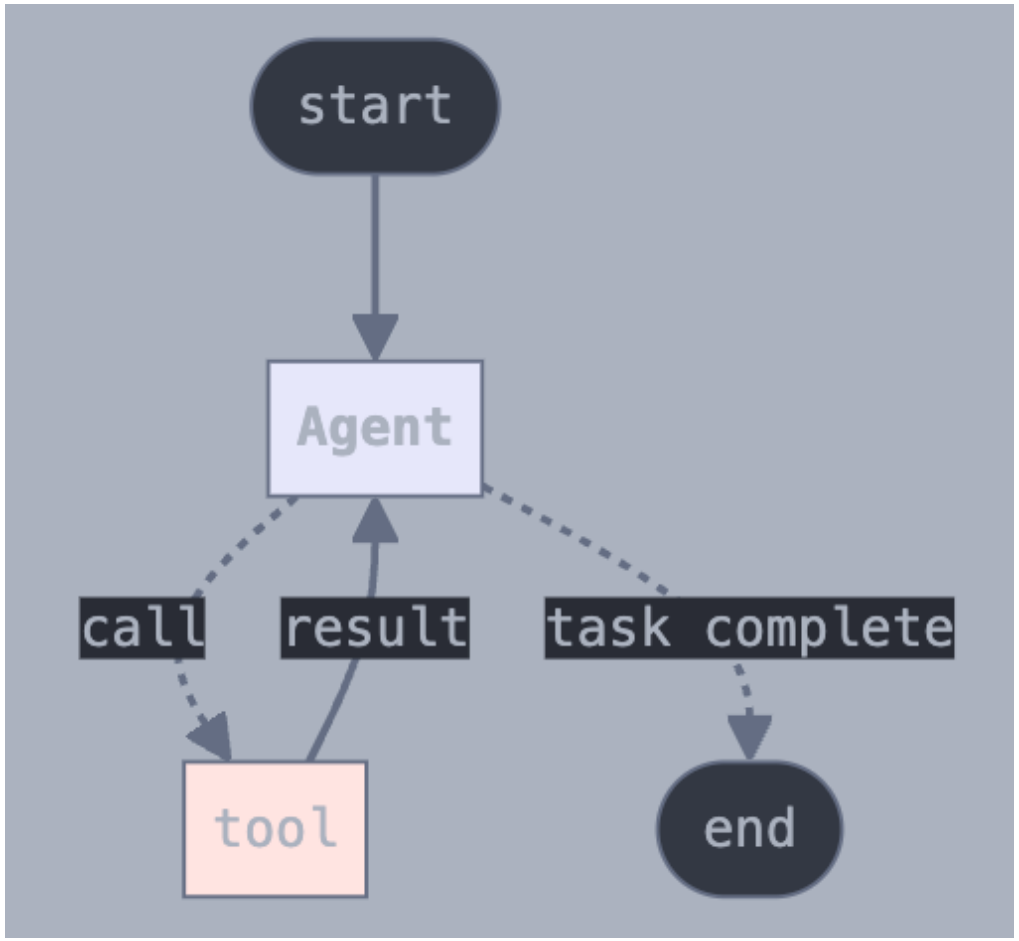
Defining agents

An agent is an AI application consisting of a **model** equipped with **instructions** that guide its behavior, access to **tools** that extend its capabilities, encapsulated in a **runtime** with a dynamic lifecycle.

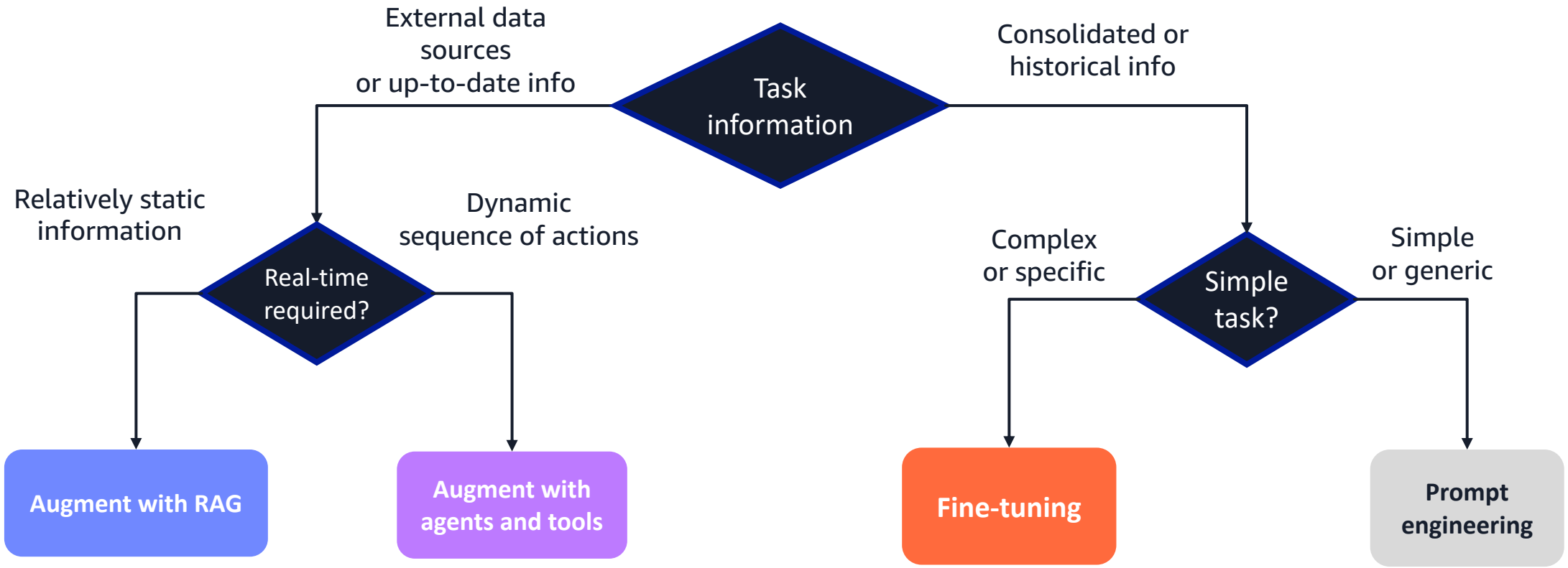
Agent =

+ Model
+ Instructions
+ Tools
+ Runtime

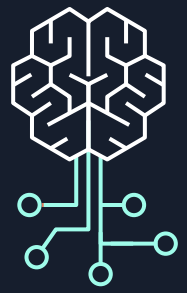
Agentic Systems



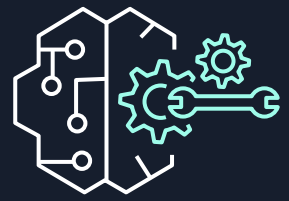
Decision Tree



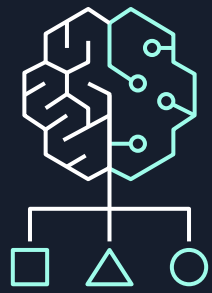
Amazon AI **simplifies**



Choice



Customization



Integration

Security and governance

Amazon Bedrock

MORE THAN JUST MODELS

Guardrails | Cost tracking and efficiency | Evaluation

Knowledge Bases | Agents | Data Automation | Prompts and Flows

Consumption options and unified API

Model Catalog | Model Customization | Bring Your Own Model

AI21labs	amazon	ANTHROPIC	cohere	Luma	Meta	MISTRAL AI_	poolside	stability.ai
Effective reasoning & rapid analysis for long context windows	Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation	Advanced reasoning & coding capabilities, including computer use skills	Multimodal search & advanced retrieval powering multilingual knowledge agents	High-quality video generation from text & images	Advanced image & language reasoning	Knowledge summarization, expert agents, & code completion	Software engineering AI for large enterprises	High-quality AI image generation, easily deployable at scale

aws

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**

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

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

Amazon Bedrock

Helps keep your data secure and private

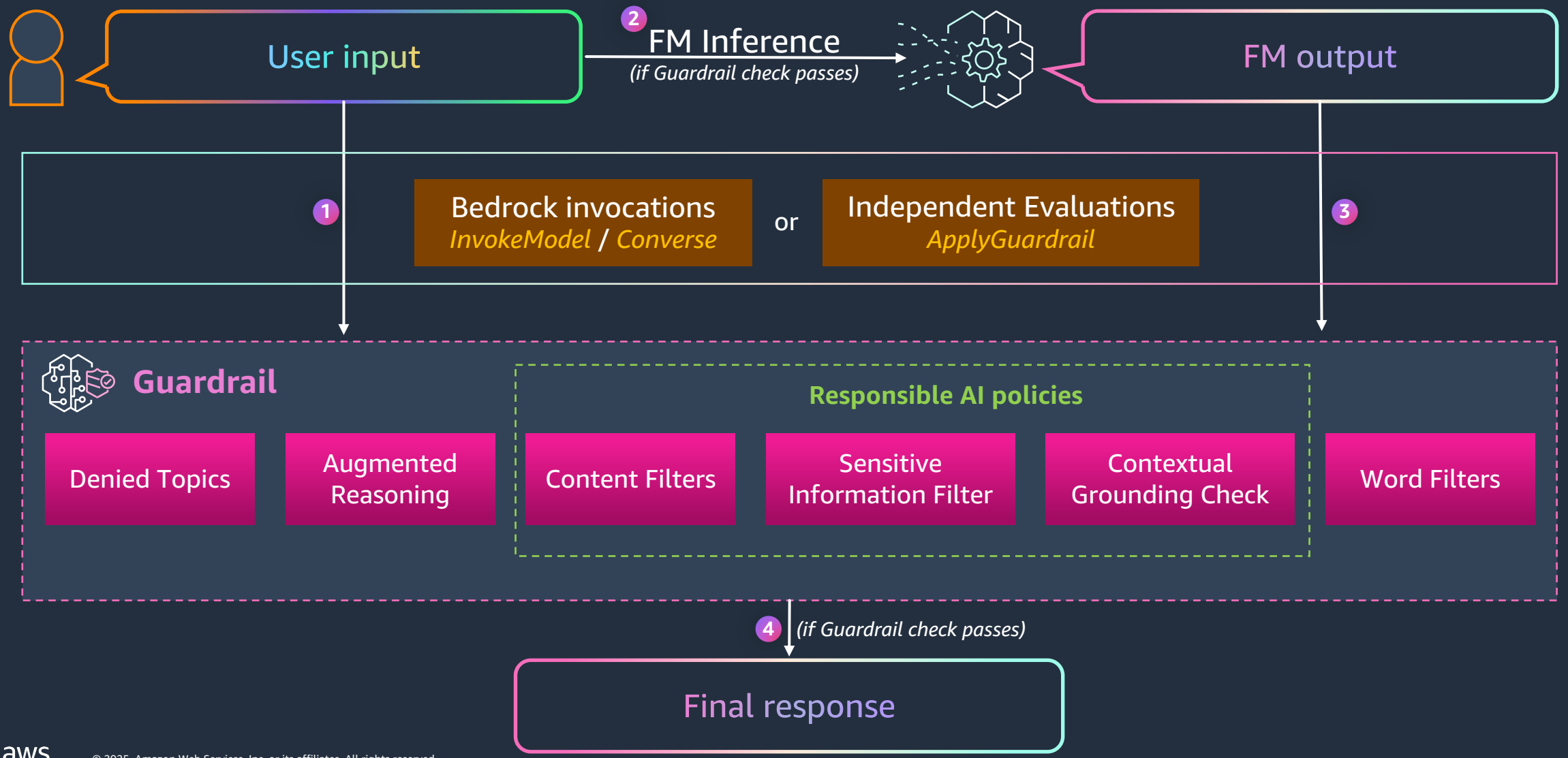
None of the customer's data is used to train the underlying model

All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC

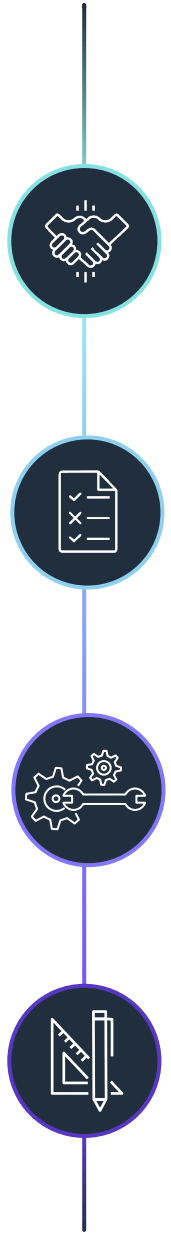
Data remains in the Region where the API is processed

Support for GDPR, SOC, ISO, CSA compliance, and HIPAA eligibility

How Guardrails work



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

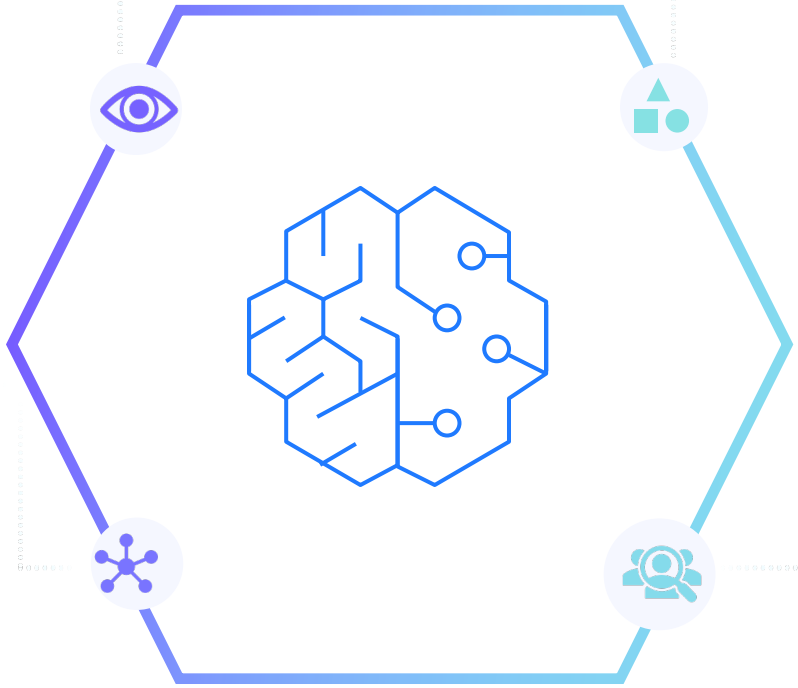
Strategic recommendations on generative AI

EYES WIDE OPEN

- Culture of continuous experimentation
- Prevent early dependencies

FLEXIBILITY IS KEY

- Innovation requires flexibility free from technical or contractual lock-ins.
- Infrastructure supporting 3rd-party Generative AI integration
- A breadth of services ensures long-term flexibility and business value



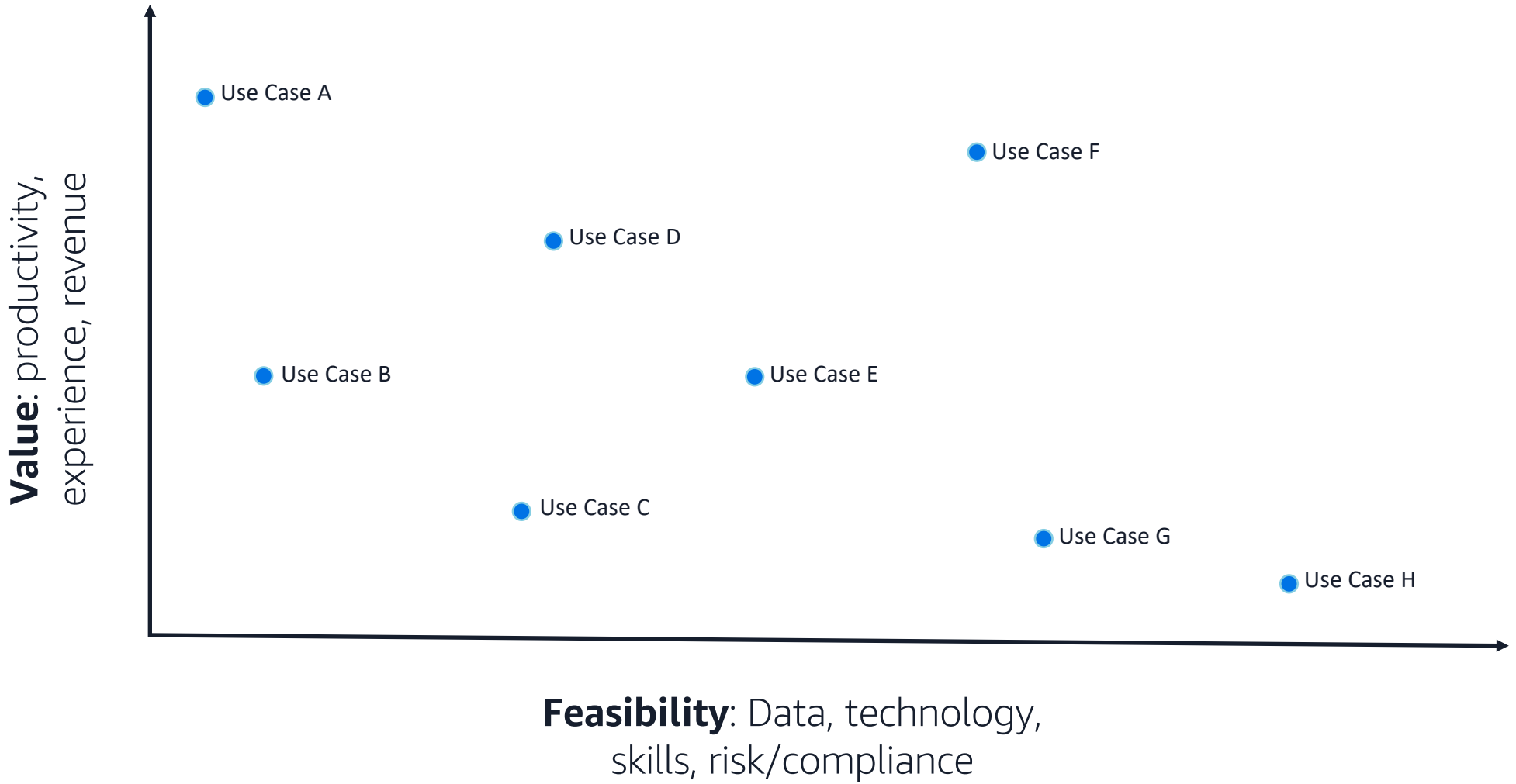
NO ONE SIZE FITS ALL

- Generative AI as an enhancement – not a replacement.
- Requirements in enterprises differ based on financial resources, security needs, governance and skills.
- Evaluate buy-vs-build

LISTEN AND ENABLE

- Work backward from your customers
- Listen to your domain experts
- Enable employees with the right set of tooling

Selecting the right use cases



Thank you!

Kai Dickman

AIML Specialist Solutions Architect
kaimatt@amazon.com

Please complete the survey for this session

Track : Artificial Intelligence and Machine Learning

Session : GenAI Master Class

Coming up NEXT

1:30pm – 3:00pm

300 level

Workshop: Building Agentic Workflows

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

aws Learning Days
State, Local, and Education