aws

AI Innovation Labs using AWS and Boomi

Venkatesh Aravamudan

Senior Partner Solutions Architect

Customer Journey

Business Use-Case

Choice:
- Models to choose
- Model Evaluation

Responsible AI:
- Data Privacy & Security
- Guardrails

Customization:
- Prompt Engineering
- RAG
- Fine-Tuning

Integration:
- Agents
- GenAIOps

Amazon Bedrock provides capabilities to accelerate generative AI application development

Amazon Bedrock

The easiest way to build and scale generative AI applications with powerful tools and foundation models
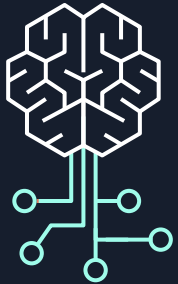
Choice of leading FMs through a single API

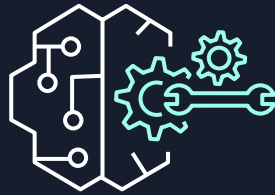Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and data governance
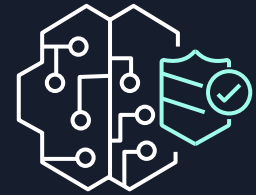
Amazon Bedrock **simplifies**

Choice

Customization

Integration

Security and governance

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

BROAD CHOICE OF MODELS

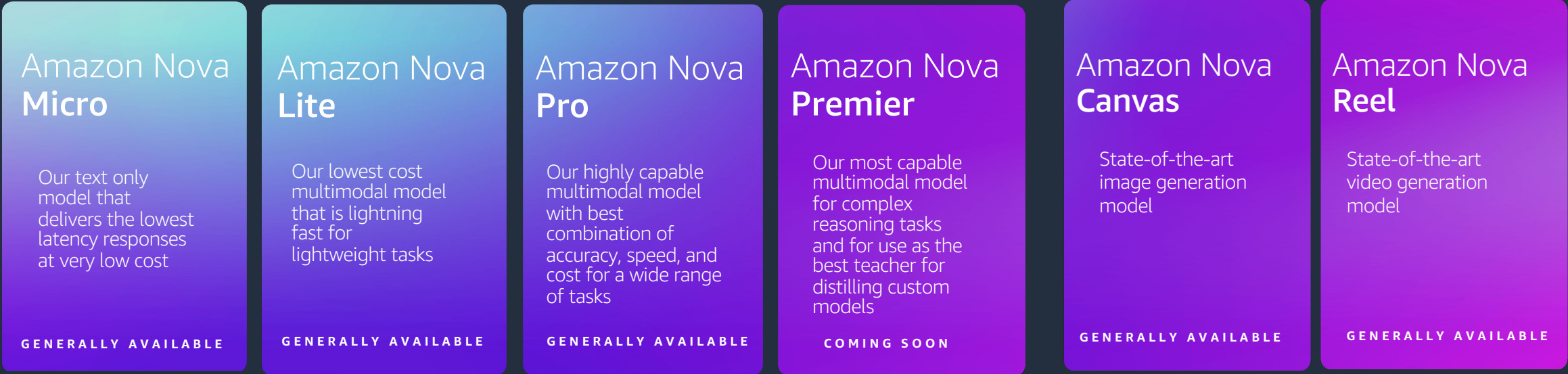| AI21 labs	amazon	ANTHROP\C	cohere	Luma	Meta	MISTRAL AI_	poolside	stability.ai
Effective reasoning & rapid analysis for long context windows	Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation	Advanced reasoning & coding capabilities, including 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
JAMBA	**AMAZON NOVA**	**CLAUDE**	**COMMAND** **EMBED** **RERANK**	**LUMA RAY 2**	**LLAMA**	**MISTRAL** **MIXTRAL**	**MALIBU** **POINT**	**STABLE DIFFUSION** **STABLE IMAGE**

Coming soon

Amazon Nova Foundation Models

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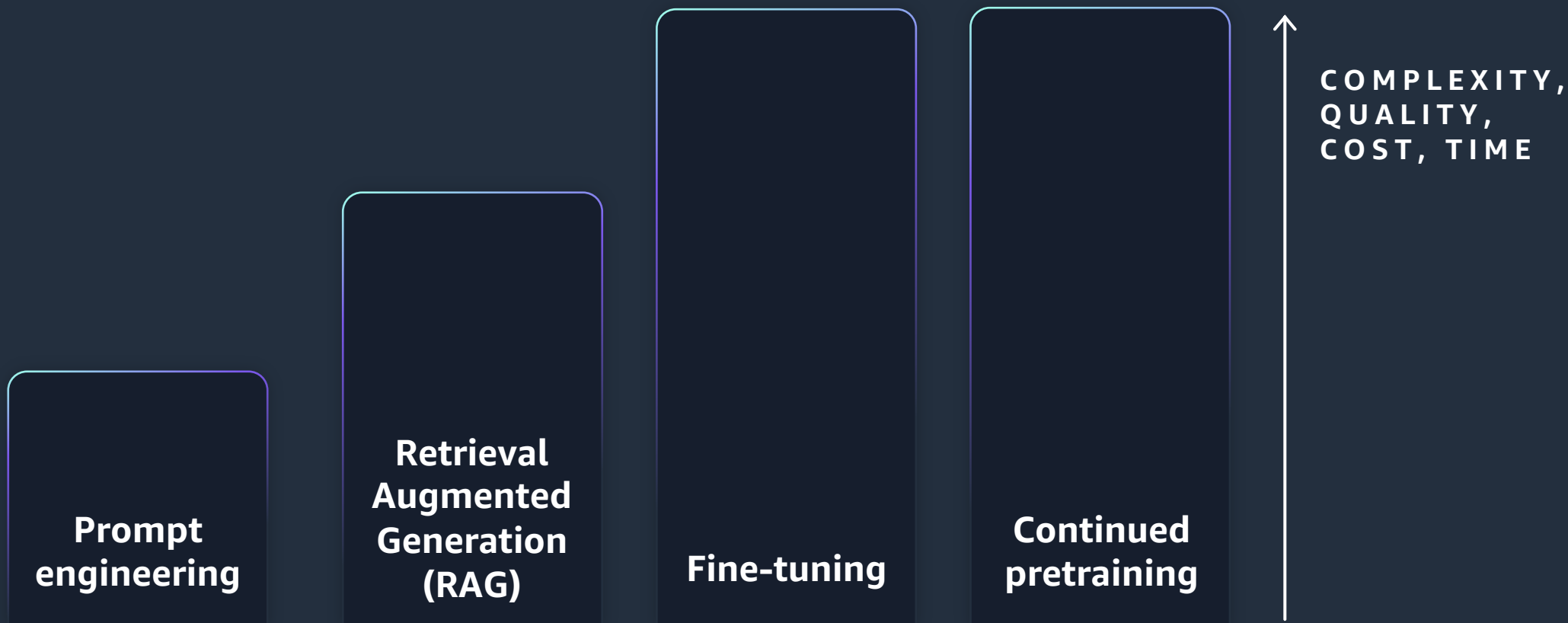
Understanding models

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

Common approaches for **customizing FMs**

Prompt engineering

Retrieval Augmented Generation (RAG)

Fine-tuning

Continued pretraining

COMPLEXITY, QUALITY, COST, TIME

What are vector embeddings?

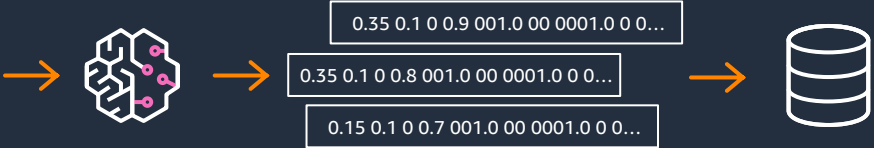| Source domain-specific data	Tokenization	Vectorization	Store in vector data store	Perform semantic similarity search	Include semantically similar context in prompt

Documents

Audio/video

Images

Semantic elements:
- Words, phrases
- Paragraphs, documents
- Scenes, song sections
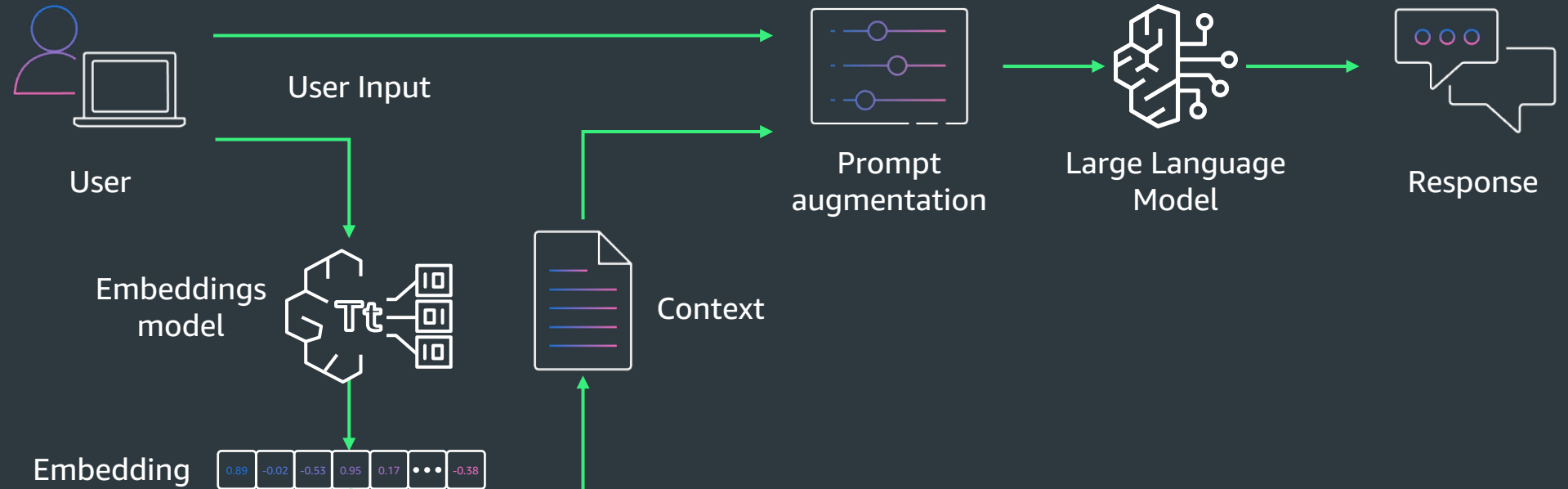- Faces, detected picture elements
- And more

0.35 0.1 0 0.9 001.0 00 0001.0 0 0...

0.35 0.1 0 0.8 001.0 00 0001.0 0 0...

0.15 0.1 0 0.7 001.0 00 0001.0 0 0...
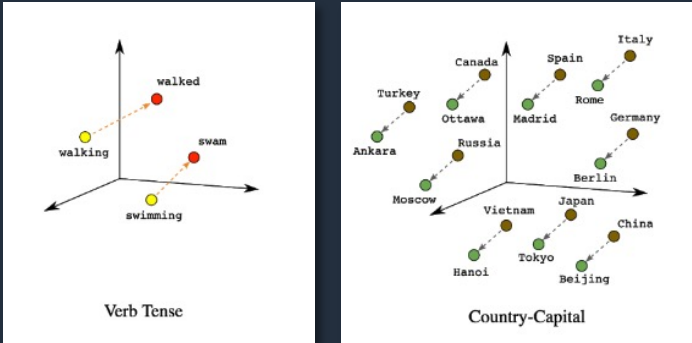
3D simplified representation. Embeddings can have thousands of dimensions. Source: https://daleonai.com/embeddings-explained

Embeddings: When vector elements are semantic, used in generative AI

RAG in Action

Text Generation Workflow

User

User Input

Embeddings model

Embedding | 0.89 | -0.02 | -0.53 | 0.95 | 0.17 | ••• | -0.38 |

Context

Prompt augmentation

Large Language Model

Response

Data Ingestion Workflow

Semantic search

Vector store

Embeddings model

Document chunks
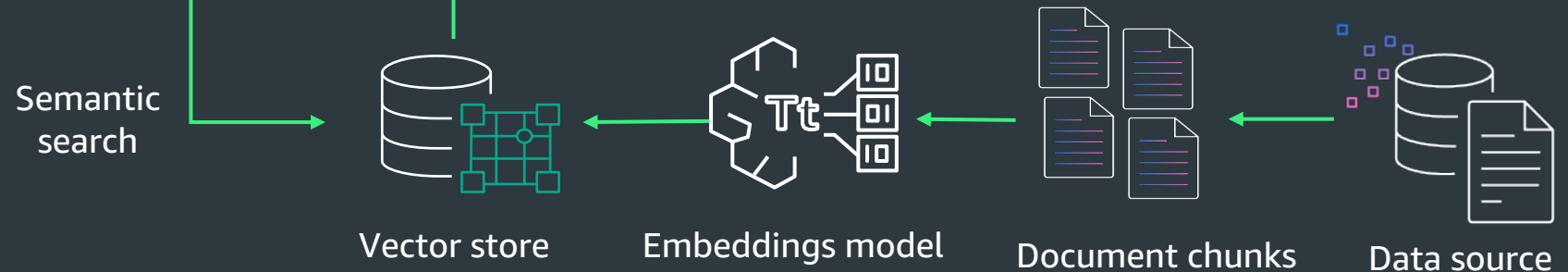
Data source

Amazon Bedrock Knowledge Bases

NATIVE SUPPORT FOR RAG

USER QUERY → 1 → AMAZON BEDROCK → 4 → AUGMENTED PROMPT → 5 → MODEL → 6 → ANSWER

2, 3 — AMAZON BEDROCK KNOWLEDGE BASES

MODEL:
- AI21 Labs—Jurassic2
- Amazon Titan Text
- Anthropic—Claude
- Meta—Llama
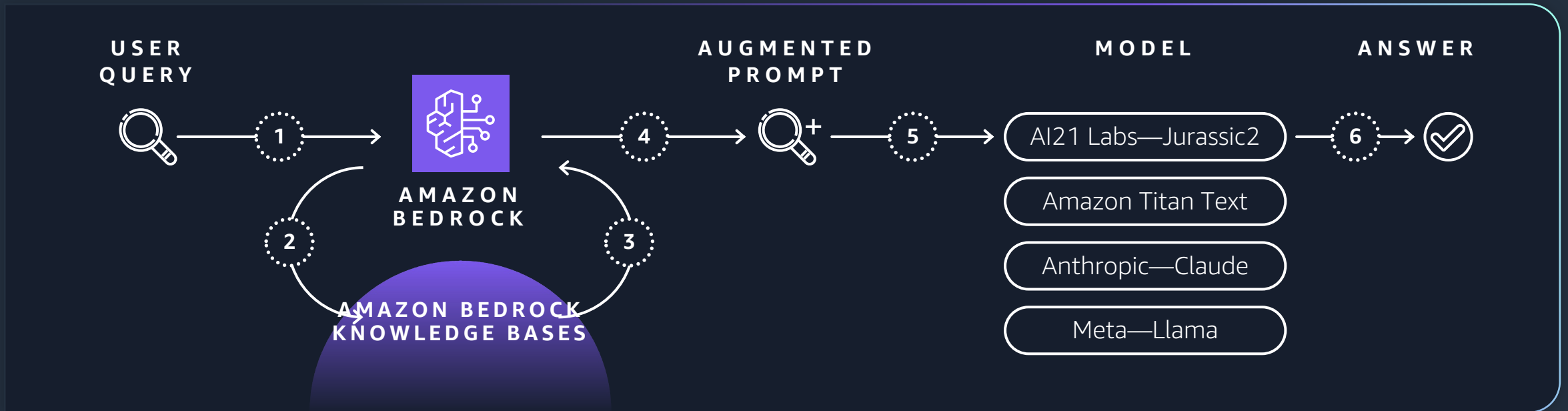
Securely connect FMs to data sources for RAG to deliver more relevant responses

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 Bedrock Model Evaluation

EVALUATE, COMPARE, AND SELECT THE BEST FM FOR YOUR USE CASE
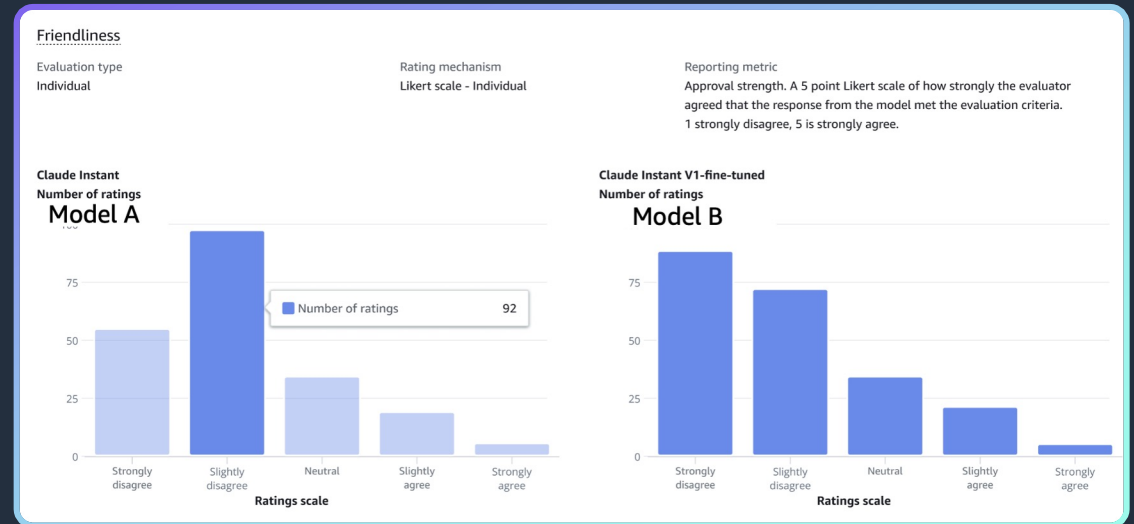
Automatic or human evaluation method

Curated datasets or bring your own

Predefined and custom metrics

Human-like evaluation quality with LLM-as-a-judge

HUMAN EVALUATION REPORT

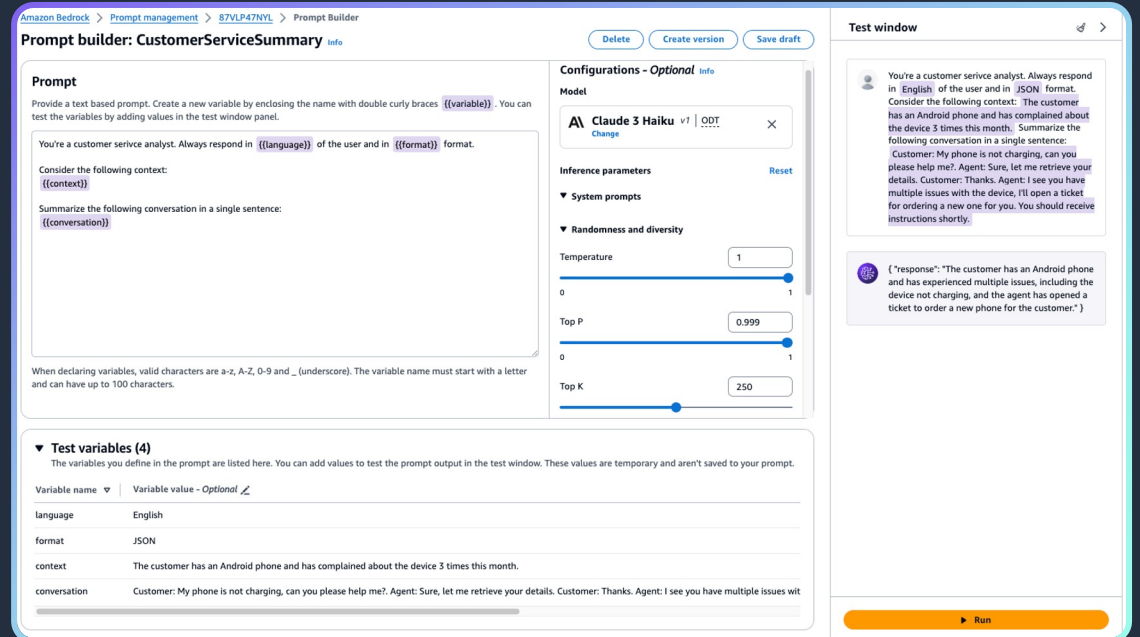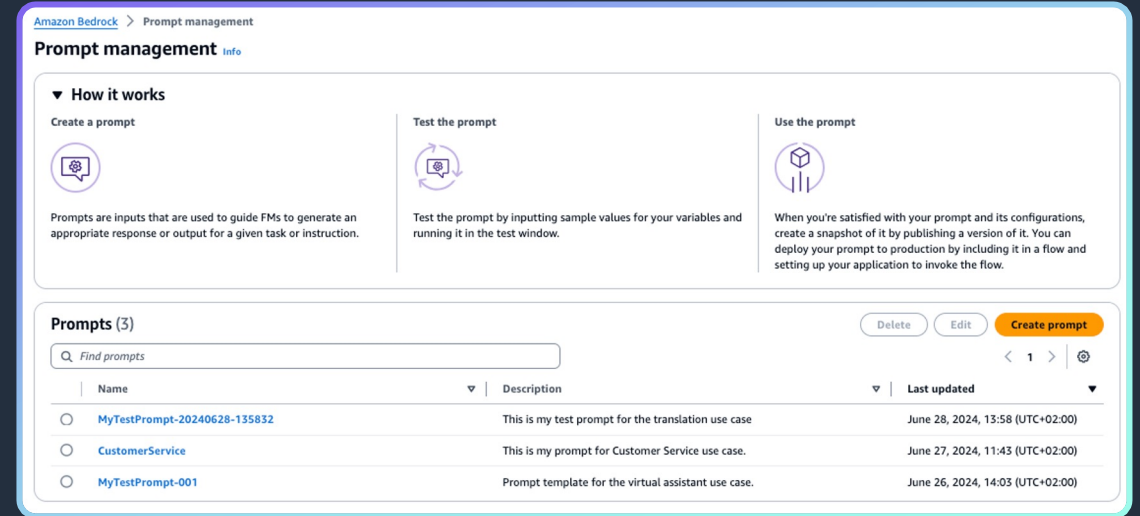## AUTOMATIC EVALUATION REPORT

Amazon Bedrock Prompt Management

STREAMLINE YOUR GENERATIVE AI APPLICATION DEVELOPMENT

Built-in Prompt Builder in Amazon Bedrock console and SDK APIs

Prompt Library for easy cataloguing and management

Seamless integration with Flows

Amazon Bedrock Guardrails

Apply safeguards customized to your gen AI application requirements and responsible AI policies

Configure thresholds to filter undesirable and potentially harmful text and image content, jailbreaks, and prompt attacks

Identify, correct, and explain factual claims in responses using Automated Reasoning
Define and disallow denied topics with short natural language descriptions

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

Define a set of words to detect and block in user inputs and model responses

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

Amazon Bedrock
Data Automation

Extract, transform and generate structured data from multi-modal content

Generate customized output based on your application's requirements and business rules

Streamline application workflows with a fully managed, single API experience

Amazon Bedrock Intelligent Prompt Routing

Automatically route prompts to different foundation models to optimize response quality and lower costs

Provides a single endpoint to efficiently route prompts

Uses advanced prompt matching techniques to meet cost and latency thresholds

Reduce costs by up to 30%

Amazon Bedrock support for prompt caching

Cache repetitive context in prompts across multiple API calls

Securely cache entire prompts

Enhance accuracy through longer, detailed prompts

Reduce costs by up to 90% and latency by up to 85% for supported models

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
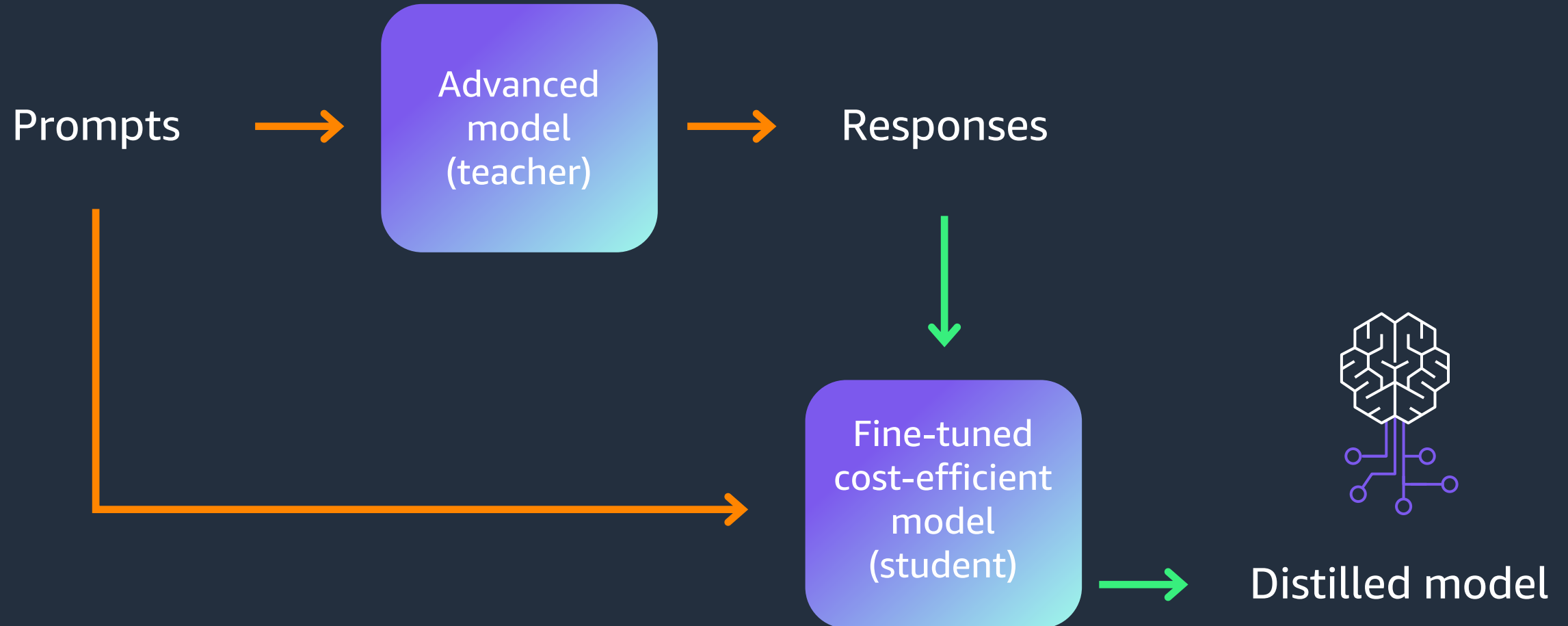
Execute action or search knowledge base

Observe results

Think about next step

Until final answer

Match the performance of advanced models with cost-efficient models for your use case with Model Distillation

Prompts → **Advanced model (teacher)** → Responses

Fine-tuned cost-efficient model (student) → Distilled model
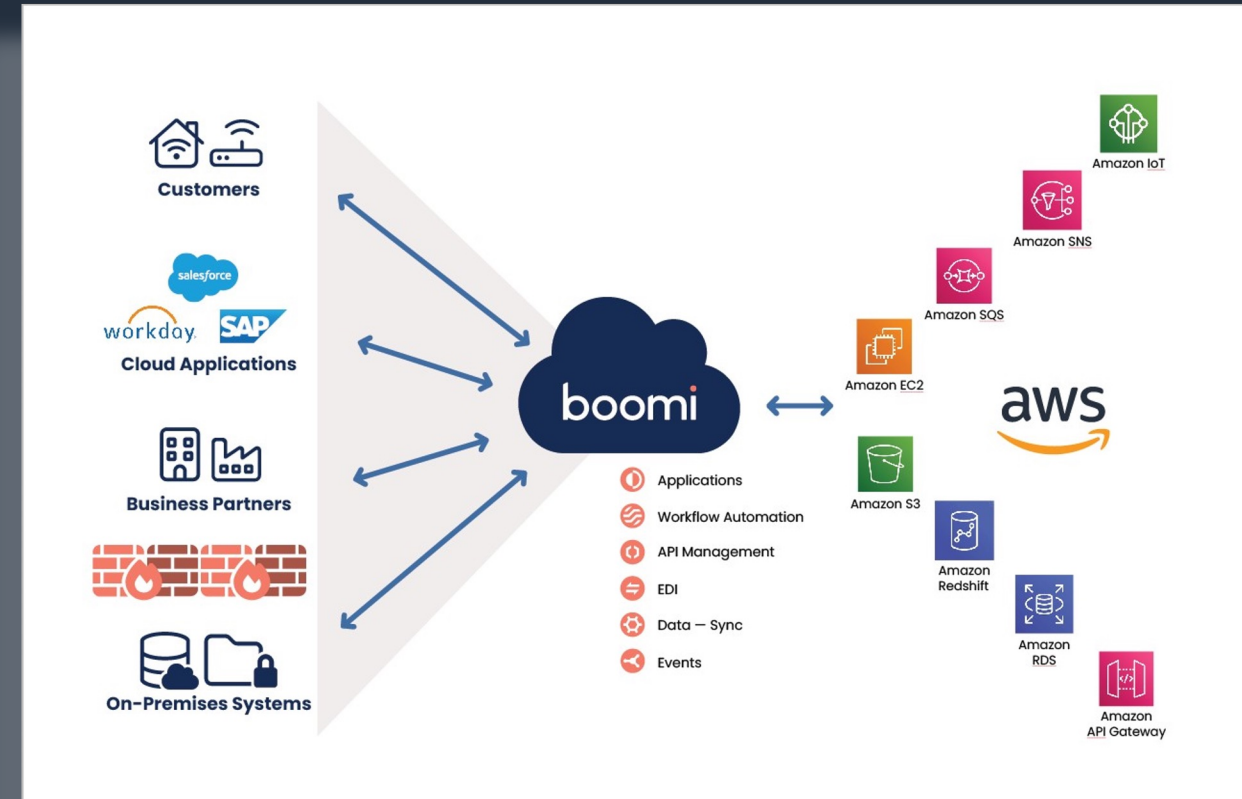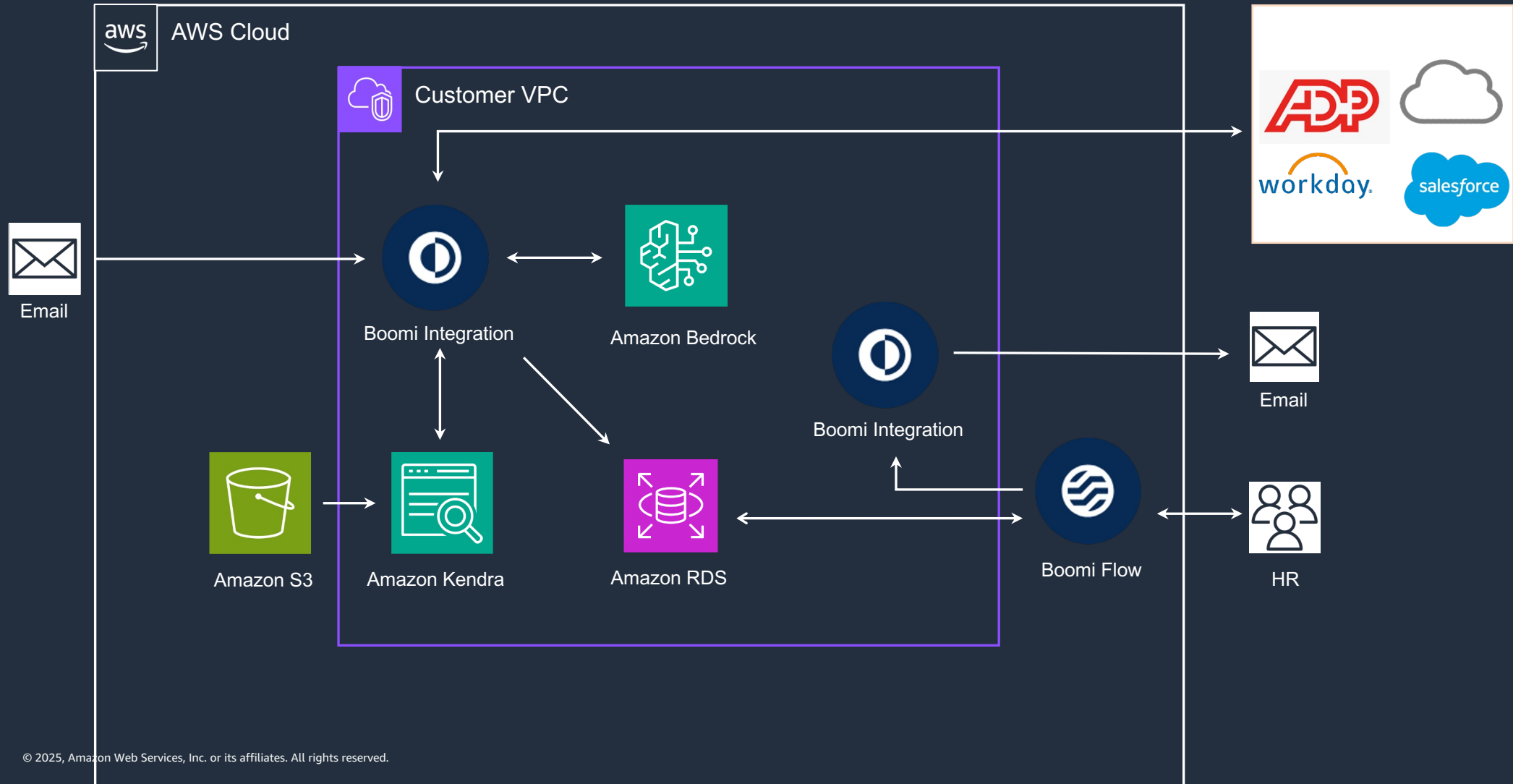
AWS – Boomi Partnership

Boomi: Intelligent Integration & Automation

- Partners since 2016

- Boomi iPaaS runs exclusively on AWS

- Boomi in AWS Marketplace since 2021. Currently offering three solutions and a Professional Services offering

- Boomi is an ISV Accelerate Partner

Demo Solution: Intelligent Email Processing

BOOMI + AWS RAG ACTION PIPELINE

AWS Cloud

Customer VPC

Email

Boomi Integration

Amazon Bedrock

Boomi Integration

Amazon S3

Amazon Kendra

Amazon RDS

Boomi Flow

Email

HR

Thank you!

Venkatesh Aravamudan
Sr Partner Solutions Architect

Amazon Web Services