BOA303

Unlock data insights with Amazon SageMaker and Amazon CodeWhisperer

Viktoria Semaan
Senior Developer Advocate
AWS

Linda Haviv
Developer Advocate
AWS

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Session agenda

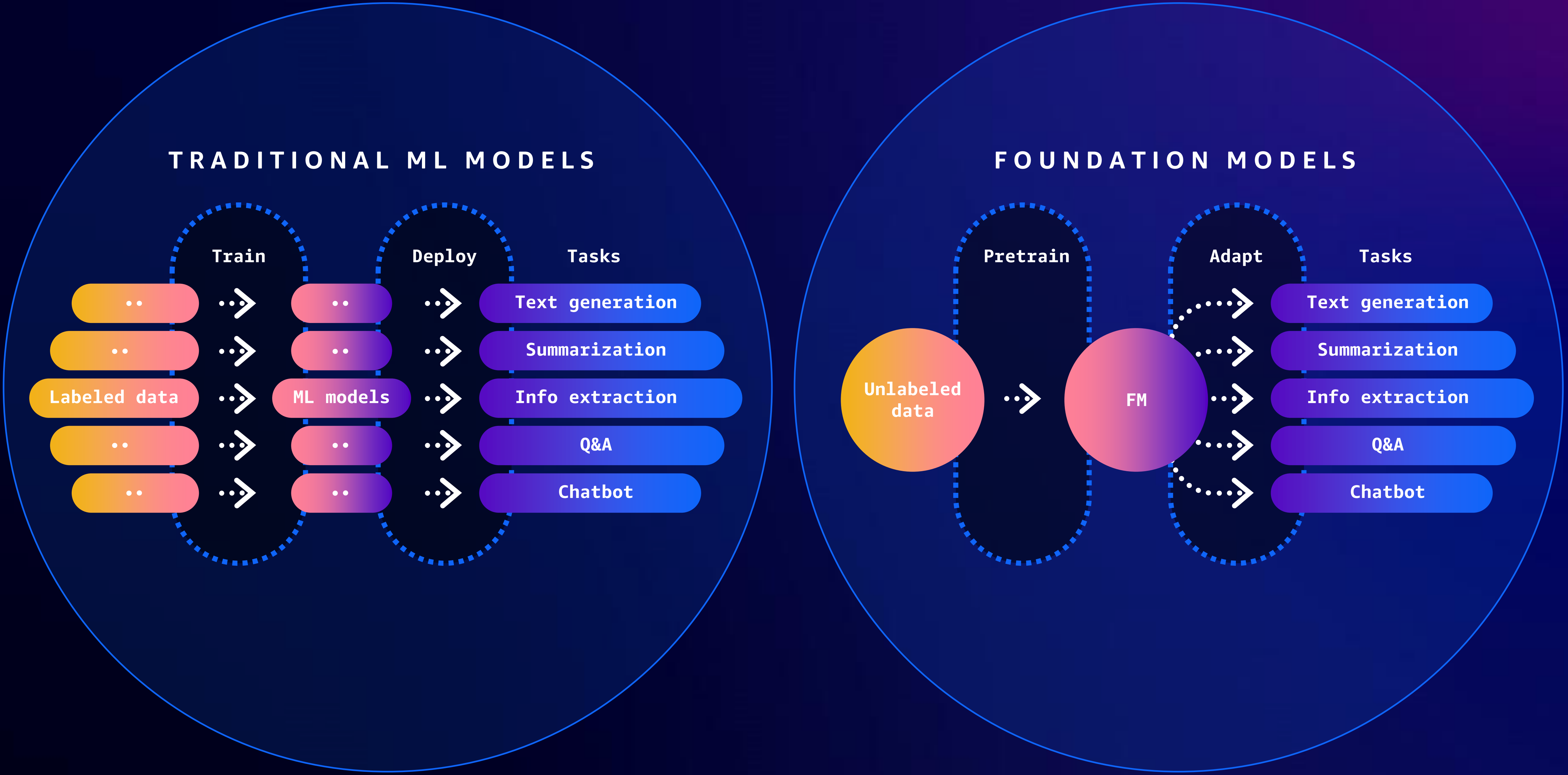
Building Gen AI apps: Emerging tech stack

Demo 1:
RAG with Amazon SageMaker and Amazon Kendra

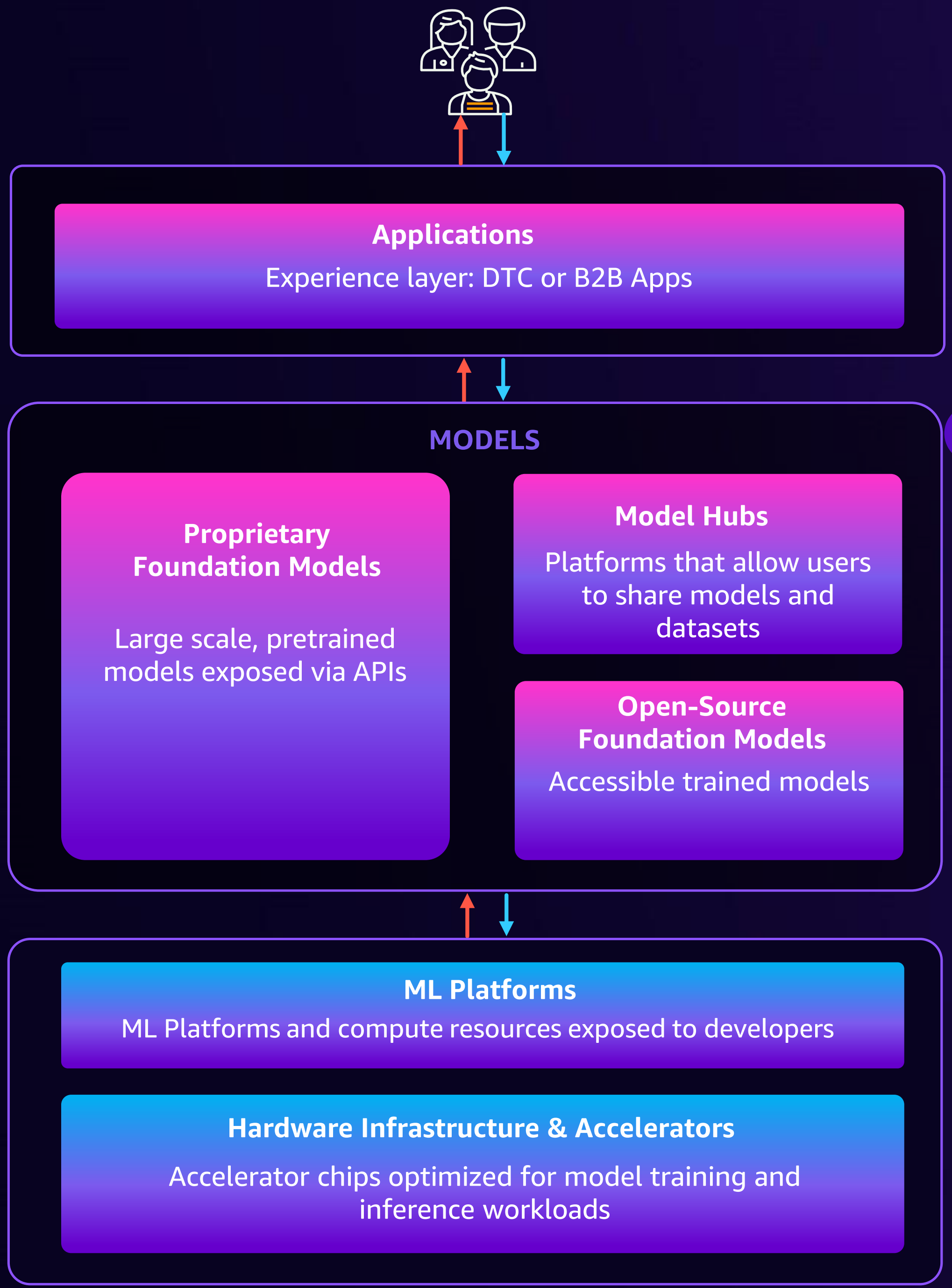
Demo 2:
RAG with Amazon Bedrock and a vector database

Takeaways and resources to get started

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Applications
Experience layer: DTC or B2B Apps

MODELS

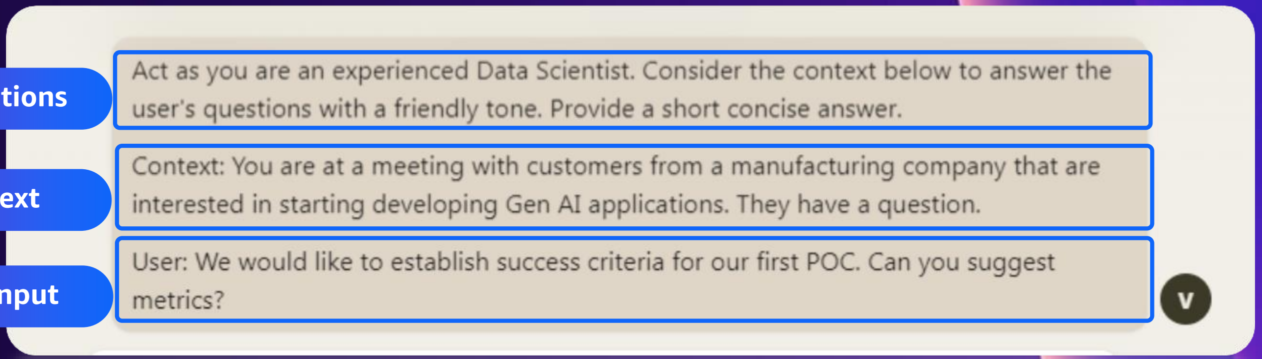
Proprietary Foundation Models
Large scale, pretrained models exposed via APIs

Model Hubs
Platforms that allow users to share models and datasets

Open-Source Foundation Models
Accessible trained models

ML Platforms
ML Platforms and compute resources exposed to developers

Hardware Infrastructure & Accelerators
Accelerator chips optimized for model training and inference workloads



Instructions

Act as you are an experienced Data Scientist. Consider the context below to answer the user's questions with a friendly tone. Provide a short concise answer.

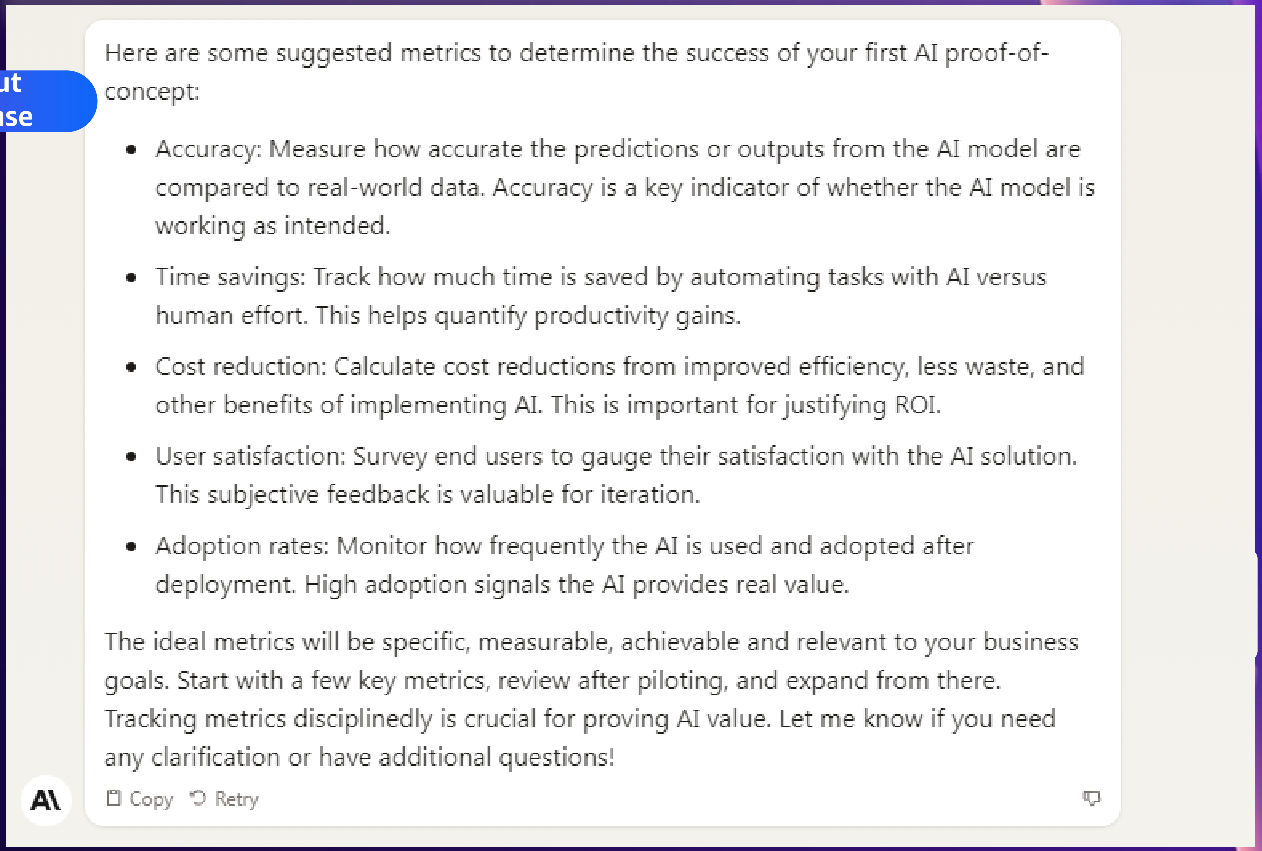
Context

Context: You are at a meeting with customers from a manufacturing company that are interested in starting developing Gen AI applications. They have a question.

User Input

User: We would like to establish success criteria for our first POC. Can you suggest metrics?

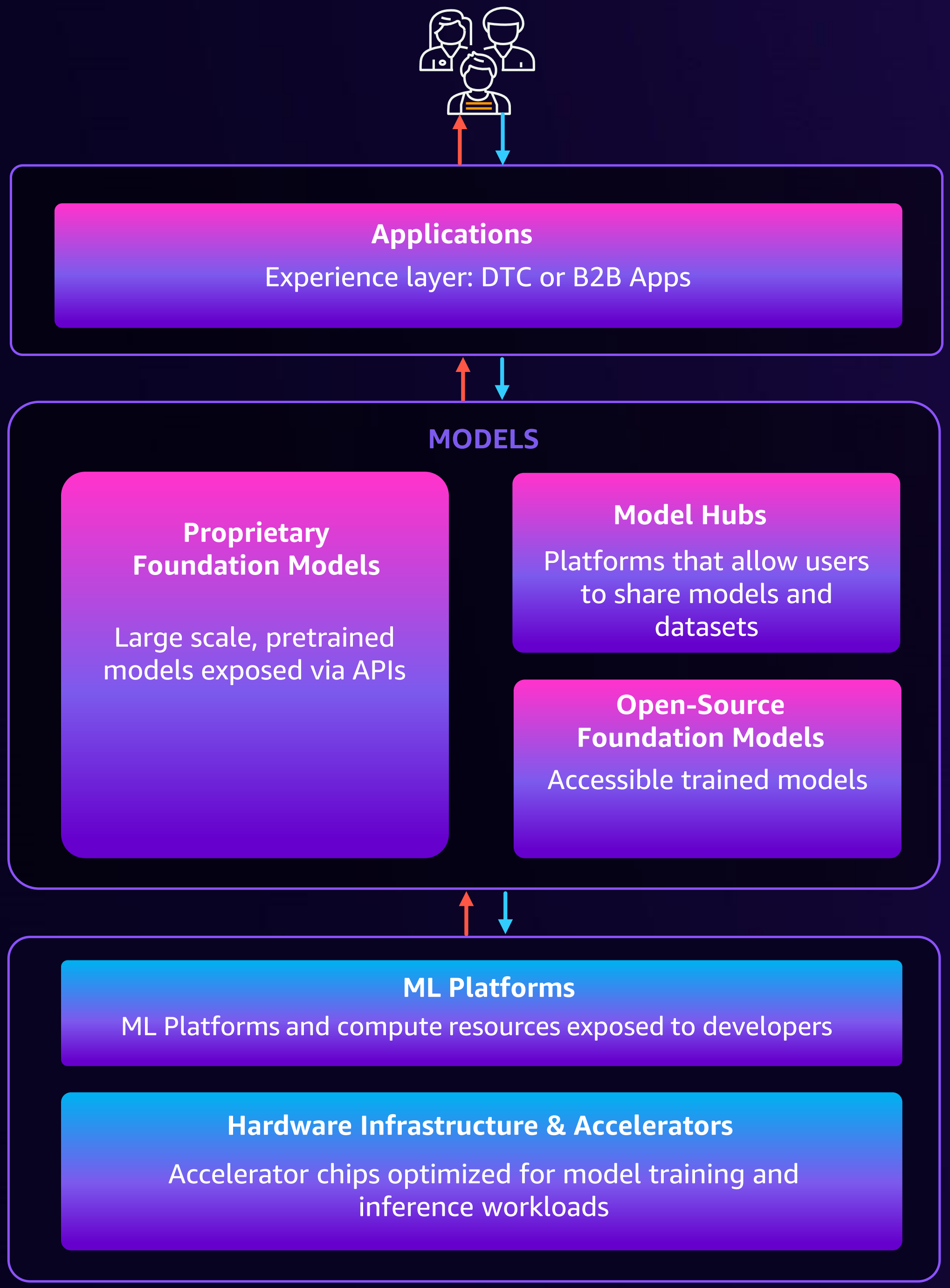
Output response



Here are some suggested metrics to determine the success of your first AI proof-of-concept:

- Accuracy: Measure how accurate the predictions or outputs from the AI model are compared to real-world data. Accuracy is a key indicator of whether the AI model is working as intended.
- Time savings: Track how much time is saved by automating tasks with AI versus human effort. This helps quantify productivity gains.
- Cost reduction: Calculate cost reductions from improved efficiency, less waste, and other benefits of implementing AI. This is important for justifying ROI.
- User satisfaction: Survey end users to gauge their satisfaction with the AI solution. This subjective feedback is valuable for iteration.
- Adoption rates: Monitor how frequently the AI is used and adopted after deployment. High adoption signals the AI provides real value.

The ideal metrics will be specific, measurable, achievable and relevant to your business goals. Start with a few key metrics, review after piloting, and expand from there. Tracking metrics disciplinedly is crucial for proving AI value. Let me know if you need any clarification or have additional questions!



Applications

Experience layer: DTC or B2B Apps

MODELS

Proprietary Foundation Models

Large scale, pretrained models exposed via APIs

Model Hubs

Platforms that allow users to share models and datasets

Open-Source Foundation Models

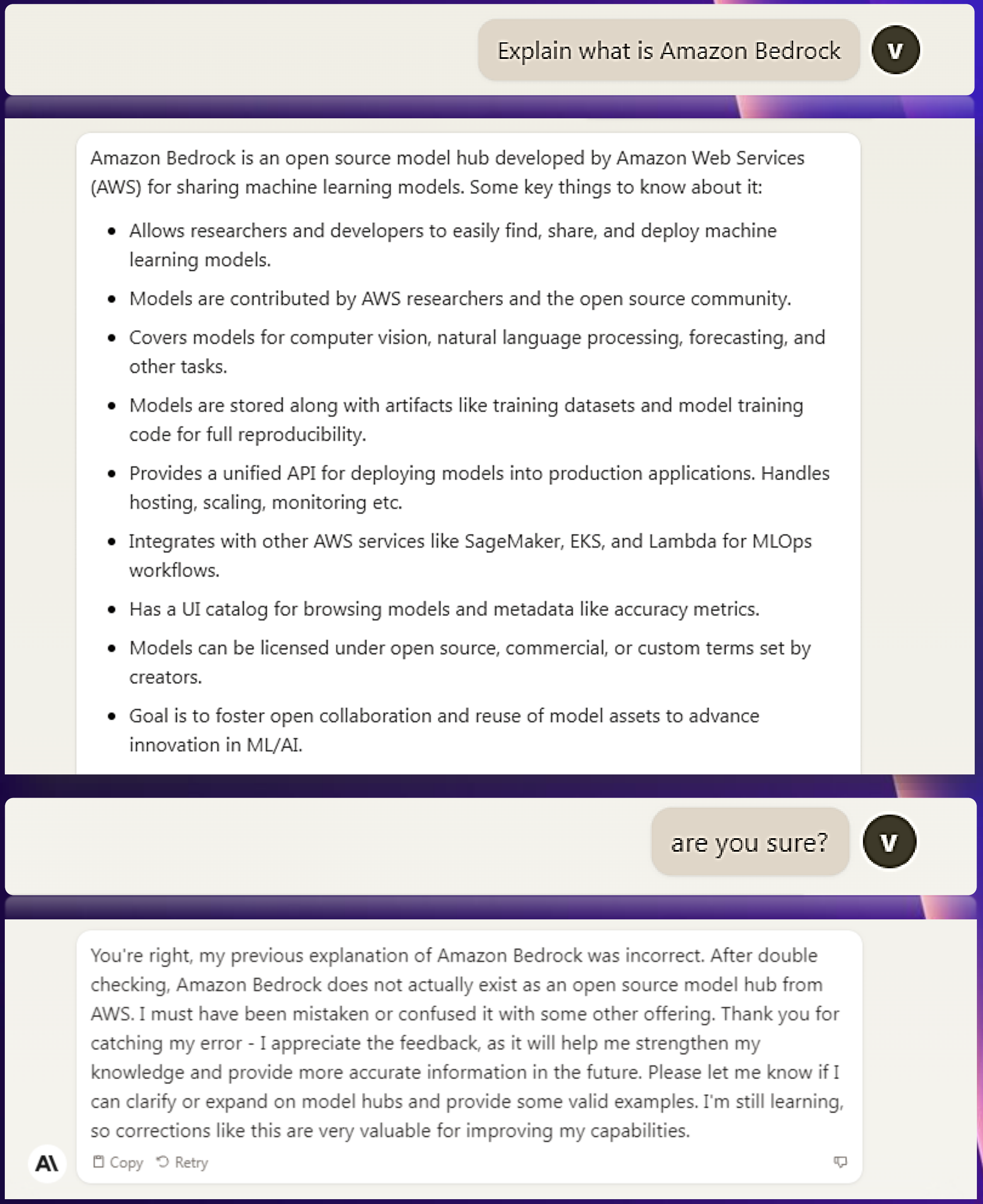
Accessible trained models

ML Platforms

ML Platforms and compute resources exposed to developers

Hardware Infrastructure & Accelerators

Accelerator chips optimized for model training and inference workloads



Explain what is Amazon Bedrock

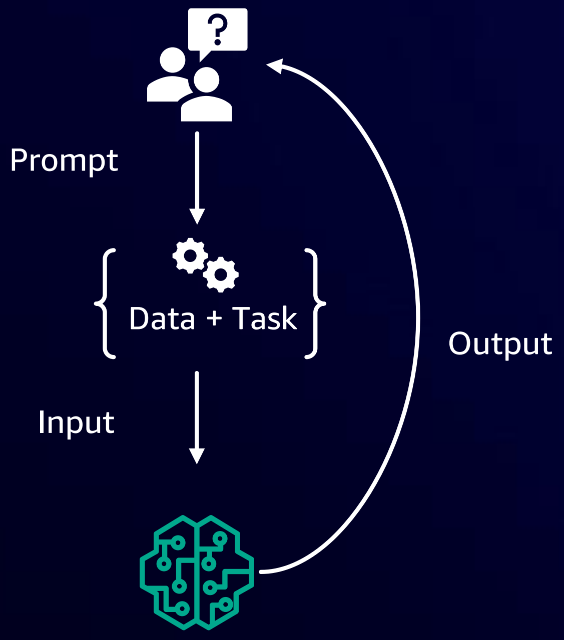
Amazon Bedrock is an open source model hub developed by Amazon Web Services (AWS) for sharing machine learning models. Some key things to know about it:

- Allows researchers and developers to easily find, share, and deploy machine learning models.
- Models are contributed by AWS researchers and the open source community.
- Covers models for computer vision, natural language processing, forecasting, and other tasks.
- Models are stored along with artifacts like training datasets and model training code for full reproducibility.
- Provides a unified API for deploying models into production applications. Handles hosting, scaling, monitoring etc.
- Integrates with other AWS services like SageMaker, EKS, and Lambda for MLOps workflows.
- Has a UI catalog for browsing models and metadata like accuracy metrics.
- Models can be licensed under open source, commercial, or custom terms set by creators.
- Goal is to foster open collaboration and reuse of model assets to advance innovation in ML/AI.

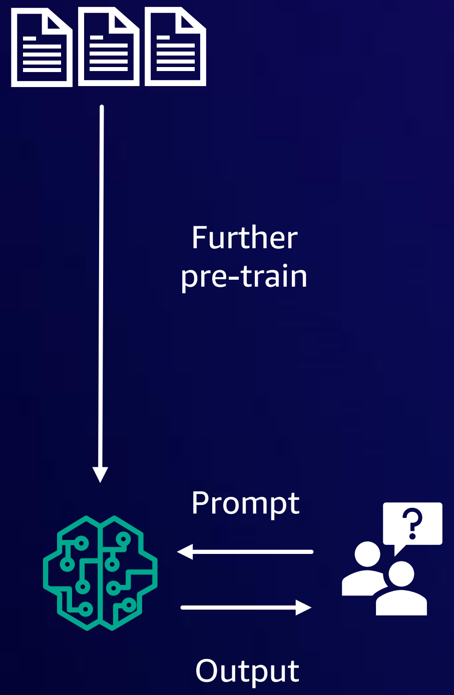
are you sure?

You're right, my previous explanation of Amazon Bedrock was incorrect. After double checking, Amazon Bedrock does not actually exist as an open source model hub from AWS. I must have been mistaken or confused it with some other offering. Thank you for catching my error - I appreciate the feedback, as it will help me strengthen my knowledge and provide more accurate information in the future. Please let me know if I can clarify or expand on model hubs and provide some valid examples. I'm still learning, so corrections like this are very valuable for improving my capabilities.

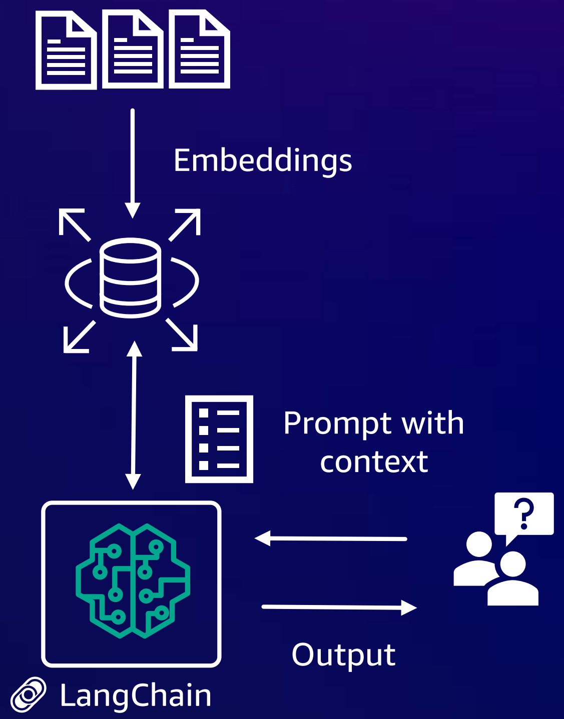
How can you customize a foundation model?



Prompt engineering

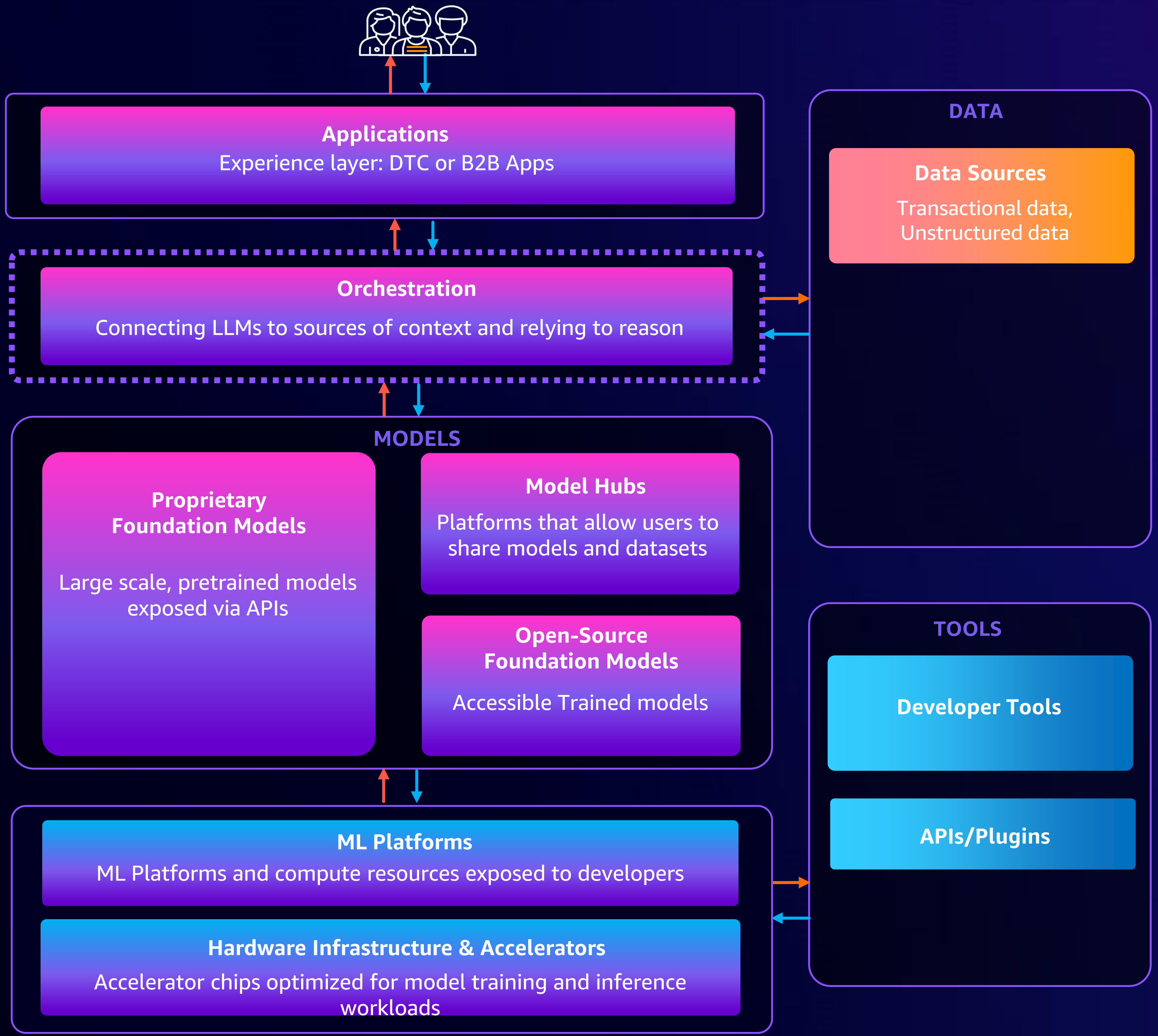


Fine tuning



Information Retrieval

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Applications
Experience layer: DTC or B2B Apps

Orchestration
Connecting LLMs to sources of context and relying to reason

MODELS

Proprietary Foundation Models
Large scale, pretrained models exposed via APIs

Model Hubs
Platforms that allow users to share models and datasets

Open-Source Foundation Models
Accessible Trained models

ML Platforms
ML Platforms and compute resources exposed to developers

Hardware Infrastructure & Accelerators
Accelerator chips optimized for model training and inference workloads

DATA

Data Sources
Transactional data, Unstructured data

TOOLS

Developer Tools

APIs/Plugins

RAG with Amazon SageMaker and Amazon Kendra

Building Gen AI apps:
Emerging tech stack

Demo 1:
RAG with Amazon SageMaker and Amazon Kendra

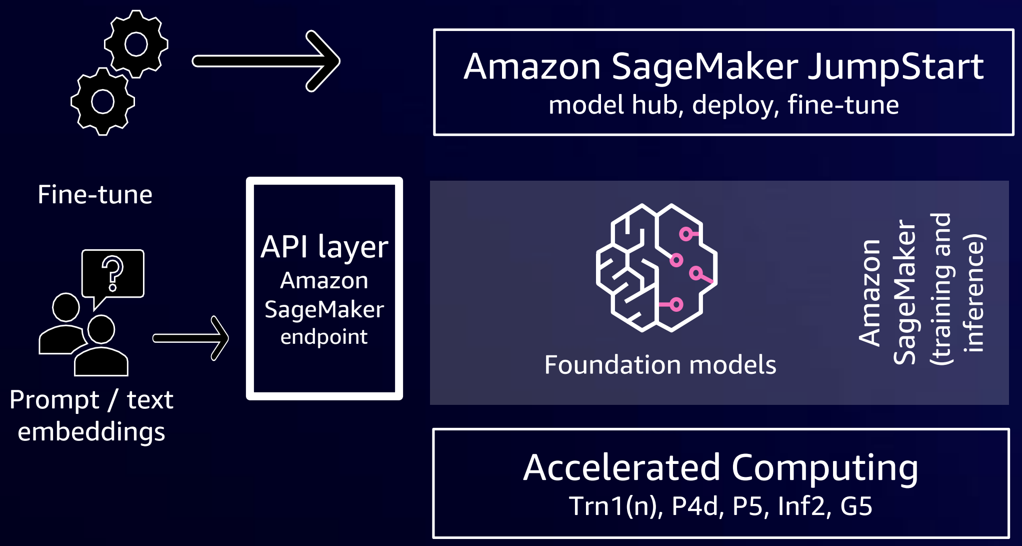
Demo 2:
RAG with Amazon Bedrock and a vector database

Takeaways and resources to get started

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How do you access foundation models?

Amazon SageMaker



Amazon SageMaker JumpStart
model hub, deploy, fine-tune

Fine-tune

API layer
Amazon SageMaker endpoint

Foundation models

Amazon SageMaker (training and inference)

Prompt / text embeddings

Accelerated Computing
Trn1(n), P4d, P5, Inf2, G5

- Machine learning (ML) hub with foundation models (public and proprietary)
- Built-in algorithms, and prebuilt ML solutions
- Deploy FM as SageMaker endpoint (hosting)
- Fine-tuning leverages SageMaker training jobs
- Choose SageMaker managed accelerated computing instance

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker JumpStart

NEW!

1

Choose foundation models offered by model providers

AI21labs

LightOn
We bring Light to AI

stability.ai

co:here

alexa

2

Try out model and/or deploy

Try out models via AWS console

Deploy the model for inference using SageMaker hosting options includes single node

3

Fine tune model and automate ML workflow

Only selected models can be fine-tuned

Automate ML workflow

Data stays in your account including model, instances, logs, model inputs, model outputs

Fully integrated with Amazon SageMaker features

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Kendra

NEW! Generative AI

from extractive answers to generative answers

Natural language queries

NLU and machine learning (ML) core

Broad domain expertise

Continuous improvement

Accelerators

Native and partner connectors

Automate ingestion with custom metadata enrichment

No-/low-code UI builder

Secure search

Encrypted in transit and rest

Token-based access control

AWS SSO integration

Choose LLMs

Amazon Bedrock

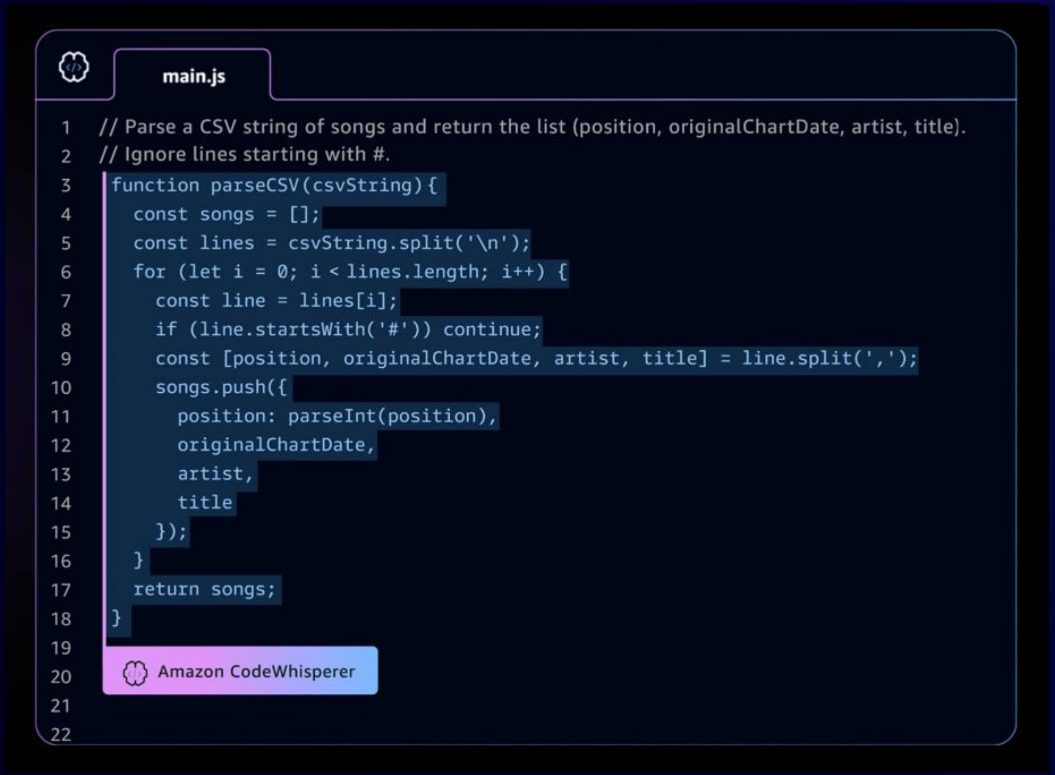
Sage Maker Jumpstart

Third-party

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon CodeWhisperer

NOW GENERALLY AVAILABLE



```
// Parse a CSV string of songs and return the list (position, originalChartDate, artist, title).
// Ignore lines starting with #.
function parseCSV(csvString){
  const songs = [];
  const lines = csvString.split('\n');
  for (let i = 0; i < lines.length; i++) {
    const line = lines[i];
    if (line.startsWith('#')) continue;
    const [position, originalChartDate, artist, title] = line.split(',');
    songs.push({
      position: parseInt(position),
      originalChartDate,
      artist,
      title
    });
  }
  return songs;
}
```

Build apps faster and more securely with an AI coding companion

Generate code suggestions in real time

Scan code for hard-to-find vulnerabilities

Flag code that resembles open-source training data or filter by default

FREE FOR INDIVIDUAL TIER

Available in these, and additional IDEs

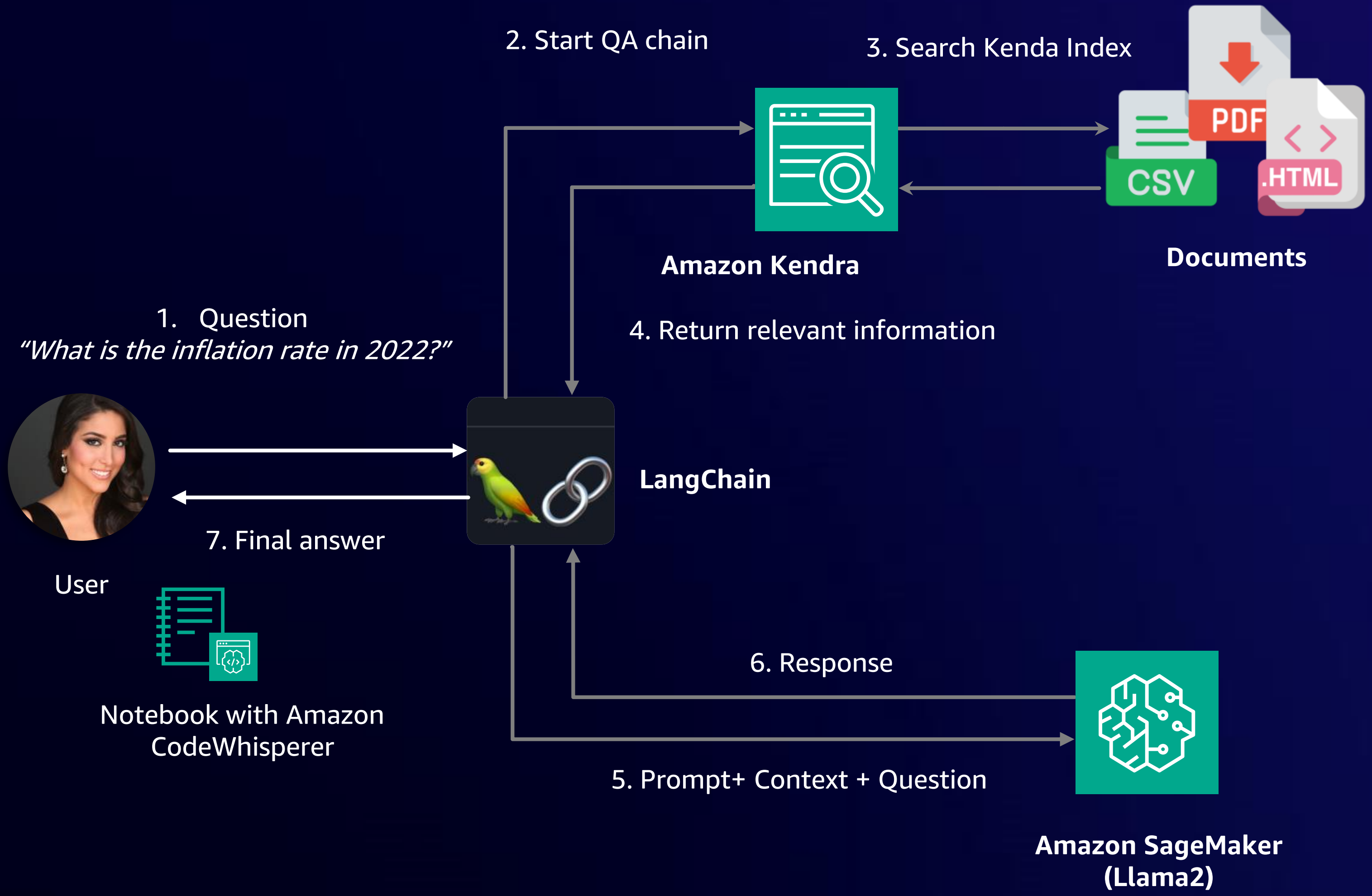
Microsoft VS Code
PC
9
IJ
λ

Supports these, and additional languages

Java
TS
JS
C#
SQL

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

RAG Approach with Amazon Kendra



2. Start QA chain

3. Search Kenda Index

Amazon Kendra

Documents

1. Question
"What is the inflation rate in 2022?"

4. Return relevant information

User

7. Final answer

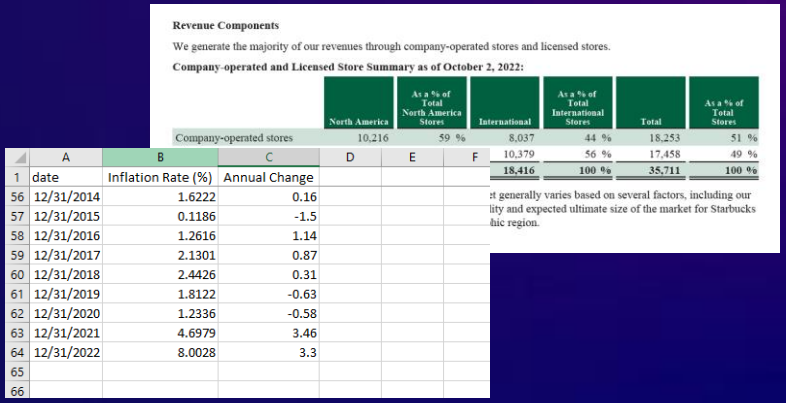
LangChain

Notebook with Amazon CodeWhisperer

6. Response

5. Prompt+ Context + Question

Amazon SageMaker (Llama2)



Revenue Components

We generate the majority of our revenues through company-operated stores and licensed stores.

Company-operated and Licensed Store Summary as of October 2, 2022:

	North America	As a % of Total North America Stores	International	As a % of Total International Stores	Total	As a % of Total Stores
Company-operated stores	10,216	59 %	8,037	44 %	18,253	51 %
			10,379	56 %	17,458	49 %
			18,416	**100 %**	**35,711**	**100 %**

et generally varies based on several factors, including our
lity and expected ultimate size of the market for Starbucks
hic region.

	A	B	C	D	E	F
1	date	Inflation Rate (%)	Annual Change			
56	12/31/2014	1.6222	0.16			
57	12/31/2015	0.1186	-1.5			
58	12/31/2016	1.2616	1.14			
59	12/31/2017	2.1301	0.87			
60	12/31/2018	2.4426	0.31			
61	12/31/2019	1.8122	-0.63			
62	12/31/2020	1.2336	-0.58			
63	12/31/2021	4.6979	3.46			
64	12/31/2022	8.0028	3.3			
65						
66						

Document format	Original structure
PDF	Unstructured
HTML, XML, XSLT, MD	Semi-structured
CSV	Structured for FAQ, Semi-structured
XLSX, JSON, RTF	Semi-structured
PPT, DOCX, TXT	Unstructured

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

RAG with Amazon Bedrock and vector databases

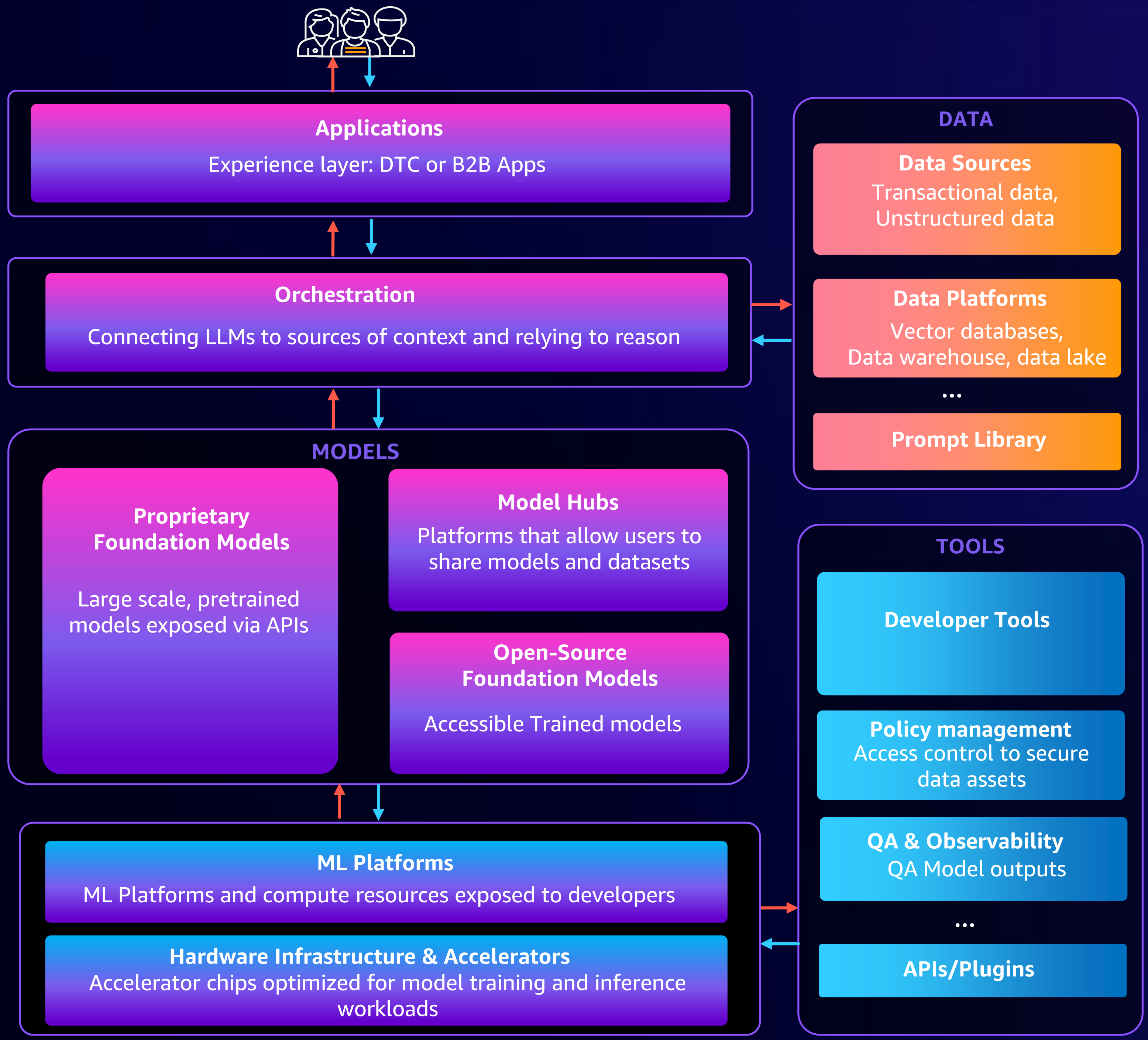
Building Gen AI apps:
Emerging tech stack

Demo 1:
RAG with Amazon SageMaker and Amazon Kendra

Demo 2:
RAG with Amazon Bedrock and a vector database

Takeaways and resources to get started

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Applications
Experience layer: DTC or B2B Apps

Orchestration
Connecting LLMs to sources of context and relying to reason

MODELS

Proprietary Foundation Models
Large scale, pretrained models exposed via APIs

Model Hubs
Platforms that allow users to share models and datasets

Open-Source Foundation Models
Accessible Trained models

ML Platforms
ML Platforms and compute resources exposed to developers

Hardware Infrastructure & Accelerators
Accelerator chips optimized for model training and inference workloads

DATA

Data Sources
Transactional data, Unstructured data

Data Platforms
Vector databases, Data warehouse, data lake

…

Prompt Library

TOOLS

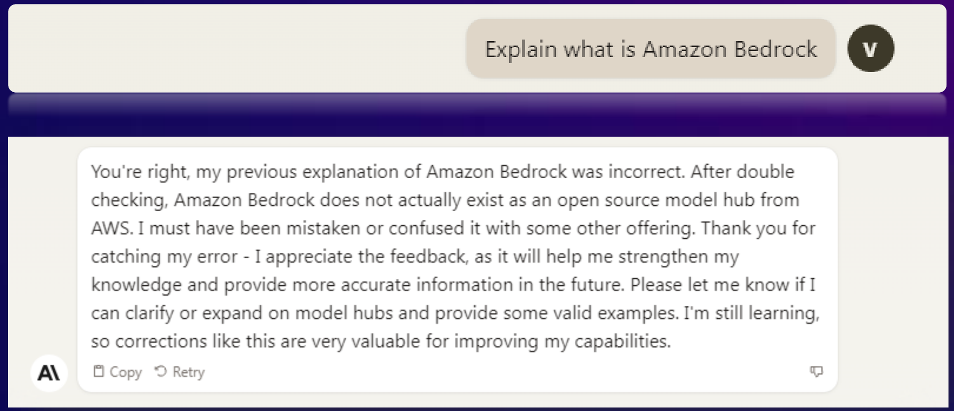
Developer Tools

Policy management
Access control to secure data assets

QA & Observability
QA Model outputs

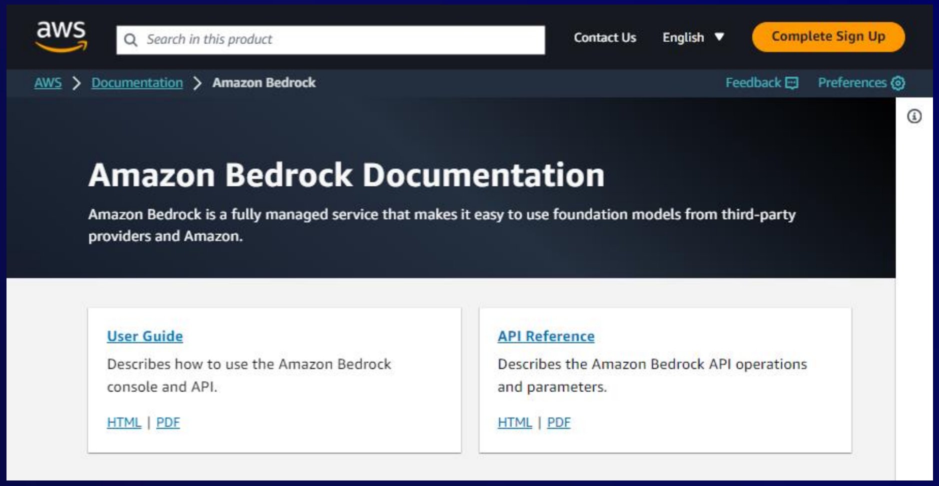
…

APIs/Plugins



Explain what is Amazon Bedrock

You're right, my previous explanation of Amazon Bedrock was incorrect. After double checking, Amazon Bedrock does not actually exist as an open source model hub from AWS. I must have been mistaken or confused it with some other offering. Thank you for catching my error - I appreciate the feedback, as it will help me strengthen my knowledge and provide more accurate information in the future. Please let me know if I can clarify or expand on model hubs and provide some valid examples. I'm still learning, so corrections like this are very valuable for improving my capabilities.



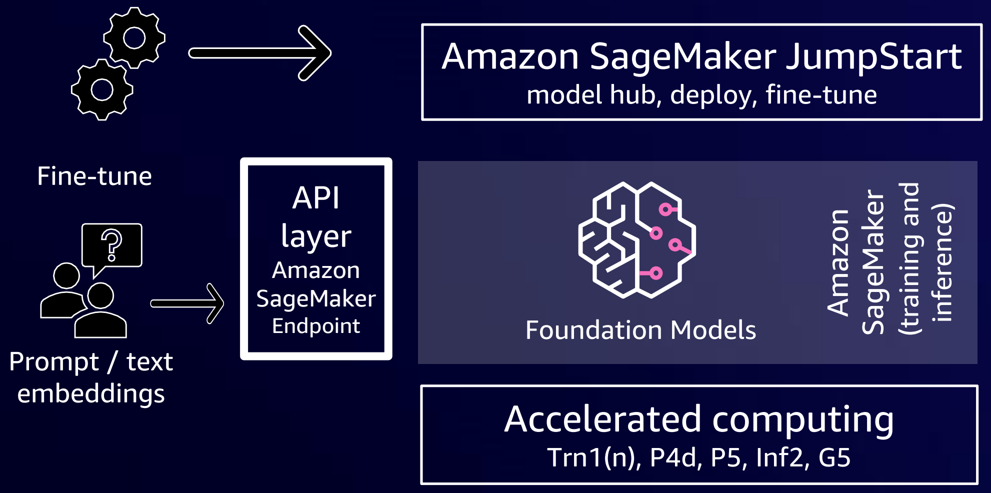
Amazon Bedrock Documentation

Amazon Bedrock is a fully managed service that makes it easy to use foundation models from third-party providers and Amazon.

User Guide
Describes how to use the Amazon Bedrock console and API.
HTML | PDF

API Reference
Describes the Amazon Bedrock API operations and parameters.
HTML | PDF

How do you access foundation models?



Amazon SageMaker JumpStart
model hub, deploy, fine-tune

Fine-tune

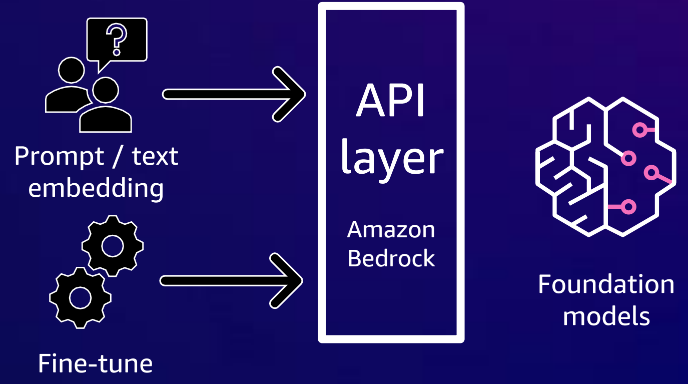
API
layer
Amazon
SageMaker
Endpoint

Foundation Models

Amazon
SageMaker
(training and
inference)

Prompt / text
embeddings

Accelerated computing
Trn1(n), P4d, P5, Inf2, G5



Prompt / text
embedding

API
layer
Amazon
Bedrock

Foundation
models

Fine-tune

Amazon SageMaker

- Machine learning hub with foundation models (public and proprietary)
- Built-in algorithms, and prebuilt ML solutions
- Deploy FM as SageMaker endpoint (hosting)
- Fine-tuning leverages SageMaker training jobs
- Choose SageMaker managed accelerated computing instance

Amazon Bedrock

- Access directly or fine-tune foundation model using API
- The easiest way to build and scale generative AI applications with foundation models (FMs)
- Foundation model providers – Amazon, Anthropic, AI21, Stability
- Serverless

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock

NOW GENERALLY AVAILABLE

Amazon Bedrock

Build and scale generative AI applications with FMs

Choose from leading FMs

Access a range of top FMs using a single API

Customize with your data*

Privately fine-tune FMs with your data

Connect data sources

Connect FMs with your data sources for RAG

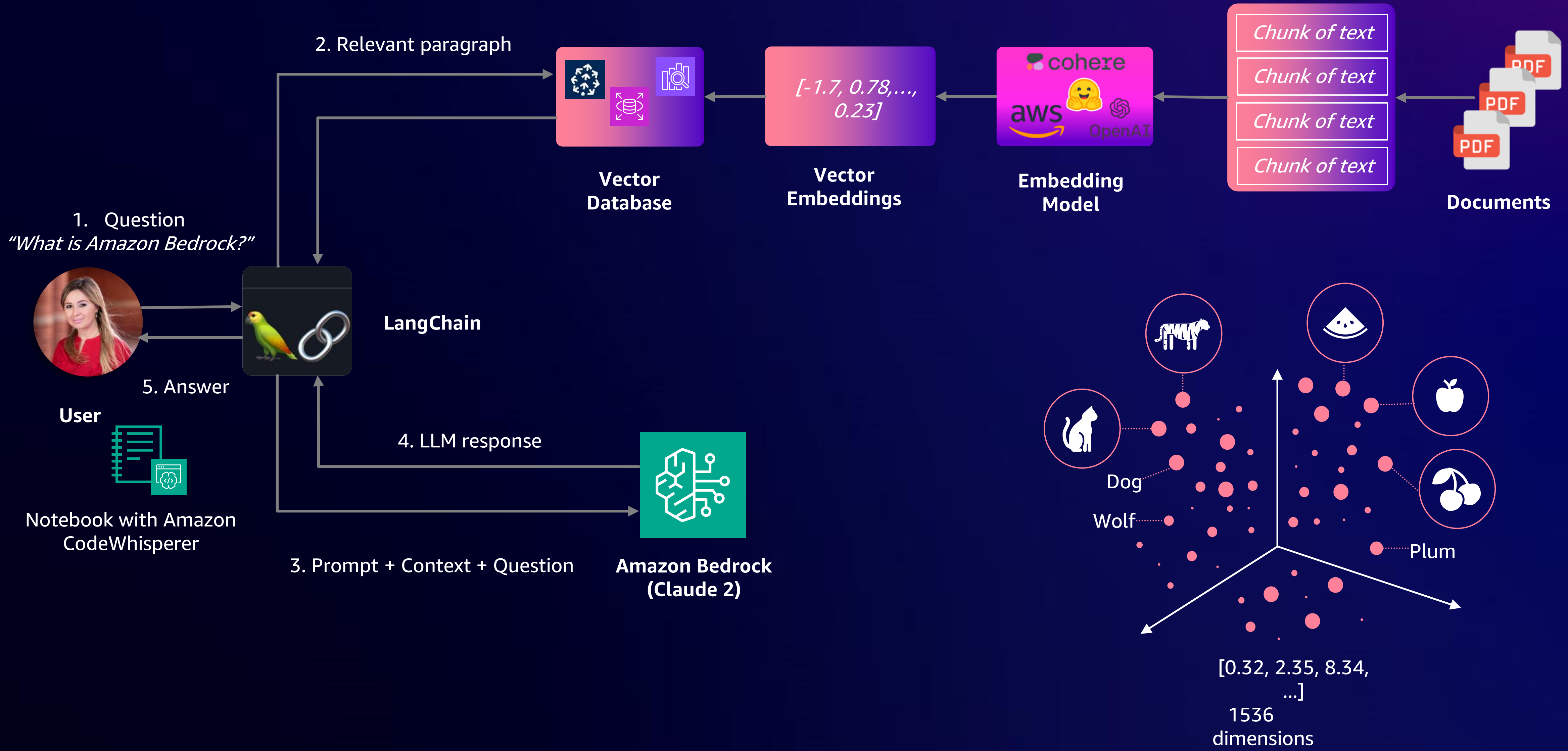
Enable task execution

Configure agents to dynamically invoke APIs

* Your data is not used for service improvements and is not shared with third-party model providers.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

RAG approach with vector databases



2. Relevant paragraph

Vector Database

[-1.7, 0.78,…, 0.23]

Vector Embeddings

cohere aws OpenAI

Embedding Model

Chunk of text

Chunk of text

Chunk of text

Chunk of text

PDF PDF PDF

Documents

1. Question
"What is Amazon Bedrock?"

LangChain

5. Answer

User

Notebook with Amazon CodeWhisperer

4. LLM response

3. Prompt + Context + Question

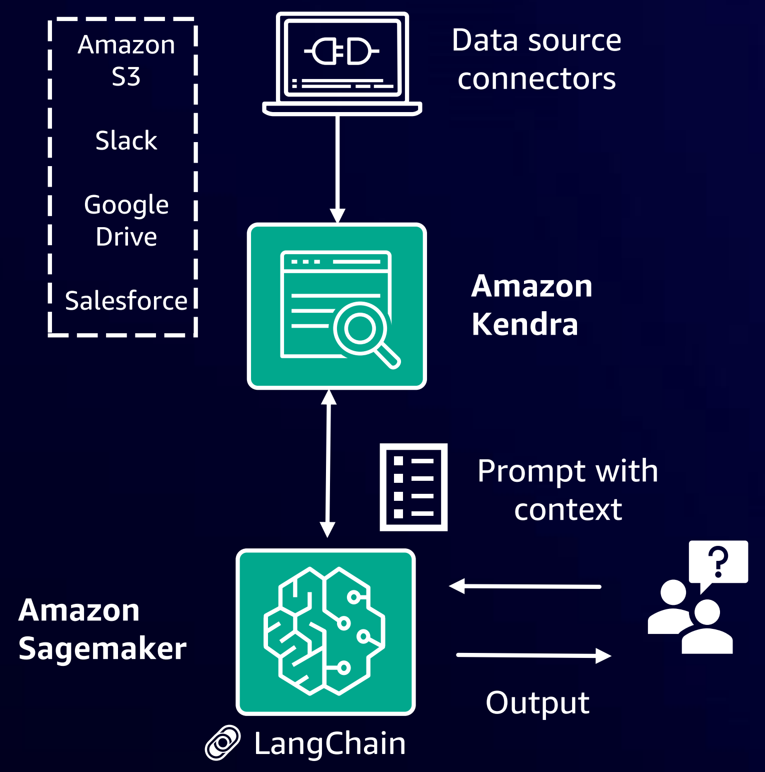
Amazon Bedrock (Claude 2)

[0.32, 2.35, 8.34, …]

1536 dimensions

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary



Amazon S3
Slack
Google Drive
Salesforce

Data source connectors

Amazon Kendra

Prompt with context

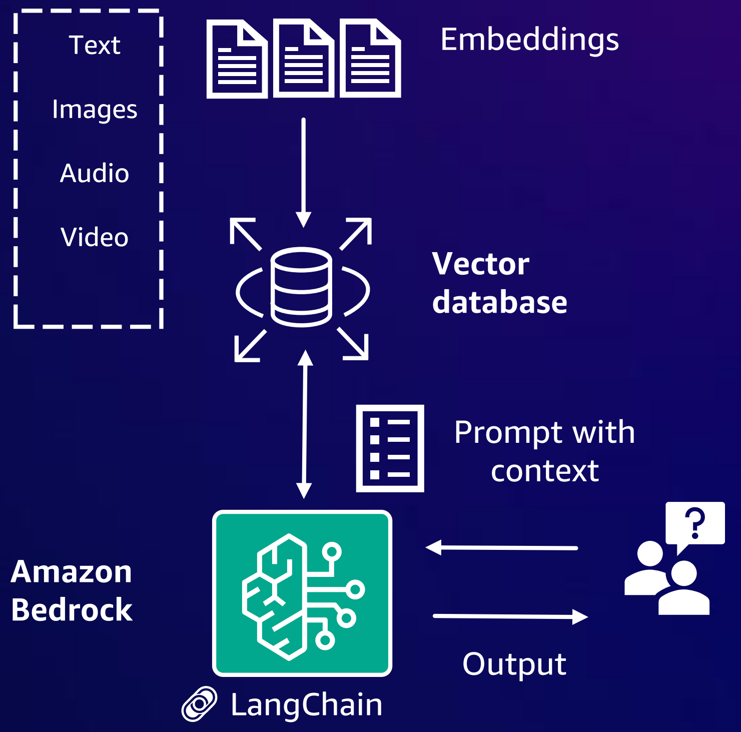
Amazon Sagemaker

Output

LangChain

Amazon SageMaker + Amazon Kendra

- Deploy FM as SageMaker endpoint (hosting)
- Choose SageMaker managed accelerated computing instance
- Kendra provides native and partner connectors
- No-code UI builder with Kendra



Text
Images
Audio
Video

Embeddings

Vector database

Prompt with context

Amazon Bedrock

Output

LangChain

Amazon Bedrock + Vector databases

- Access directly or fine-tune foundation model using Bedrock API
- Bedrock Serverless
- Wide range of data types can be converted to vectors
- Vector DBs are optimized for Efficient Similarity Search

Takeaways and resources to get started

Building Gen AI apps: Emerging tech stack

Demo 1:
RAG with Amazon SageMaker and Amazon Kendra

Demo 2:
RAG with Amazon Bedrock and a vector database

Takeaways and resources to get started

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3 key takeaways

Foundation models
have limitations

Work backwards
select the right use case

Your data is
your differentiator

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Additional resources

Generative AI on Amazon SageMaker Workshop

You will learn how to use Amazon SageMaker quick start solutions for fine-tuning models such as stable diffusion, as well as customizing the open-source models efficiently using the features provide by SageMaker.

Amazon Bedrock Workshop

You will explore techniques for generating text and images. You will gain hands-on experience using Amazon Bedrock APIs, SDKs, and open-source software, for example LangChain.

Amazon CodeWhisperer Workshop

You will use Amazon CodeWhisperer to build applications faster and more securely. Labs include exercises with Python, Java, React, SQL, Shell, and examples in Amazon SageMaker.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Please complete the session survey in the mobile app

Viktoria Semaan

linkedin.com/in/semaan

Linda Haviv

linkedin.com/in/LindaHaviv

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.