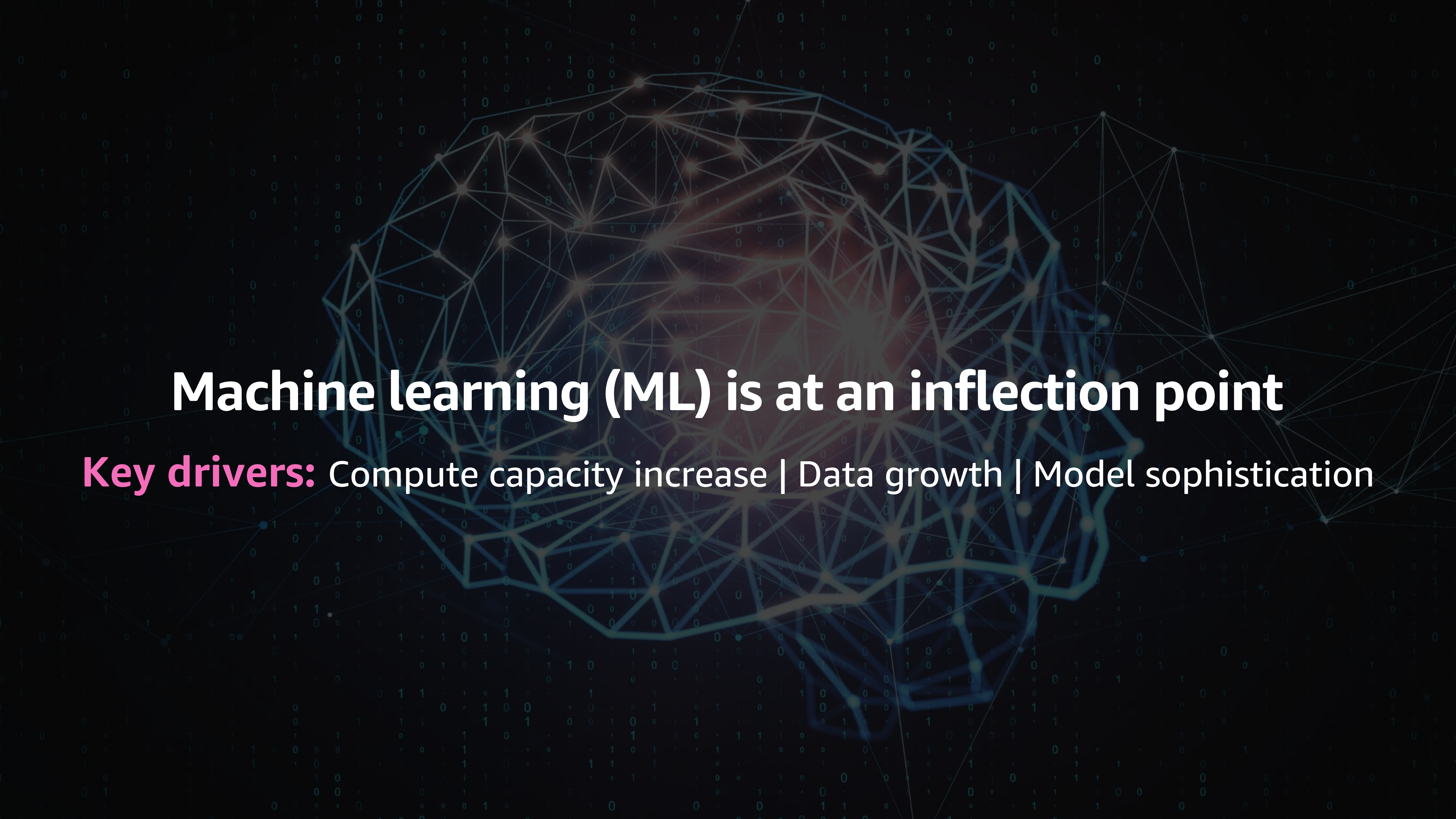
Enhancing AI Search with Machine Learning

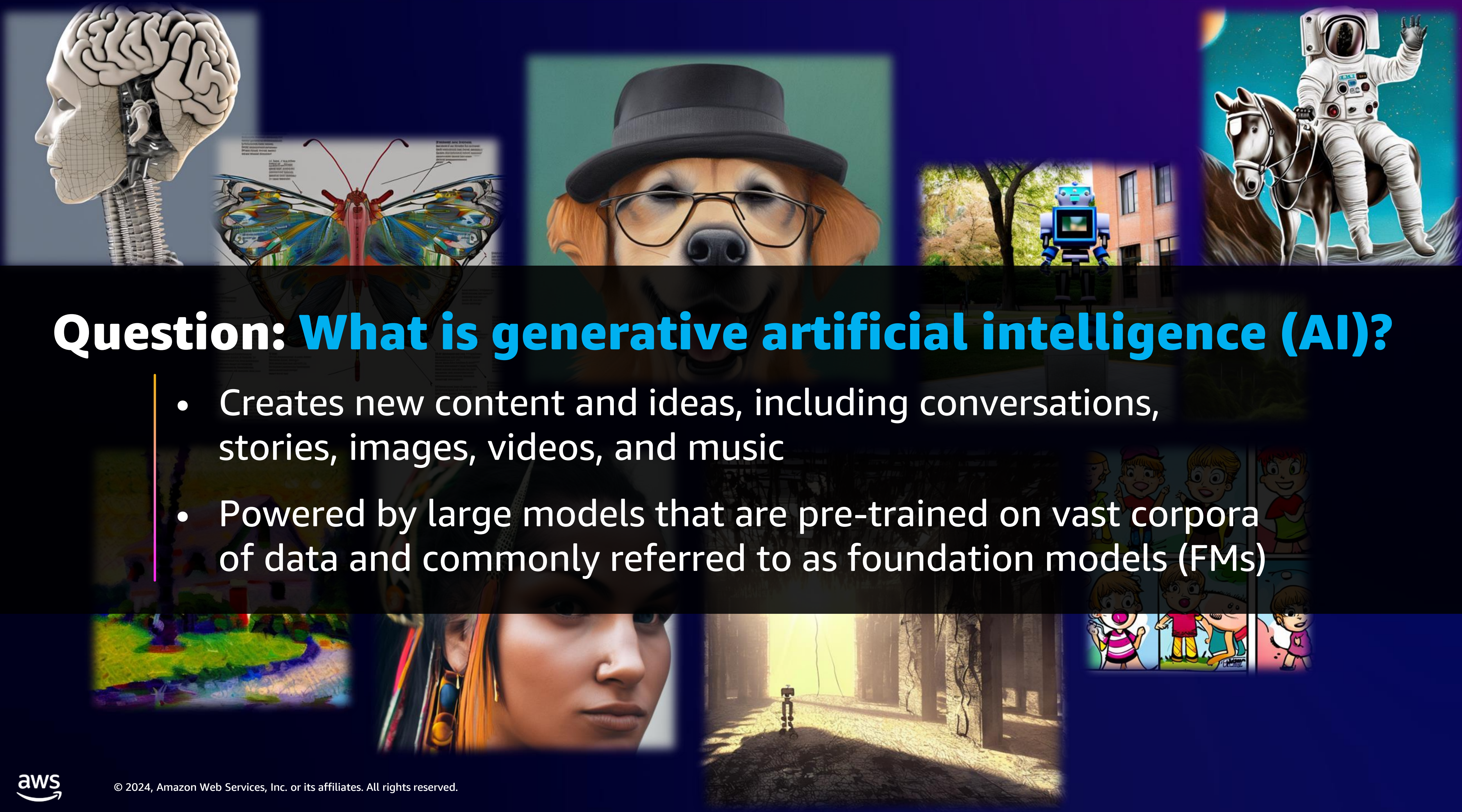
Suman Debnath

Developer Advocate, Data and Machine Learning
Amazon Web Services



Machine learning (ML) is at an inflection point

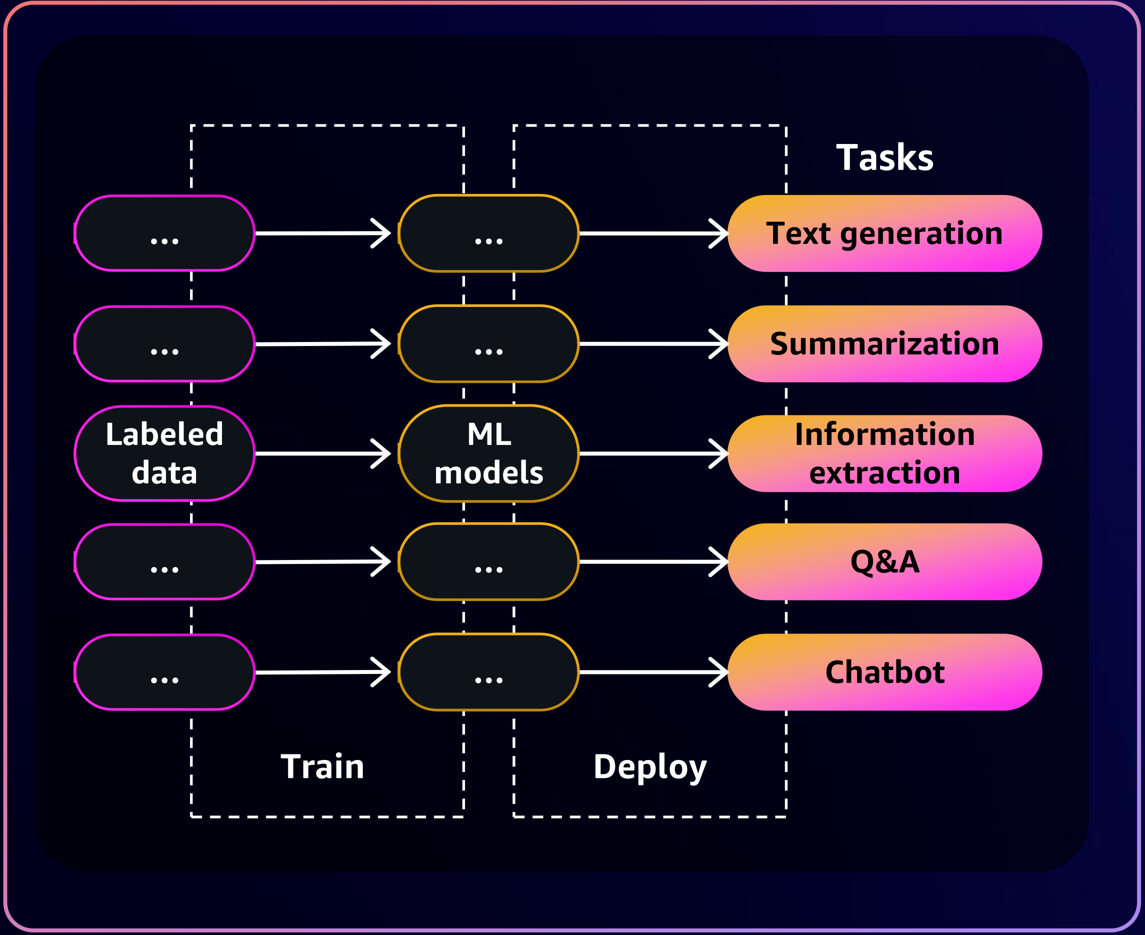
Key drivers: Compute capacity increase | Data growth | Model sophistication



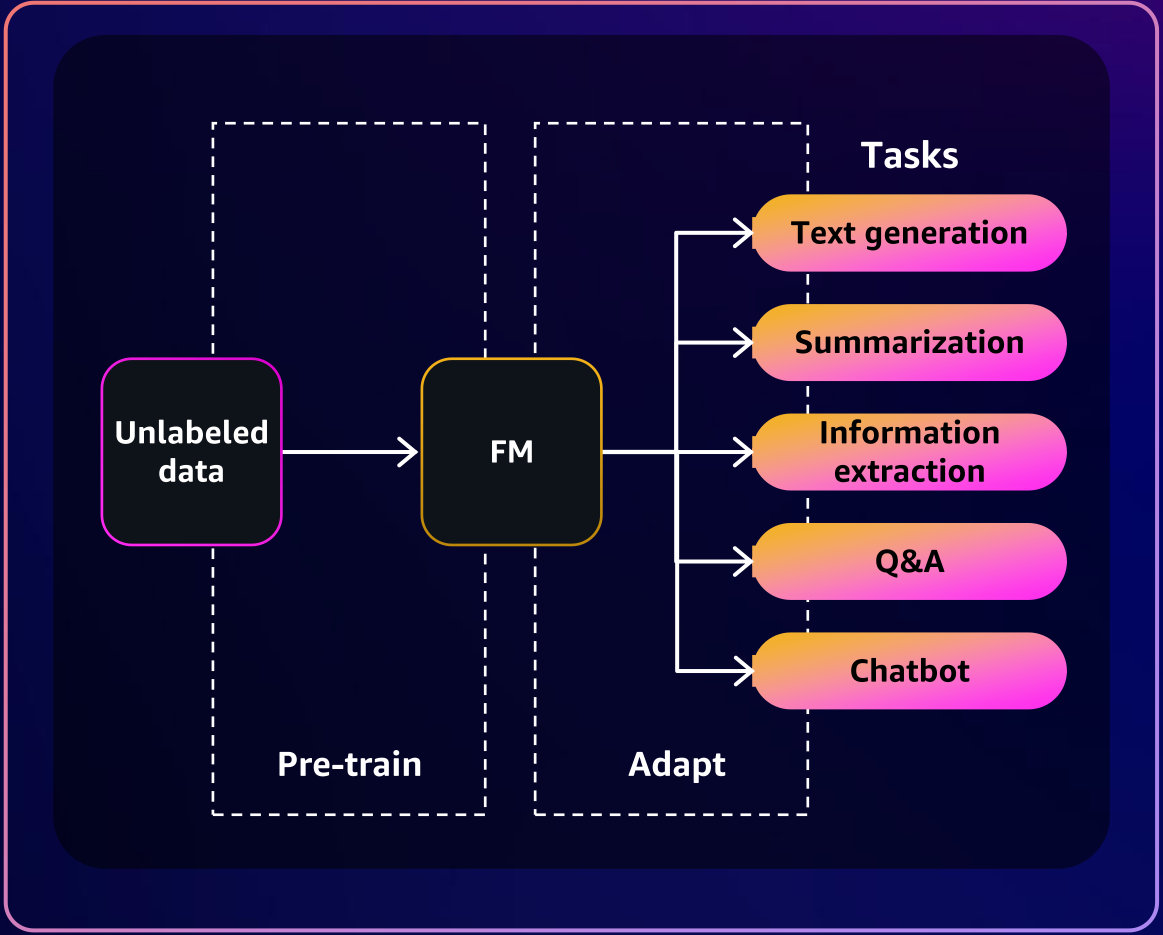
Question: What is generative artificial intelligence (AI)?

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

How foundation models differ from other ML models ?



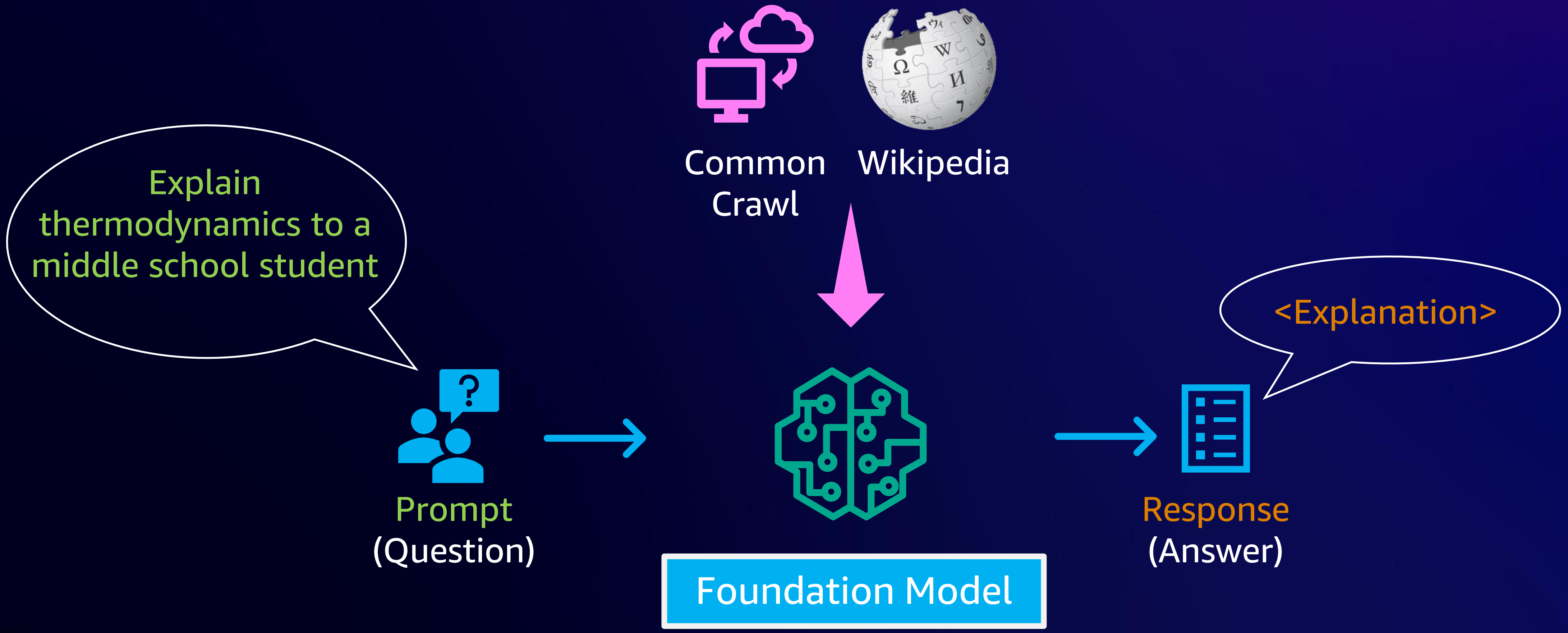
Traditional ML models



Foundation models

What are inputs & outputs of foundation models ?

Initial pre-training



Common Crawl

Wikipedia

Explain thermodynamics to a middle school student

Prompt
(Question)

Foundation Model

<Explanation>

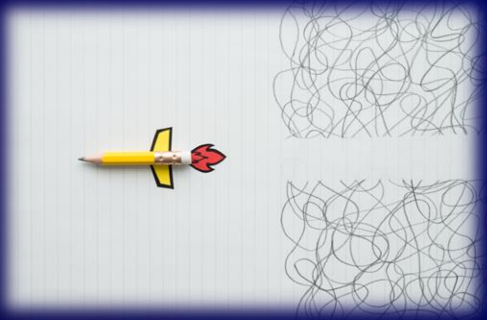
Response
(Answer)

Let’s say we want to know…

- Which are the products that got the best reviews on XYZ platform in last 15 days ?

Hallucination

- Who won the India vs Afghanistan 2024 T20 championship ?



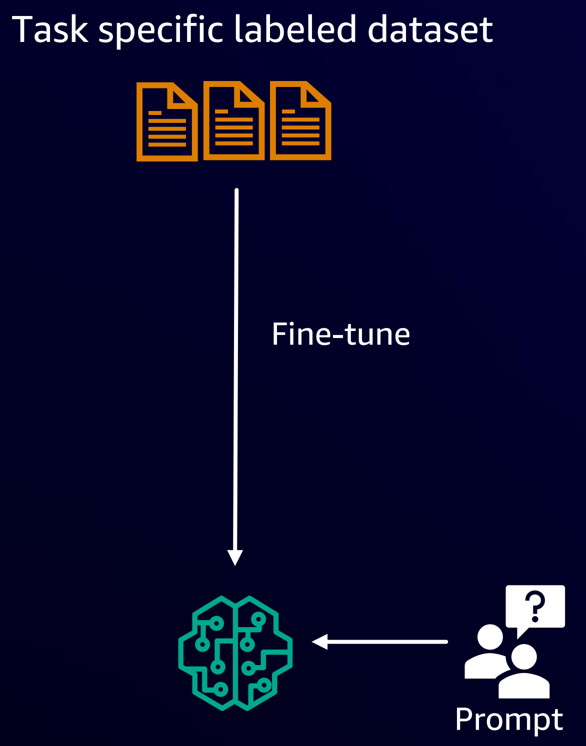
Knowledge Cutoff

Large Language Model Limitations

Photo License: @AdobeStock_{428564002, 630708219}

Photo License: @AdobeStock_{56886440}

How can we customize a foundation model ?

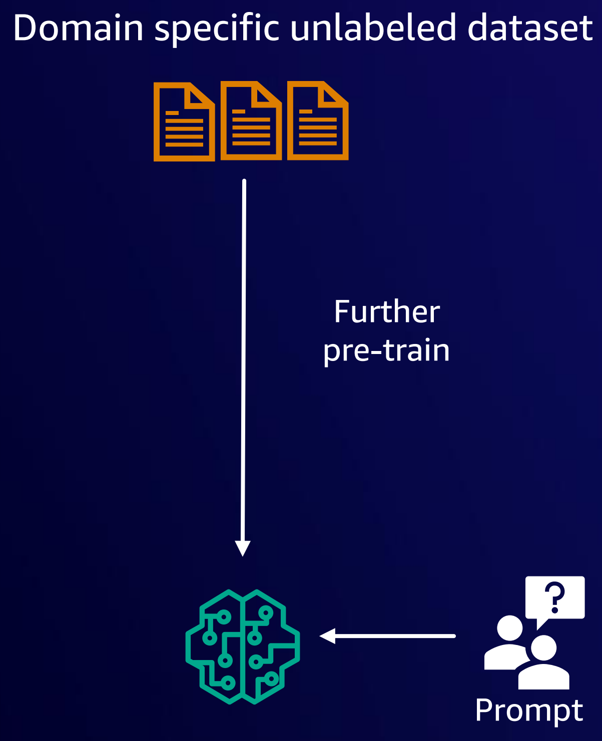


Task specific labeled dataset

Fine-tune

Prompt

Instruction-based fine-tuning

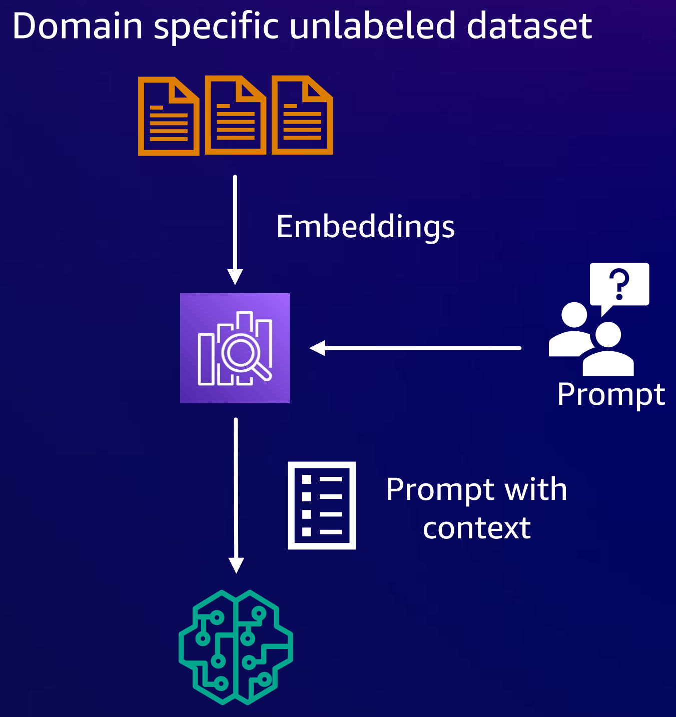


Domain specific unlabeled dataset

Further
pre-train

Prompt

Domain adaptation



Domain specific unlabeled dataset

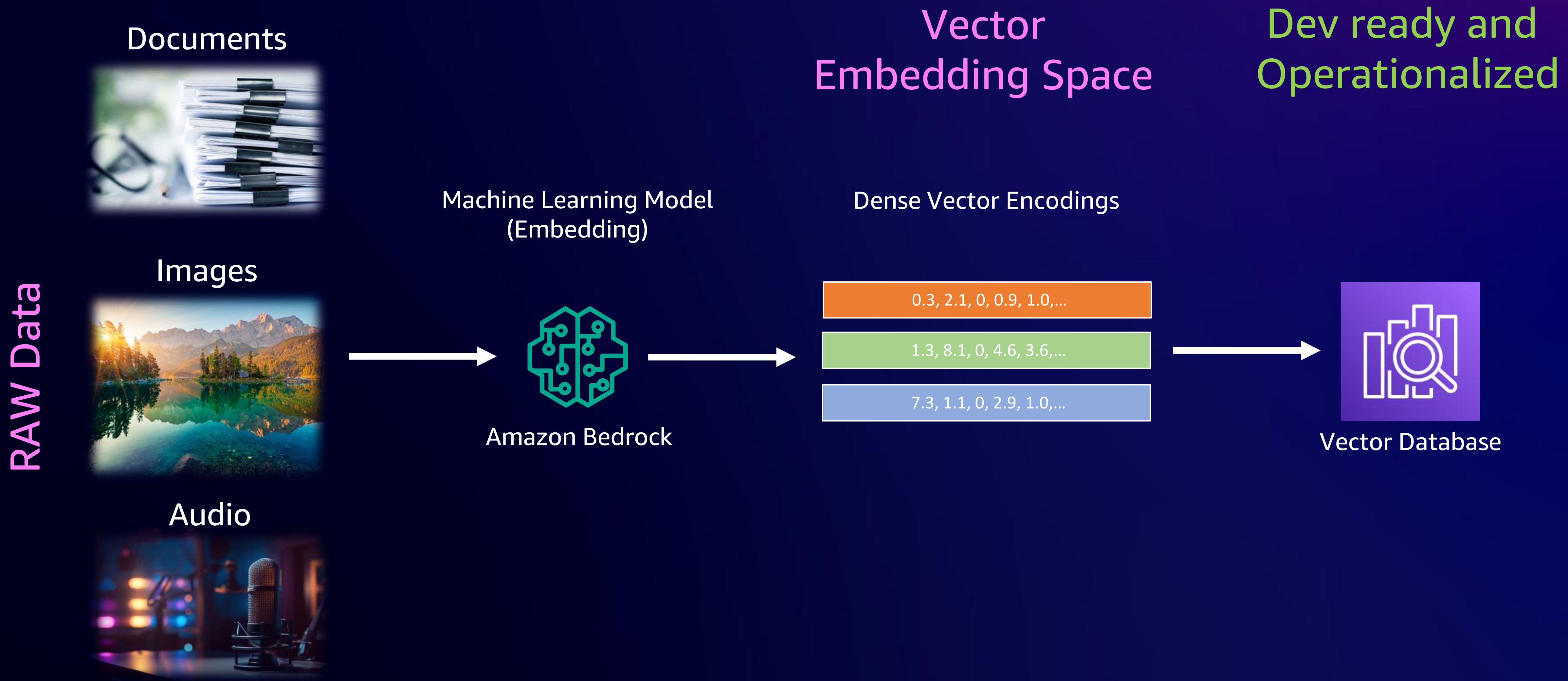
Embeddings

Prompt

Prompt with
context

Information Retrieval

Vector store



RAW Data

Documents

Images

Audio

Machine Learning Model
(Embedding)

Amazon Bedrock

Vector
Embedding Space

Dense Vector Encodings

0.3, 2.1, 0, 0.9, 1.0,...

1.3, 8.1, 0, 4.6, 3.6,...

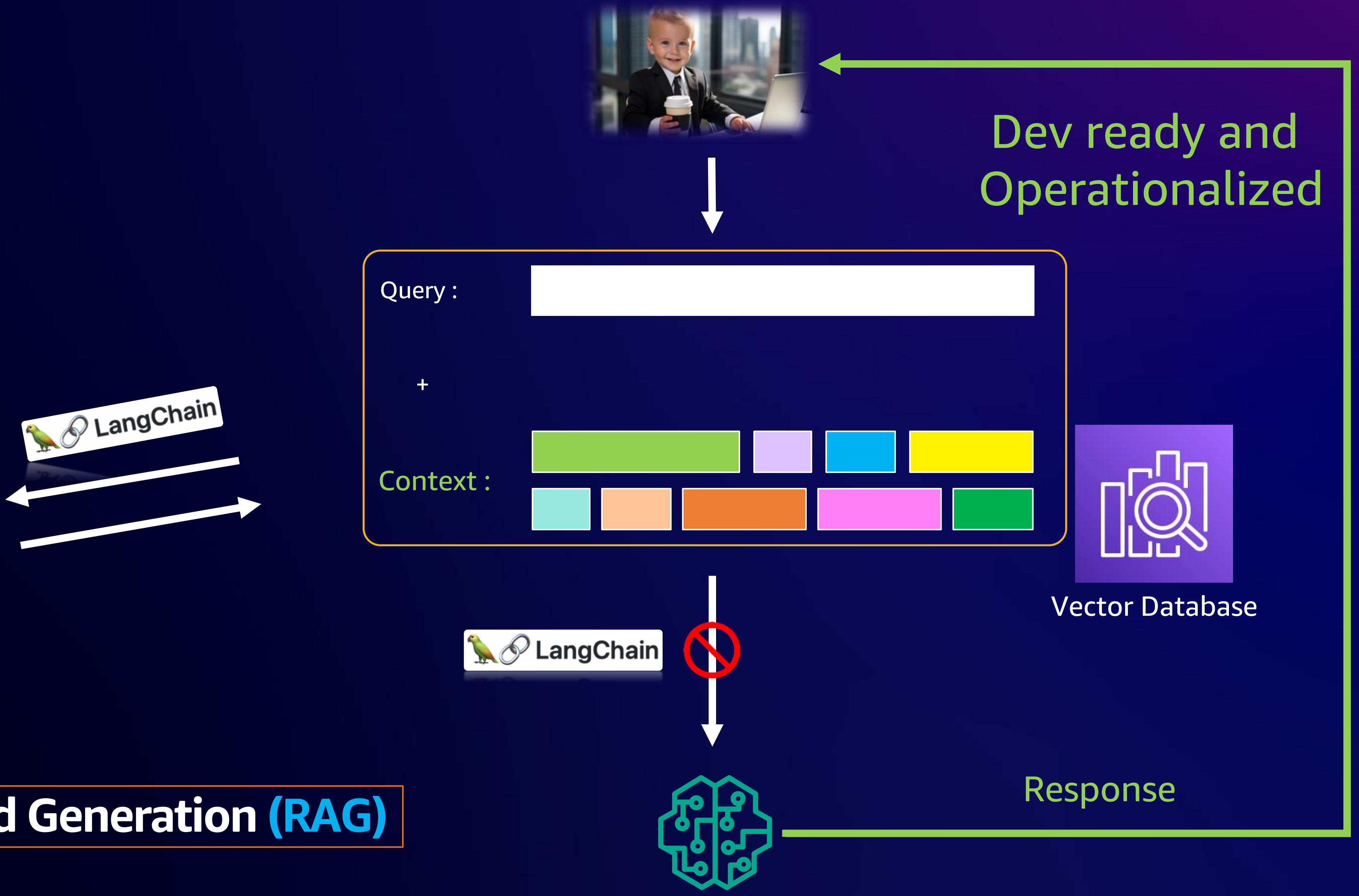
7.3, 1.1, 0, 2.9, 1.0,...

Dev ready and
Operationalized

Vector Database

Photo License: @AdobeStock_{170110000, 610223671, 270350073}

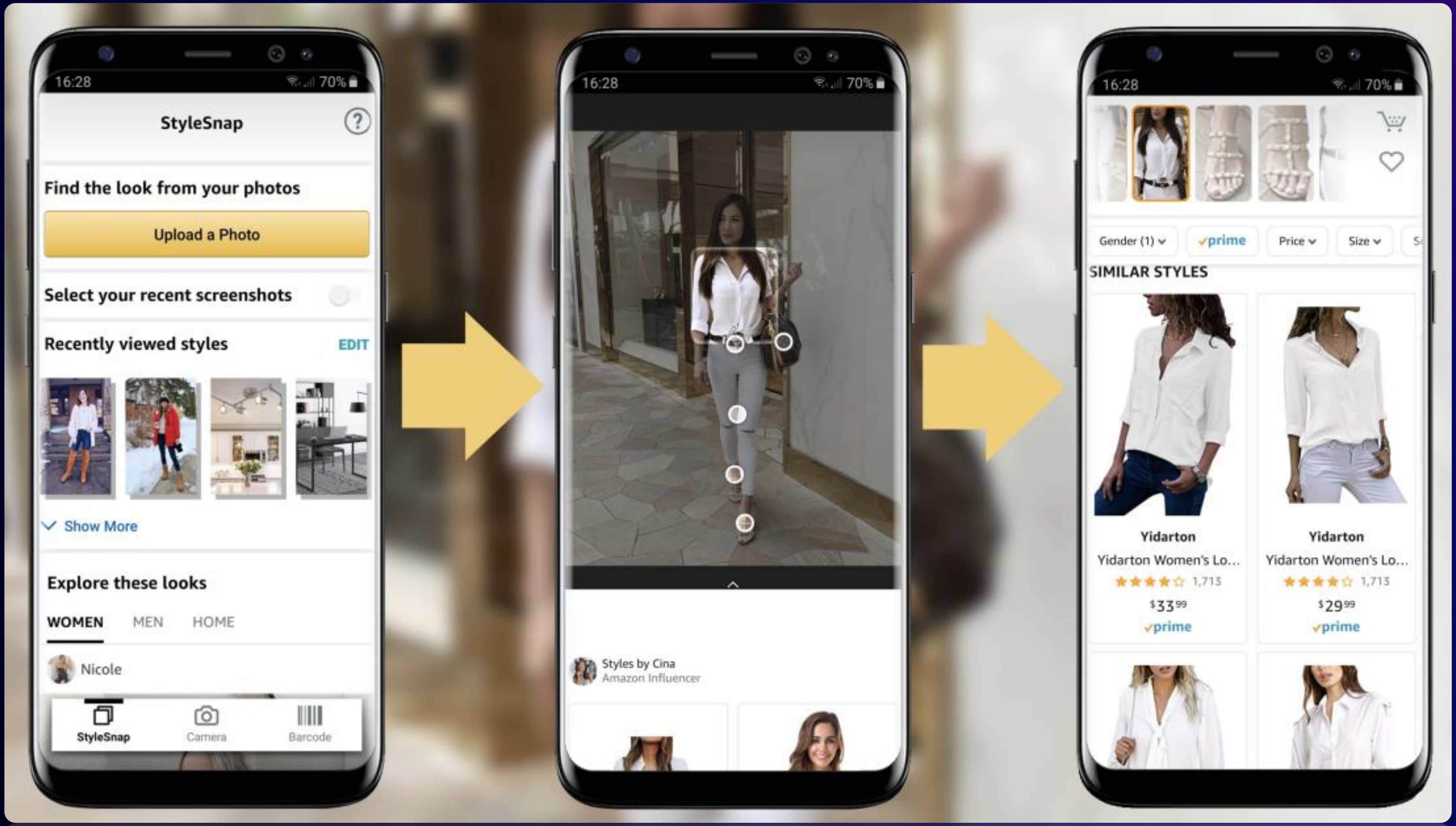
Vector store



Retrieval-Augmented Generation (RAG)

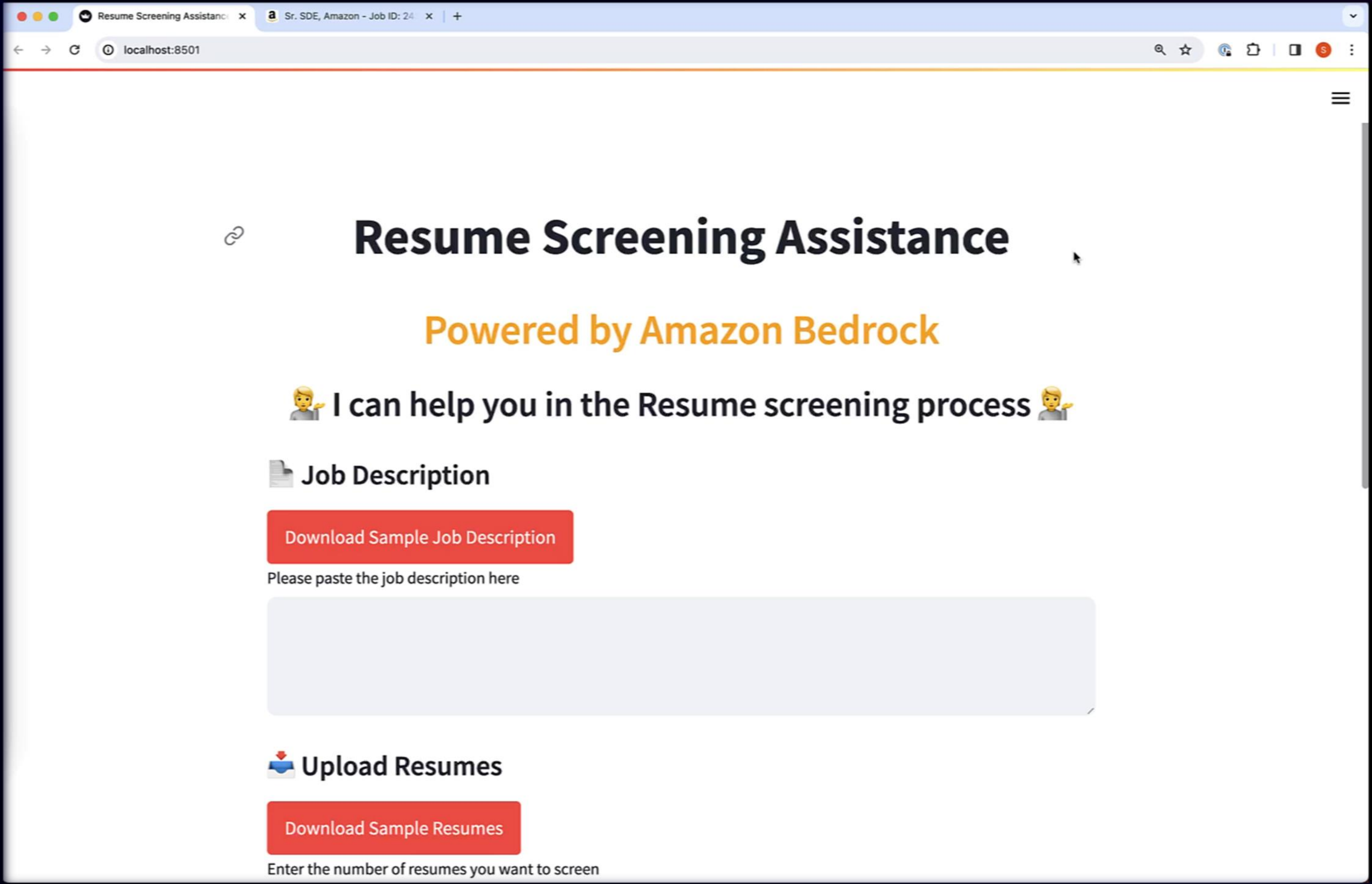
Photo License: @AdobeStock_{631479565}

Visual Search



Let’s build it : Resume Screening

Photo License: @AdobeStock_{651077888}



Resume Screening Assistance

Powered by Amazon Bedrock

💁‍♂️ I can help you in the Resume screening process 💁‍♂️

📄 Job Description

Download Sample Job Description

Please paste the job description here

📥 Upload Resumes

Download Sample Resumes

Enter the number of resumes you want to screen

Let’s build it : Resume Screening

Input

Job Description

No. of resumes you want ?

```
relavant_docs = vectorstore.similarity_search_with_score(job_description,
                                                          resume_count)
```

Amazon Bedrock

```
from langchain.llms.bedrock import Bedrock
from langchain.chains.summarize import load_summarize_chain

llm = Bedrock()
chain = load_summarize_chain(llm,..)

summary = chain.run(relavant_docs)
```

Bunch of Resumes

Amazon Bedrock
(Embedding)

0.3, 2.1, 0, 0.9, 1.0,...

1.3, 8.1, 0, 4.6, 3.6,...

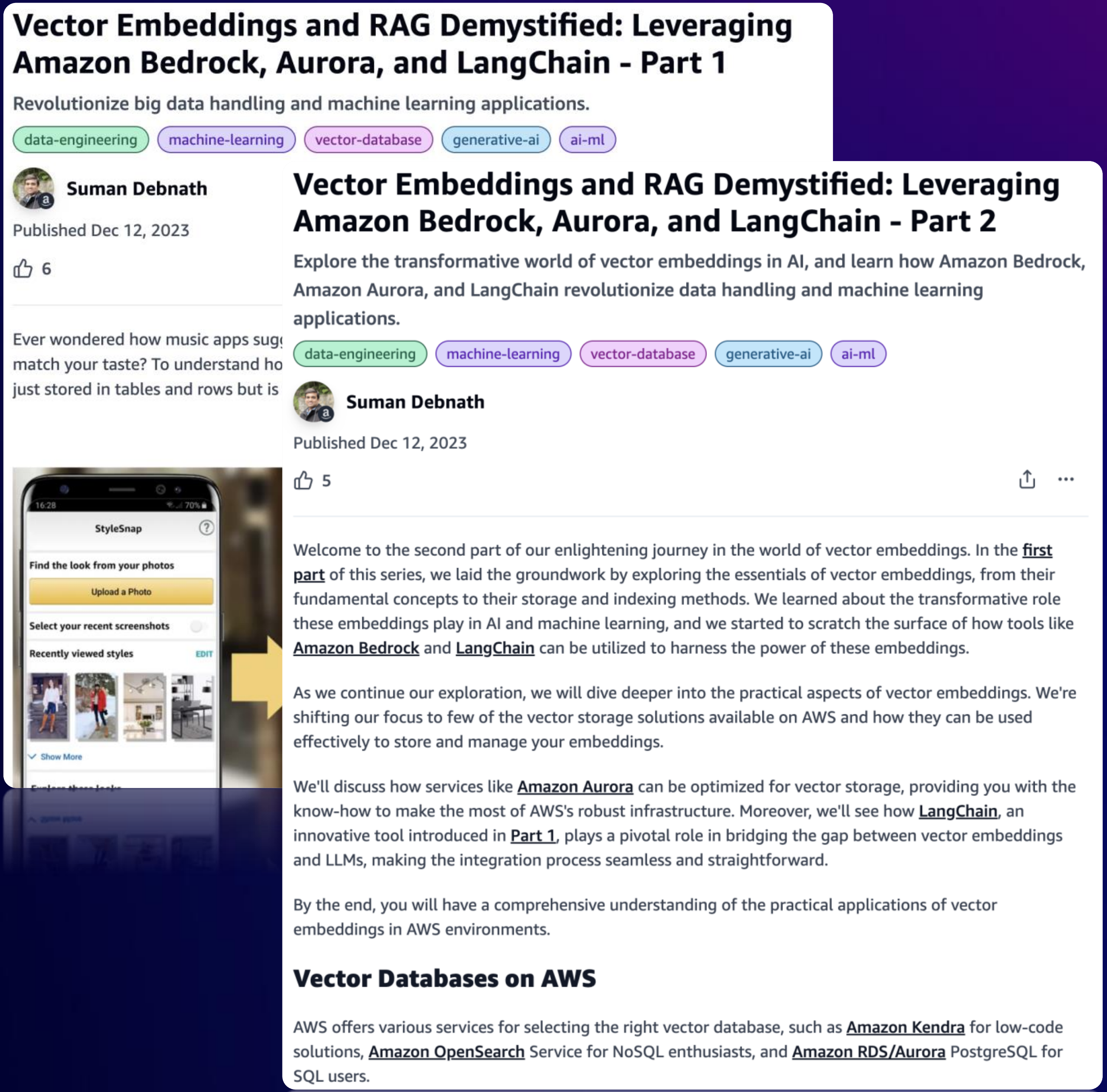
7.3, 1.1, 0, 2.9, 1.0,...

Dense Vector Encodings

Vector Database

```
from langchain.vectorstores.pgvector import PGVector
vectorstore = PGVector.from_documents(…)
```

https://bit.ly/3SmHjbk



Vector Embeddings and RAG Demystified: Leveraging Amazon Bedrock, Aurora, and LangChain - Part 1

Revolutionize big data handling and machine learning applications.

data-engineering machine-learning vector-database generative-ai ai-ml

Suman Debnath

Published Dec 12, 2023

6

Ever wondered how music apps sug
match your taste? To understand ho
just stored in tables and rows but is

Vector Embeddings and RAG Demystified: Leveraging Amazon Bedrock, Aurora, and LangChain - Part 2

Explore the transformative world of vector embeddings in AI, and learn how Amazon Bedrock, Amazon Aurora, and LangChain revolutionize data handling and machine learning applications.

data-engineering machine-learning vector-database generative-ai ai-ml

Suman Debnath

Published Dec 12, 2023

5

Welcome to the second part of our enlightening journey in the world of vector embeddings. In the first part of this series, we laid the groundwork by exploring the essentials of vector embeddings, from their fundamental concepts to their storage and indexing methods. We learned about the transformative role these embeddings play in AI and machine learning, and we started to scratch the surface of how tools like Amazon Bedrock and LangChain can be utilized to harness the power of these embeddings.

As we continue our exploration, we will dive deeper into the practical aspects of vector embeddings. We're shifting our focus to few of the vector storage solutions available on AWS and how they can be used effectively to store and manage your embeddings.

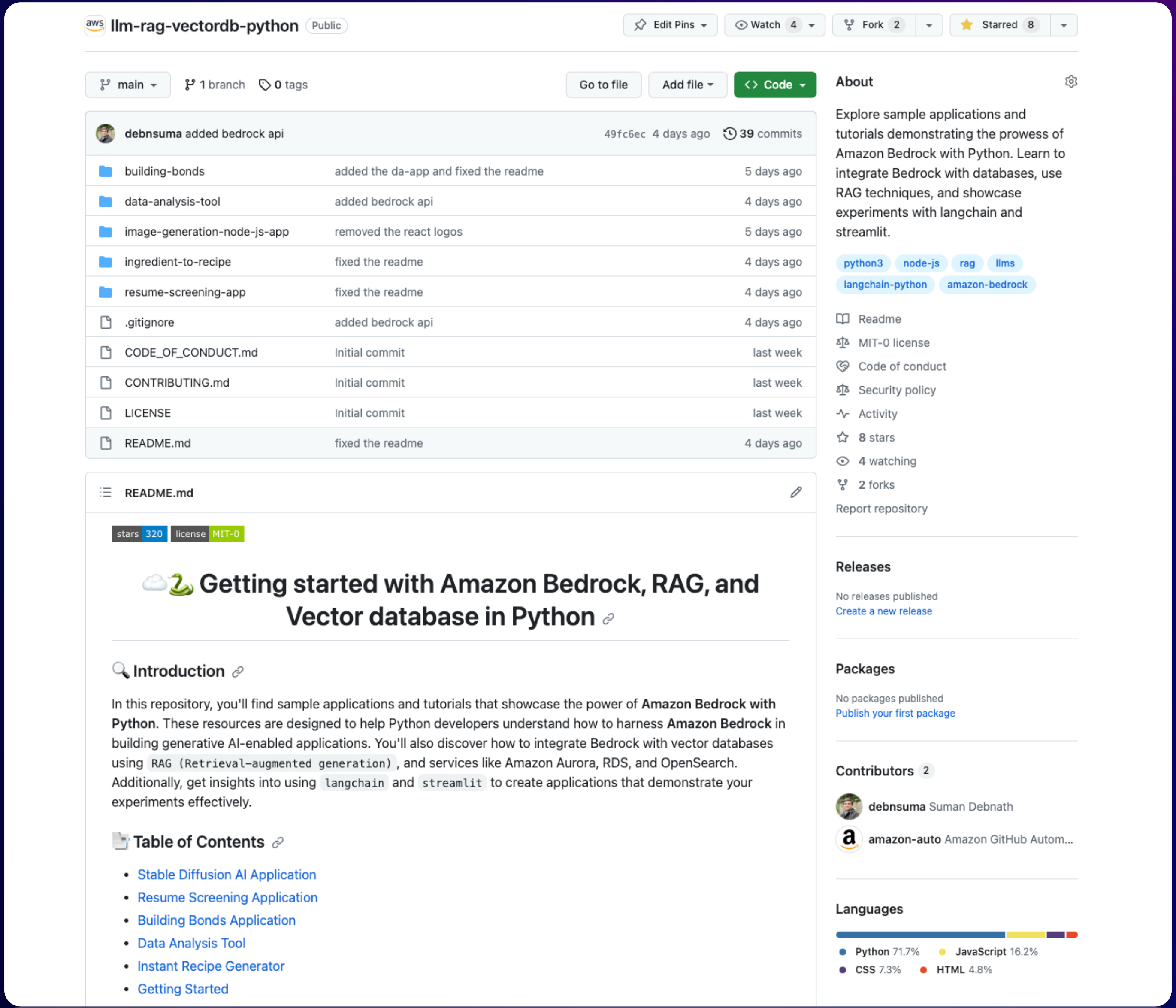
We'll discuss how services like Amazon Aurora can be optimized for vector storage, providing you with the know-how to make the most of AWS's robust infrastructure. Moreover, we'll see how LangChain, an innovative tool introduced in Part 1, plays a pivotal role in bridging the gap between vector embeddings and LLMs, making the integration process seamless and straightforward.

By the end, you will have a comprehensive understanding of the practical applications of vector embeddings in AWS environments.

Vector Databases on AWS

AWS offers various services for selecting the right vector database, such as Amazon Kendra for low-code solutions, Amazon OpenSearch Service for NoSQL enthusiasts, and Amazon RDS/Aurora PostgreSQL for SQL users.

https://bit.ly/3Q3amy0



debnsuma added bedrock api 49fc6ec 4 days ago 39 commits

building-bonds	added the da-app and fixed the readme	5 days ago
data-analysis-tool	added bedrock api	4 days ago
image-generation-node-js-app	removed the react logos	5 days ago
ingredient-to-recipe	fixed the readme	4 days ago
resume-screening-app	fixed the readme	4 days ago
.gitignore	added bedrock api	4 days ago
CODE_OF_CONDUCT.md	Initial commit	last week
CONTRIBUTING.md	Initial commit	last week
LICENSE	Initial commit	last week
README.md	fixed the readme	4 days ago

README.md

☁️🐍 Getting started with Amazon Bedrock, RAG, and Vector database in Python

🔍 Introduction

In this repository, you'll find sample applications and tutorials that showcase the power of **Amazon Bedrock with Python**. These resources are designed to help Python developers understand how to harness **Amazon Bedrock** in building generative AI-enabled applications. You'll also discover how to integrate Bedrock with vector databases using `RAG (Retrieval-augmented generation)`, and services like Amazon Aurora, RDS, and OpenSearch. Additionally, get insights into using `langchain` and `streamlit` to create applications that demonstrate your experiments effectively.

📑 Table of Contents

- Stable Diffusion AI Application
- Resume Screening Application
- Building Bonds Application
- Data Analysis Tool
- Instant Recipe Generator
- Getting Started

About

Explore sample applications and tutorials demonstrating the prowess of Amazon Bedrock with Python. Learn to integrate Bedrock with databases, use RAG techniques, and showcase experiments with langchain and streamlit.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2

debnsuma Suman Debnath
amazon-auto Amazon GitHub Autom...

Languages

● Python 71.7% ● JavaScript 16.2%
● CSS 7.3% ● HTML 4.8%

GENERATIVE AI STACK

APPLICATIONS THAT LEVERAGE FMs

- Amazon Q
- Amazon Q in Amazon QuickSight
- Amazon Q in Amazon Connect
- Amazon CodeWhisperer

TOOLS TO BUILD WITH LLMs & OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

SageMaker

INFRASTRUCTURE FOR FM TRAINING & INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

THANK YOU

Suman Debnath

linkedin.com/in/suman-d

debnsuma@amazon.com