Generative AI in 2024

AI Engineering

Mike Chambers (he/him)

Snr. Developer Advocate Specialist
Amazon Web Services

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

👋 hello

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Where are we now?

Image

Generated with Amazon Titan Image Generator G1

“A happy golden retriever wearing glasses and a hat in a portrait painting”

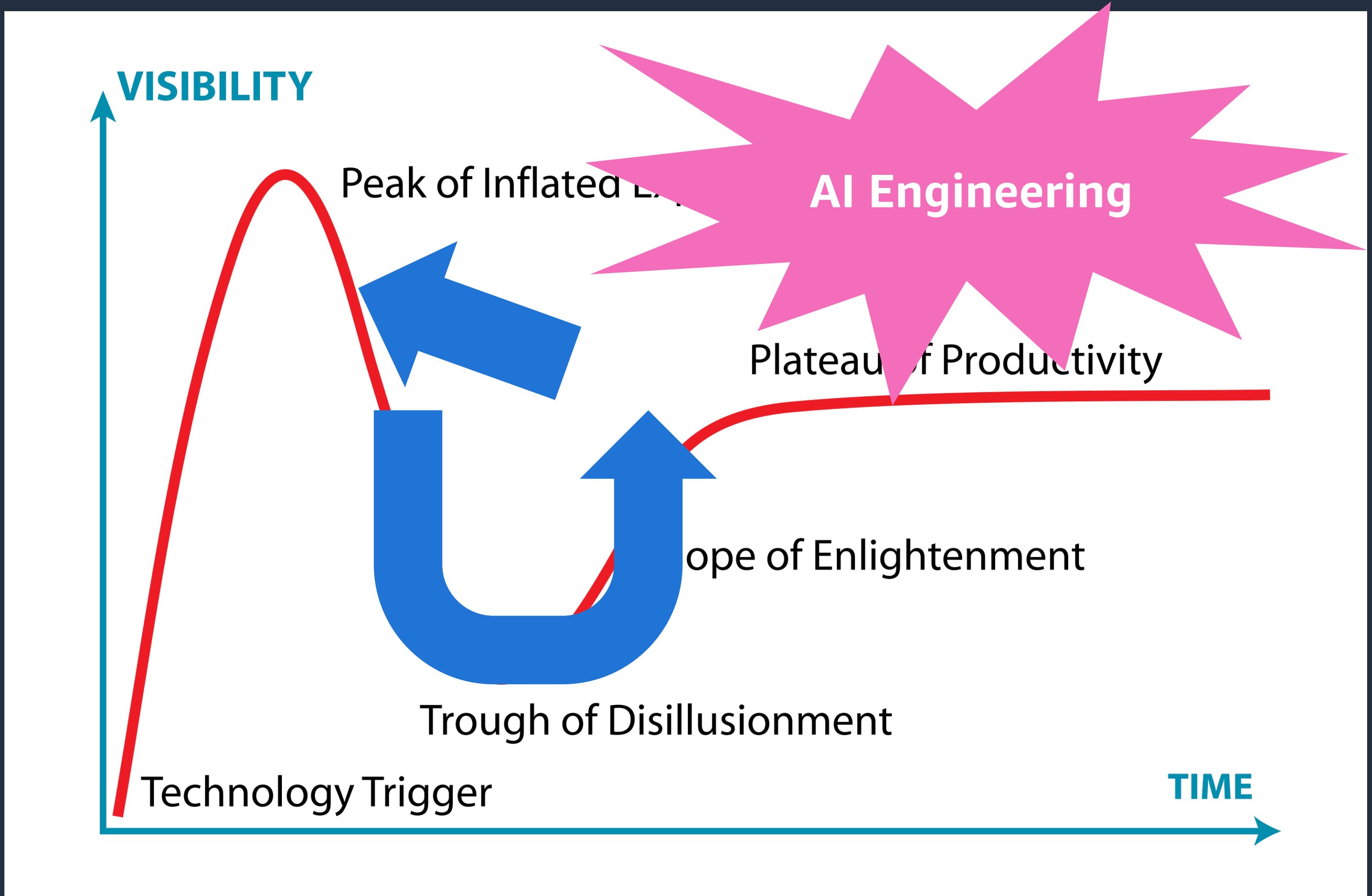
Text/Chat

```
> Question: What is generative AI?
Generative AI refers to a type of
artificial intelligence that is capable
of creating new and original content.
It uses algorithms and machine learning
to generate text, images, music, and
other forms of media based on patterns
and rules learned from existing data.
```

Generated with Amazon Titan Text G1 - Express

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Where are we now?



Gartner Research's Hype Cycle diagram

Jeremykemp at English Wikipedia

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI Engineering

- Foundation Models
- Prompting engineering (yes it’s still important)
- Agents
- Retrieval Augmented Generation
- Fine-tuning and Continuation training

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Foundation Models

Jurassic-2 series
By AI21 Labs

Titan
By Amazon

Claude
By Anthropic

Command
By Cohere

Llama 2
By Meta

Stable Diffusion
By Stability AI

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt Engineering

Text generation

Text summarization

Translation

Sentiment analysis

More …

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt Engineering

“

```
Summarise the following text:
<text>
…
</text>
```

”

In context learning

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt Engineering

> “You are a personal assistant. You are friendly, polite and casual. You help with…”

> “You are a classifying agent that filters user inputs into categories. Your job is to sort these inputs before they are passed along to our function calling agent. The purpose of our function calling agent is to call functions in order to answer user's questions.”

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prompt Engineering

```
The user input is
<question>
What is the weather in Oslo right now?
</question>
```

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agents

- Foundation models have a learning cut off.
- LLMs will “hallucinate” if given the chance.
- What do we want from an LLM?
- World knowledge…

Up to up…

Down is down…

The sky is blue…

What is the weather in Oslo right now?

Answer: As of 12:00 PM (noon) local time in Oslo, Norway, the current weather conditions are:

* Temperature: 4°C (39°F)
* Humidity: 84%
* Wind: Northwest at 12 km/h (7 mph)
* Conditions: Overcast with light rain showers

Llama 2 Chat 13B

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agents

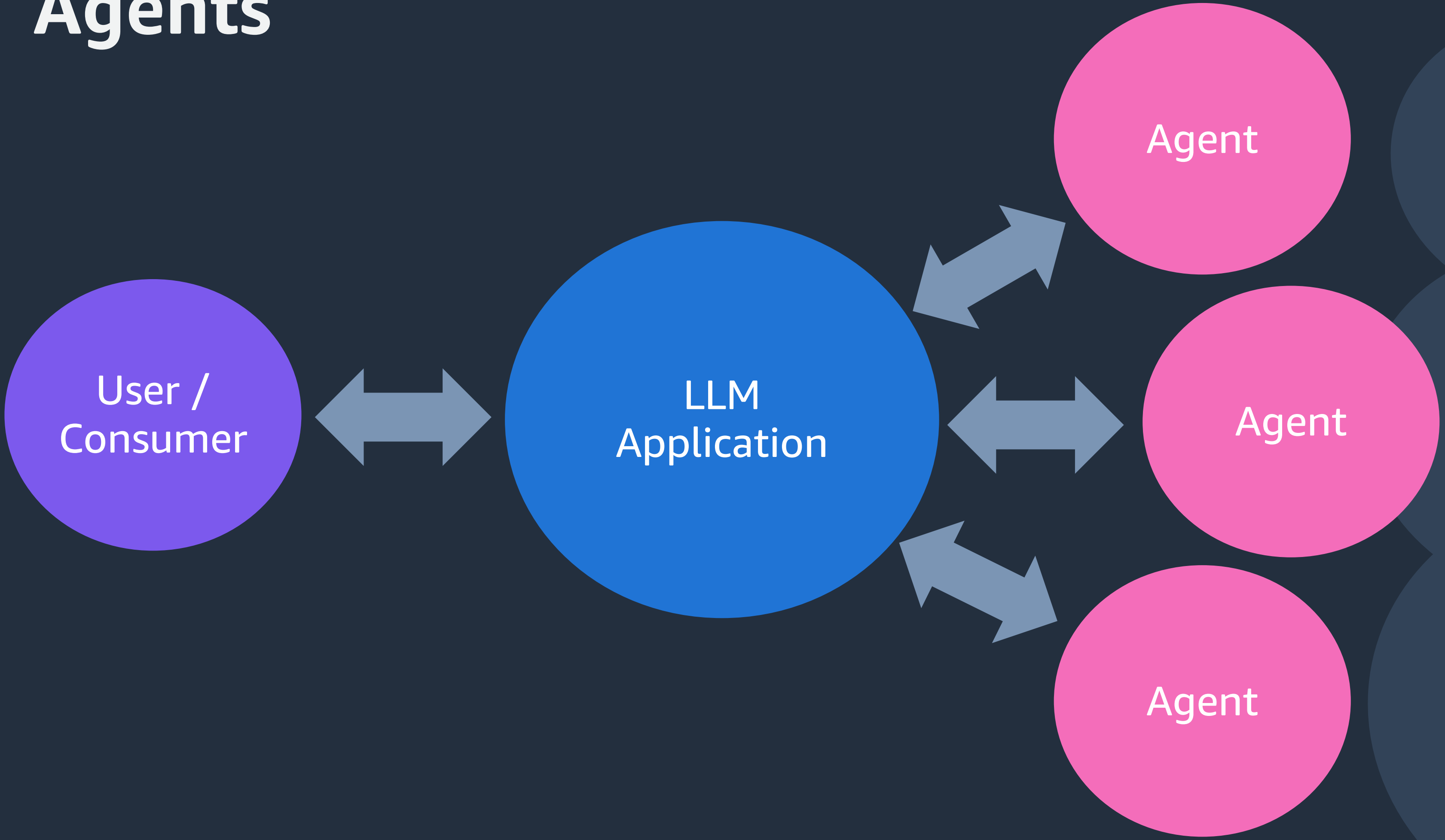
Give LLM applications agents.

agency
/ˈeɪdʒ(ə)nsi/

2. action or intervention producing a particular effect.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agents



Request Data

Retrieve Data

Trigger Events

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agents

What *is* the weather in Oslo right now?

Demo time…

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

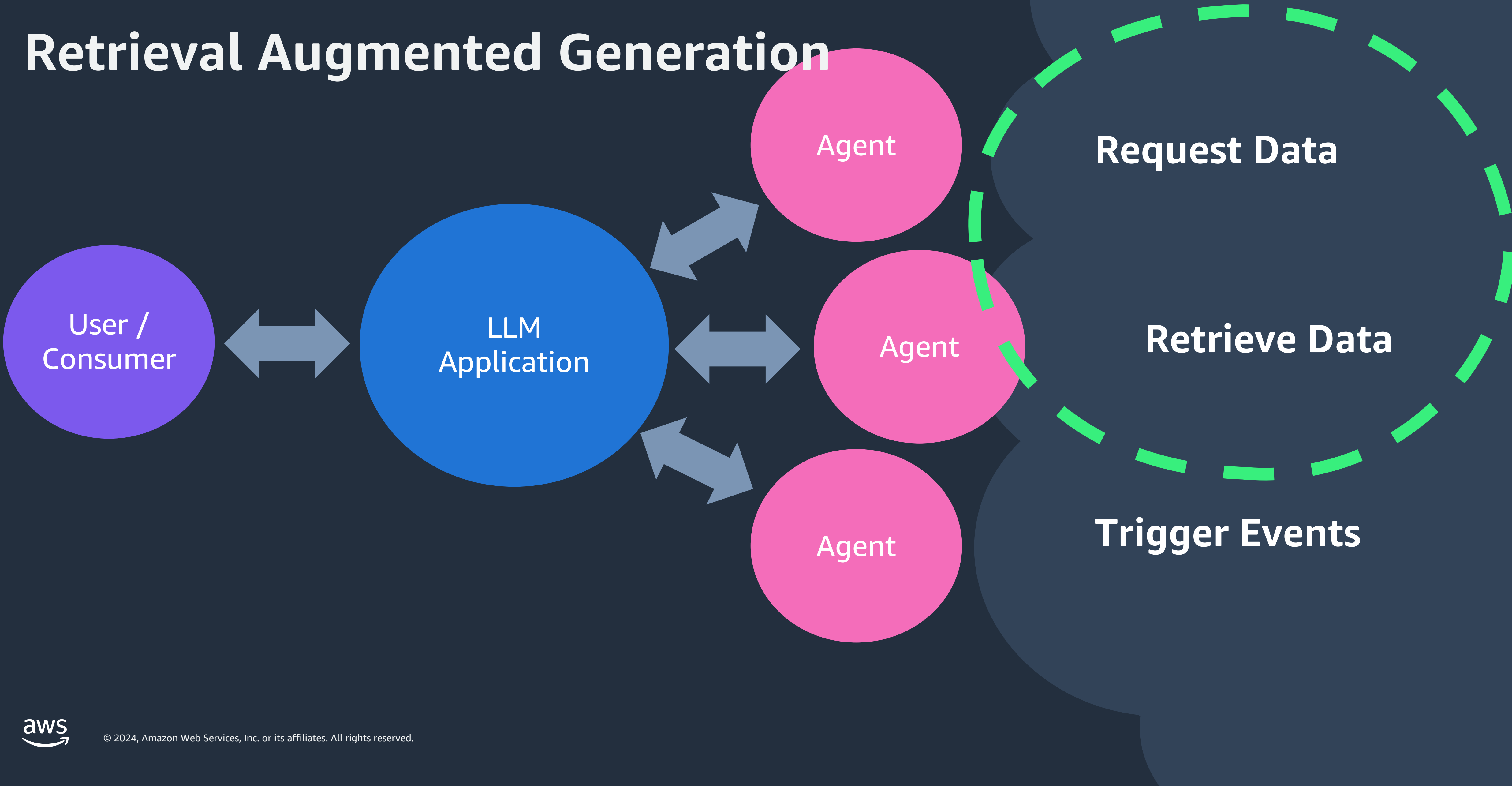
Agents

What *is* the weather in Oslo right now?

(Preview)

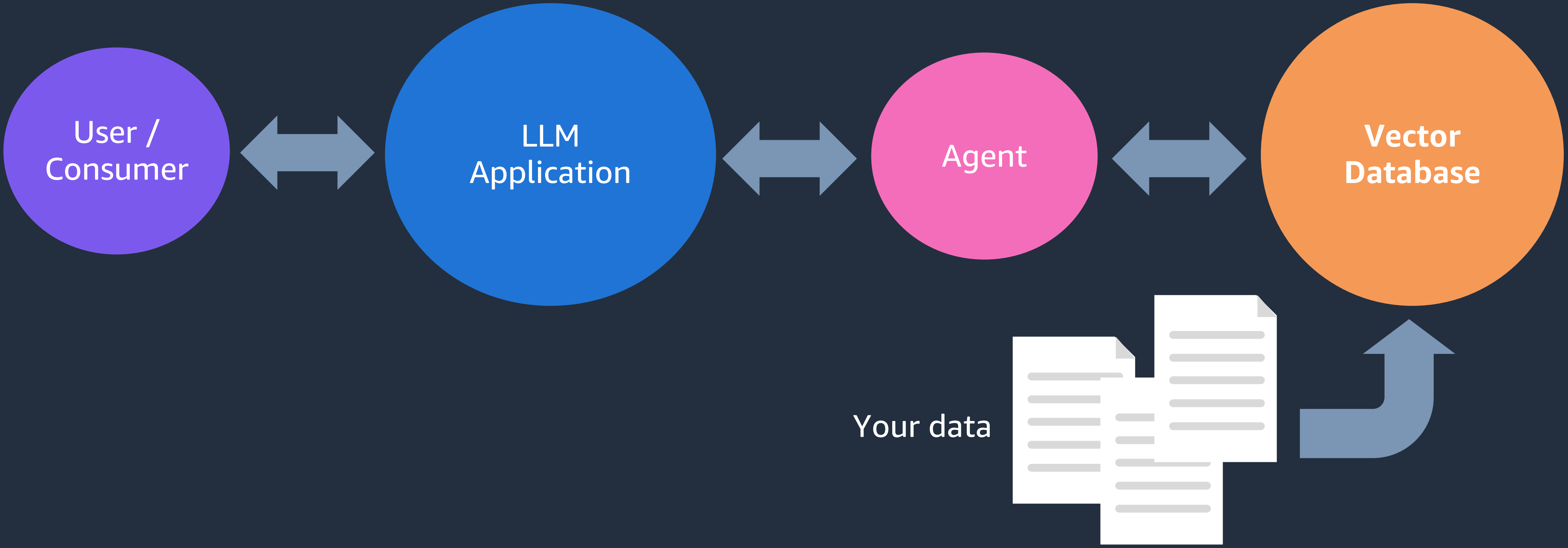
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval Augmented Generation



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval Augmented Generation



User / Consumer

LLM Application

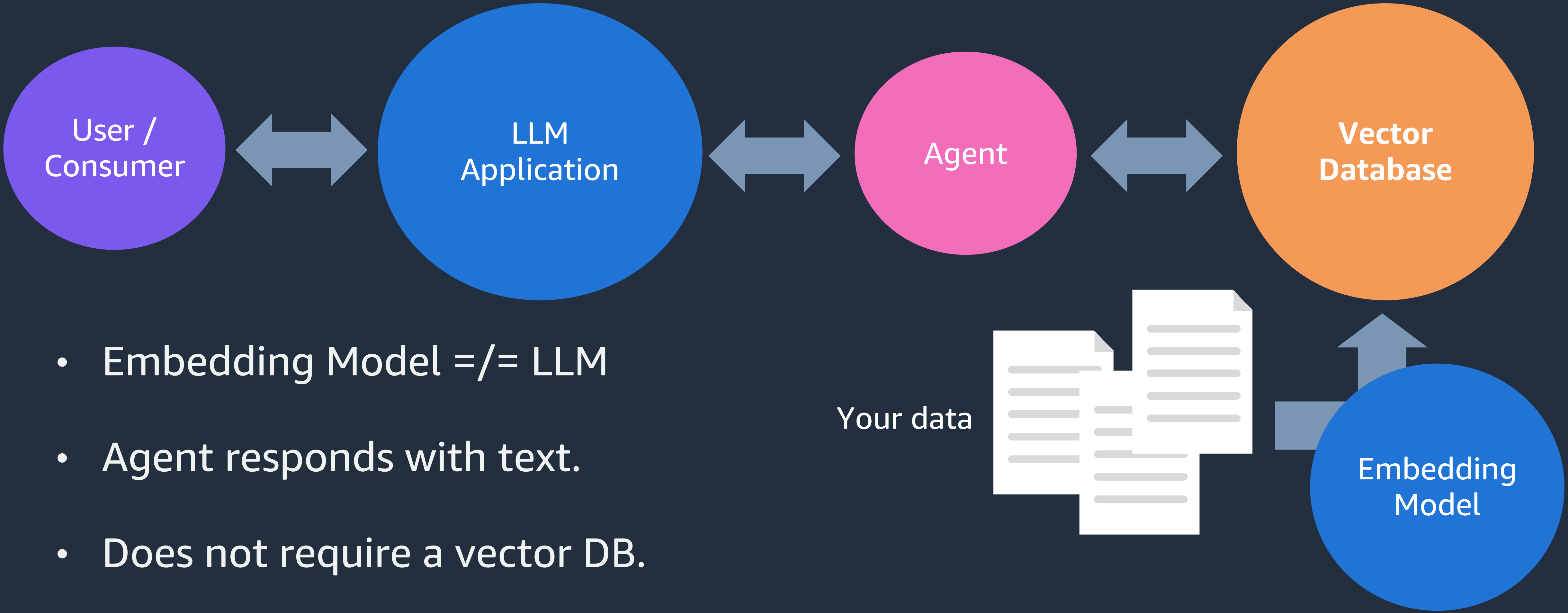
Agent

Vector Database

Your data

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval Augmented Generation

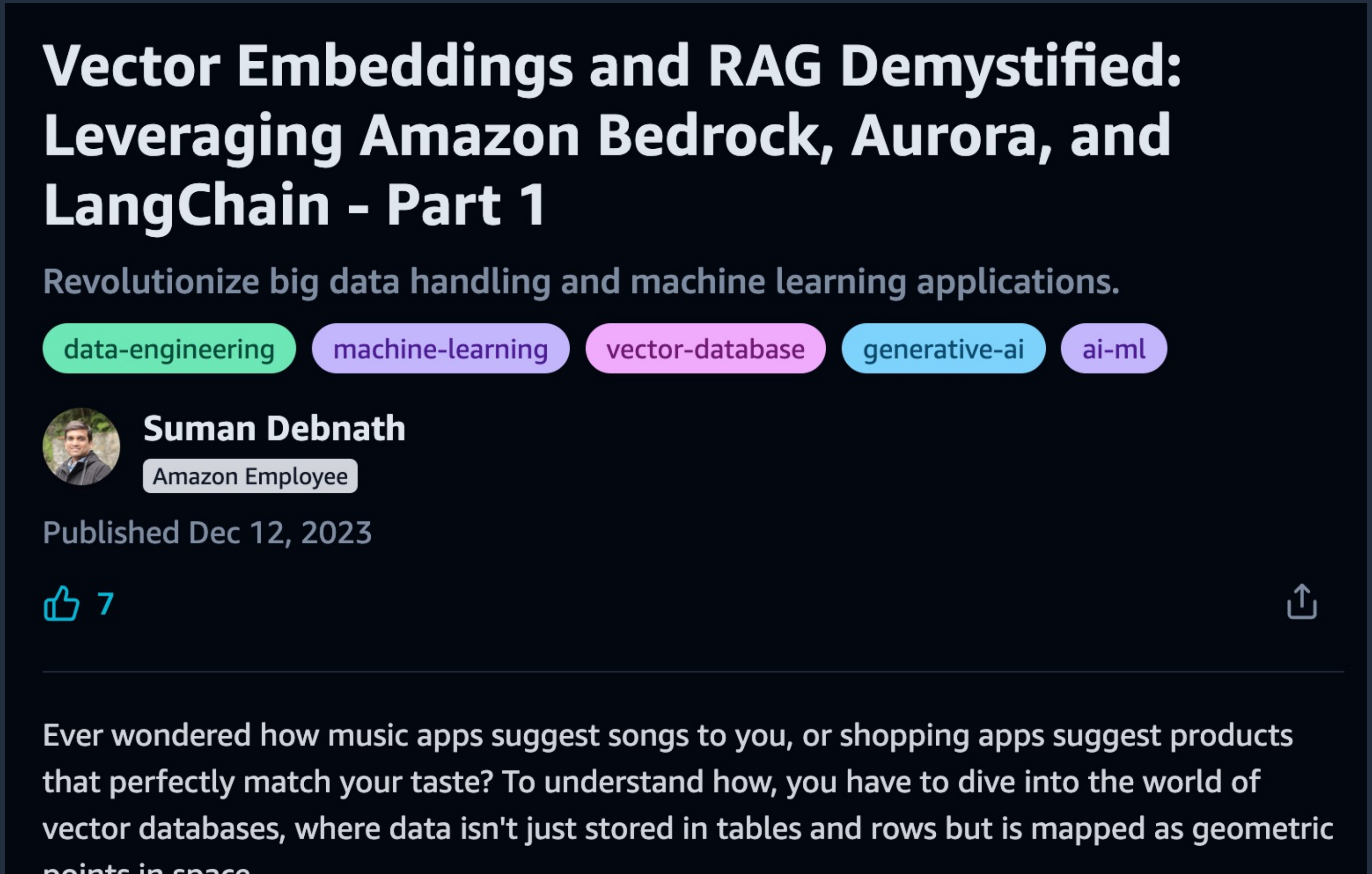


- Embedding Model =/= LLM
- Agent responds with text.
- Does not require a vector DB.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retrieval Augmented Generation

aws COMMUNITY



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
Amazon Employee

Published Dec 12, 2023

7

Ever wondered how music apps suggest songs to you, or shopping apps suggest products that perfectly match your taste? To understand how, you have to dive into the world of vector databases, where data isn't just stored in tables and rows but is mapped as geometric

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fine-tuning and Continuation training

“I have experimented with different models and different prompts, and I am not getting the generation I want.”

“Gather some examples and provide them to the model. You can do this in the prompt with in-context-learning… Or if you have enough, you can **fine-tune** the model.”

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fine-tuning and Continuation training

“I work in a specialized field, and the model does not seem to understand the terminology we use.”

“Gather as much domain specific example text that you can, and perform **continuation training** on the model.”

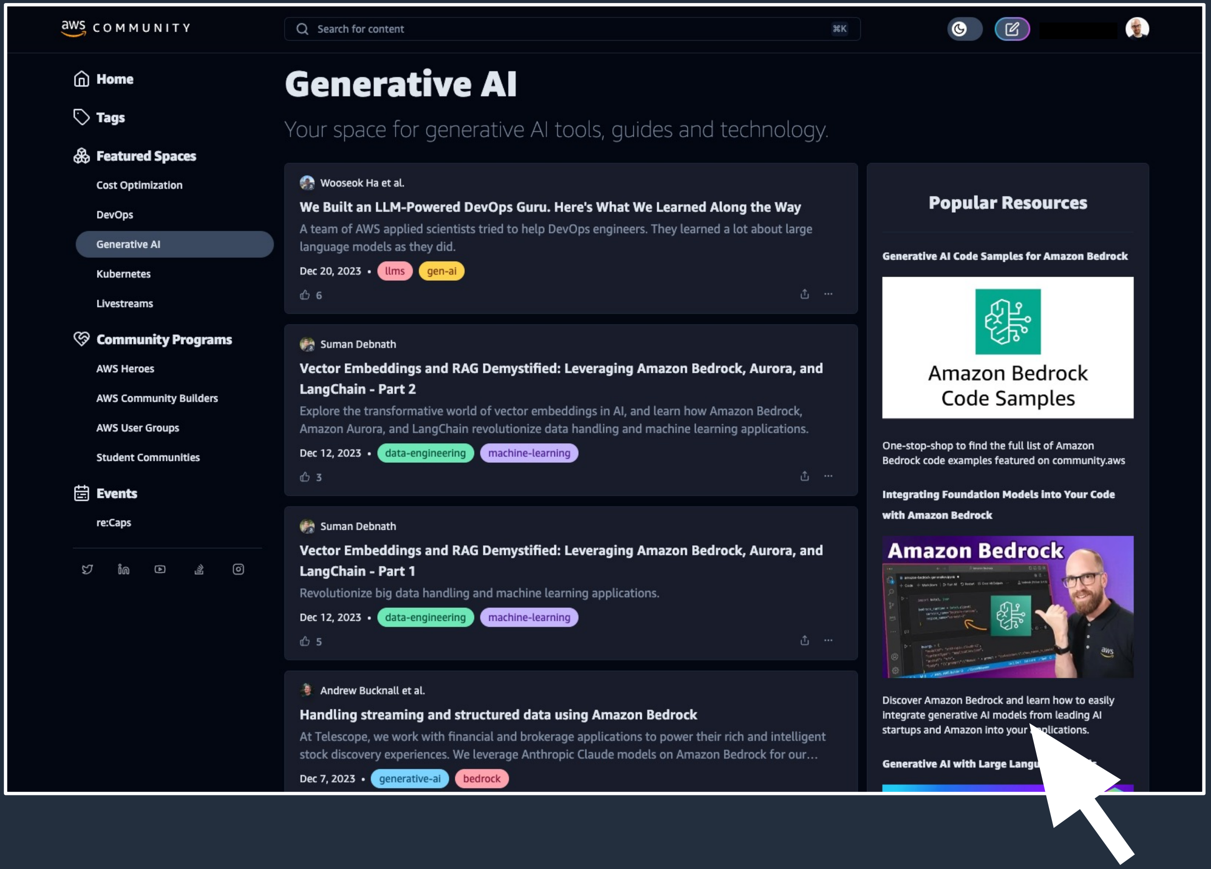
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Questions

- Please tell us what you think
- Your generative ai journey
- What should we do next?

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Find out more



Generative AI

Your space for generative AI tools, guides and technology.

Wooseok Ha et al.

We Built an LLM-Powered DevOps Guru. Here's What We Learned Along the Way

A team of AWS applied scientists tried to help DevOps engineers. They learned a lot about large language models as they did.

Dec 20, 2023 • llms gen-ai

6

Suman Debnath

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.

Dec 12, 2023 • data-engineering machine-learning

3

Suman Debnath

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

Revolutionize big data handling and machine learning applications.

Dec 12, 2023 • data-engineering machine-learning

5

Andrew Bucknall et al.

Handling streaming and structured data using Amazon Bedrock

At Telescope, we work with financial and brokerage applications to power their rich and intelligent stock discovery experiences. We leverage Anthropic Claude models on Amazon Bedrock for our…

Dec 7, 2023 • generative-ai bedrock

Popular Resources

Generative AI Code Samples for Amazon Bedrock

Amazon Bedrock Code Samples

One-stop-shop to find the full list of Amazon Bedrock code examples featured on community.aws

Integrating Foundation Models into Your Code with Amazon Bedrock

Amazon Bedrock

Discover Amazon Bedrock and learn how to easily integrate generative AI models from leading AI startups and Amazon into your applications.

Generative AI with Large Langu

aws COMMUNITY

Generative AI

Your space for generative AI tools, guides and technology.

https://community.aws/generative-ai

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

👋 thanks

linkedin.com/in/mikegchambers

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.