Experts Live Kenya

EXPERTS LIVE KENYA

26TH JULY 2024
NAIROBI, KENYA

Experts Live Kenya

Build RAG Chat App using Azure Cosmos DB and Azure OpenAI

Farah Abdirahman
Research Scientist Intern @IBM

Agenda

- Introduction
- Azure and Microsoft Technology
- Demo

Experts Live Kenya

Microsoft

Prerequisites:

- Azure subscription (aka.ms/azure4student)
- Azure OpenAI Access (aka.ms/oaiapply)
- GitHub Account

CODE

What Is RAG?

Retrieval Augmented Generation (RAG) is a technique in NLP that allows LLMs like LLMs to generate customized outputs that are outside the scope of the data it was trained on.

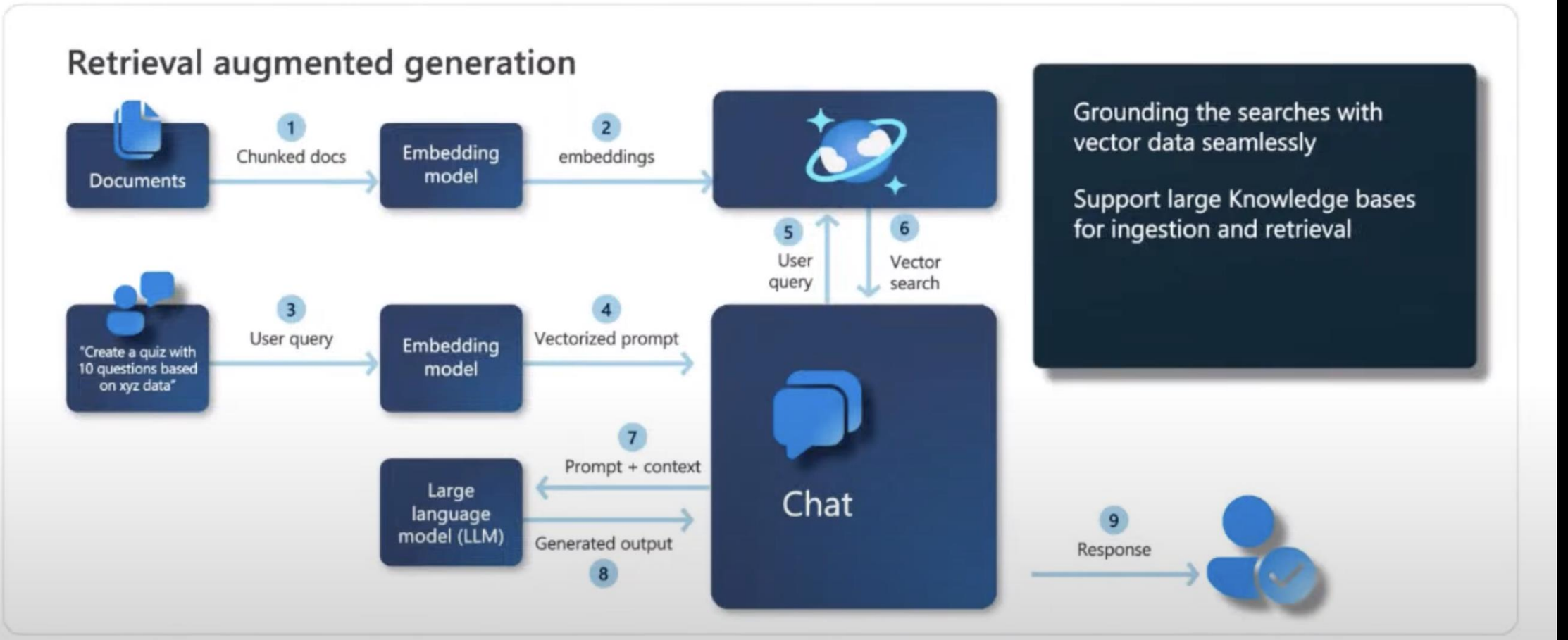
CODE

Experts Live Kenya

Lets Jump Right In

Retrieval Augmented Generation

Empower LLMs with Operational Data context



Lets Jump Right In

RAG... Why?

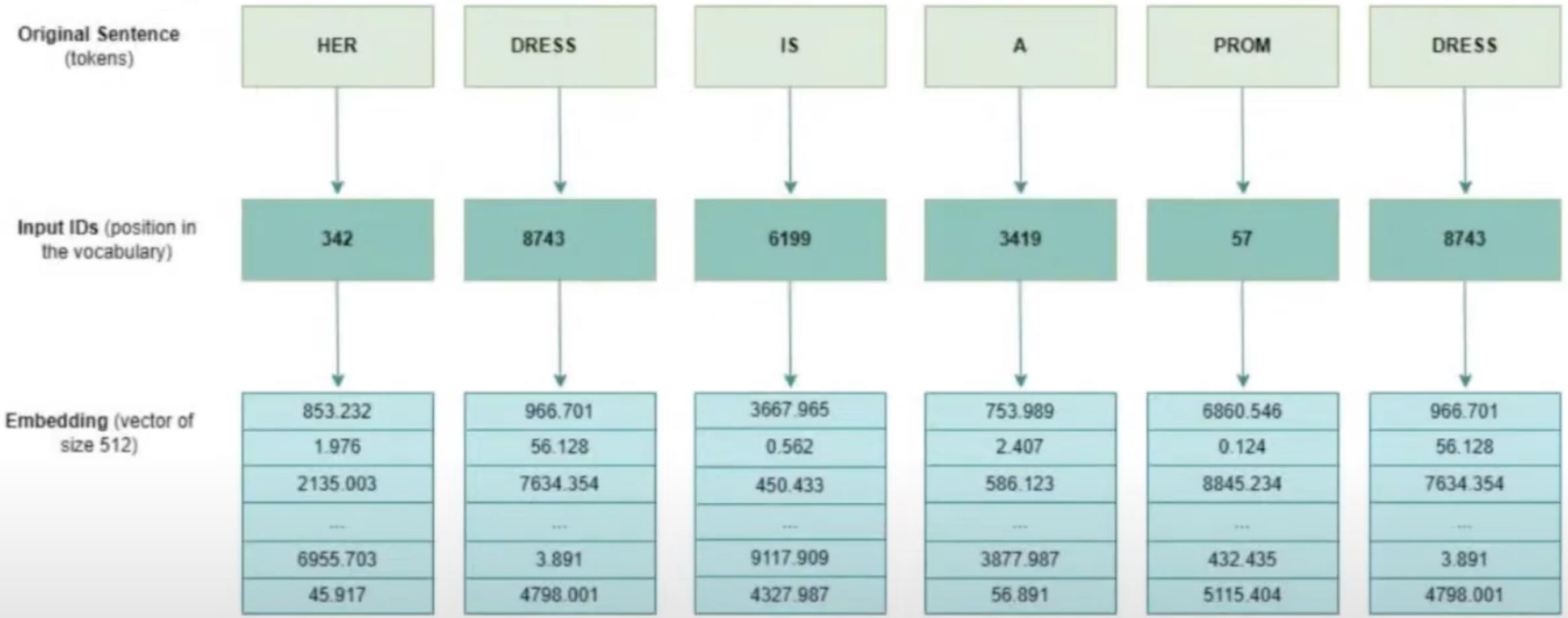
- Provide **Grounding** and **Context** for the Large Language Model.
- Overcome the outdated training data limitation.
- Low cost compared to other solutions like fine tuning.
- Supercharge data retrieval with a powerful generative model.

What Is Embeddings?

Floating point vectors that represents text or other data
They capture semantic meanings and context which results in text with similar meanings having closer embeddings

CODE

Vector Embeddings

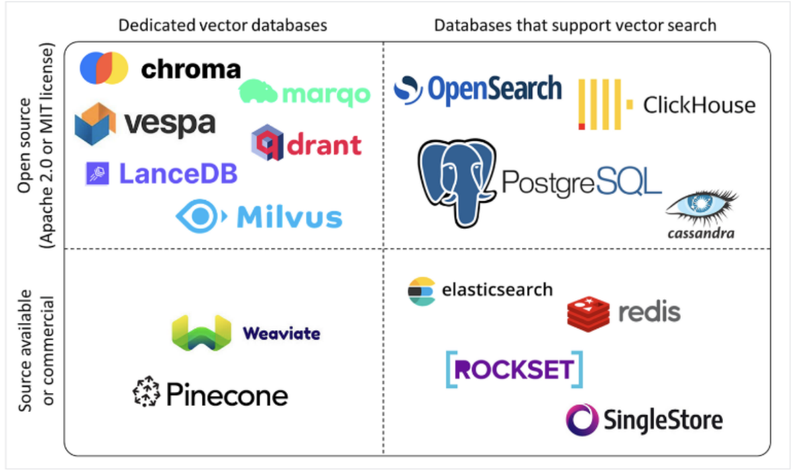


What Is Embeddings?

Vector Database

- A **vector database** is a type of database that **stores data as high-dimensional vectors**, which are mathematical representations of features or attributes.

- Each vector has a certain number of dimensions, which can range from tens to thousands, depending on the complexity and granularity of the data.



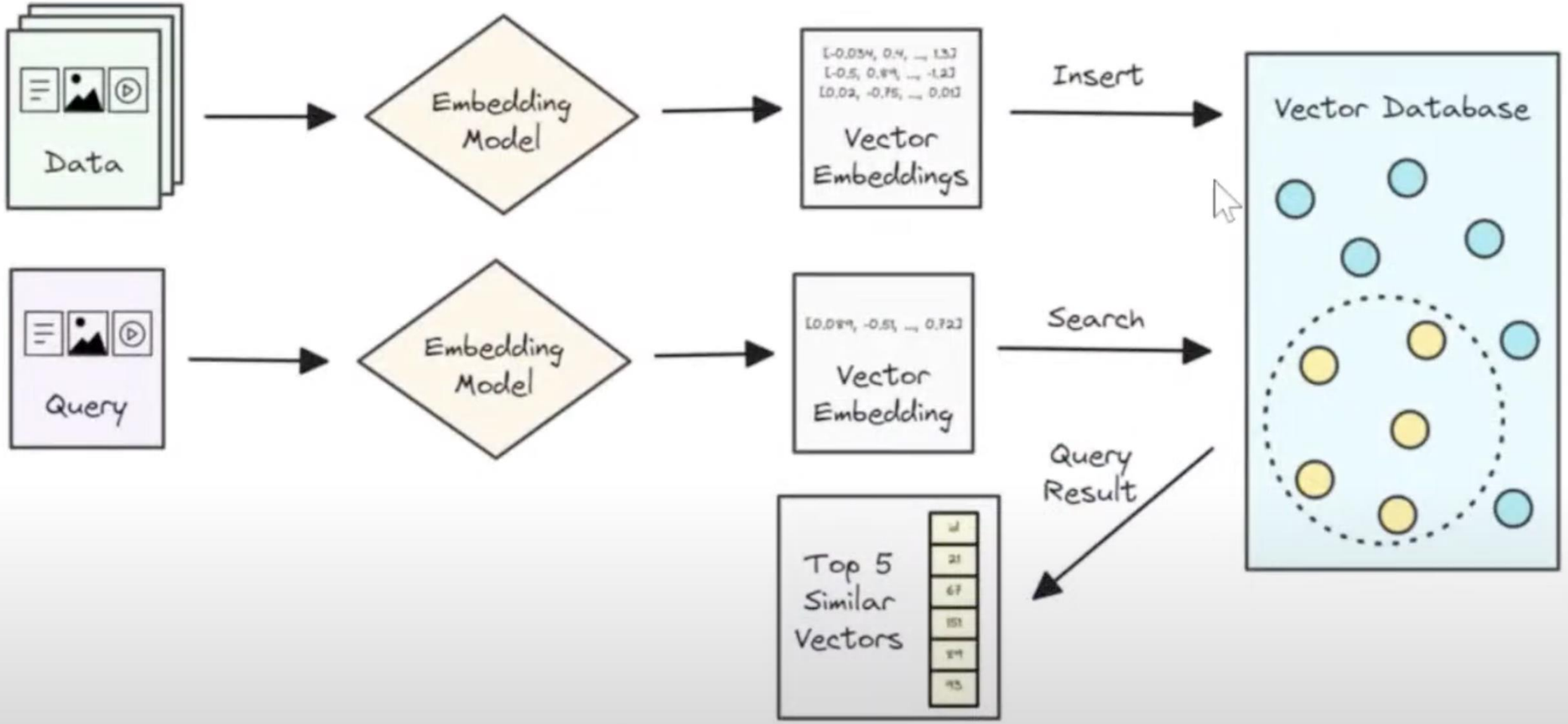
What Is Embeddings?

While both semantic and similarity search deal with finding related information, they approach it in slightly different ways:

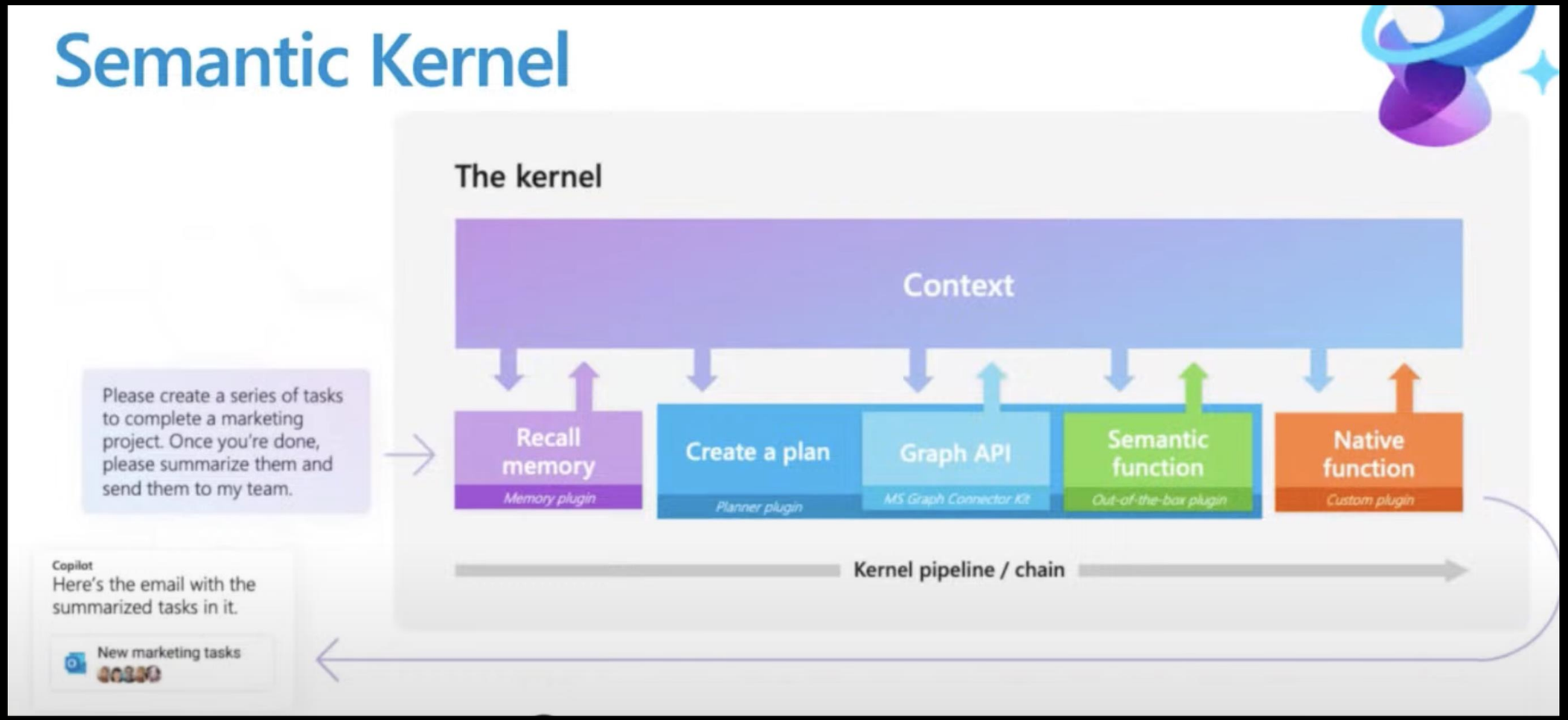
- Semantic Search: focuses on meaning and tries to understand the intent/meaning behind the query. Eg. the best laptop for a student: affordability, portability, battery life etc
- Similarity Search: focuses on similar features. Eg. picture of a cat: find similar species, similar settings etc

Objective: Semantic search aims to find relevant information, while similarity search aims to find similar items.

Vector Similarity Search Flow



Semantic Kernel



Resources

- **Blog**
  https://techcommunity.microsoft.com/t5/educator-developer-blog/build-rag-chat-app-using-azure-cosmos-db-for-mongodb-vcore-and/ba-p/4055852
- Github Repo
- https://github.com/john0isaac/rag-semantic-kernel-mongodb-vcore

Experts Live Kenya

THANK YOU

Speaker Name
Speaker Title
Speaker Social Media/links

Experts Live Kenya

THANK YOU TO OUR SPONSORS!

NTT DATA

gebeya

Microsoft

FEITIAN
WE BUILD SECURITY

ADC AFRICA DEVELOPMENT CENTER

USIU AFRICA
United States International University-Africa
Education to take you places

2LINKIT

Computer Revolution Africa Group
Technology redefined

NiCE
IT Management Solutions