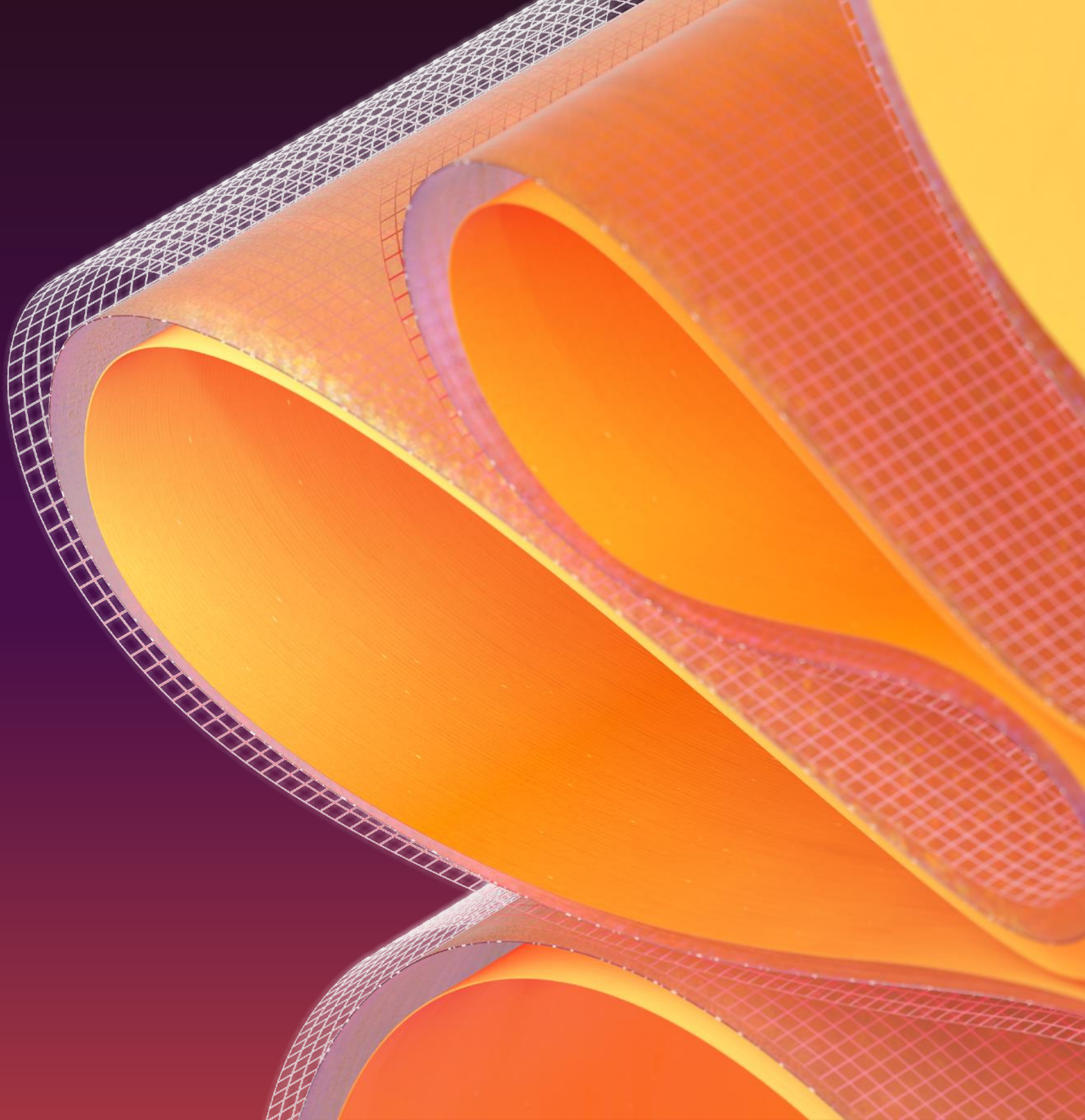
Microsoft

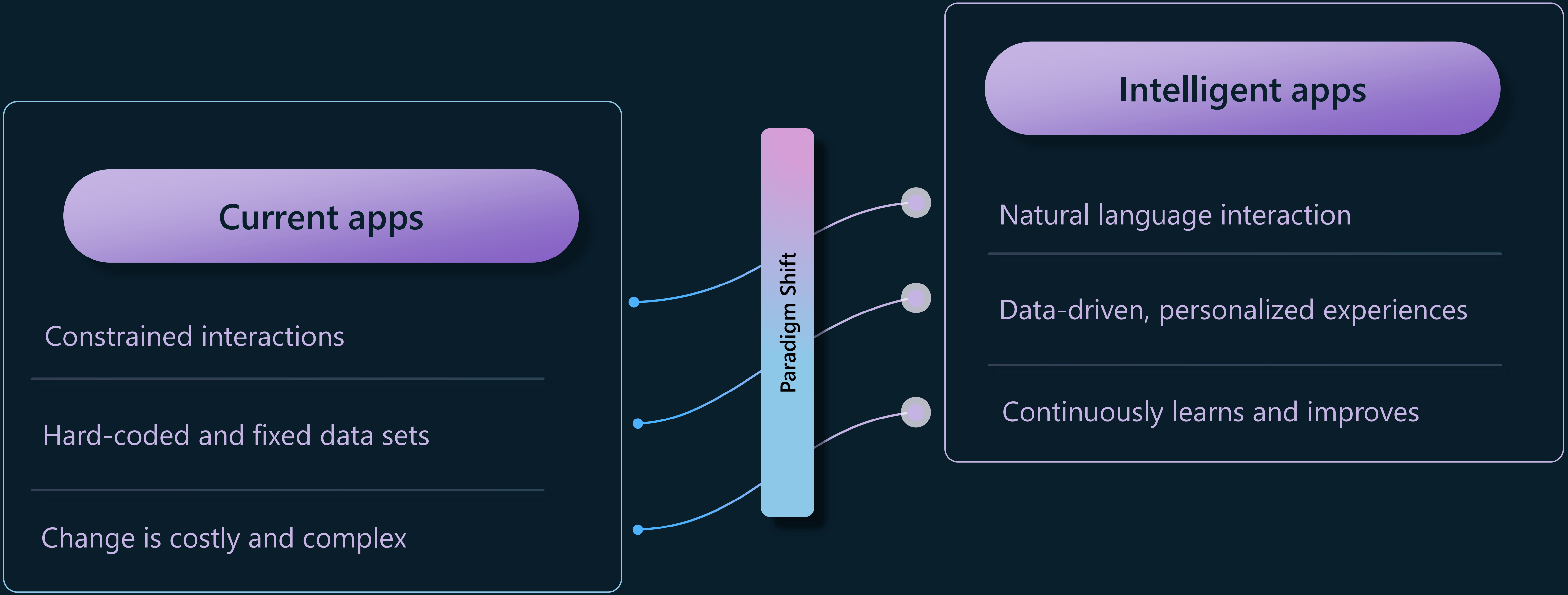
Build Your Own Copilot on Azure

Build and modernize intelligent apps on Azure

October 5, 2023



Generative AI makes apps truly intelligent



Current apps

- Constrained interactions
- Hard-coded and fixed data sets
- Change is costly and complex

Paradigm Shift

Intelligent apps

- Natural language interaction
- Data-driven, personalized experiences
- Continuously learns and improves

Essential elements of intelligent applications

Pre-trained models

State of the art pre-trained AI models that are easy to discover, customize, and integrate into new and existing enterprise applications.

Scalability and high performance

Ability to handle high volumes of unstructured data, in real time, from disparate sources

App platform that can scale based on the app's demand and ensure reliable performance.

Simplified app delivery

Developer-ready environments to ship apps securely, and quickly in their language of choice.

Enable frequent iteration by streamlining costly and time-consuming app delivery.

A common platform with the technology you need

Application Services

- Azure Kubernetes Service
- Azure App Service
- Azure Container Apps
- Azure Functions

+

AI portfolio

- Azure AI Services (including Azure OpenAI Service)
- Azure AI Studio

+

Managed databases

- Azure SQL family
- Azure PostgreSQL
- Azure Cosmos DB

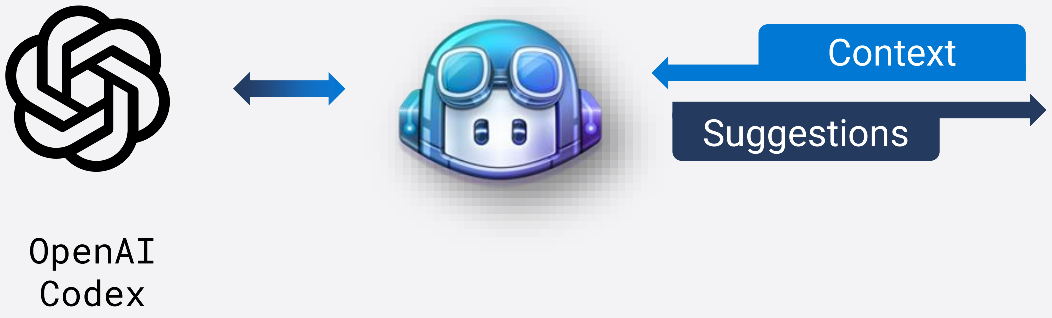
What is a copilot?

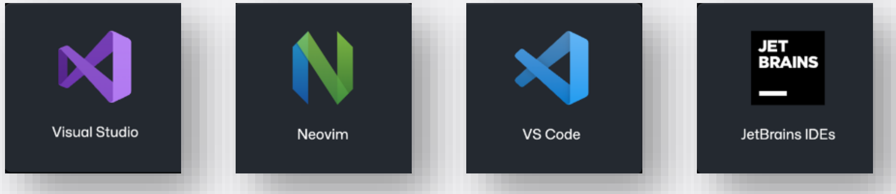
Copilots are intelligent apps that enable the use of natural language to find better and more relevant answers to questions.

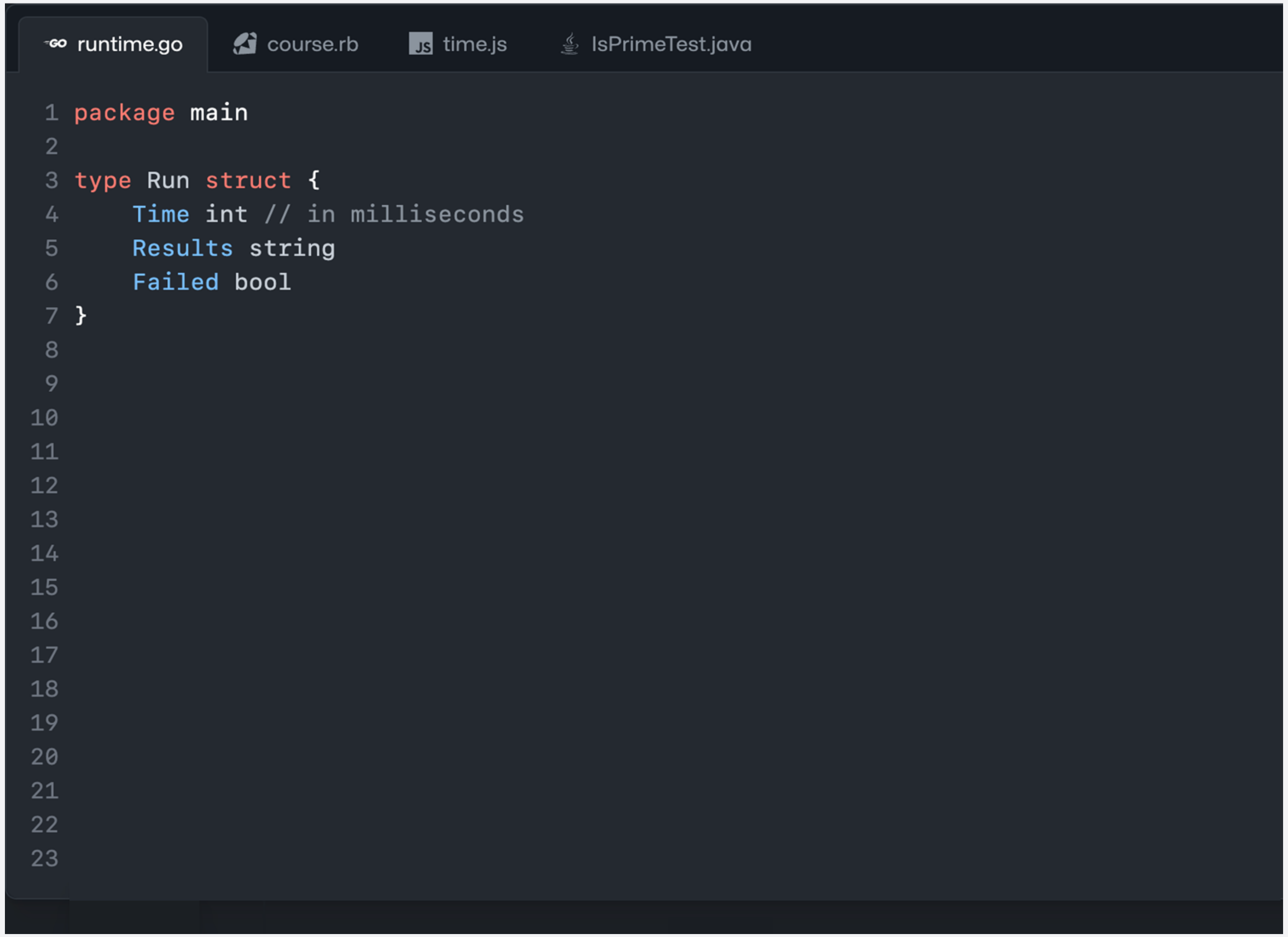
Streamline the developer experience

GitHub Copilot

Development: Faster, more productive, and satisfying

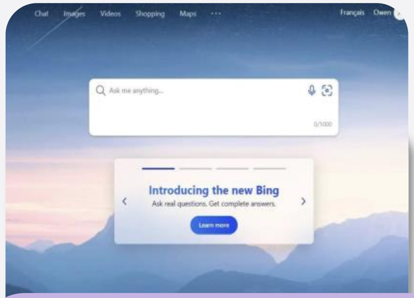


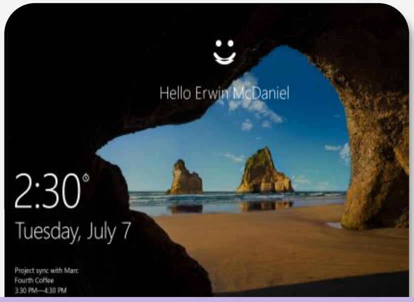


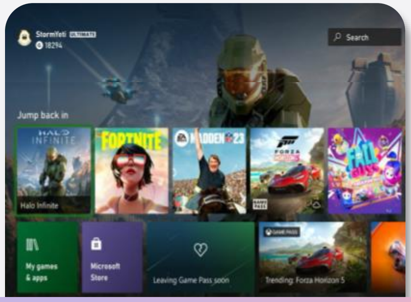


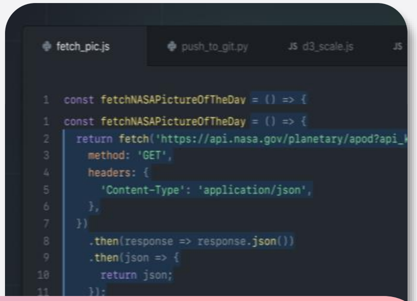
```go
package main

type Run struct {
    Time int // in milliseconds
    Results string
    Failed bool
}
```

Microsoft copilots offer differentiation









```
const fetchNASAPictureOfTheDay = () => {
const fetchNASAPictureOfTheDay = () => {
  return fetch('https://api.nasa.gov/planetary/apod?api_
    method: 'GET',
    headers: {
      'Content-Type': 'application/json',
    },
  })
    .then(response => response.json())
    .then(json => {
      return json;
    });
```

Microsoft modernized and augmented flagship products with copilots

AI-based search with ChatGPT

Bing

Copilot for Work across Office 365

Office

Biometric identity verification

Windows

Personalized recommendations

XBOX

Copilot assisted coding

GitHub

Build your own copilot: solution architecture

What are we building?

User experience

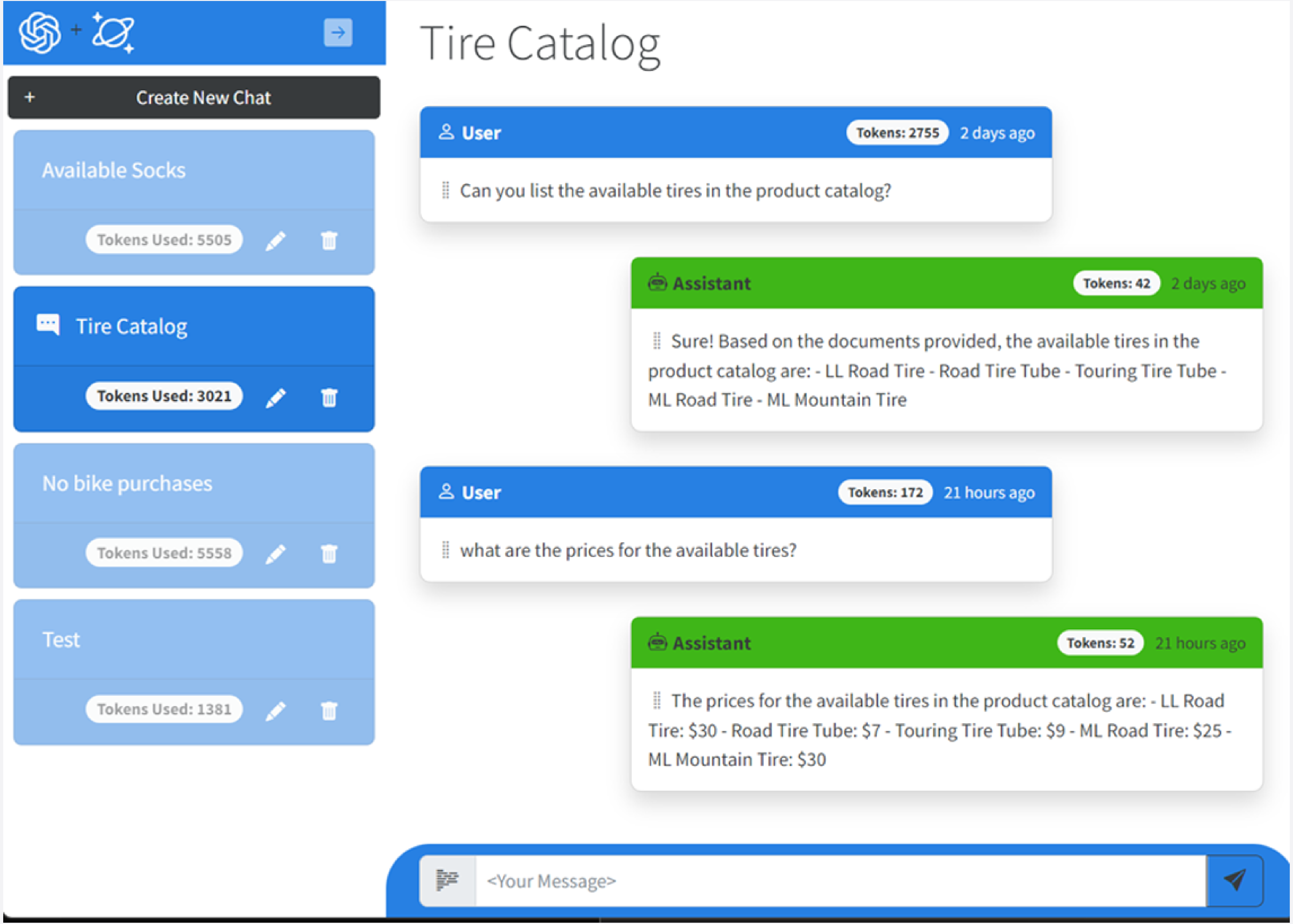
- ChatGPT-like interface
- Find product information from the inventory data of a fake retailer
- Users interact by asking questions and having a conversation

Tech requirements

- Fast performance
- Connect to existing business data (product, customer, order, etc)
- Manage conversational context
- Manage and store search result history
- Build and leverage custom analytics

Demo and Code

https://aka.ms/2023DriveInConf



Tire Catalog

User Tokens: 2755 2 days ago

Can you list the available tires in the product catalog?

Assistant Tokens: 42 2 days ago

Sure! Based on the documents provided, the available tires in the product catalog are: - LL Road Tire - Road Tire Tube - Touring Tire Tube - ML Road Tire - ML Mountain Tire

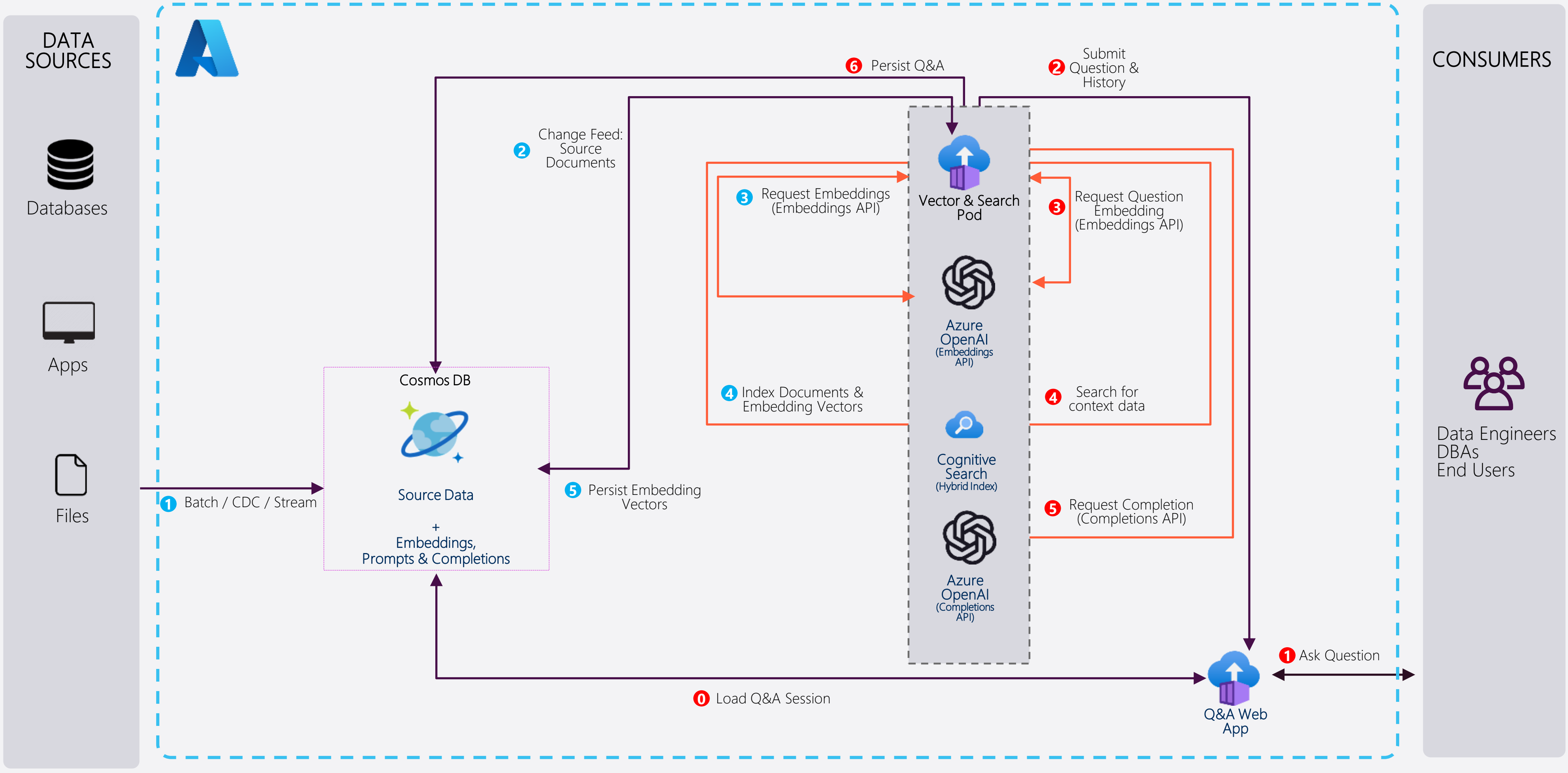
User Tokens: 172 21 hours ago

what are the prices for the available tires?

Assistant Tokens: 52 21 hours ago

The prices for the available tires in the product catalog are: - LL Road Tire: $30 - Road Tire Tube: $7 - Touring Tire Tube: $9 - ML Road Tire: $25 - ML Mountain Tire: $30

Solution architecture: BYO copilot



DATA SOURCES
Databases
Apps
Files
1 Batch / CDC / Stream
Cosmos DB
Source Data
+
Embeddings,
Prompts & Completions
2 Change Feed: Source Documents
5 Persist Embedding Vectors
6 Persist Q&A
3 Request Embeddings (Embeddings API)
4 Index Documents & Embedding Vectors
Vector & Search Pod
Azure OpenAI (Embeddings API)
Cognitive Search (Hybrid Index)
Azure OpenAI (Completions API)
2 Submit Question & History
3 Request Question Embedding (Embeddings API)
4 Search for context data
5 Request Completion (Completions API)
0 Load Q&A Session
1 Ask Question
Q&A Web App
CONSUMERS
Data Engineers
DBAs
End Users

Legend: Source Data — Semantic Kernel Orchestrated Call — Other Service Call — User Interaction

1 Source data loading

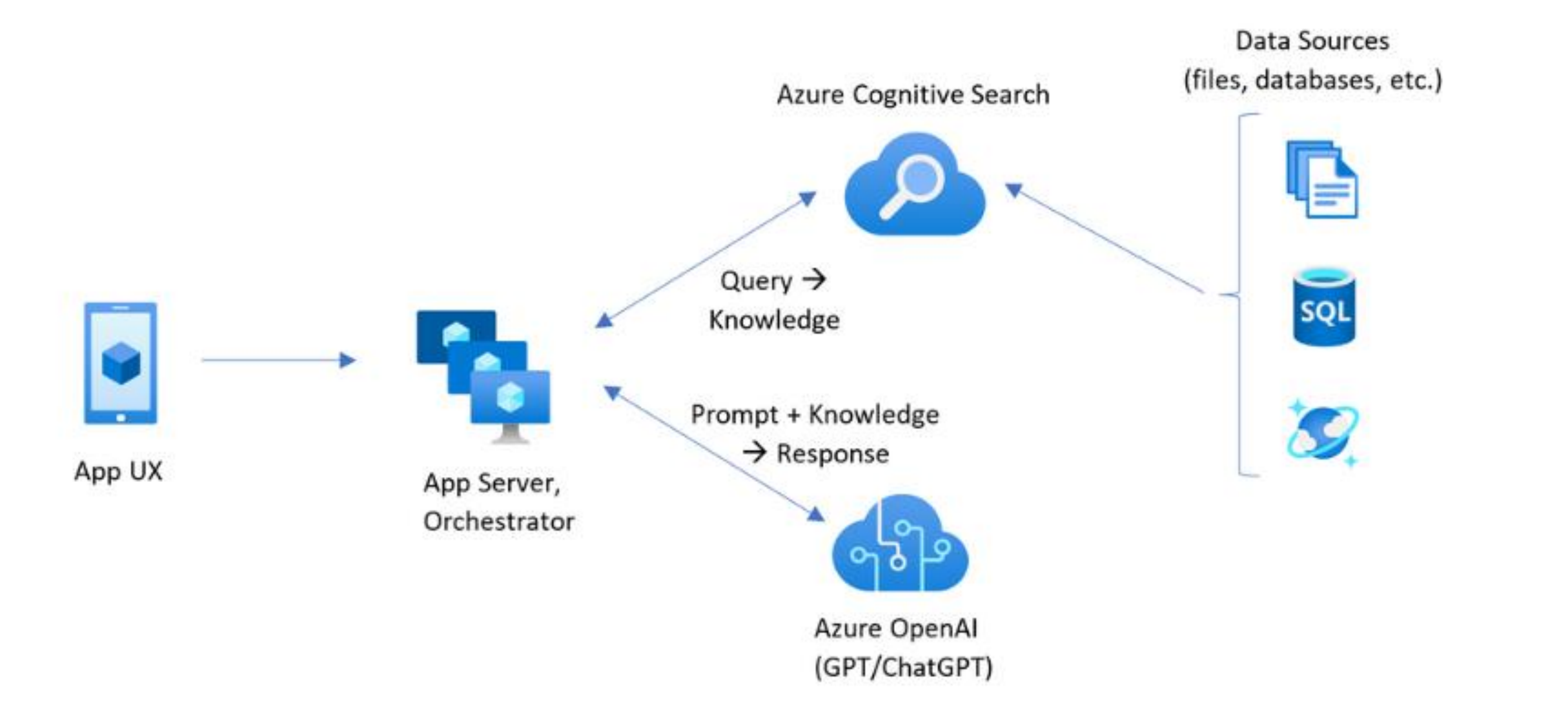
1 Q&A interaction

Hmmmmm?

The question?

How is it that a LLM trained on data and fixed in time,
know about my corporate data?

Retrieval Augmented Generation (RAG)



Architecture that augments the capabilities LLM adding an information retrieval system that provides the data.

Hmmmmm v 2.0?

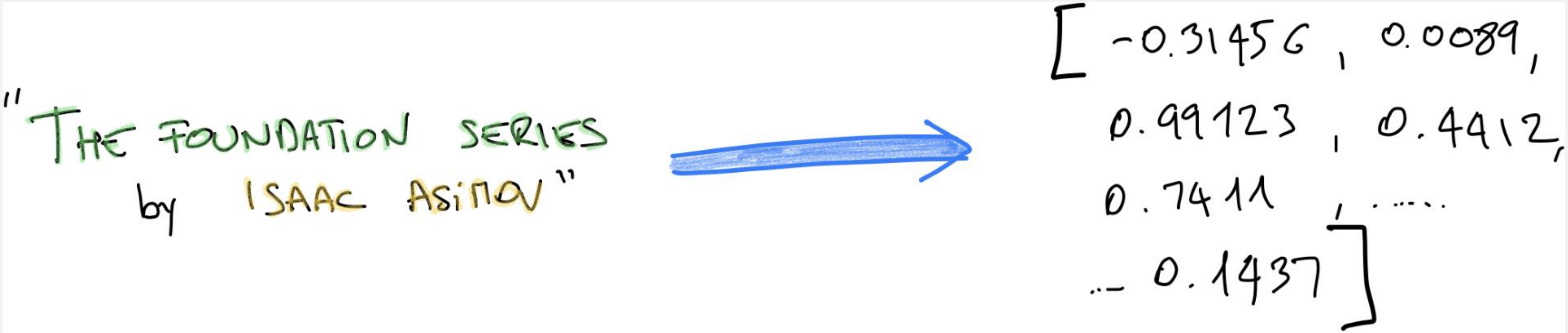
The question?

Prompts are limited in size. I can’t supply all my data in the prompt.
What’s the secret?

Prompt engineering

1. Tokens
2. Embeddings
3. Vectors

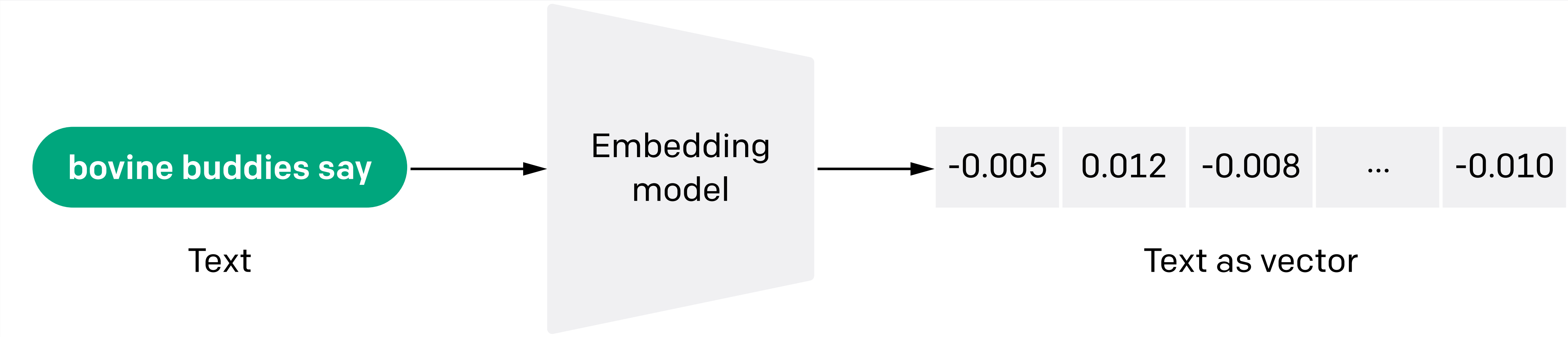
Tokens



Tokens are the basic units of text or code that an LLM AI uses to process and generate language.

OpenAI and Azure OpenAI uses a subword tokenization method called "Byte-Pair Encoding (BPE)" for its GPT-based models.

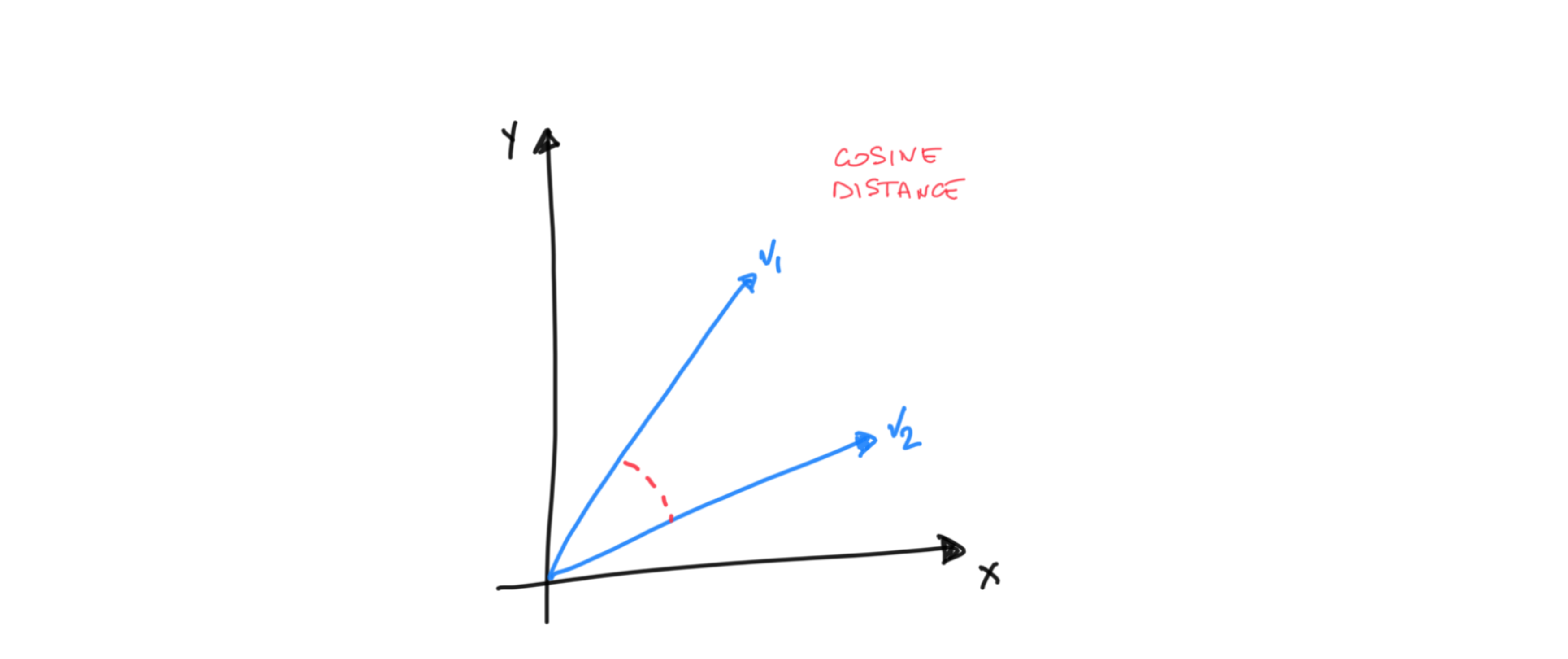
Embeddings



An embedding is a special format of data representation that machine learning models and algorithms can easily use.

The embedding is an information dense representation of the semantic meaning of a piece of text.

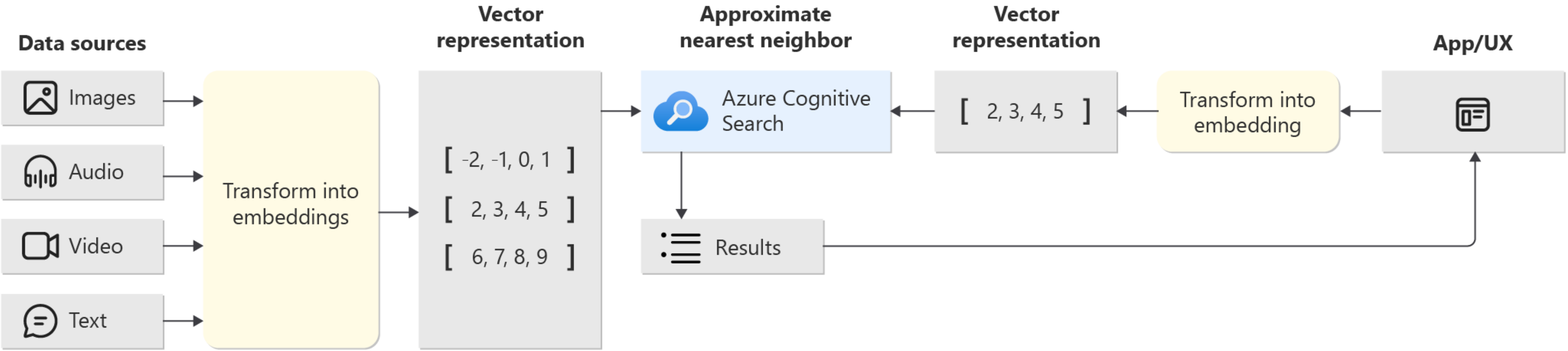
Vectors



More specifically, embeddings are vectors…hence the great interest for vector databases.

Vectors represent similar object is as easy as calculating the distance between the vectors.

Vector Search



Indexing, storing, and retrieving vector embeddings from a search index. You can use it to power similarity search, multi-modal search, recommendations engines, or applications implementing the Retrieval Augmented Generation (RAG) architecture.

Text similarity models

Embeddings from the text-similarity-babbage-001 model, applied to the DBpedia dataset.

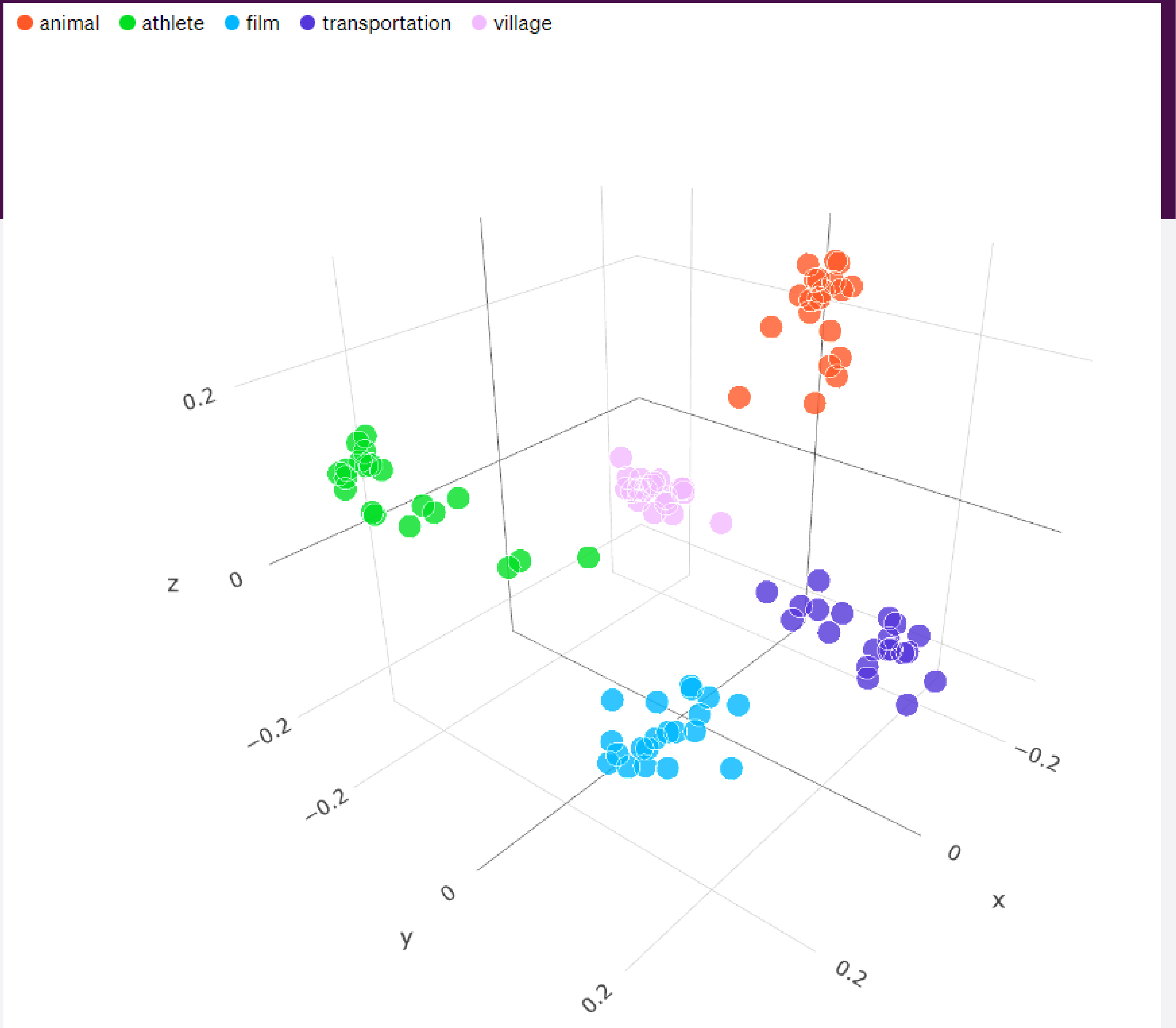
We randomly selected 100 samples from the dataset covering 5 categories and computed the embeddings via the /embeddings endpoint.

The different categories show up as 5 clear clusters in the embedding space.

To visualize the embedding space, we reduced the embedding dimensionality from 2048 to 3 using PCA.

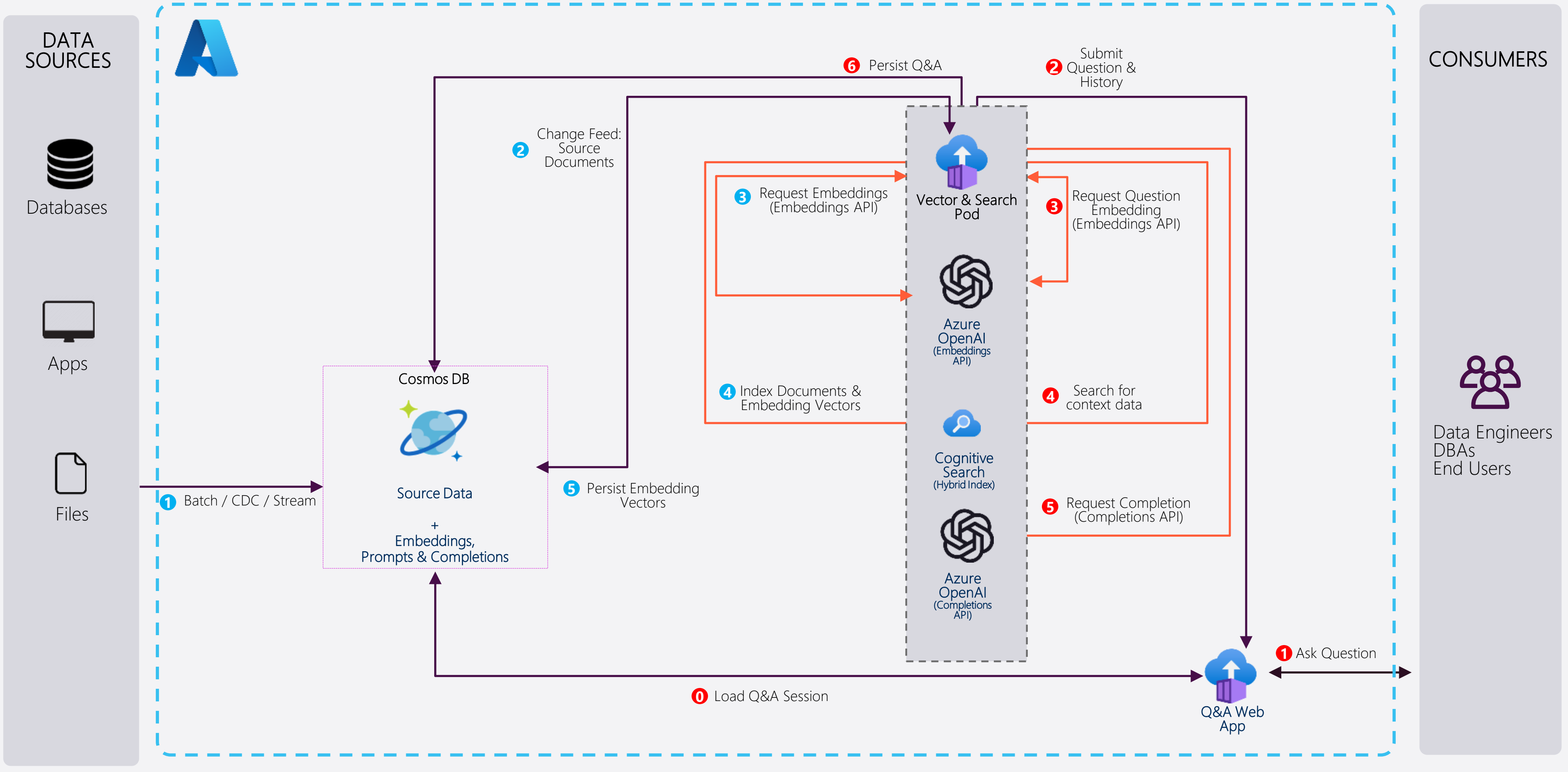
The code for how to visualize embedding space in 3D dimension is available here.

https://openai.com/blog/introducing-text-and-code-embeddings



Principal Component Analysis - Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space.

Solution architecture: BYO copilot


DATA SOURCES
Databases
Apps
Files
1 Batch / CDC / Stream
Cosmos DB
Source Data
+
Embeddings,
Prompts & Completions
2 Change Feed: Source Documents
5 Persist Embedding Vectors
6 Persist Q&A
2 Submit Question & History
3 Request Embeddings (Embeddings API)
4 Index Documents & Embedding Vectors
Vector & Search Pod
Azure OpenAI (Embeddings API)
Cognitive Search (Hybrid Index)
Azure OpenAI (Completions API)
3 Request Question Embedding (Embeddings API)
4 Search for context data
5 Request Completion (Completions API)
0 Load Q&A Session
1 Ask Question
Q&A Web App
CONSUMERS
Data Engineers
DBAs
End Users

Legend: Source Data — Semantic Kernel Orchestrated Call — Other Service Call — User Interaction

1 Source data loading

1 Q&A interaction

Solution architecture: BYO copilot hero products

Azure AI Services

(Azure OpenAI Service & Azure Cognitive Search)

- Generates ChatGPT responses to natural language questions
- Creates embeddings that describe similarities between data
- Indexes data to make it more easily searchable

Azure Cosmos DB

- Stores transactional data to be queried
- Generates vectors on the stored data
- Stores user prompt and completion history
- Enables conversational memory: follow-up questions and user conversations
- Can dynamically add and remove data to enable real-time AI

Azure Container Apps

Azure Kubernetes Service

- Receives questions, and routes to Azure OpenAI Service
- Initiates searches for contextual data

Microsoft

Thank you