Microsoft Reactor

Build a Production-Scale Conversation AI Assistant With RAG

Cihan Cinar
Software & AI Architect

Microsoft

Reactor

Cihan Cinar

Software & AI Architect

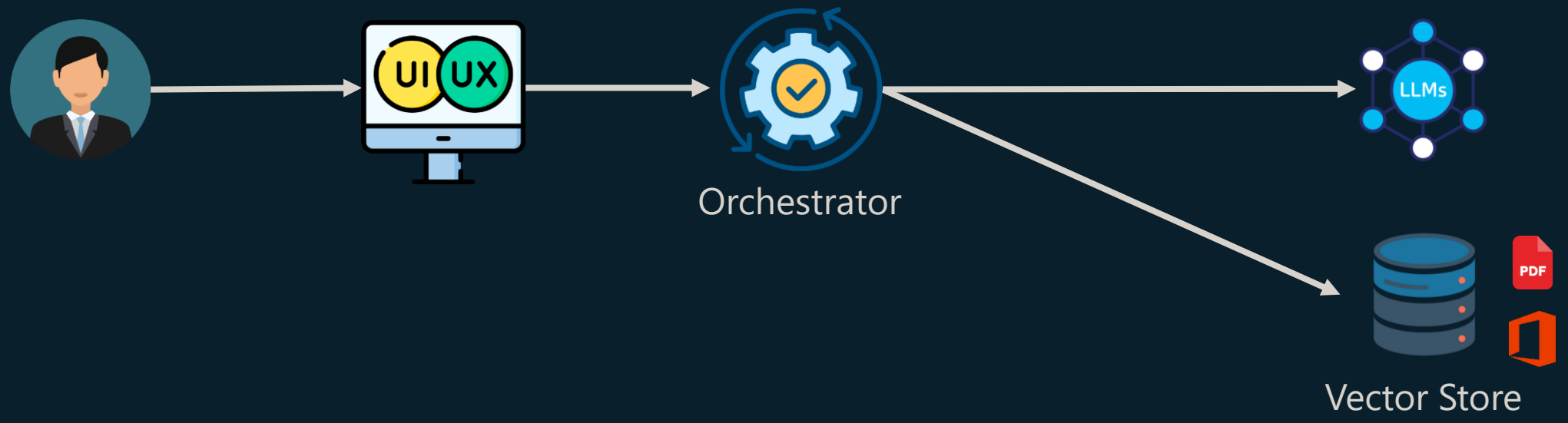
Givaudan

Linkedin: @cihancinar

Github: @cihancinar

https://blog.cihan-cinar.com/

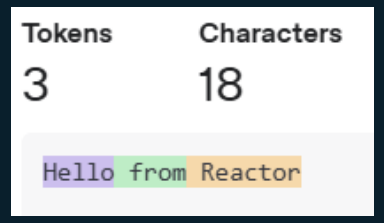
Build a Production-Scale Conversation AI Assistant With RAG



Microsoft Reactor

Chat Completion

- conversation-in and message-out
- LLM based on token



- Context window
  - 4k input & 4k output tokens for common models
  - 128k+ input & 16k+ output tokens for newer models

Microsoft
Reactor

Chat Completion

Request

```
POST
https://{endpoint}/openai/deployments/{deployment-id}/chat/completions?api-version=2024-10-21
```

```
{
    "model": "gpt-4o",
    "messages": [
      {
        "role": "system",
        "content": "you are a helpful assistant that talks like a pirate. Generate short answer only."
      },
      {
        "role": "user",
        "content": "can you tell me how to care for a parrot?"
      }
    ],
    "temperature": 0.7,
    "stream": false,
    "max_tokens": 2000
  }
```

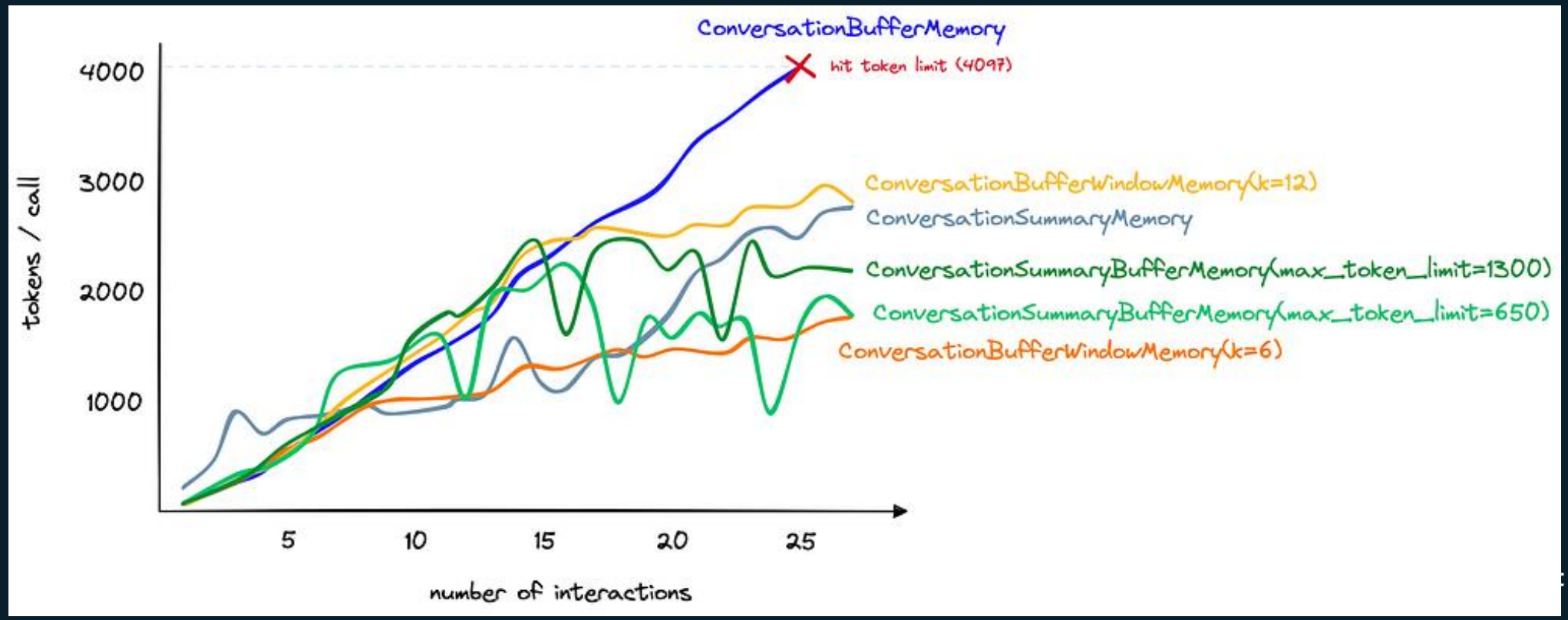
Response

```
{
    "choices": [
      {
        "finish_reason": "stop",
        "index": 0,
        "message": {
          "content": "Aye! Provide fresh food n' water, a roomy cage, toys fer enrichment, regular vet check-ups, and plenty o' social interaction. Arr!",
          "role": "assistant"
        }
      }
    ],
    "created": 1740803441,
    "id": "chatcmpl-B68mnFNQafMCjwosWz8zPS1ZtcRVO",
    "model": "gpt-4o-2024-05-13",
    "object": "chat.completion",
    "system_fingerprint": "fp_65792305e4",
    "usage": {
      "completion_tokens": 32,
      "prompt_tokens": 39,
      "total_tokens": 71
    }
  }
```

Microsoft Reactor

How to Manage Conversation Memory

- With Conversational Buffer or Summary



source

Reactor

Create Image

Request

```
POST
https://{endpoint}/openai/deployments/{deployment-id}/image/generations?api-version=2024-02-01
```

```
{
    "prompt": "A multi-colored umbrella on the beach, disposable camera",
    "size": "1024x1024",
    "n": 1,
    "quality": "standard",
    "style": "vivid",
    "response_format": "b64_json"
  }
```

Response

```
{
    "created": 1740805367,
    "data": [
      {
        "content_filter_results": { ... },
        "prompt_filter_results": { ... },
        "revised_prompt": "A multi-colored umbrella situated on a sandy beach with its vibrant colors contrasting the blue sky. Nearby on the sand, there's a disposable camera ready to capture beautiful memories.",
        "url":
"https://dalleprodsec.blob.core.windows.net/private/images/1352de45-3e6d-418e-b6f1-0d4ed592a78c/generated_00.png?se=2025-03-02T05%3A02%3A57Z&sig=H5eZKnOOOA%2B1lYFpZYwnDGhVgVgA7JwSG0Q76ZvTuvA%3D&ske=2025-03-07T13%3A56%3A15Z&skoid=e52d5ed7-0657-4f62-bc12-7e5dbb260a96&sks=b&skt=2025-02-28T13%3A56%3A15Z&sktid=33e01921-4d64-4f8c-a055-5bdaffd5e33d&skv=2020-10-02&sp=r&spr=https&sr=b&sv=2020-10-02"
      }
    ]
  }
```

Microsoft

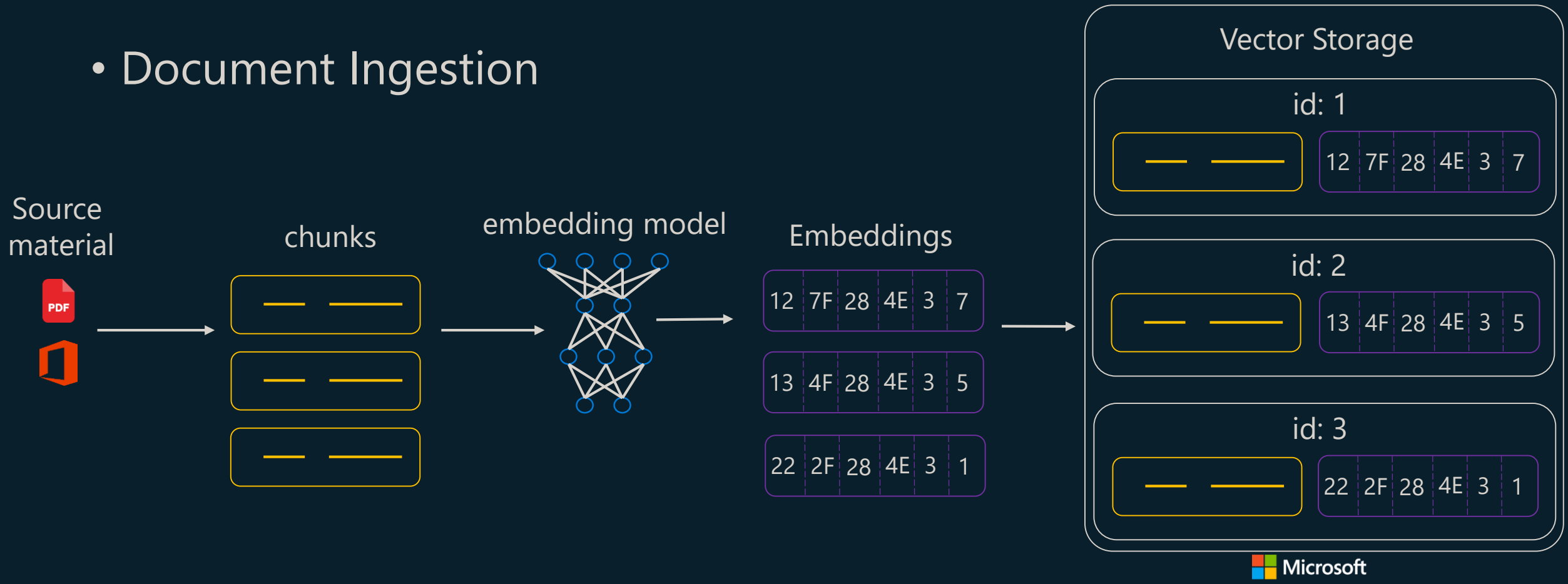
Reactor

How RAG Works (Retrieval-Augmented Generation)

- Document ingestion
- Transform, the user query into a vector
- Search in vector DB for relevant documents
- Add found documents to the context

RAG (Retrieval-Augmented Generation)

- Document Ingestion

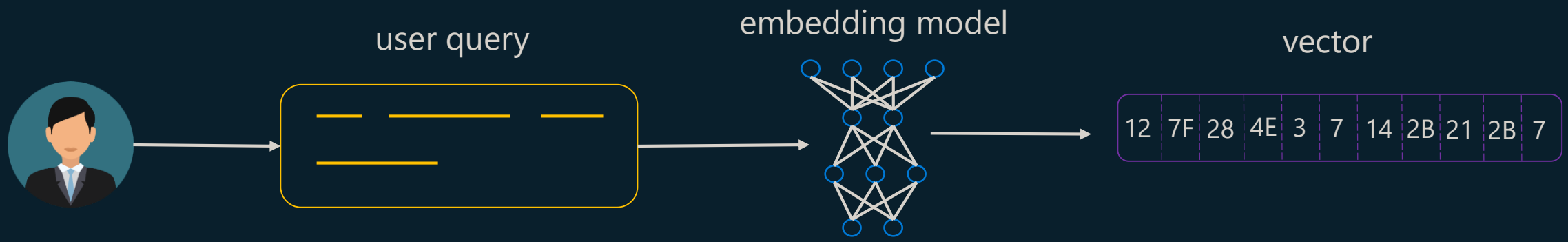


Microsoft

Reactor

RAG (Retrieval-Augmented Generation)

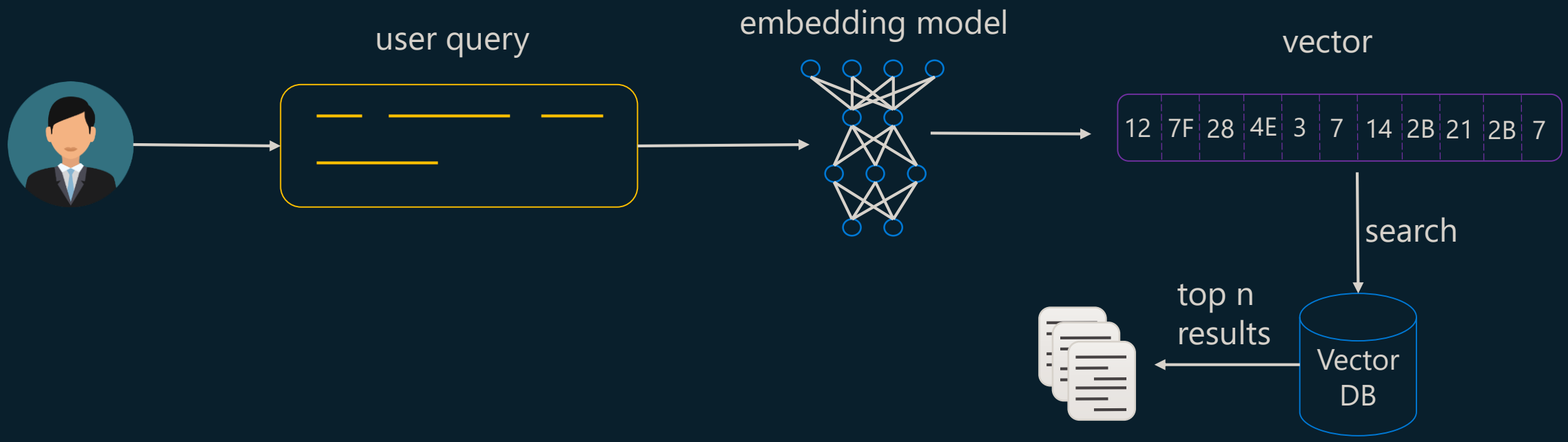
- Transform, the user query into a vector



Microsoft Reactor

RAG (Retrieval-Augmented Generation)

- Search in vector DB for relevant documents

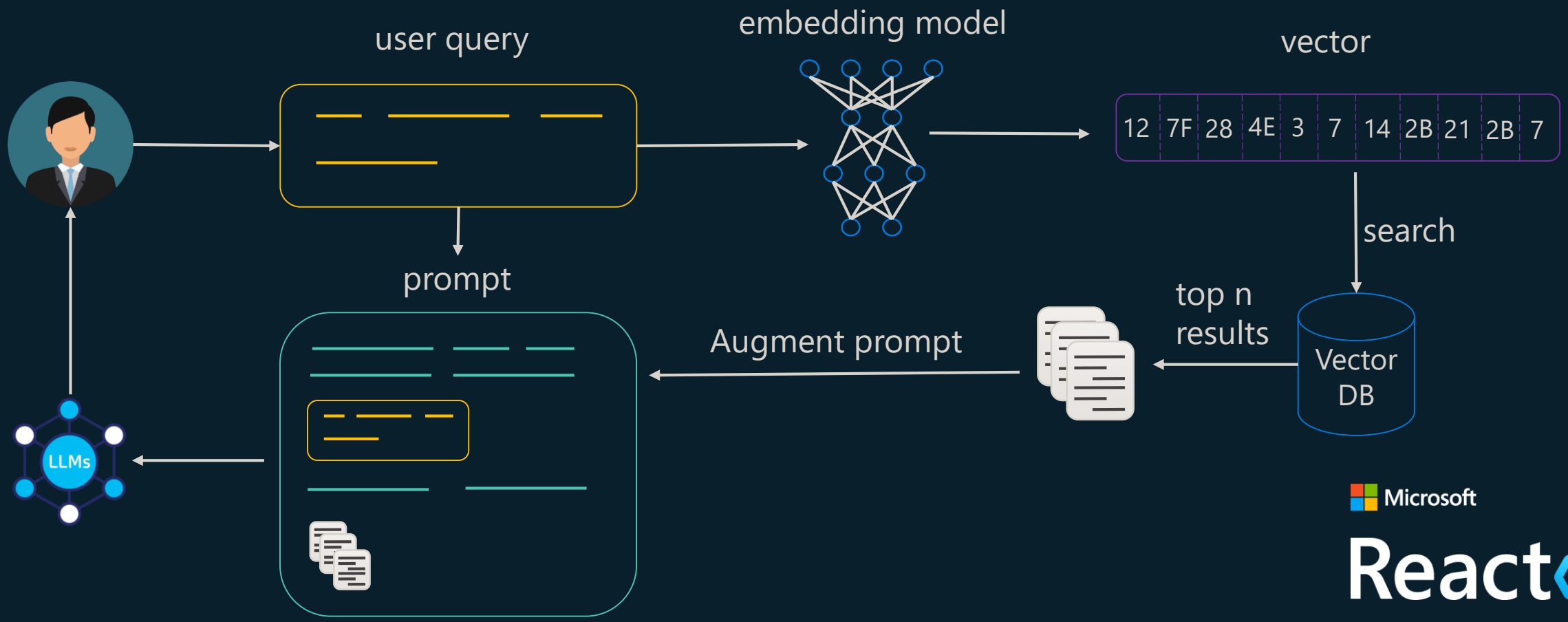


Microsoft

Reactor

RAG (Retrieval-Augmented Generation)

- Add found documents to the context

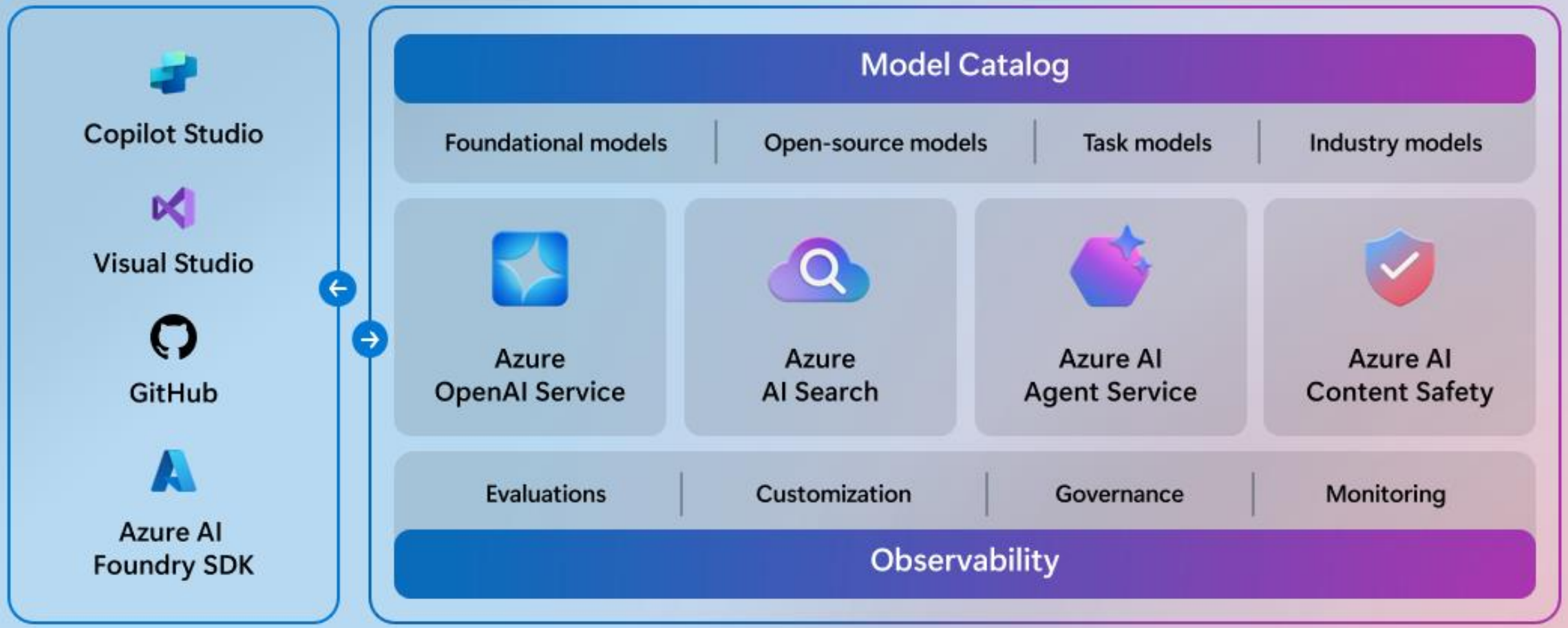


Microsoft

Reactor

Announcing

Azure AI Foundry



Microsoft
Reactor

Demo

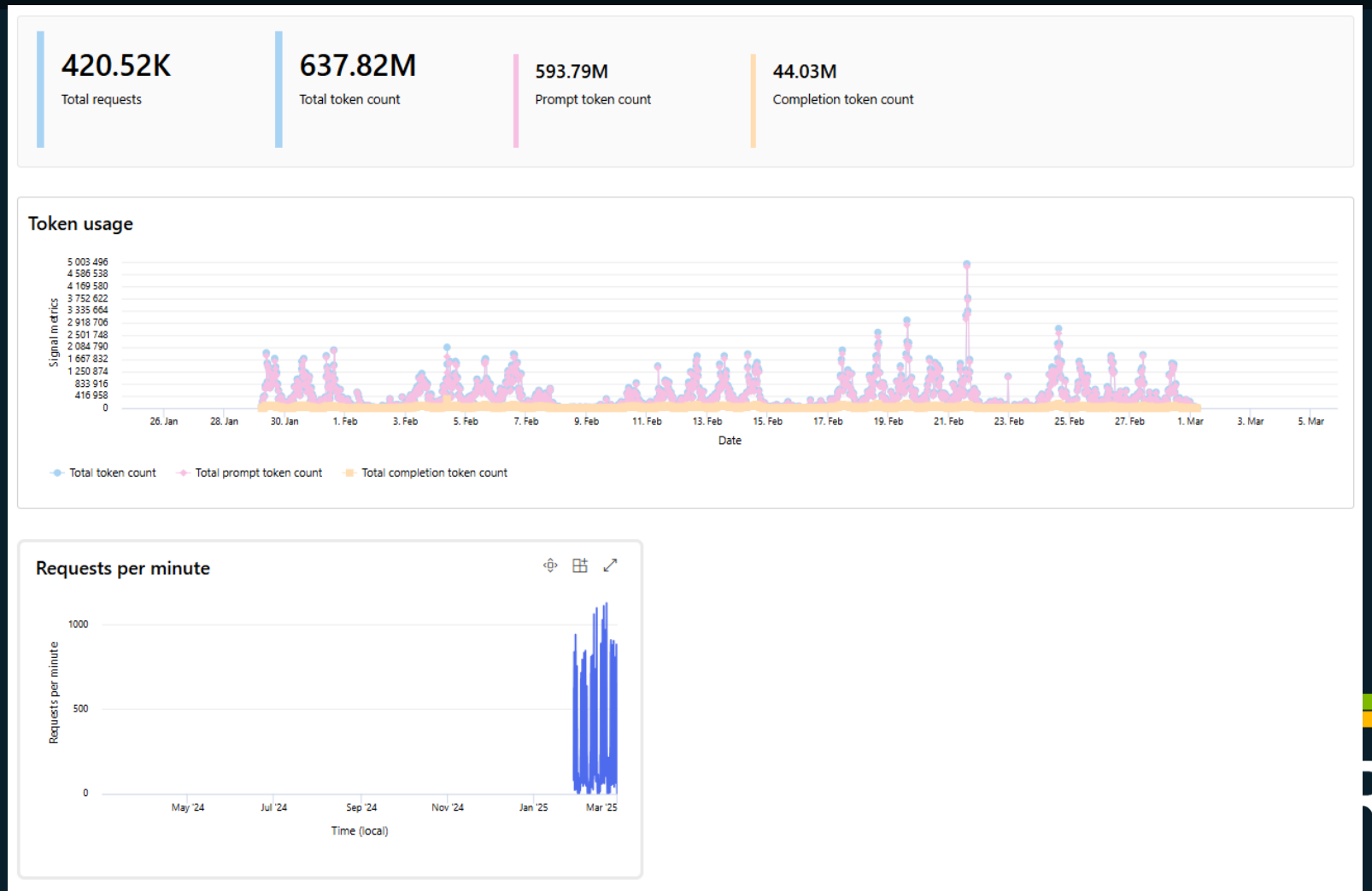
Azure

Microsoft Reactor

Demo

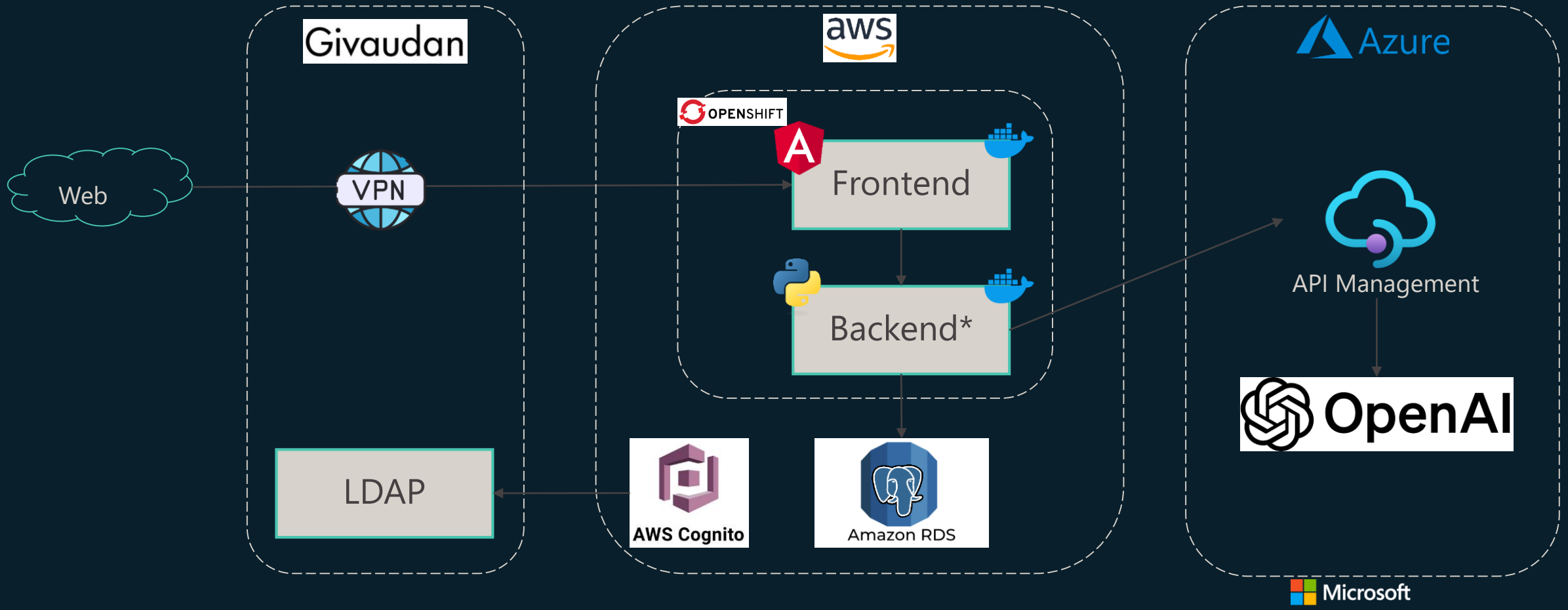
Alfred – Givaudan

Alfred Demo – Azure OpenAI Monitor



Microsoft Reactor

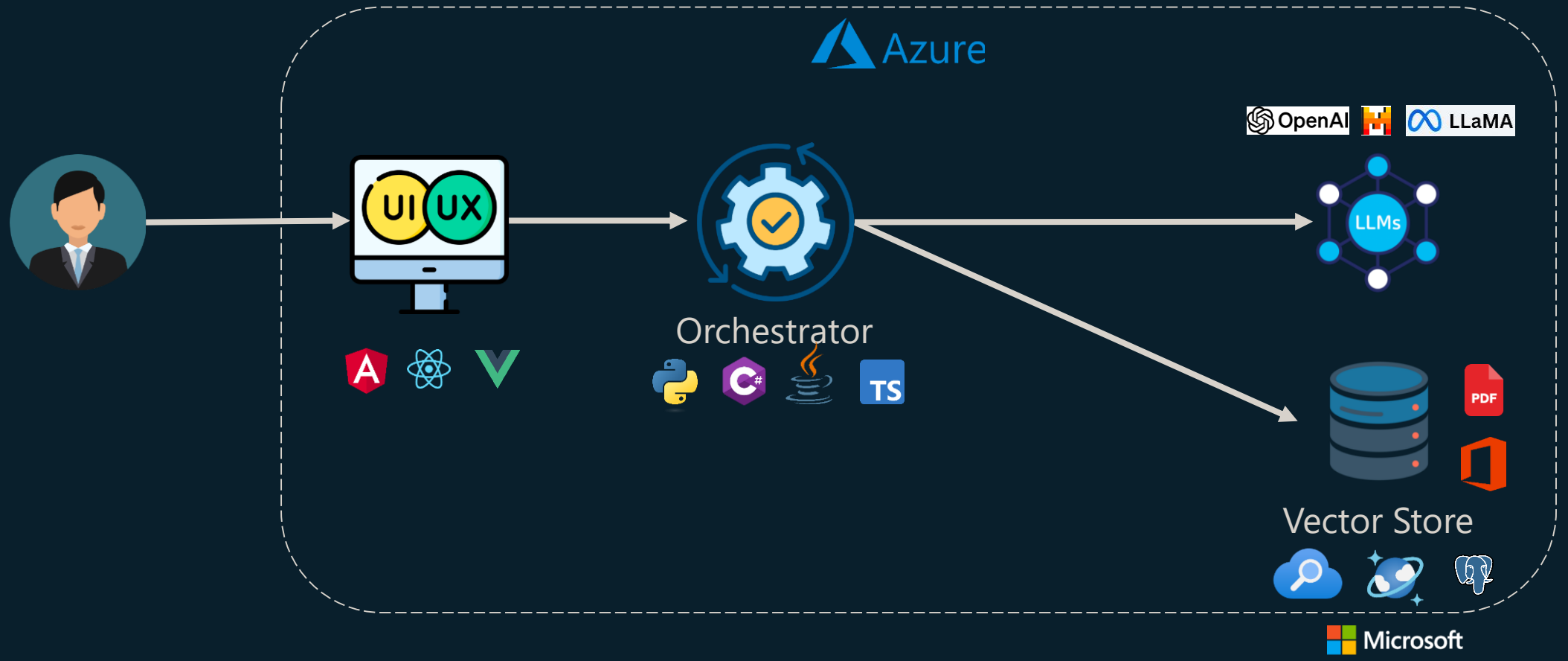
Architecture – Technical Stack



*Langchain library

Reactor

Build a Production-Scale Conversation AI Assistant With RAG



Reactor

SDK

Language	SDK
Java	Azure SDK for Java, Semantic Kernel, Spring AI, LangChain4j
Python	Azure SDK for Python, Semantic Kernel, LangChain
C#	Azure SDK for .NET, Semantic Kernel
JS	Azure SDK for Js, LangChainJs, Vercel AI SDK

Microsoft Reactor

Resources

- https://aka.ms/Apr7AzureAIServicesOpenAI1
- https://aka.ms/Apr7AIFoundryConceptsRAG1
- https://github.com/Azure-Samples/
- https://github.com/Azure/GPT-RAG

Microsoft Reactor

Microsoft

Reactor

Thank You

Linkedin: @cihancinar
Github: @cihancinar

https://blog.cihan-cinar.com/

Cihan Cinar
Software & AI Architect