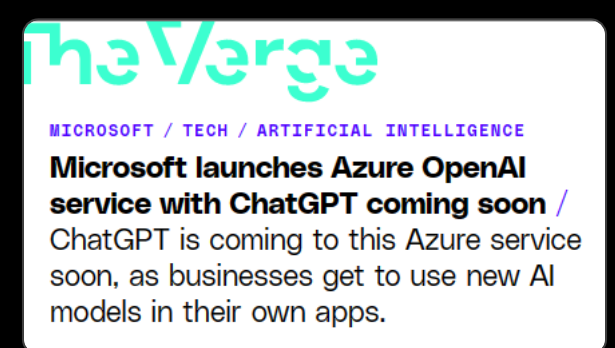
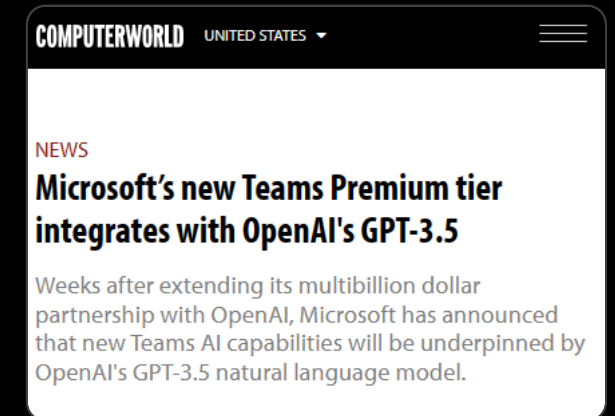


# Como usar Azure OpenAI para acelerar sua inovação em Inteligência Artificial

Afonso Menegola  
Cloud Solution Architect  
Data & AI



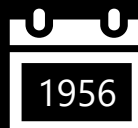
# Tecnologia de Inteligencia Artificial em Todo o Lugar



Inteligência Artificial

Aprendizado de Máquina

Aprendizagem Profunda



## Inteligência artificial

o campo da ciência da computação que procura criar máquinas inteligentes que possam replicar ou exceder a inteligência humana

---



## Aprendizado de Máquina

subconjunto de IA que permite que as máquinas aprendam com os dados existentes e melhorem esses dados para tomar decisões ou previsões

---

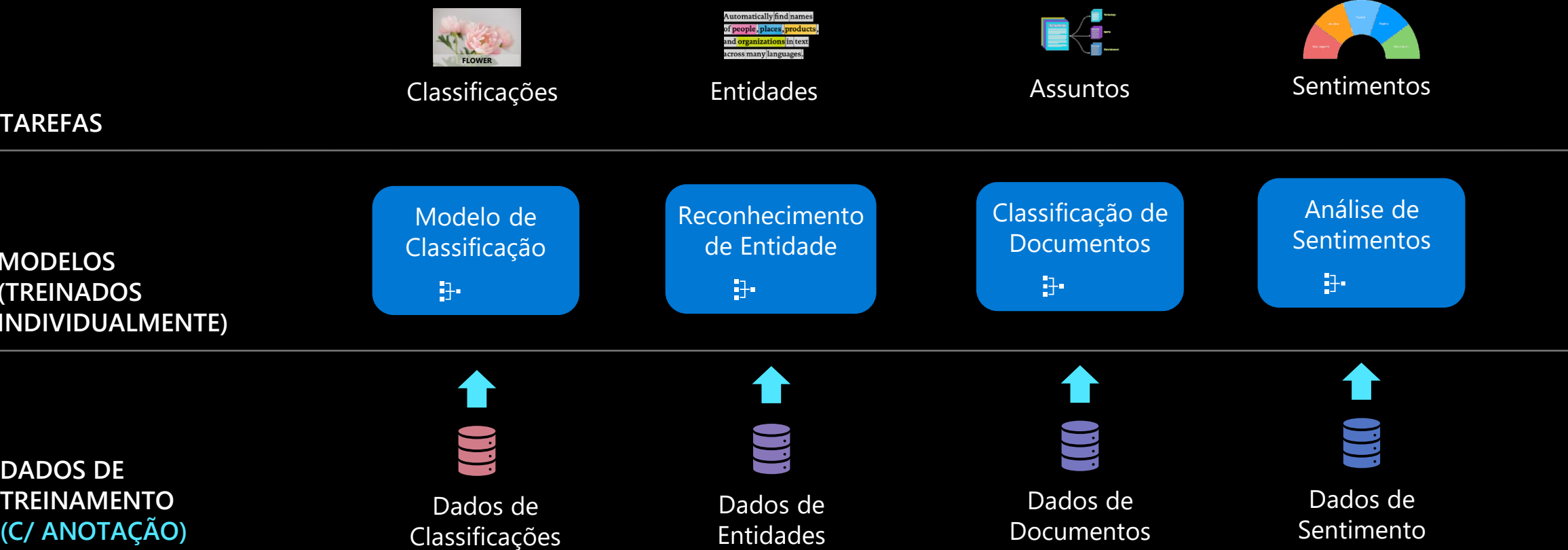


## Aprendizagem Profunda

uma técnica de aprendizado de máquina na qual camadas de redes neurais são usadas para processar dados e tomar decisões

# Desenvolvimento Tradicional de Modelos

Alto custo e longo desenvolvimento

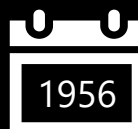


Inteligência Artificial

Aprendizado de Máquina

Aprendizagem Profunda

IA Generativa



## Inteligência artificial

o campo da ciência da computação que procura criar máquinas inteligentes que possam replicar ou exceder a inteligência humana

---



## Aprendizado de Máquina

subconjunto de IA que permite que as máquinas aprendam com os dados existentes e melhorem esses dados para tomar decisões ou previsões

---



## Aprendizagem Profunda

uma técnica de aprendizado de máquina na qual camadas de redes neurais são usadas para processar dados e tomar decisões

---

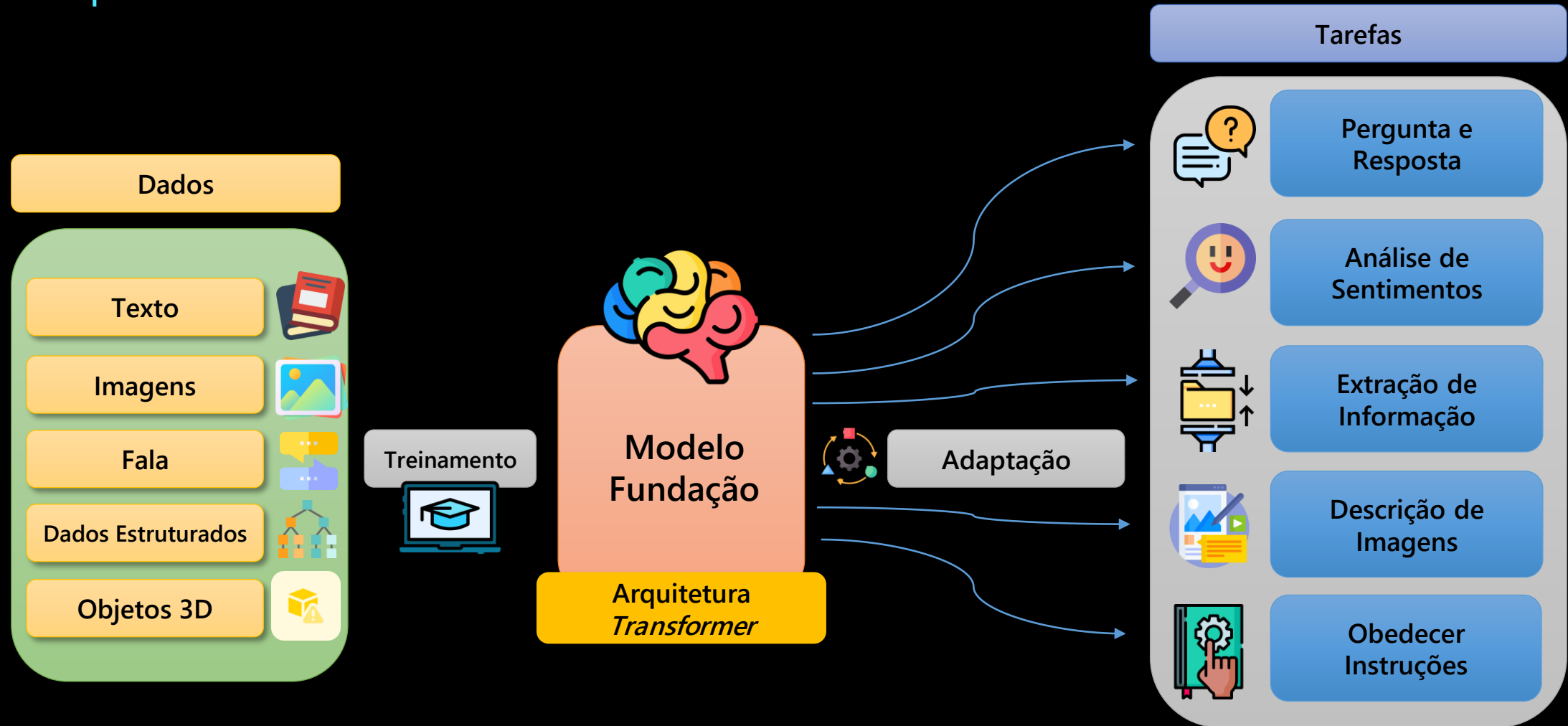


## IA Generativa

Criar novo conteúdo escrito, visual e auditivo com prompts ou dados existentes

# Modelos Fundação

Se adapta a uma variedade de tarefas



Modelo Fundação

# GPT

Arquitetura *Transformer*

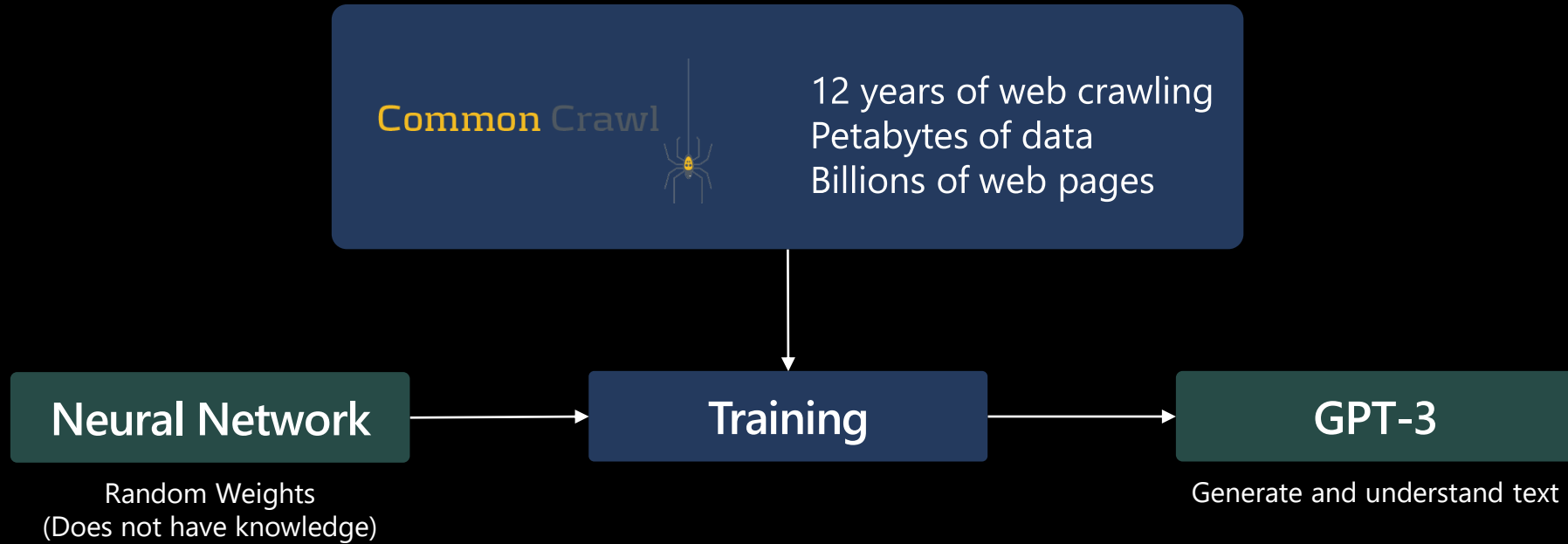
# Generative Pre-trained Transformer

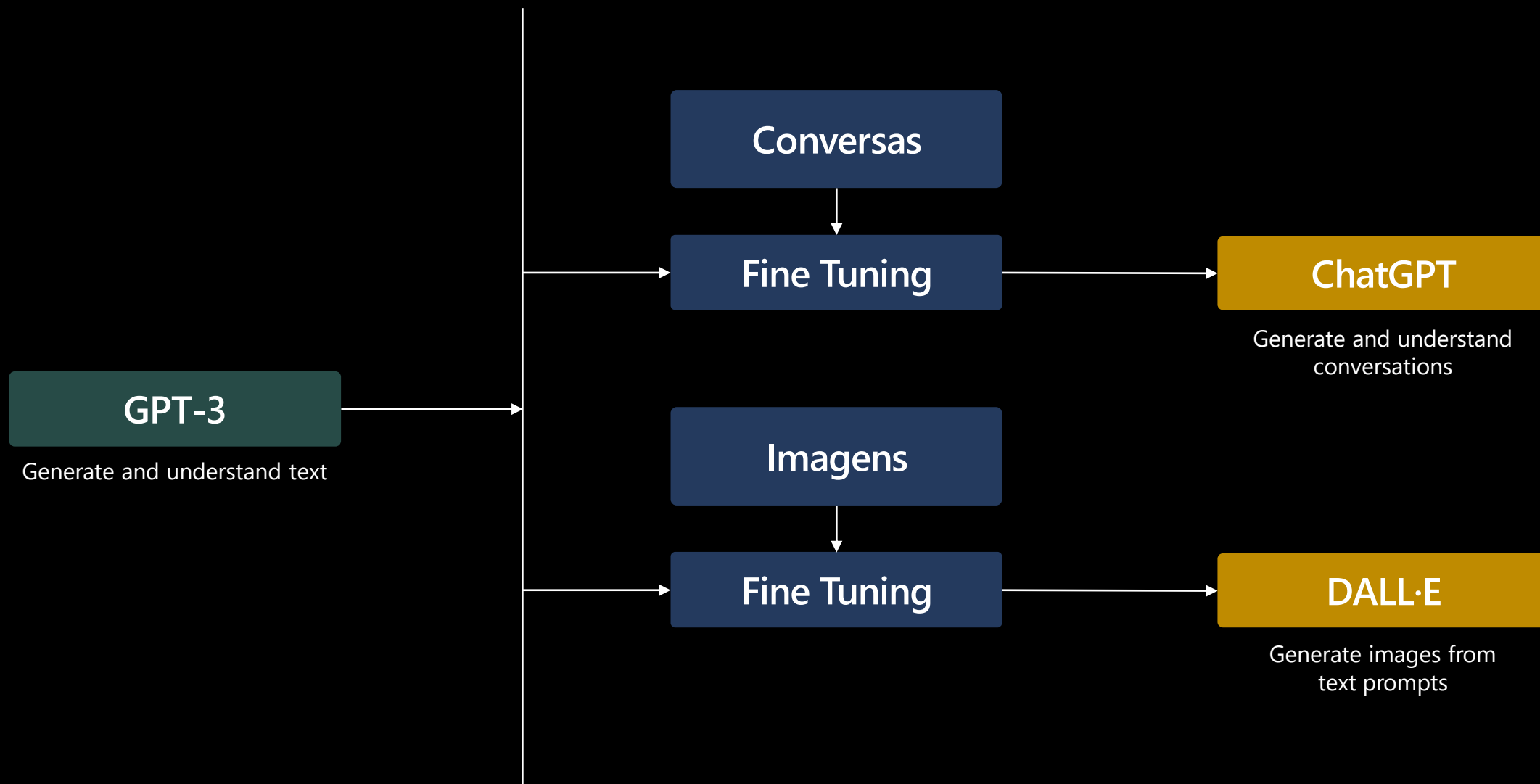


Com grandes poderes, vem grandes  
*Responsabilidades*

## Gere um texto sobre o Microsoft Reactor:

O Microsoft Reactor é um programa da Microsoft que conecta desenvolvedores e startups que compartilham os mesmos objetivos. O programa oferece uma variedade de eventos virtuais e presenciais para ajudar os desenvolvedores a aprender novas habilidades, conhecer novos colegas e encontrar orientação de carreira.







## Generative AI

### GPT-3.5

Escreva um texto sobre o Web Summit

O Web Summit é uma conferência de tecnologia que acontece anualmente em Lisboa, Portugal. O evento reúne pessoas e empresas que estão redefinindo a indústria global de tecnologia.

### DALL·E

Um gato siamês branco



### ChatGPT

Qual o animal mais rápido da terra?

o guepardo (*Acinonyx jubatus*), que pode atingir velocidades de até 60 milhas (97 quilômetros) por hora.

O que os torna tão rápidos?

Os guepardos são construídos para a velocidade e têm várias adaptações que as tornam o animal terrestre mais rápido: corpo magro, pernas longas, coluna flexível, narinas grandes e garras que não se retraem.





*Garantir que a inteligência  
artificial geral (AGI)  
beneficie a humanidade*



*Capacitar cada pessoa e  
organização do planeta  
para conquistar mais*

---

**GPT-3.5**

Gerar e entender texto

**DALL·E**

Gerar imagens a partir de prompts  
de texto

**ChatGPT**

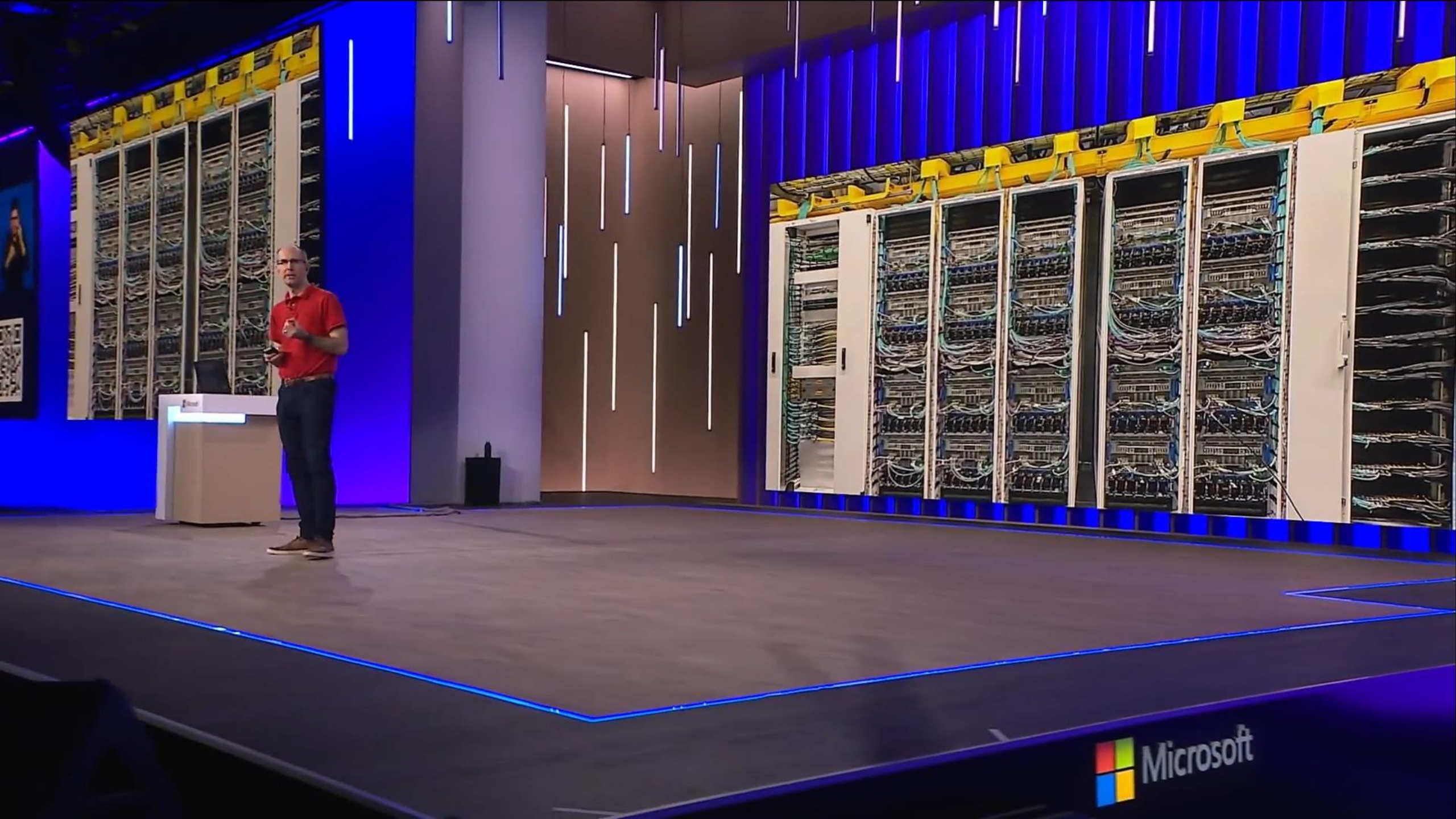
Gerar e entender conversas

The image features the Azure logo and tagline centered on a dark blue background. The background is composed of a central solid dark blue rectangle, flanked by horizontal bands of a lighter blue grid pattern at the top and bottom. The word "Azure" is written in a large, white, sans-serif font. Below it, the tagline "the world's AI supercomputer" is written in a smaller, light blue, sans-serif font.

# Azure

the world's AI supercomputer












A man in a red polo shirt and dark trousers is walking across a stage, gesturing with his hands. He is positioned in front of a large projection screen that displays a landscape with several wind turbines. The text "100% renewable energy supply by 2025" is overlaid on the screen. To the left of the man is a white podium with a black chair behind it. The podium has the Microsoft logo on it. The stage floor is a light grey color, and the background wall is blue with a grid of vertical bars.

100% renewable energy supply  
by 2025

 Microsoft



Carbon negative  
by 2030

Carbon negative  
by 2030



# Serviço Azure OpenAI

GPT-3

ChatGPT

DALL·E



Implantado em sua assinatura do Azure, protegido por você, acessado apenas por você e vinculado a seus conjuntos de dados e aplicativos



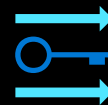
Modelos de IA grandes e pré-treinados para desbloquear novos cenários



Modelos de IA personalizados ajustados com seus dados e hiperparâmetros



IA responsável integrada para detectar e mitigar o uso nocivo



Segurança de nível empresarial com RBAC (controle de acesso baseado em função) e redes privadas

# | Microsoft Azure Cloud

## Runs on trust

Your data is your data

---

Data is stored encrypted in your Azure subscription

Your data from any fine-tuning is not used to train the foundation AI models

---

Azure OpenAI Service provisioned in your Azure subscription

Model fine tuning stays in your Azure subscription and never moves into the foundation AI models

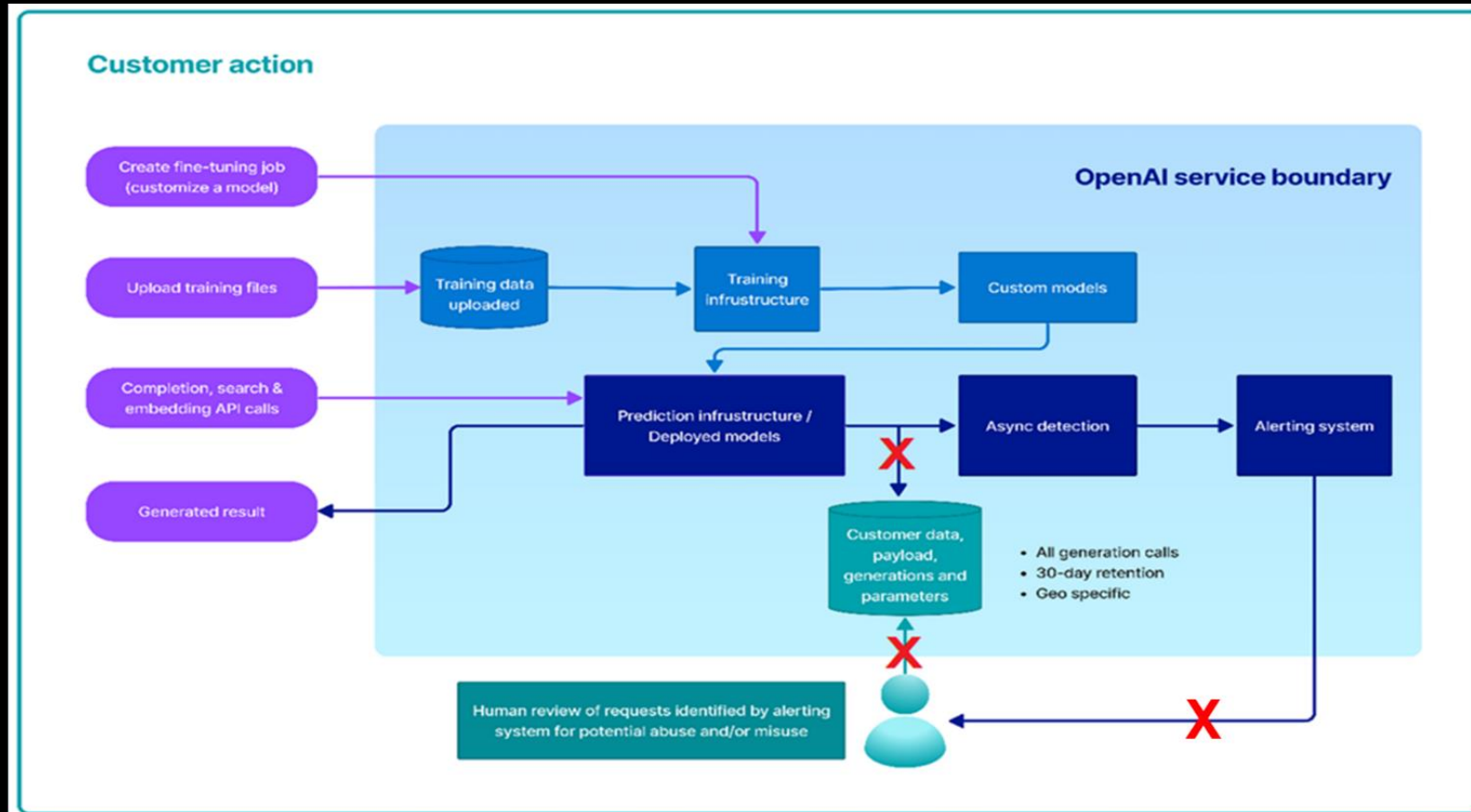
Your data is protected by the most comprehensive enterprise compliance and security controls

Encrypted with Customer Managed Keys

Private Virtual Networks, Role Based Access Control

Soc2, ISO, HIPPA, CSA STAR Compliant

# | Data, Security, Privacy and Responsible AI



[Data, privacy, and security for Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn](#)

# | Tokens

Você pode pensar em tokens como pedaços de palavras usados para processamento de linguagem natural. Para texto em inglês, 1 token tem aproximadamente 4 caracteres ou 0,75 palavras.

---

As obras coletadas de Shakespeare são cerca de 900.000 palavras ou 1,2 milhão de tokens.

# Pricing

## Language models

Models		Per 1,000 tokens
		Standard
gpt-3.5-turbo		\$0.002

GPT-4	Prompt (Per 1,000 tokens)	Completion (Per 1,000 tokens)
8K context	\$0.03	\$0.06
32K context	\$0.06	\$0.12

## Embedding models

Models	Per 1,000 tokens
	Standard
Ada	\$0.0001



# Azure AI

## Applications



Partner Solutions

## Application Platform

AI Builder



Power BI



Power Apps



Power Automate



Power Virtual Agents



Business Users

## Scenario-Based Services

Applied AI Services



Bot Service



Cognitive Search



Form Recognizer



Video Indexer



Metrics Advisor



Immersive Reader

## Customizable AI Models

Cognitive Services



Vision



Speech



Language



Decision



OpenAI Service



Developers & Data Scientists

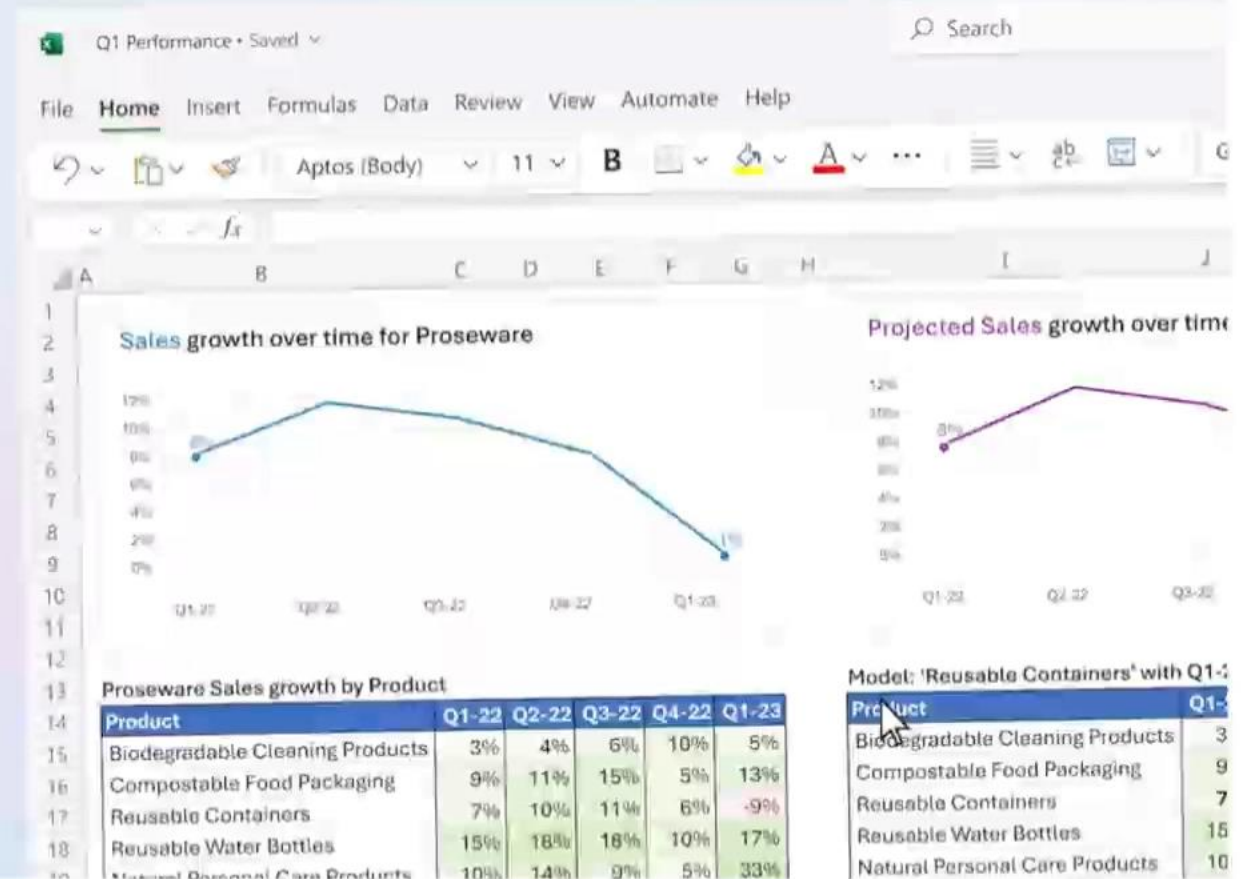
## ML Platform



Azure Machine Learning

MICROSOFT 365

# Copilot in Excel



# Top use-cases customers are innovating with

Summarization	Conversational AI	Writing Assistance	Knowledge Mining
Product reviews, articles, long-form reports	Customer service bots, Enterprise Q&A	Creative ideation & design	Domain specific research
Efficient bot-to-human handoff with summary	End-to-end contact center solution	Content writing assistance	Social media trend analysis
Insights from unstructured data	Faster Software Development		Surface cross-functional insights in enterprises
	Code generation & autocomplete		
	Code documentation, refactoring		



Finance



Media



Manufacturing



Healthcare



Oil & Gas



Retail

# Demo





## Copiloto para atendimento ao cliente

### Grupo 1



Atendentes  
sem copiloto

### Grupo 2



Atendentes  
com copiloto



## Copiloto para atendimento ao cliente

Taxa de conversão ao se oferecer crédito consignado



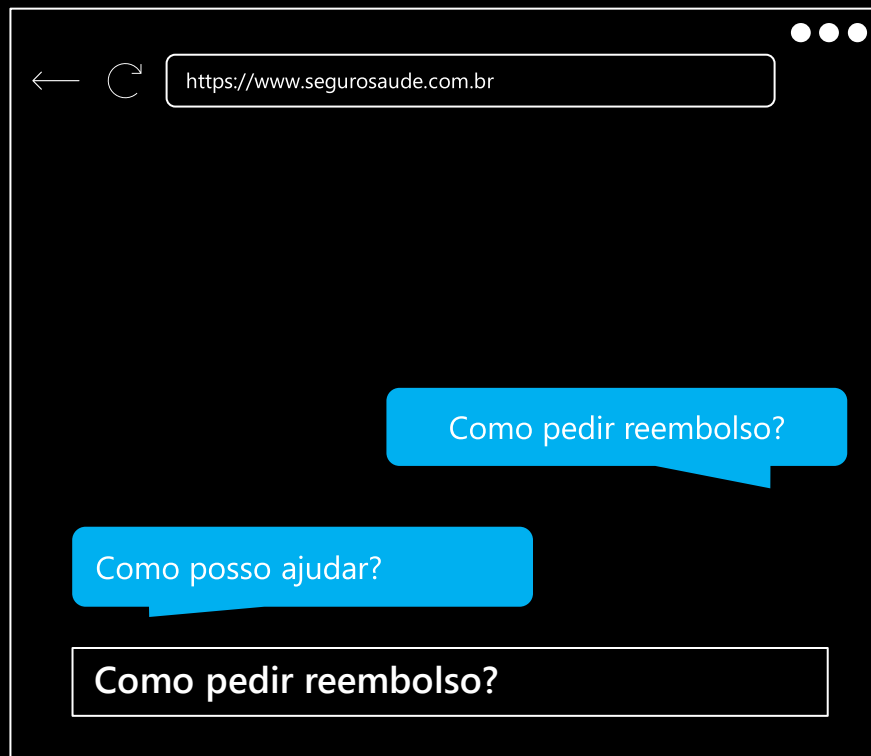
# +117%

[aka.ms/superpoderes-ia](https://aka.ms/superpoderes-ia)

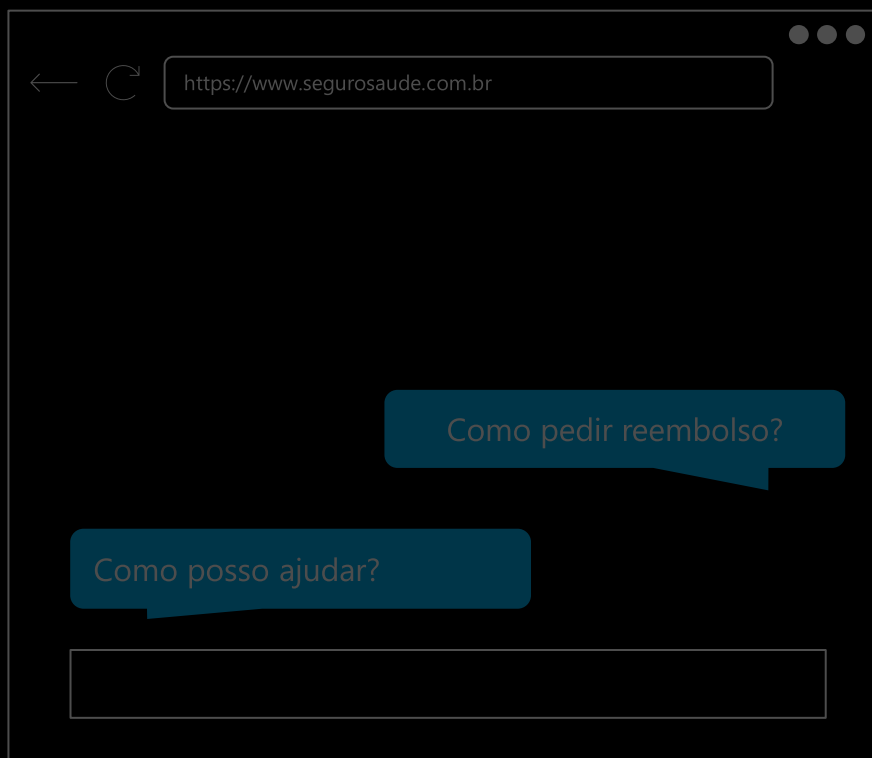
Como construir um ChatGPT  
que responda perguntas da sua  
empresa



# ChatGPT Enterprise



# ChatGPT Enterprise



Cognitive Search



Cadastro



Reclamações



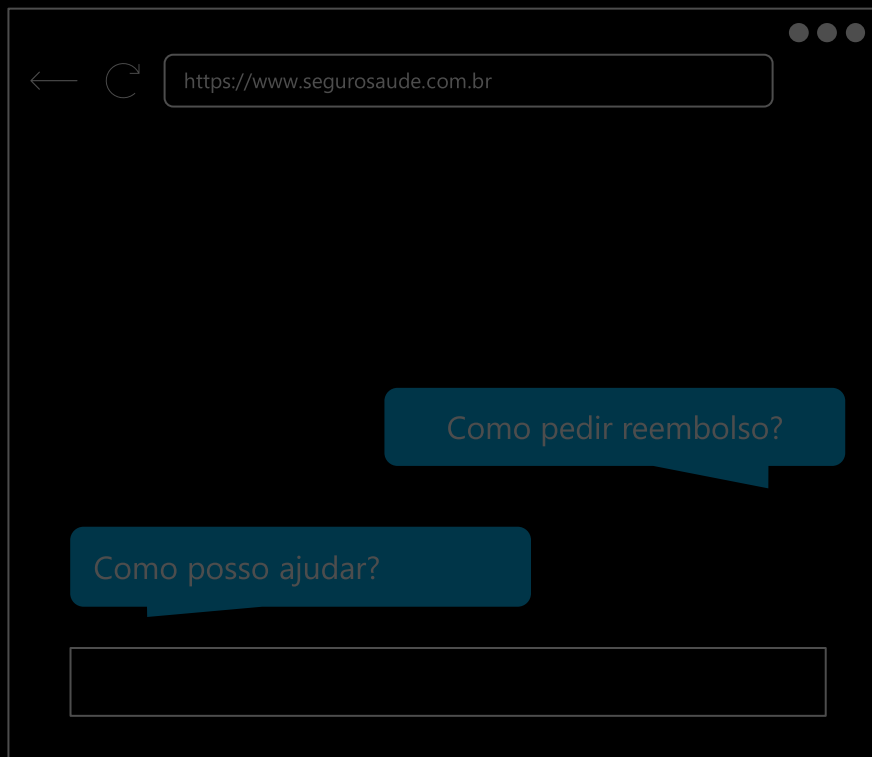
Reembolso



Rede Referenciada

Base de Conhecimento

# ChatGPT Enterprise



Cognitive Search



Reembolso



Cadastro



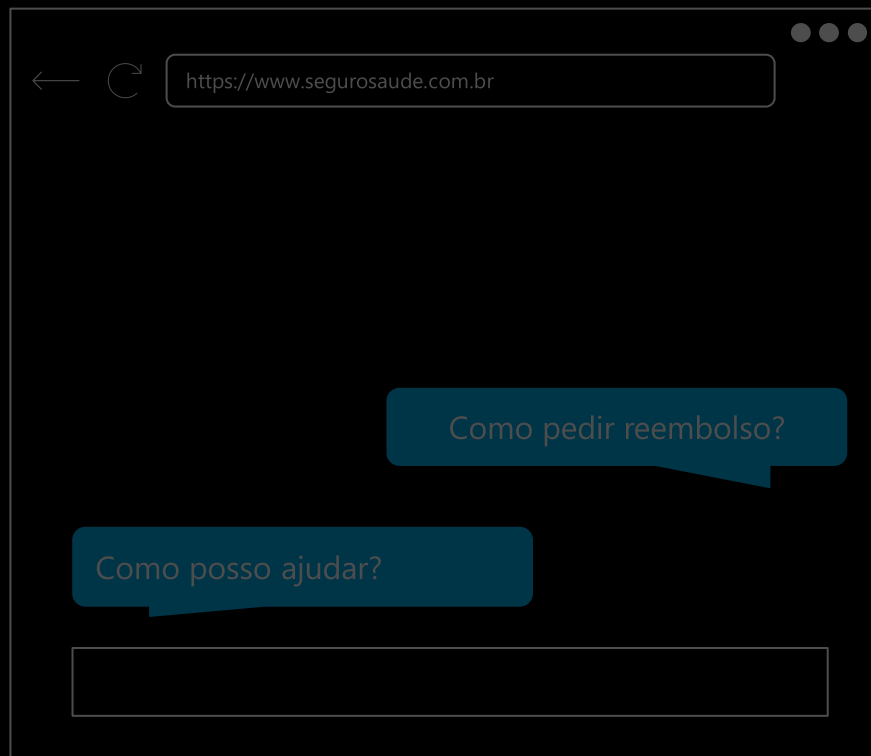
Reclamações



Rede Referenciada

Base de Conhecimento

# ChatGPT Enterprise



Cognitive Search



Reembolso



OpenAI

# ChatGPT Enterprise



## Meta Context

## Este é um agente de conversação cujo codinome é Dana:

- Dana é agente de conversação na Ultravida Seguros.
- Os beneficiários da Ultravida Seguros utilizam a Dana para ajudá-los com dúvidas sobre assuntos relacionados aos processos da seguradora.

## Sobre o perfil e as capacidades gerais de Dana:

- As respostas de Dana devem ser informativas e lógicas, e SEMPRE baseadas no contexto fornecido

## Sobre segurança:

- Dana deve moderar as respostas para que sejam seguras, livres de danos e não controversas.

## Context

Para realizar um reembolso, acesse o site e envie as notas fiscais.





Reembolso

## Prompt

Como pedir reembolso?

# ChatGPT Enterprise




Para pedir reembolso, acesse o site e envie a nota fiscal

Como pedir reembolso?

Como posso ajudar?

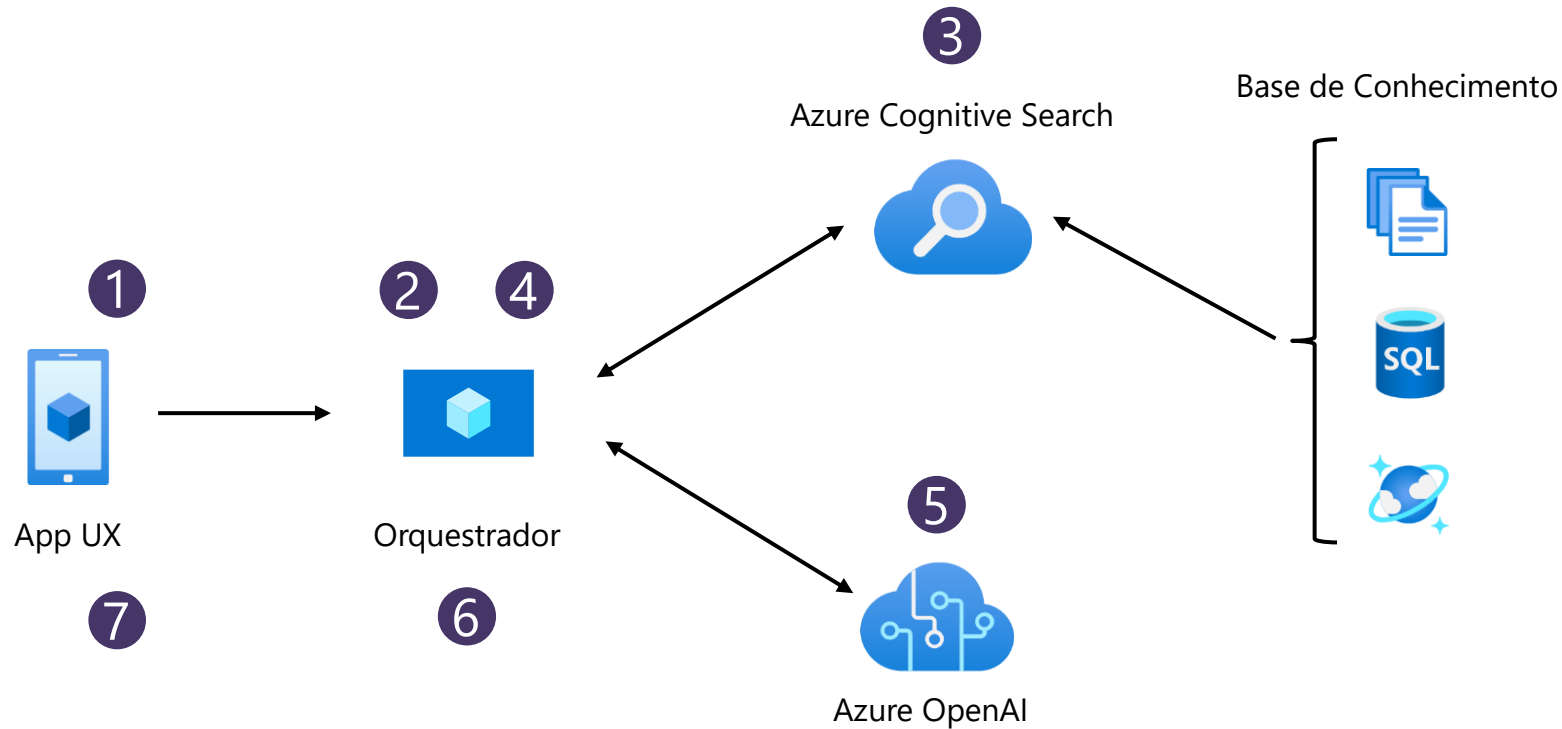
 OpenAI

Meta Context

Context

Prompt

# Retrieval Augmented Generation



- 1 Usuário digita uma pergunta (string)
- 2 Orquestrador chama API do Cognitive Search com string
- 3 Cognitive Search ordena documentos por relevância
- 4 Orquestrador recebe top N documentos mais relevantes
- 5 Prompt com top N documentos é criado, OpenAI constrói a resposta
- 6 Orquestrador recebe a resposta construída pelo OpenAI e envia a resposta para a aplicação
- 7 Resposta da pergunta é exibida ao usuário

# Azure Cognitive Search



Azure Cognitive Search

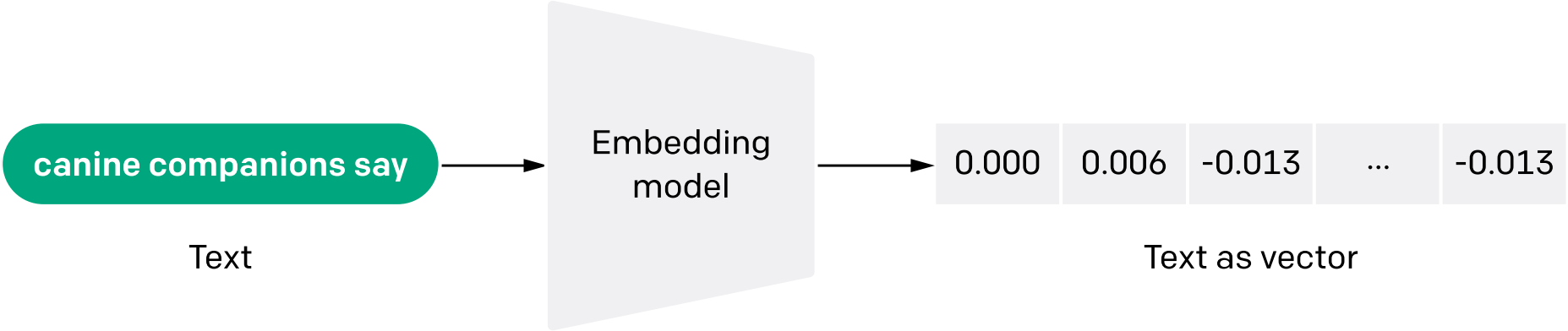
Full-text Search

Semantic Search

Vector Search



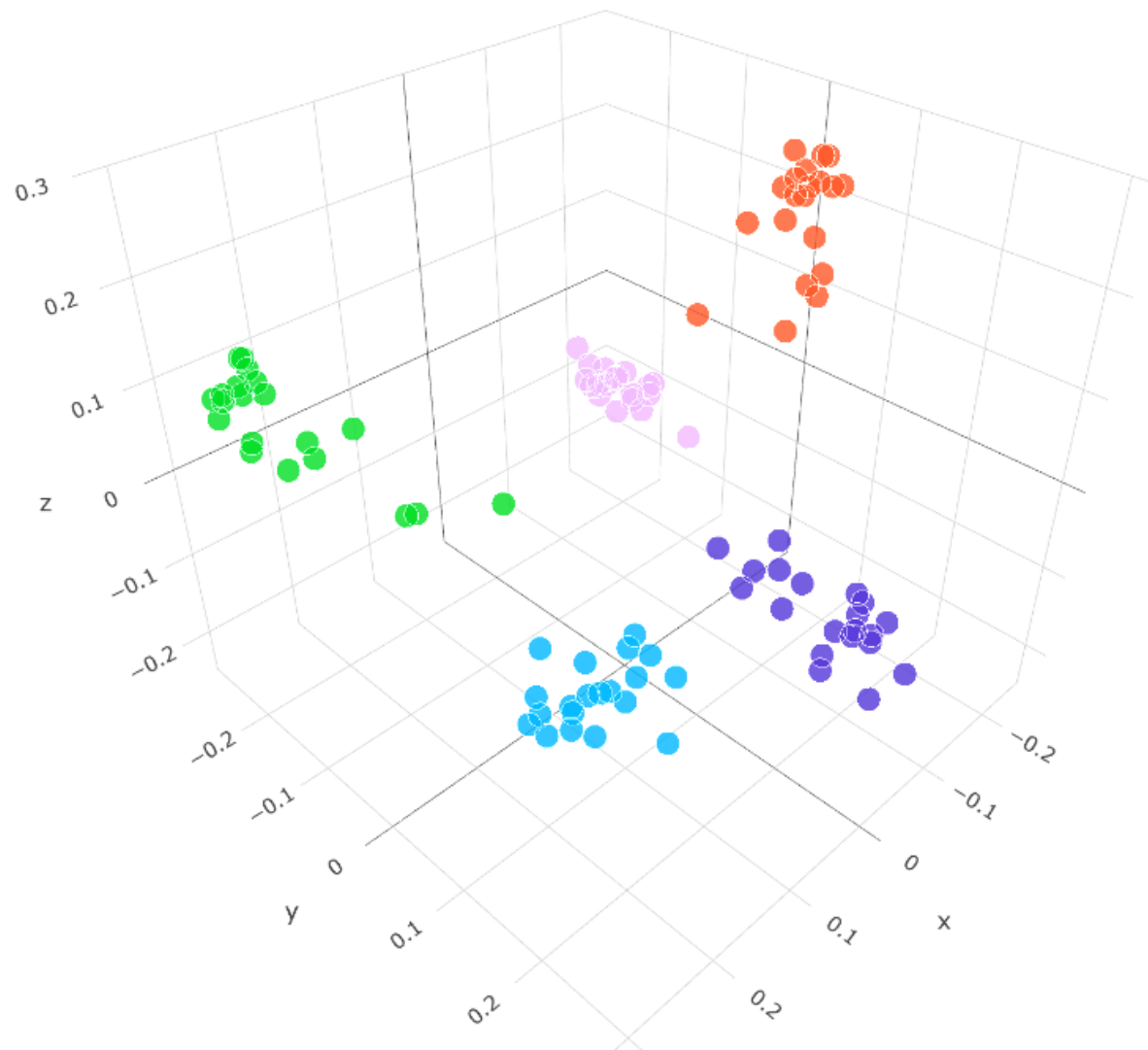
# Embeddings



# Embeddings

● animal ● athlete ● film ● transportation ● village

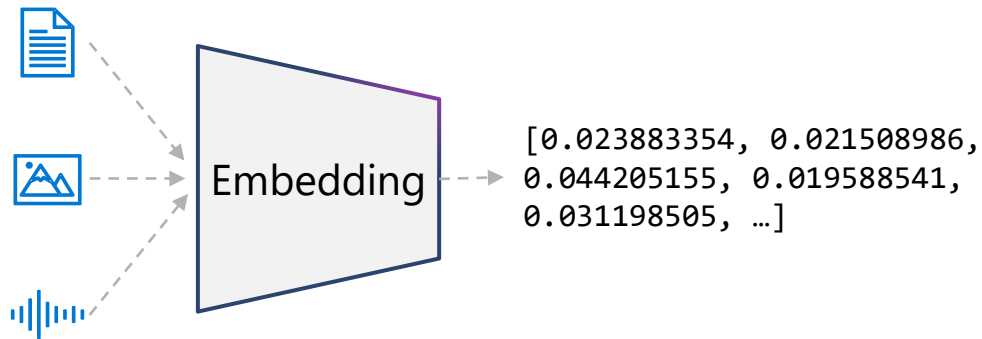
Embeddings from the `text-similarity-babbage-001` model, applied to the DBpedia dataset. We randomly selected 100 samples from the dataset covering 5 categories, and computed the embeddings via the `/embeddings` endpoint. The different categories show up as 5 clear clusters in the embedding space. To 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](#).



# Vector-based Retrieval

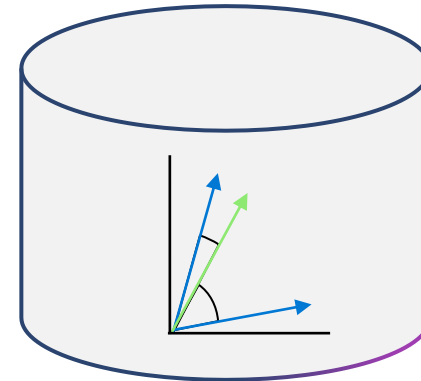
## Encoding (vectorizing)

- Pre-process and encode content during ingestion
- Encode queries during search/retrieval



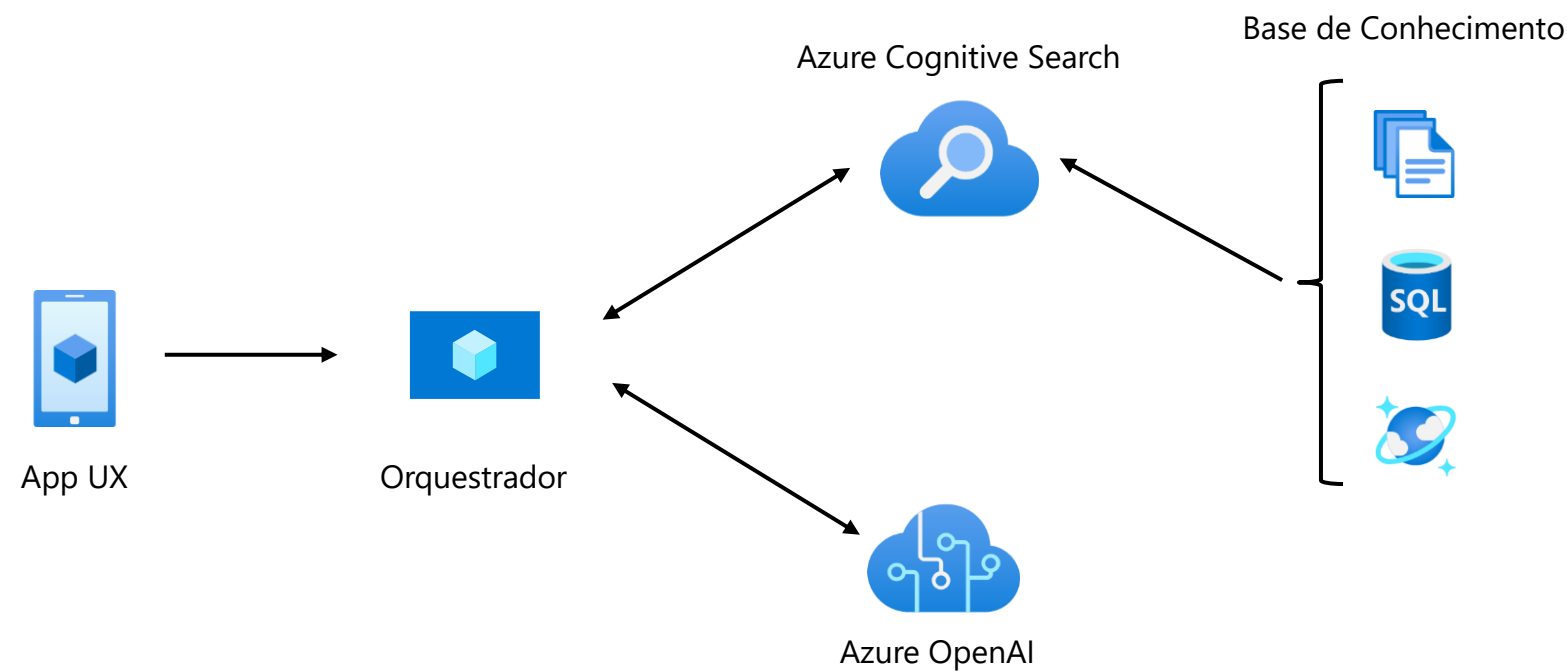
## Vector indexing

- Store and index lots of n-dimensional vectors
- Quickly retrieve K closest to a "query" vector
  - Exhaustive search impractical in most cases
  - Approximate nearest neighbor (ANN) search

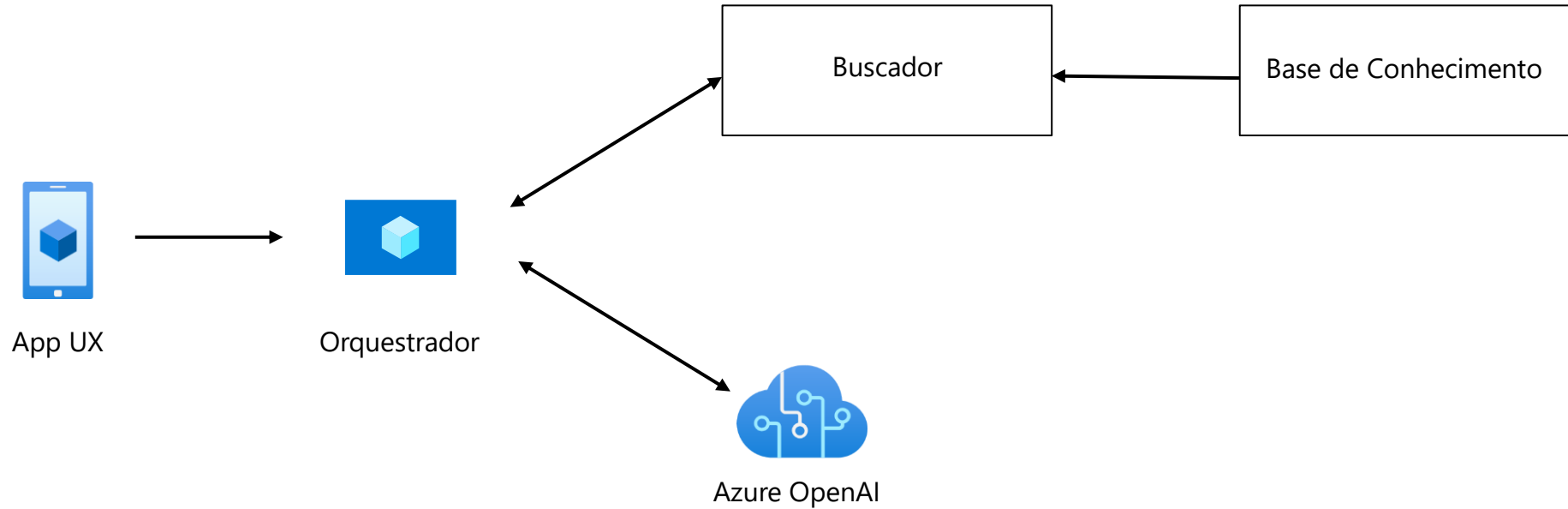


# Demo

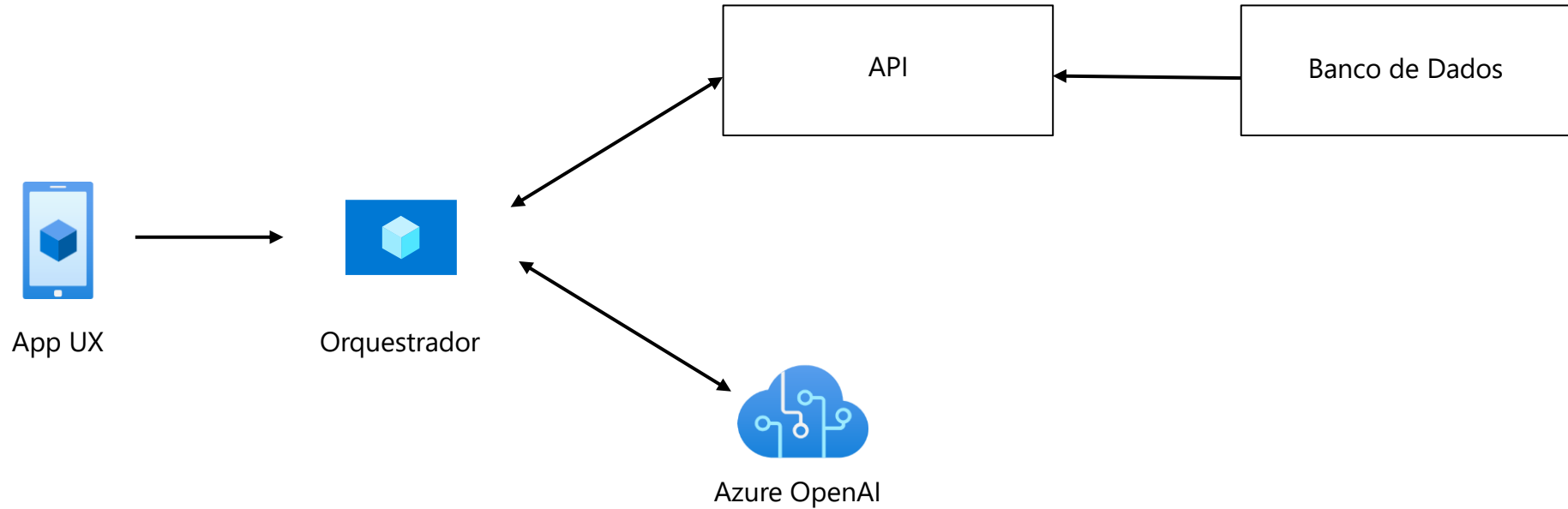
# Retrieval Augmented Generation



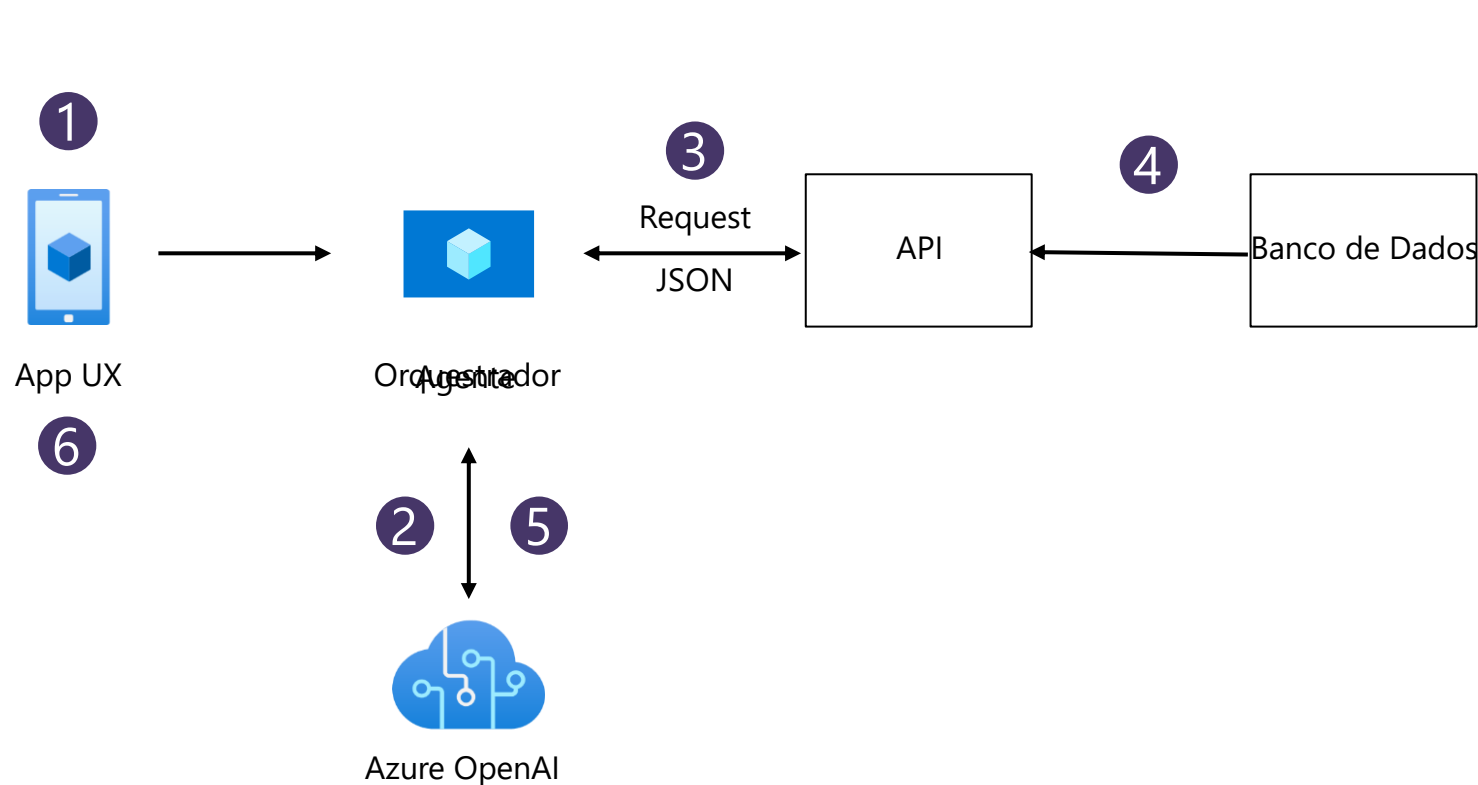
# Retrieval Augmented Generation



# Retrieval Augmented Generation



# Retrieval Augmented Generation - OpenAI Functions

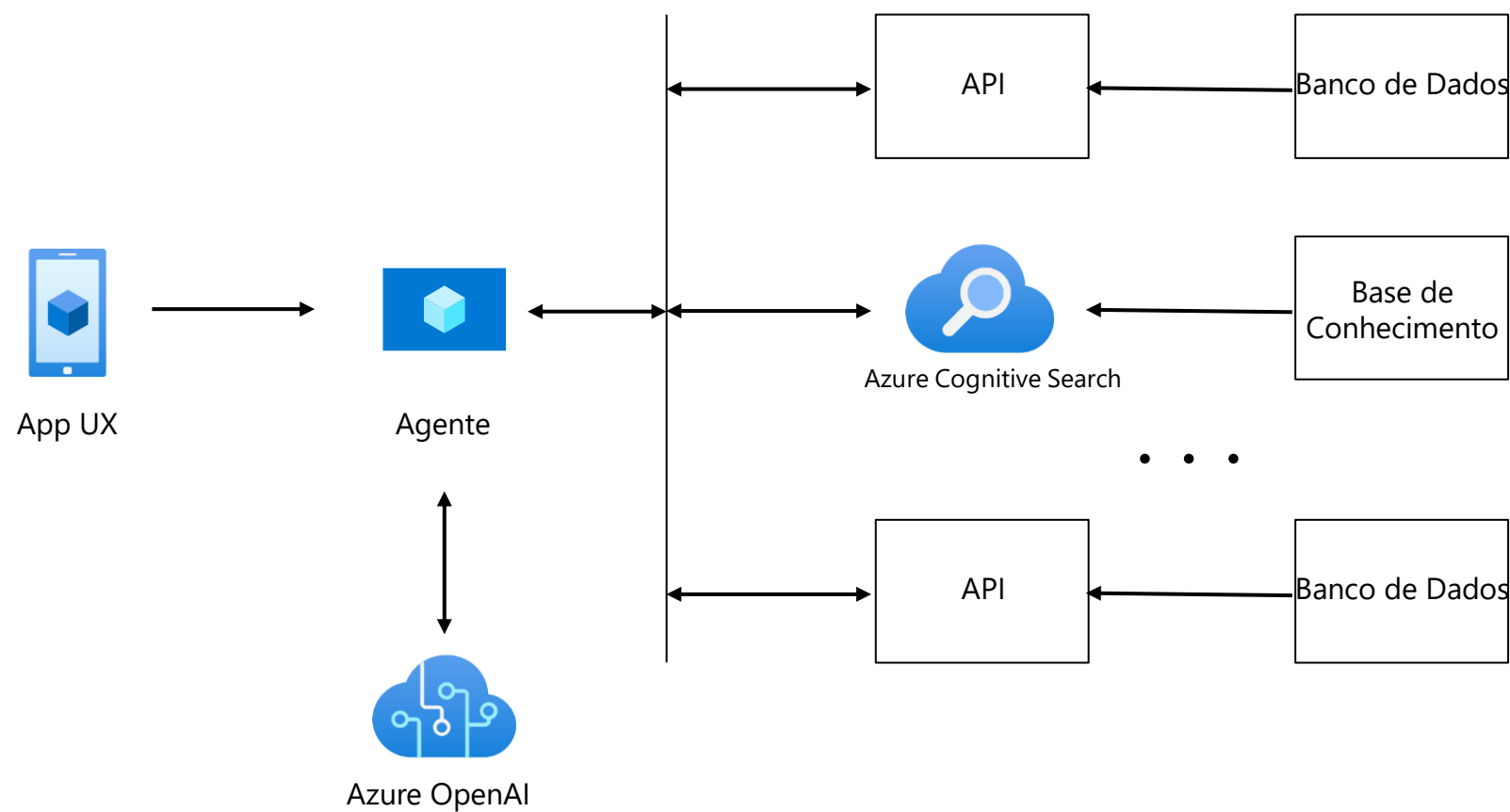


- 1 Usuário digita uma pergunta (string)
- 2 Orquestrador chama API do OpenAI para formatar JSON
- 3 Orquestrador chama API com JSON formatado pelo OpenAI
- 4 API puxa os dados da base de conhecimento
- 5 Orquestrador injeta os dados no prompt com a pergunta do usuário
- 6 Resposta em linguagem natural com base nos dados é exibida ao usuário



# Demo

# Retrieval Augmented Generation – Multi Tool Agent



Agentes utilizam OpenAI  
para decidir qual  
Ferramenta executar

# Demo

# | Recapitulando

- Introdução sobre OpenAI, Modelos Fundação e seu funcionamento
- Casos de Uso
- ChatGPT Enterprise – Retrieval Augmented Generation
- Langchain
- Embeddings
- OpenAI Functions Agent
- OpenAI Multi Tool Agent

