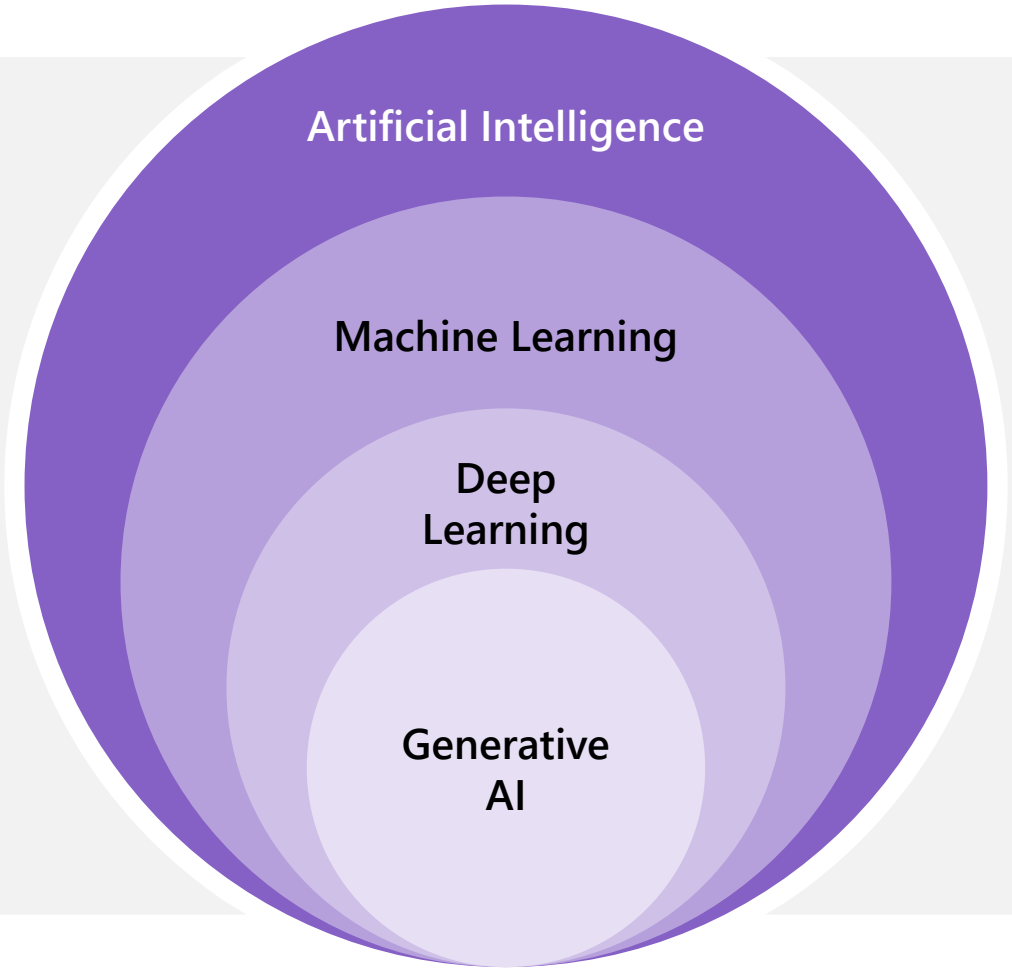
Azure Open AI

Ujjwal Kumar
Lead Architect, Emerging Tech & Sustainability
Microsoft Technology Center (MTC) APAC, Singapore

Agenda

Where we are Today	Introduction to generative AI, Azure AI portfolio, our commitment to Responsible AI and Protecting your Data
Models & Capabilities	What is behind the scenes? Overview of Azure OpenAI Service and its cutting-edge models, features and solutions with other Azure AI Services
Concepts	High-level overview of fundamental concepts like Tokens, Chunking, Embedding
Sample Architectures	Example diagrams showcasing select use cases and scenarios featuring Azure OpenAI Service integrated with other Azure AI Services
Success Stories	Publicly available Azure OpenAI Service customer stories featuring a wide range of use cases and company types

The journey continues with generative AI



1956

Artificial Intelligence

The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997

Machine Learning

Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012

Deep Learning

A machine learning technique in which layers of neural networks are used to process data and make **decisions.**

2021

Generative AI

Create new written, visual, and auditory content given prompts or existing data.

Announced at Ignite..

Azure AI Studio

Copyright Commitment

Improvements to Provisioned Throughput Units

Microsoft and OpenAI partnership

OpenAI

Ensure that artificial general intelligence (AGI) benefits humanity

Microsoft

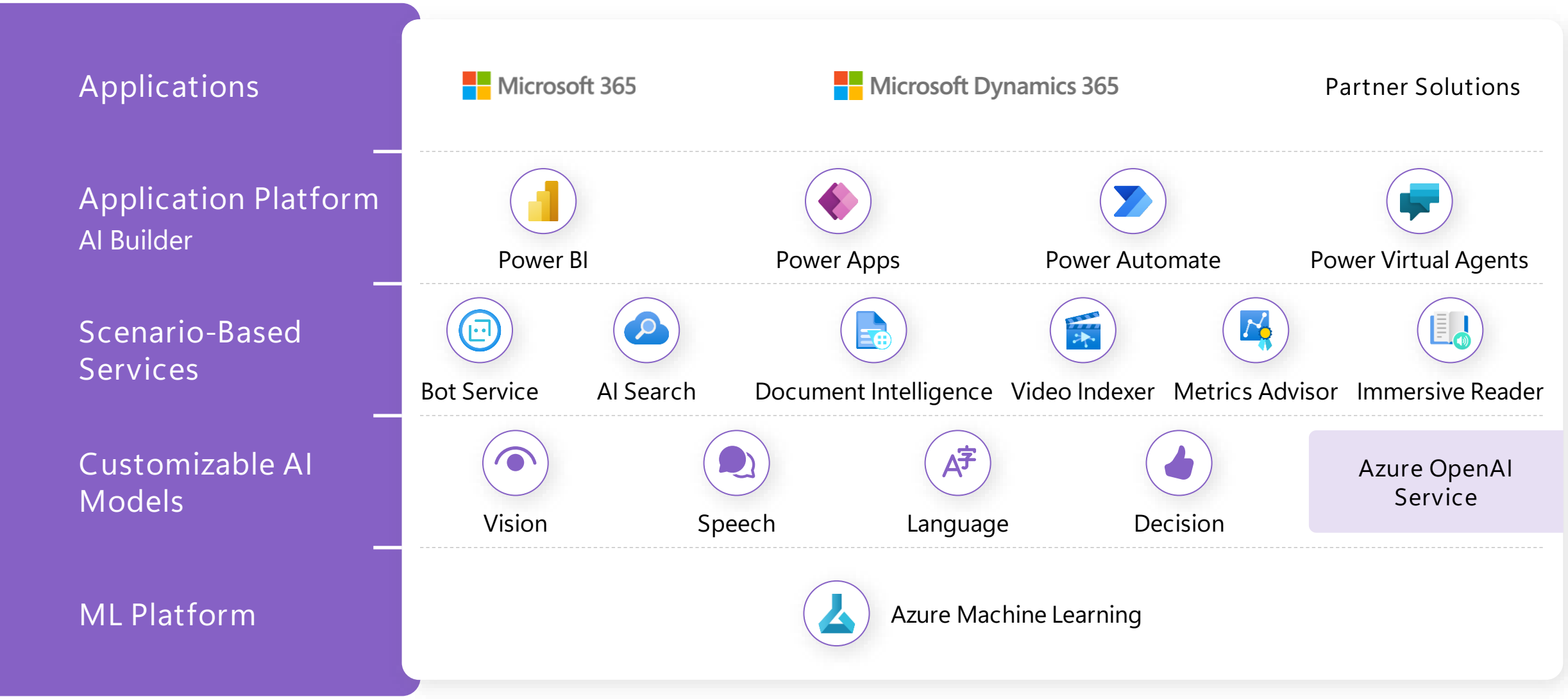
Empower every person and organization on the planet to achieve more

Azure OpenAI Service – as of November 15, 2023

GPT-4, GPT-4-Turbo, GPT-3.5-Turbo	GPT-4-Turbo with Vision	Babbage, Davinci, GPT-3.5-Turbo	DALL·E 3	Whisper
Language	Multi-Modal	Fine Tuning	Images	Transcription & Translation

Azure AI Studio

Microsoft is powered by Azure AI



From Copilot to enterprise scale AI

Built-in

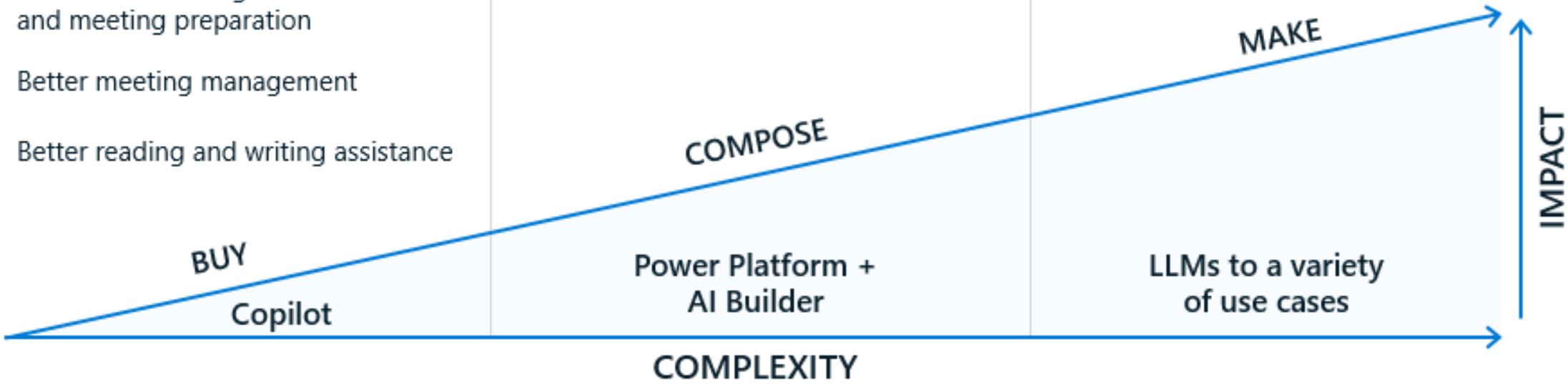
- Search and Q&A
- Better customer engagement and support
- Create and tailor reports in seconds
- Better email management and meeting preparation
- Better meeting management
- Better reading and writing assistance

AI-Builder

- Enriching self-service with conversational AI
- Automate answering questions, summarizing documents, and more.
- Intuitive low-code development for Everyone

AI-Models

- Azure OpenAI Language Models
- Model Catalog / Prompt flow



Models and Capabilities

A copilot for every Microsoft Cloud experience

Microsoft 365 Copilot

Empower everyone with a copilot that works alongside you

Dynamics 365 Copilot

Specialized copilots for every role and function

Copilot in Power Platform

Imagine it, describe it, and Power Platform builds it

Microsoft Security Copilot

Defend at machine speed with Microsoft Security Copilot

Windows Copilot

The first centralized AI assistance on a platform

GitHub Copilot

Increase developer productivity to accelerate innovation

Models

GPT-3.5

Prompt:

Write a tagline for an ice cream shop.

Response:

We serve up smiles with every scoop!

GPT-4

The next level in text generation with improved alignment

- Generate complex documents
- Steer with nuanced instructions
- Instruct and annotate in any language, slang, dialect

Azure OpenAI Service

GPT-4

GPT-4-Turbo

GPT-3.5-Turbo

New: GPT-4 for Vision

Generative Text Models, with varying capabilities and uses

DALL·E 3

Generative Image Model

Deploy on your own data

Provisioned throughput units (PTUs)

Functions and Plugins

Understanding the GPT3.5 format

System Message:

The system message is included at the beginning of the prompt and is used to prime the model and you can include a variety of information in the system message including:

- A brief description of the assistant
- The personality of the assistant
- Instructions for the assistant
- Data or information needed for the model

User and assistant messages:

After the system message, you can include a series of messages between the *user* and the *assistant*. You denote who the message is from by setting the role to user or assistant.

```
{
    “role”: “user”,
    “content”: “What is a garbanzo bean?”
}
```

Example Prompt:

{ “role”: “system”, “content”: “You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.” },

{ “role”: “user”, “content”: “Why won’t my Xbox turn on?”},

{ “role”: “assistant”, “content”: “There could be a few reasons why your Xbox isn't turning on….”},

{ “role”: “user”, “content”: “I confirmed the power cord is plugged in but it’s still not working” }

Public preview

Announcing GPT-4 Turbo with Vision

Unlock new scenarios with GPT-4-Turbo, Azure Open AI Service and Azure AI Vision integration

Add images to retrieval augment generation (RAG) patterns

Prompt with video, images, and text

What GPT-4V Offers:

GPT-4 with Vision (GPT-4V) is a multimodal model developed by OpenAI that accepts both image and text inputs and generates text outputs.

Note: GPT-4V doesn't generate image outputs

DALL·E 3

In Preview: Azure OpenAI Service

DALL-E 3 is an image generation model that allows you to generate images from text prompts






Use Cases for DALL·E 3

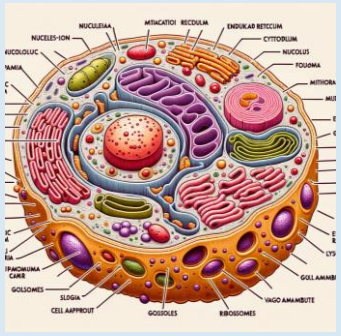
LOGO & BRANDING: QUICK CONCEPT GENERATION.

CREATIVE INSPIRATION: OVERCOME DESIGN BLOCKS.

CONTENT ILLUSTRATIONS: UNIQUE IMAGES FOR BLOGS/ARTICLES.

AD CAMPAIGNS: VISUALIZE MARKETING CONCEPTS.

PRODUCT VISUALIZATION: GAUGE INTEREST & FEEDBACK.



EDUCATION: CUSTOM IMAGERY FOR COURSES.

FASHION DESIGN: VISUALIZE CLOTHING PATTERNS.

GAMING: CHARACTER & ENVIRONMENT CONCEPTS.

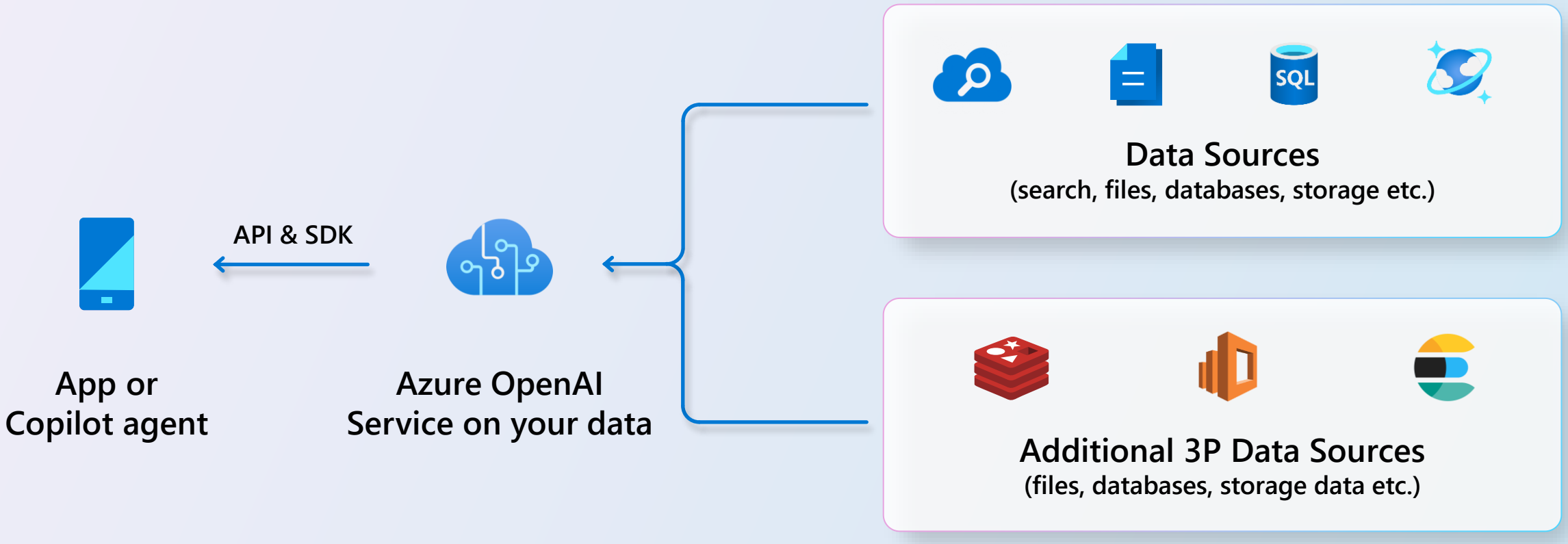
Whisper

The next level
in transcription and
translation

Transcribe

Translate

Azure OpenAI Service on *your* data



Concepts

Azure AI Studio

Build and train your own models

Ground Azure OpenAI Service and OSS models using your data

Built-in vector indexing

Retrieval augmented generation made easy

Create prompt flows

AI safety built-in

What is Prompt Engineering?

Prompt engineering is a concept in Natural Language Processing (NLP) that involves embedding descriptions of tasks in input to prompt the model to output the desired results.

How to adapt GPT-3 model for your task

Zero-Shot	One-Shot	Few-Shot

Custom-Tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

Prepare and upload training data → Train a new custom-tuned model → Use your custom-tuned model

1. Higher quality results than prompt design

2. Ability to train on more examples than can fit in a prompt

3. Token savings due to shorter prompts

4. Lower latency requests

Understanding Prompts

Understanding tokens and possibilities

Tokens:

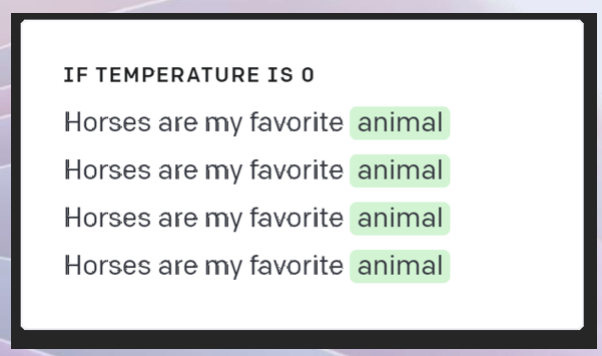
I have an orange cat named Butterscotch.

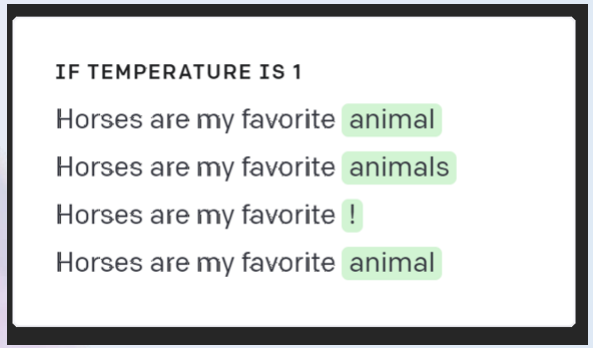
I have an orange cat named Butterscotch.

Horses are my favorite

animal	49.65%
animals	42.58%
\n	3.49%
!	0.91%

Probabilities:





Prompt Instruction

Suggest three names for an animal that is a superhero.

Animal: Cat
Names: Captain Sharpclaw, Agent Fluffball, The Incredible Feline
Animal: Dog
Names: Ruff the Protector, Wonder Canine, Sir Barks-a-Lot
Animal: Horse
Names:

Completion Temperature 0 (always the same)

Mighty Equine, The Great Galloper, Thunderhoof

Completion Temperature 1 (often different)

Blaze the Miracle Mare, Pegasus the Winged Warrior, Secretariat the Superhorse

Completion Temperature 1 (often different)

Blaze of Glory, Sterling Silver, Thunderbolt

Chunking

Chunking

The process of breaking up large documents into smaller chunks.

Chunking Techniques:

1. Fixed-size chunks
2. Variable-sized chunks
3. Content Overlap

Embeddings

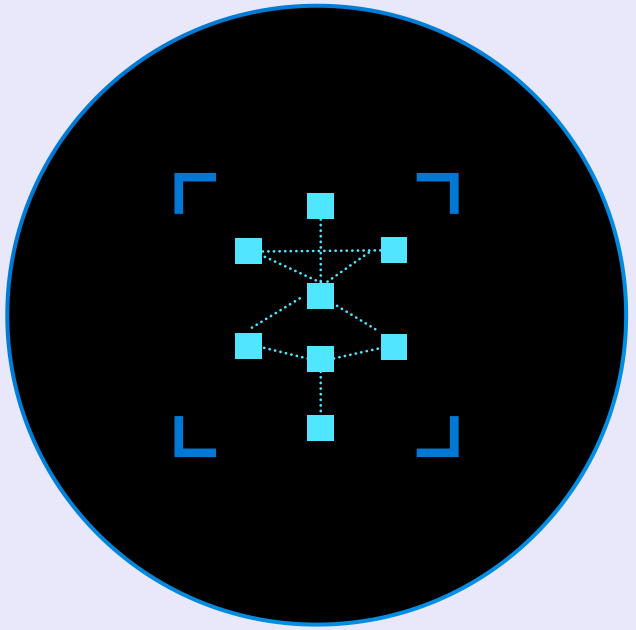
Embeddings

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

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

Each embedding is a vector of floating-point numbers, such that the distance between two embeddings in the vector space is correlated with semantic similarity between two inputs in the original format.

For example, if two texts are similar, then their vector representations should also be similar.



Embeddings make it possible to map content to a “semantic space”

A neutron star is the collapsed core of a massive supergiant star

[15 34 24 13 ...]

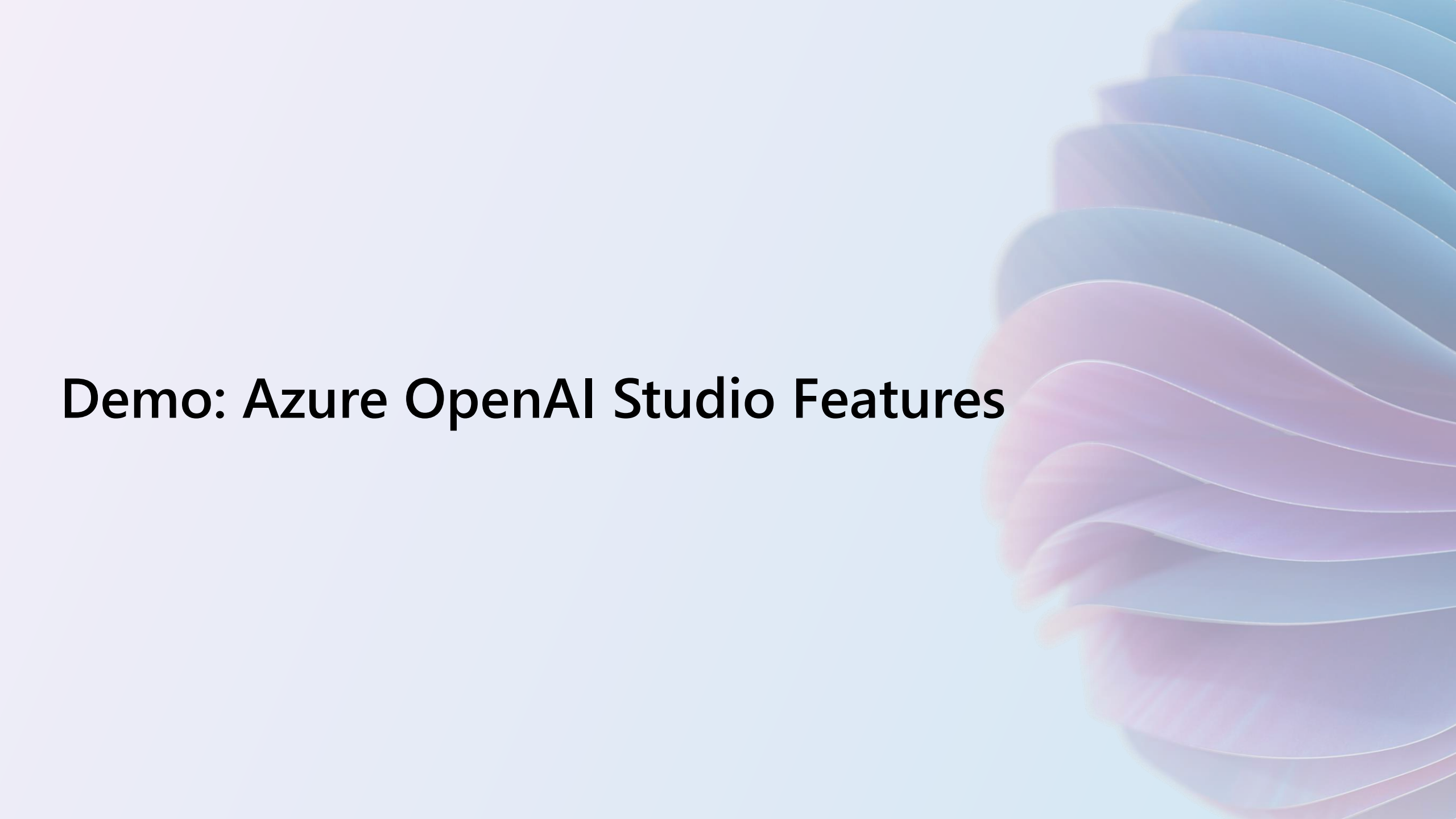
A star shines for most of its active life due to thermonuclear fusion

[16 22 89 26 ...]

The presence of a black hole can be inferred through its interaction with other matter

[20 13 31 89 ...]

Demo: Azure OpenAI Studio Features



Best For You Organics Company

localhost:3000

Best For You Organics - Chat

Sources | Prompt | Content

Prompt Template

Default Prompt

```
# Context
The following is an excellent demonstration of a customer interaction with {name} who is {age} years old and lives in the {location} timezone.

# Task
John, the agent, answers questions briefly, succinctly, and in a personable manner using markdown and even adds some personal flair with appropriate emojis. John also uses the following documentation to inform his response:

# Documentation
{documentation}
```

Customer Context

Seth

```
{ "name": "Seth", "image": "/images/sethjuarez.jpg", "age": 40, "location": "Pacific" }
```

Company Context

None Selected

How can we help you?

Function Calling and Plugins

Introducing:

Azure OpenAI Service Plugins (coming soon)

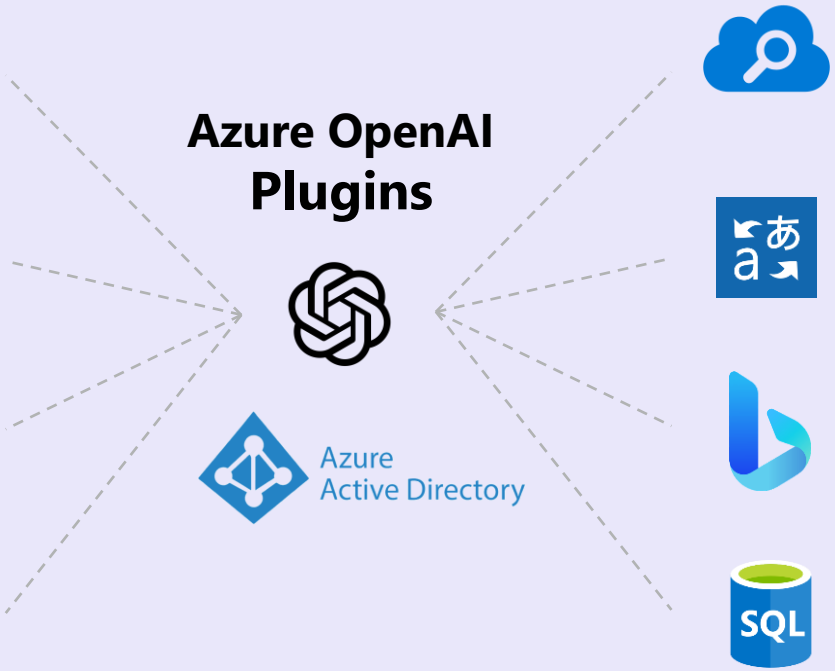
Build powerful AI Copilots with secure access to Microsoft services

Retrieve data with Azure Cognitive Search

Translate >100 languages with Azure Translator

Ground with recent info with Bing Search

Extract structured data from Azure SQL



- Securely access your data in various data stores, vector databases and the web
- Data path access controlled via Azure AD and Managed Identities
- Admin roles to choose what plugins to enable

Introducing:

Azure OpenAI Service Plugins (coming soon)

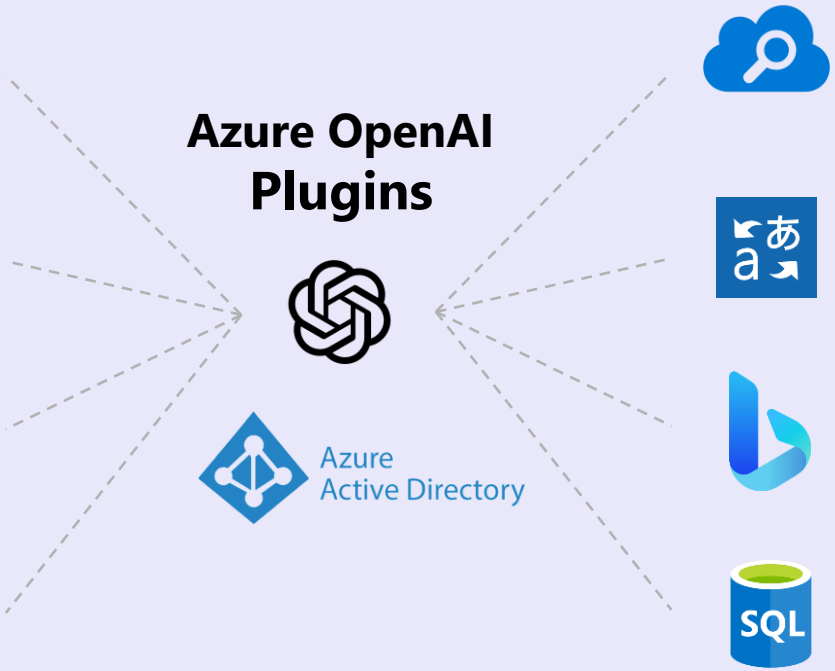
Build powerful AI Copilots with secure access to Microsoft services

Retrieve data with Azure Cognitive Search

Translate >100 languages with Azure Translator

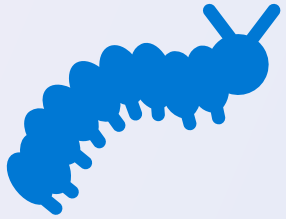
Ground with recent info with Bing Search

Extract structured data from Azure SQL

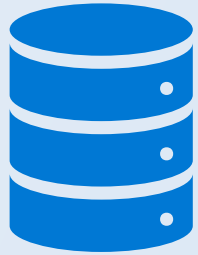


- Securely access your data in various data stores, vector databases and the web
- Data path access controlled via Azure AD and Managed Identities
- Admin roles to choose what plugins to enable

Azure OpenAI Service Function Calling



Latest 0613 versions of gpt-35-turbo and gpt-4



Use Cases

Retrieve data from data sources and APIs

Integrating with APIs or tools

Creating structured outputs

You can have confidence when using Azure OpenAI Service

When you use Azure OpenAI Service, your prompts (inputs) and completions (outputs), your embeddings, and your training data

Are NOT available to other customers.

ARE NOT available to OpenAI.

Are NOT used to improve OpenAI models.

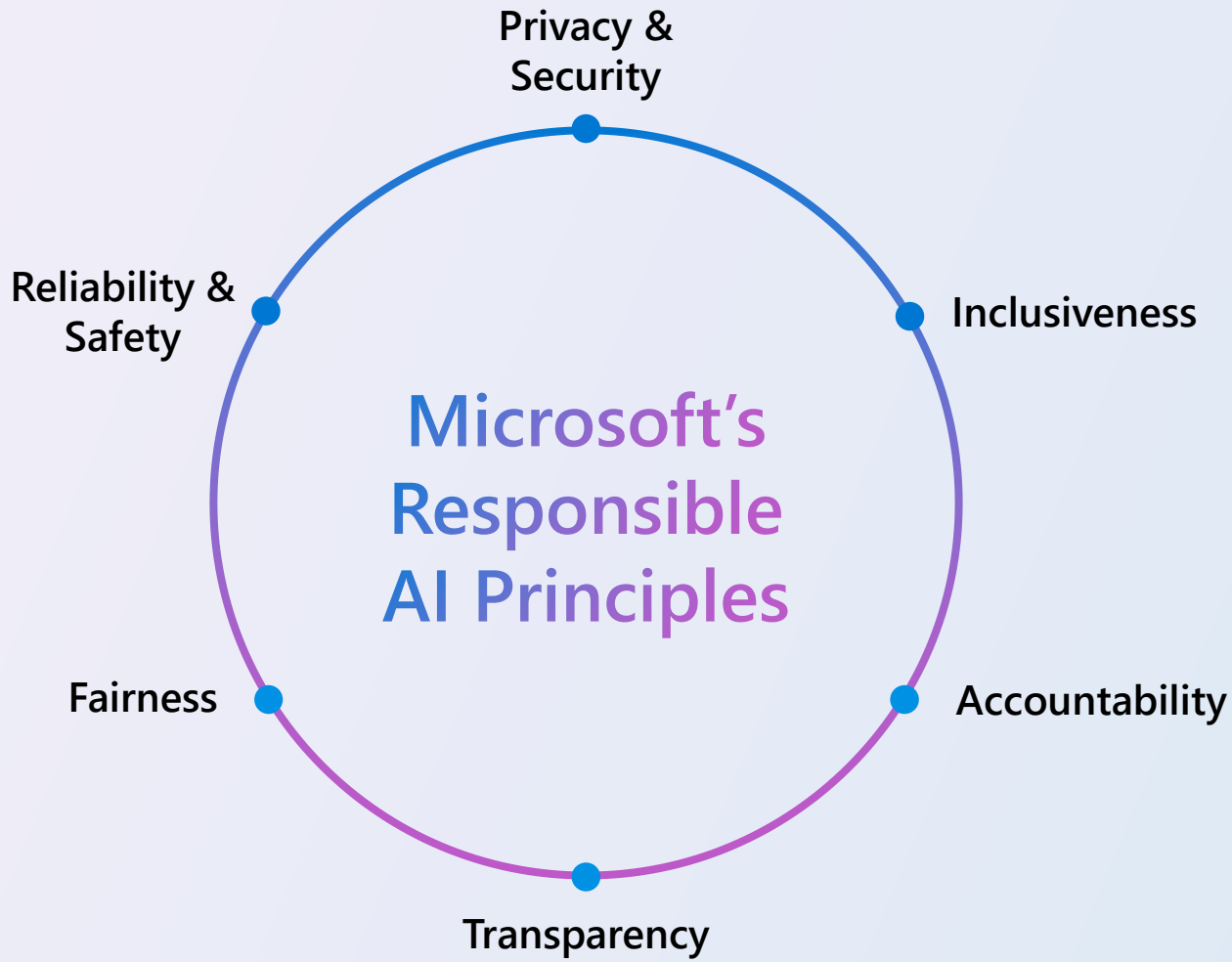
Are NOT used to improve any Microsoft or 3rd party products or services.

Are NOT used for automatically improving Azure OpenAI models for your use in your resource (The models are stateless, unless you explicitly fine-tune models with your training data).

Your fine-tuned Azure OpenAI models are available exclusively for your use.

The Azure OpenAI Service is fully controlled by Microsoft; Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does NOT interact with any services operated by OpenAI (e.g., ChatGPT, or the OpenAI API).

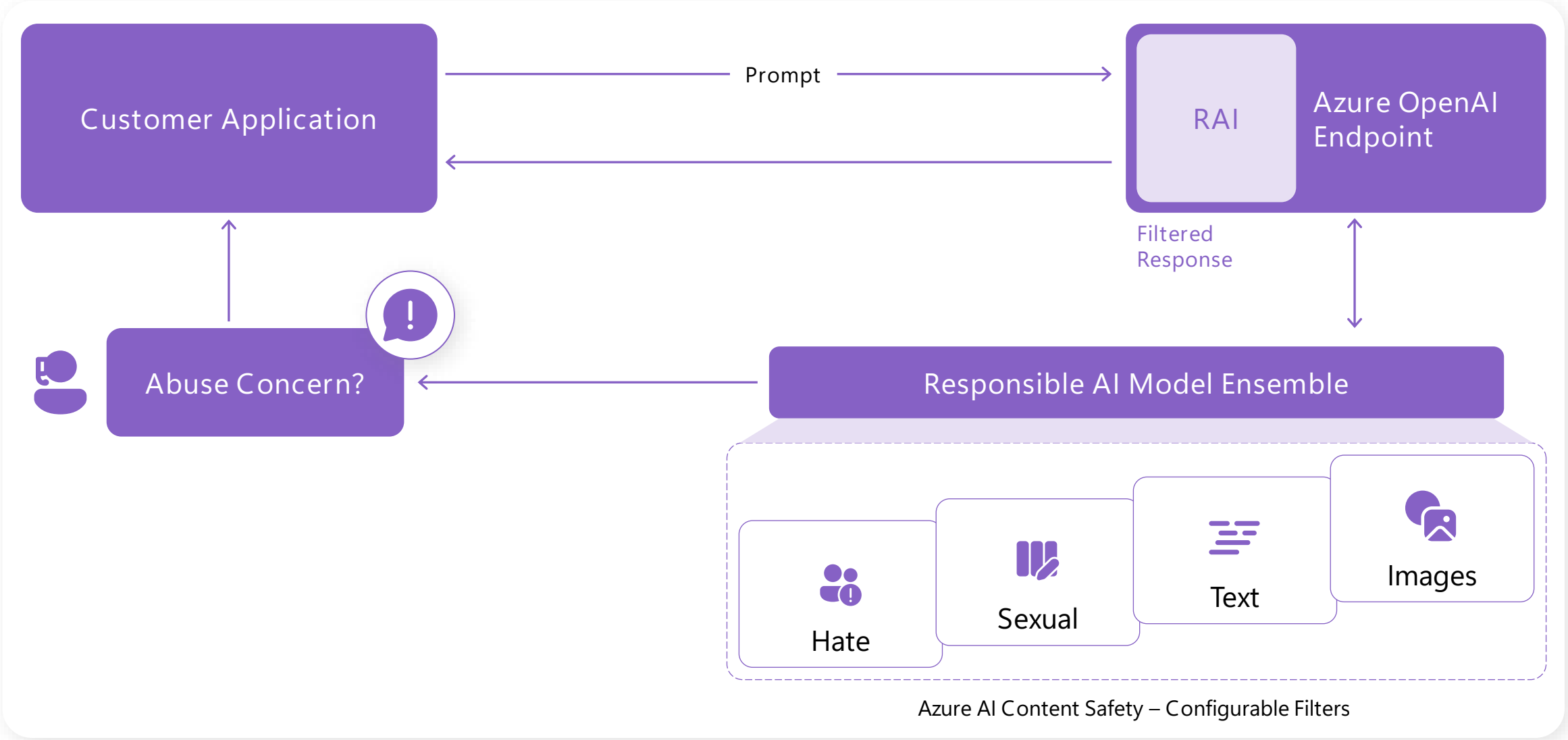
Microsoft's Responsible AI Principles



Building blocks to enact principles

- Tools and processes
- Training and practices
- Rules
- Governance

Responsible AI in Azure OpenAI Service



Introducing
Azure AI Content Safety

Azure AI Content Safety uses AI to help you create safer online spaces.

- With cutting edge AI models, it can detect hateful, violent, sexual, and self-harm content and assign it a **severity score,** allowing businesses to prioritize what content moderators review.
- **Azure AI Content Safety can handle nuance and context**, which eases the load on human content moderator teams.
- Azure AI Content Safety isn't one-size-fits-all—**it can be customized to help businesses implement their policies.** Plus, its multi-lingual models enable it to **understand many languages simultaneously.**

1 **Azure AI Content Safety classifies harmful content into four categories:**

Hate Sexual Self-harm Violence

2 **Next, it returns a four or eight severity level for each category:**

Hate: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7
Sexual: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7
Self-harm: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7
Violence: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7

3 **Then, users take actions based on the severity levels:**

Auto allowed
Auto rejected
Send to human moderator

Introducing
Azure Content Safety

AI offers more sophisticated approach t
o content moderation

Understands Nuance

An AI model can better understand context and nuance than traditional content moderation tools

Responsible AI

Microsoft places responsible AI at the heart of our innovation

Azure OpenAI Service content filtering

The service includes a content filtering system that works alongside core models. This system works by running both the prompt and completion through an ensemble of classification models aimed at detecting and preventing the output of harmful content.

Supported languages: English, German, Japanese, Spanish, French, Italian, Portuguese, and Chinese

1

Classifies harmful content into four categories via Azure OpenAI API response

- Hate
- Sexual
- Violence
- Self-harm

2

Returns a severity level score for each category from 0 to 6

- 2
- 0
- 4
- 6

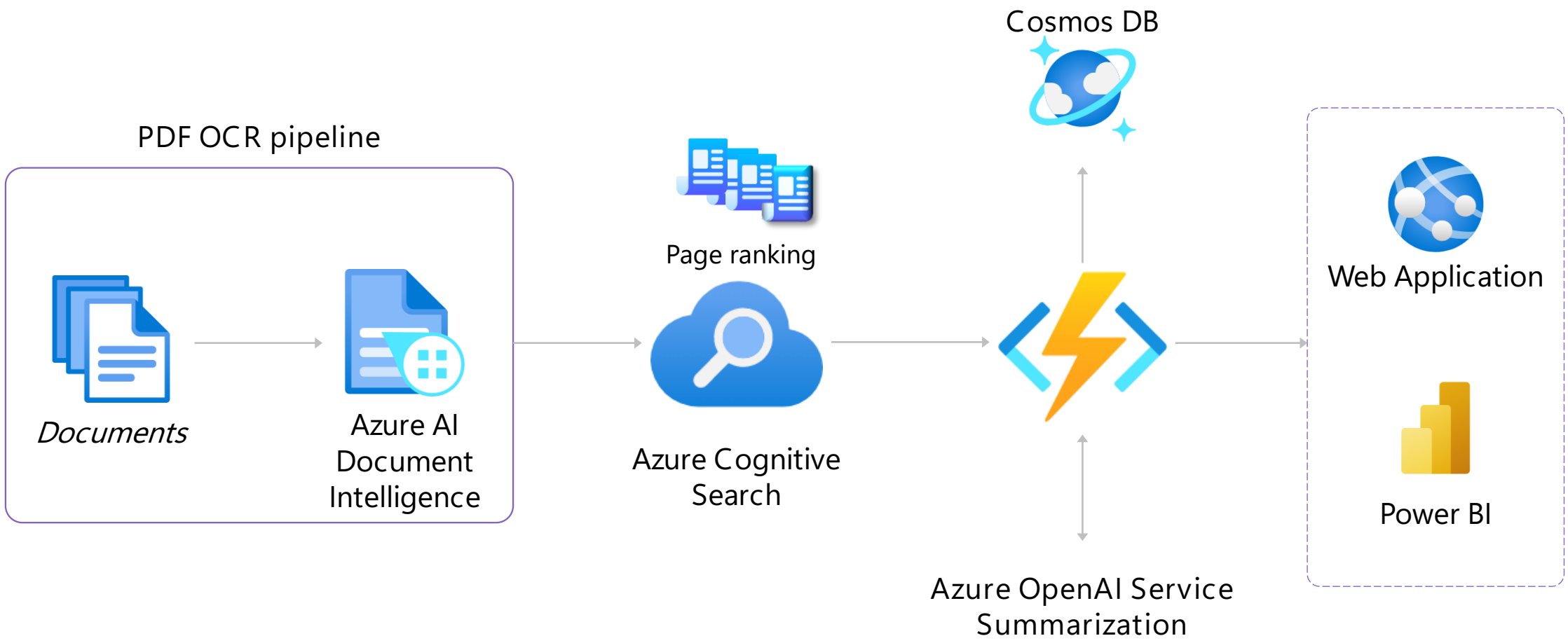
3

Surfaces content based on the severity level

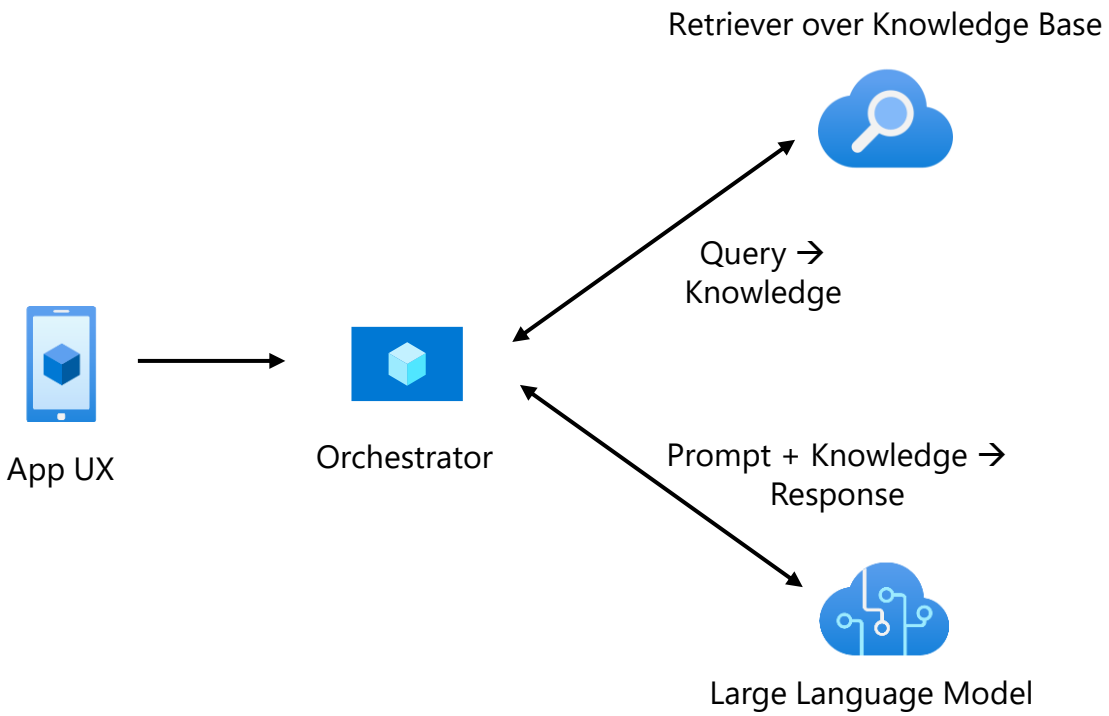
High risk: Auto blocked
Medium risk: Auto blocked
Low risk: Auto approved

Architecture Examples

Document processing and summarization



Anatomy of a RAG app



Build your own experience

UX, orchestration, calls to retriever and LLM

e.g., Copilots, in-app chat

Extend other app experiences

Plugins for retrieval, symbolic math, app integration, etc.

e.g., plugins for OpenAI ChatGPT

Customer Story – Case Study/Talking Points

CARMAX

Content Generation & Summarization

Customer:
CarMax

Industry:
Retailer

Size:
10,000+ employees

Country:
United States

Products and services:
Azure AI
Azure OpenAI Service

[Read full story here]

"With the help of Azure OpenAI Service, we're disrupting our industry for a second time by delivering cutting-edge digital tools and capabilities and becoming a true omnichannel retailer."

— Shamim Mohammad, Executive Vice President and Chief Information and Technology Officer, CarMax

Situation:

With 45,000 cars in its inventory, CarMax needed a fast and efficient way to analyze customer reviews and provide brief, meaningful summaries for each model that would aid potential purchasers and boost the pages' search engine rankings.

Solution:

After choosing to work with OpenAI, CarMax migrated to OpenAI Service to take advantage of the scalability, security, and Responsible AI features it provides.

Impact:

CarMax was able to produce the equivalent of 11 years' worth of car summaries in a matter of months, freeing editorial staff to focus on more substantive content, providing customers with valuable insights, and successfully boosting search rankings.

Microsoft

Customer Experience

Customer:
Take Blip

Industry:
Professional Services

Size:
Large (1,000 – 9,999 employees)

Country:
Brazil

Products and services:
Azure OpenAI Service
Azure AI

Read full story here

"By using Azure OpenAI Service [...] Take Blip is taking a leading-edge approach to AI that makes us future-ready and gives us a competitive edge in the marketplace for customer experience technology."

— Milton Stiilpen, Innovation & Research Director, Take Blip

Situation:

Take Blip wanted to pursue an AI-first approach to messaging between brands and customers using the latest language models, like GPT-4.

Solution:

Take Blip began using Azure OpenAI Service and other Azure Cognitive Services to develop a robust, multichannel, AI-driven customer conversation platform backed a highly secure and scalable cloud infrastructure.

Impact:

Using Microsoft AI technologies has boosted developer productivity, accelerating time to market for new campaigns. Clients are excited about the new AI-driven features and have enthusiastically embraced GPT capabilities.

Trelent

Customer Experience

Customer:
Trelent

Industry:
Professional Services

Size:
1-49 employees

Country:
Canada

Products and services:
Azure OpenAI Service

[Read full story here]

“With a product like OpenAI Service behind you, you can focus a lot more on what really matters, which is delivering a great experience, a great product, and a lot of value to your customers.”

— Calum Bird, CEO, Trelent

Situation:

Trelent, a pre-seed code documentation startup that uses the OpenAI Codex algorithm, found its audience in high-growth tech companies with large or distributed engineering teams. The problem: How does a two-person team provide enterprise-ready service?

Solution:

A Microsoft for Startups webinar provided the answer: Microsoft Azure OpenAI Service paired the powerful OpenAI algorithms Trelent was already using with Azure security, safety controls, and global availability—and worked with the startup’s existing solution.

Impact:

With Azure OpenAI Service, Trelent benefits from content filtering, increased security, and faster response times (from 1-3s down to an average of 750ms). Those enterprise-ready features free Trelent engineers up to focus on their core innovation.

Microsoft

Thank you