

# Azure and Open AI:

Partners in transforming the  
world with AI

Technical Bootcamp  
06 June 2023





# Microsoft Team



Ben Rohling  
CTO -GPS  
CEMA



Alex Peles  
CEMA - SE  
Cloud Solution  
Architect



Jurgens De Bruin  
CEMA - Africa  
Cloud Solution  
Architect



Andrey Vykhodtsev  
CEMA - SE  
AI Cloud Solution  
Architect



Haroon Rashid  
CEMA - Middle East  
Cloud Solution  
Architect



Mahmoud Moussa  
CEMA - Middle East  
Cloud Solution  
Architect



# Today's Opportunity – Your Opportunity



**Ben Rohling**  
**CTO - CEMA**

**Today**

**Momentum**

**THE Platform for AI**

**Your Opportunity**

# Agenda

Topic	Speaker
<b>Opening</b>	<b>Ben</b>
<b>Azure OpenAI Service Overview</b>	
<b>Introduction to AOAI Models</b>	
Overview of ChatGPT (Large Conversational Foundation Model)	<b>Haroon</b>
Overview of GPT-3 / GPT-4 (Large Language Foundation model)	
Overview of DALL-E (Large Image to Text Foundation Model)	
Overview of Codex (Large Code Foundation Model)	
<b>Responsible AI</b>	
<b>Why OAI vs. AOAI</b>	<b>Alex</b>
<b>Uses and Capabilities</b>	
<b>Demo</b>	
<b>Deep Drive Architecture for Azure OpenAI</b>	
Different Azure Open AI Architecture	<b>Andrey</b>
Document Process Automation	
Q&A with Semantic Answering with Azure OpenAI Service	
Contact Center Analytics using Speech API & Azure OpenAI Service	
AI-Powered Q&A over Enterprise Data Sources	
<b>Available Azure OpenAI Accelerators</b>	
<b>Overview and Technical deep dive of GPT -3/4 models</b>	
Model naming convention	<b>Moussa/Jurgens</b>
Understanding Tokens and Probabilities	
Key concepts: Natural Language Understanding and Generation examples (Zero-Shot, One-Shot, Few-Shot)	
ChatGPT Prompt Examples	
Model Adaptation (Fine-Tuning)	
Embeddings	
Prompt Engineering Guide and Techniques	



# Azure OpenAI Service and models

Haroon Rashid  
Alex Peles  
Cloud Solution Architect



# Agenda

- Azure OpenAI service overview
- Introduction to AOAI model
  - Overview of GPT-3/4 (Large Language Foundation model)
  - Overview of DALL-E (Large Image to Text Foundation Model)
  - Overview of ChatGPT -3/4 (Large Conversational Foundation Model)
- Responsible AI
- Why OAI vs. AOAI
- Data Privacy
- Uses and capabilities
- Demo

## Artificial Intelligence

### Machine Learning

### Deep Learning

### Generative AI



## Artificial Intelligence

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

---



## Machine Learning

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

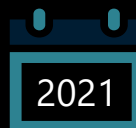
---



## Deep Learning

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

---



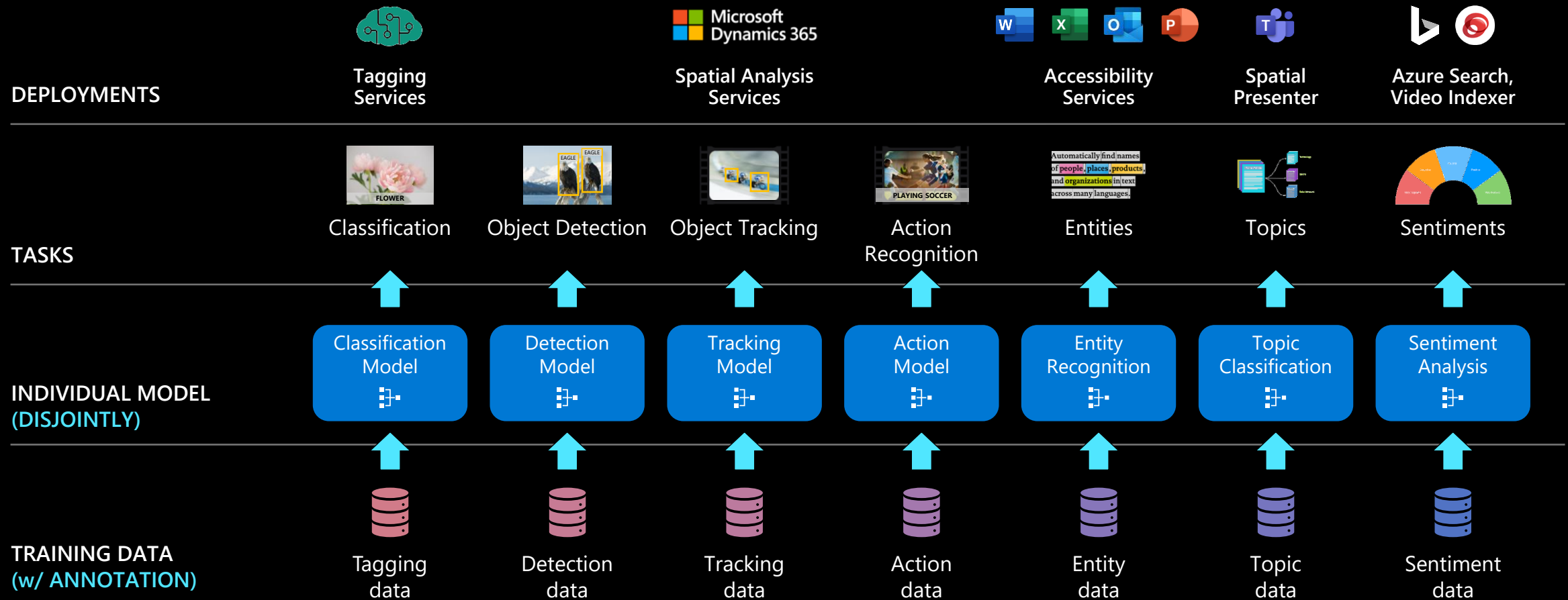
## Generative AI

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

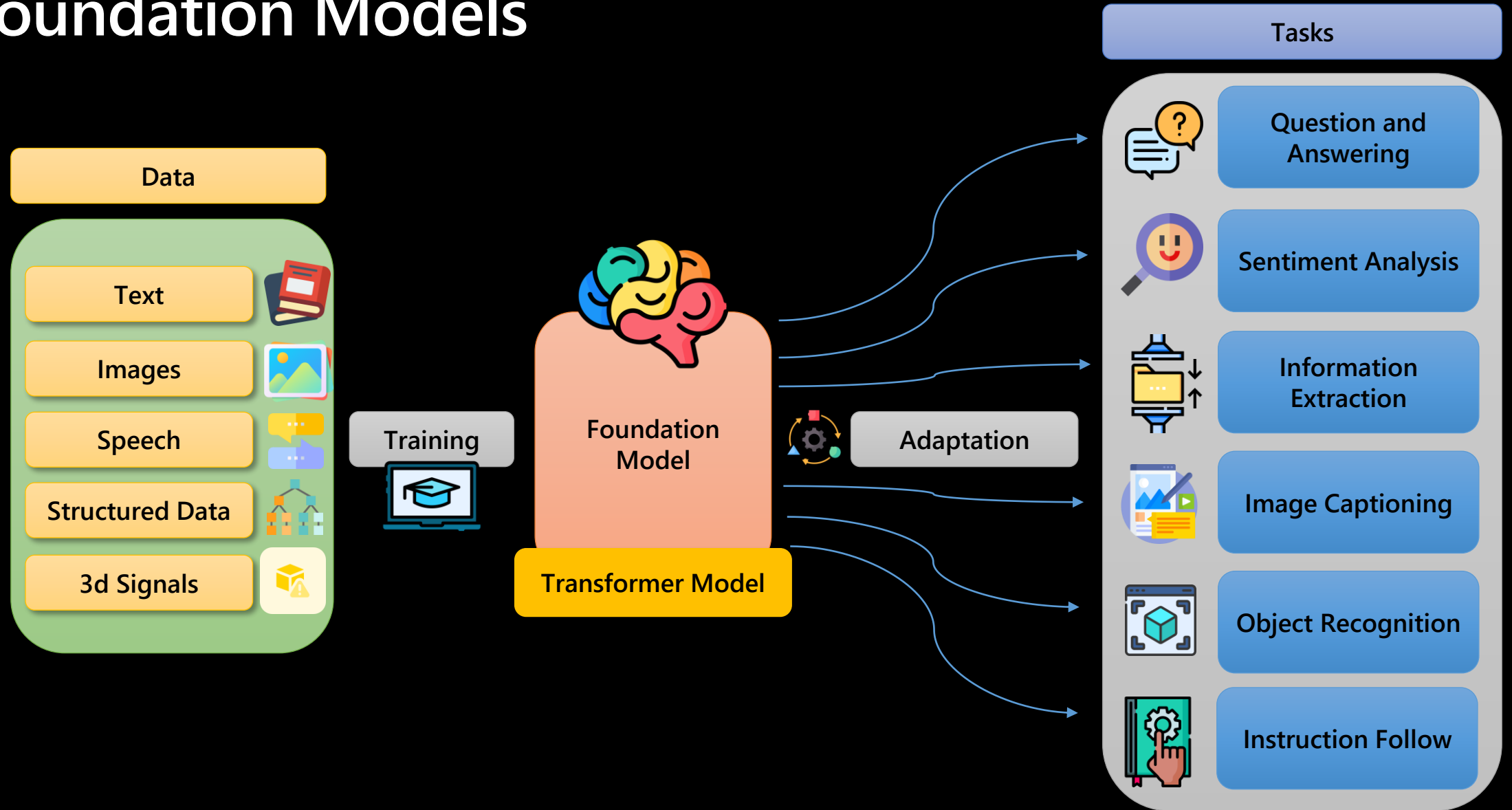


# Traditional model development

High cost and slow deployment—each service is trained disjointly



# Foundation Models



# Large Language Foundation models

Microsoft AI at Scale

## Project Turing

**17B**  
Turing-NLG | 2020

**340M**  
BERT-Large | 2018



Our mission is to expand the boundaries of natural language understanding.

machine reading comprehension, question answering, transfer learning, reinforcement learning, computer vision, and even building interpretable models

Turing-NLG

Megatron NLG Turing

T-ULRv5

DeepSpeed

ZeRO

**175B**  
GPT-3 | 2020

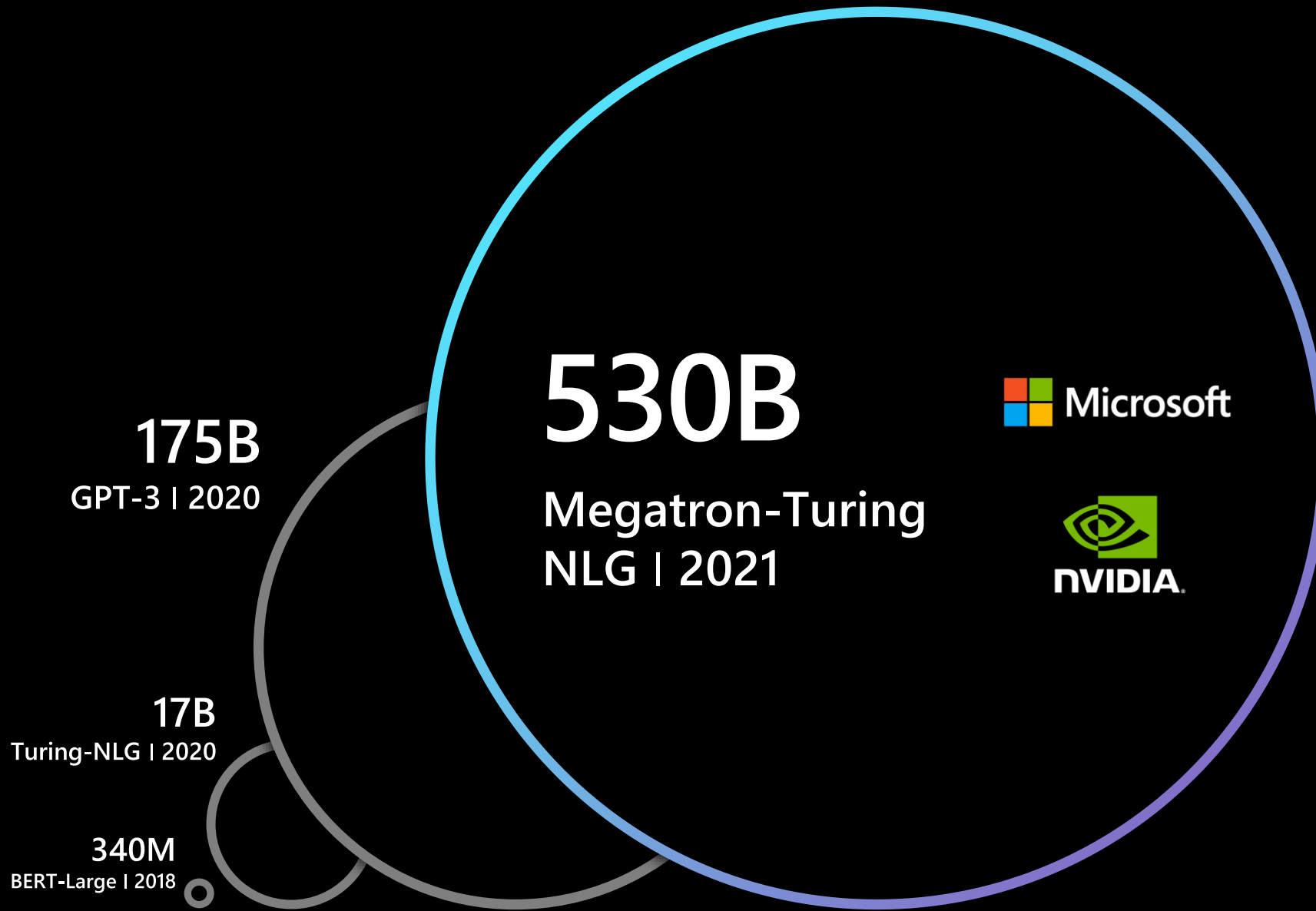
**17B**  
Turing-NLG | 2020

**340M**  
BERT-Large | 2018



**OpenAI**





**175B**  
GPT-3 | 2020

**17B**  
Turing-NLG | 2020

**340M**  
BERT-Large | 2018

**530B**

Megatron-Turing  
NLG | 2021



# Our partnership with OpenAI



*Ensure that artificial general intelligence (AGI) benefits humanity.*



*Empower every person and organization on the planet to achieve more*

---

**GPT-3x & 4x**

Generate and Understand Text

**DALL·E**

Generate images from text prompts

**ChatGPT**

Generate and Understand Code

# Generative AI

## Content creation by API

GPT-4

Prompt:

Write a tagline for an ice cream shop.

Response:

We serve up smiles with every scoop!

DALL·E

Prompt: A white Siamese cat

Response:



ChatGPT

What is the fastest animal on land?

the cheetah (*Acinonyx jubatus*), which can reach speeds of up to 60 miles (97 kilometers) per hour.

What makes them so fast?

Cheetahs are built for speed and have several adaptations that make them the fastest land animal: lean body, long legs, flexible spine, large nostrils and claws that don't retract.



# Azure OpenAI Service

GPT-4

DALL·E

ChatGPT



Deployed in your Azure subscription, secured by you, and tied to your datasets and applications



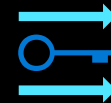
Large, pretrained AI models to unlock new scenarios



AI models, some custom-tunable with your data and hyperparameters



Built-in responsible AI to detect and mitigate harmful use



Enterprise-grade security with role-based access control (RBAC) and private networks



# OpenAI GPT-3

*Why is the sky blue?*

**OpenAI GPT-3 answer:**

The sky is blue because the Earth's atmosphere scatters sunlight in all directions and blue light is scattered more than the other colors because it travels as shorter, smaller waves

# OpenAI

## DALL·E 2

An astronaut riding a horse in a photorealistic style



Teddy bear working on new AI research on the moon in 1980



A bowl of soup that looks like a monster knitted out of wool





Text to image

Image to video

## NUWA | Infinity

A cloudy morning  
on the beach with  
the tide coming in



Mountains  
with clouds



# | Azure OpenAI | GPT-3 Family of Models

Model	Max # Tokens per Request	Description, performance, cost	Use cases
Davinci	4,096 tokens	<b>Most capable</b> GPT-3 model. Can do any task the other models can do, often with <i>higher quality, longer output and better instruction-following</i> .	Complex intent, cause and effect, summarization for audience
Curie	2048 tokens	<b>Very capable</b> , but <i>faster and lower cost</i> than Davinci.	Language translation, complex classification, text sentiment, summarization
Babbage	2048 tokens	<b>Capable</b> of straightforward tasks, <i>very fast, and lower cost</i> .	Moderate classification, semantic search classification
Ada	2048 tokens	<b>Capable</b> of very simple tasks, usually the <i>fastest</i> model in the GPT-3 series, and <u>lowest cost</u> .	Parsing text, simple classification, address correction, keywords

# | Azure OpenAI | Comparing the GPT-3.5 models

Model	Description
text-davinci-002	A GPT-3.5 model that was custom-tuned on natural language instructions and can perform a variety of tasks including summarization, question answering, classification, and more.
text-davinci-003	An improvement over the text-davinci-002 model. The model is similar to its predecessor but generally more capable across all tasks.
ChatGPT model (gpt-3.5-turbo)	A model custom-tuned for working with dialogue. ChatGPT is a great model to use for conversational tasks.

# Model Use Cases

## GPT-3

Large, pretrained Language Models that use deep learning to generate content



### Use cases:

- Generating natural language for chatbots and virtual assistants with awareness of the previous history of chat
- Power chatbots that can handle customer inquiries, provide assistance, and converse but doesn't have memory of conversations
- Automatically summarize lengthy texts
- Assist writers by suggesting synonyms, correcting grammar and spelling errors, and even generating entire sentences or paragraphs
- Help researchers by quickly processing large amounts of data and generating insights, summaries, and visualizations to aid in analysis
- Generate good quality code based on natural language

## ChatGPT (preview)

Customize generation by adding your data as context to the AI model



### Use cases:

- Generating natural language for chatbots and virtual assistants that has memory of the previous history of chat
- Improving the accuracy and efficiency of automated customer service interactions
- Enhancing language translation software for more accurate and natural translations
- Generating more human-like content for social media and marketing campaigns
- Improving the effectiveness of natural language processing for tasks such as sentiment analysis and text classification

## GPT-4 (preview)

Built-in responsible AI to detect and mitigate harmful use



### Use cases:

- Generating and understanding natural language for customer service interactions, chatbots, and virtual assistants – doesn't have memory of conversations
- Generating high-quality code for programming languages based on natural language input.
- Providing accurate translations between languages
- Improving text summarization and content generation
- Provides for multi-modal interaction (text and images)
- Substantial reduction in Hallucinations
- Consistency between different runs is high

# Azure OpenAI | Overview of GPT-3

## Generative pre-trained transformer 3 (GPT-3)

Autoregressive language model that uses deep learning to produce human-like text

Pre-trained on trillions of words

Predicts the most likely next word based on input text

General text-in/text-out interface



# Azure OpenAI | GPT-3 Models

Powerful language models accessible to all skill levels



General purpose text-in/text-out interface—flexibility



Simple UX—validate proof of concepts fast



Built in ML science intuition for everyone, with deeper controls for ML practitioners



# Azure OpenAI | GPT-3 Prompt Design

Extract the mailing address from this email:

Hi John Doe,

It was great to meet up at Build earlier this week. I thought the AI platform talk was great and I really enjoyed it.

I appreciate the offer for the book. If you are OK, you can mail it to me at home, or 123 Microsoft Way, Bellevue WA 92004.

Regards,

Chris

**Prompt**—Text input that provides some context to the engine on what is expecting.

**Completion**—Output that GPT-3 generates based on the prompt.

# Azure OpenAI | GPT-3 Ideate, Experiment and Fine-Tuning

Iterate on ideas with a general-purpose text-in/text-out interface

## Prompt

Summarize game commentary into highlights:

Shey Peddy is applying ball pressure at the top against Sabrina Ionescu. At 7:48 remaining in the quarter; Peddy

What are the main highlights of the game so far?

## Sample response

The game has been close with Phoenix leading New York 7-5. Shey Peddy has been key for Phoenix.

Refine with examples ('few shot learning') with a simple UX

## Prompt

Turn game commentary into highlights:

Commentary: What a pickup she has  
Main highlights: New York has domina  
###

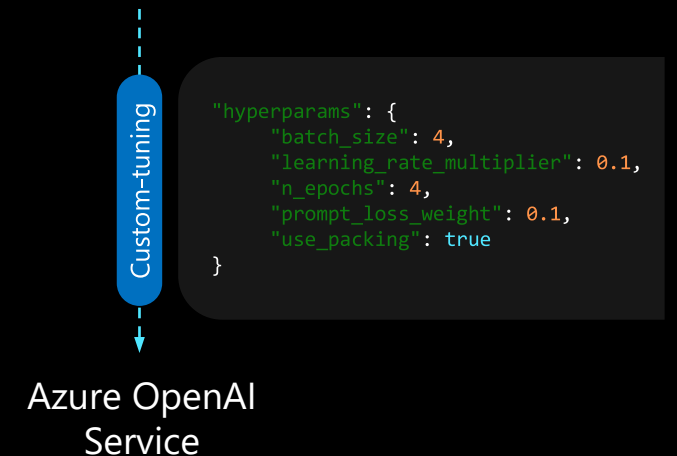
Commentary:  
1. Turner is so important defensively to  
2. Griner pulled way out, Hartley with  
3. At 1:54 remaining in the quarter, Pho

## Sample response

Main highlights:  
1. New York has had a strong run in the  
2. Phoenix leading by 1 point, 24-23  
3. New York Liberty's comeback has be

Optimize accuracy and latency to validate proof of concept fast

## Prompt and completion examples



# Azure OpenAI GPT-4 Overview

GPT-4 is a large **multimodal** model (**accepting text inputs, emitting text outputs**) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.

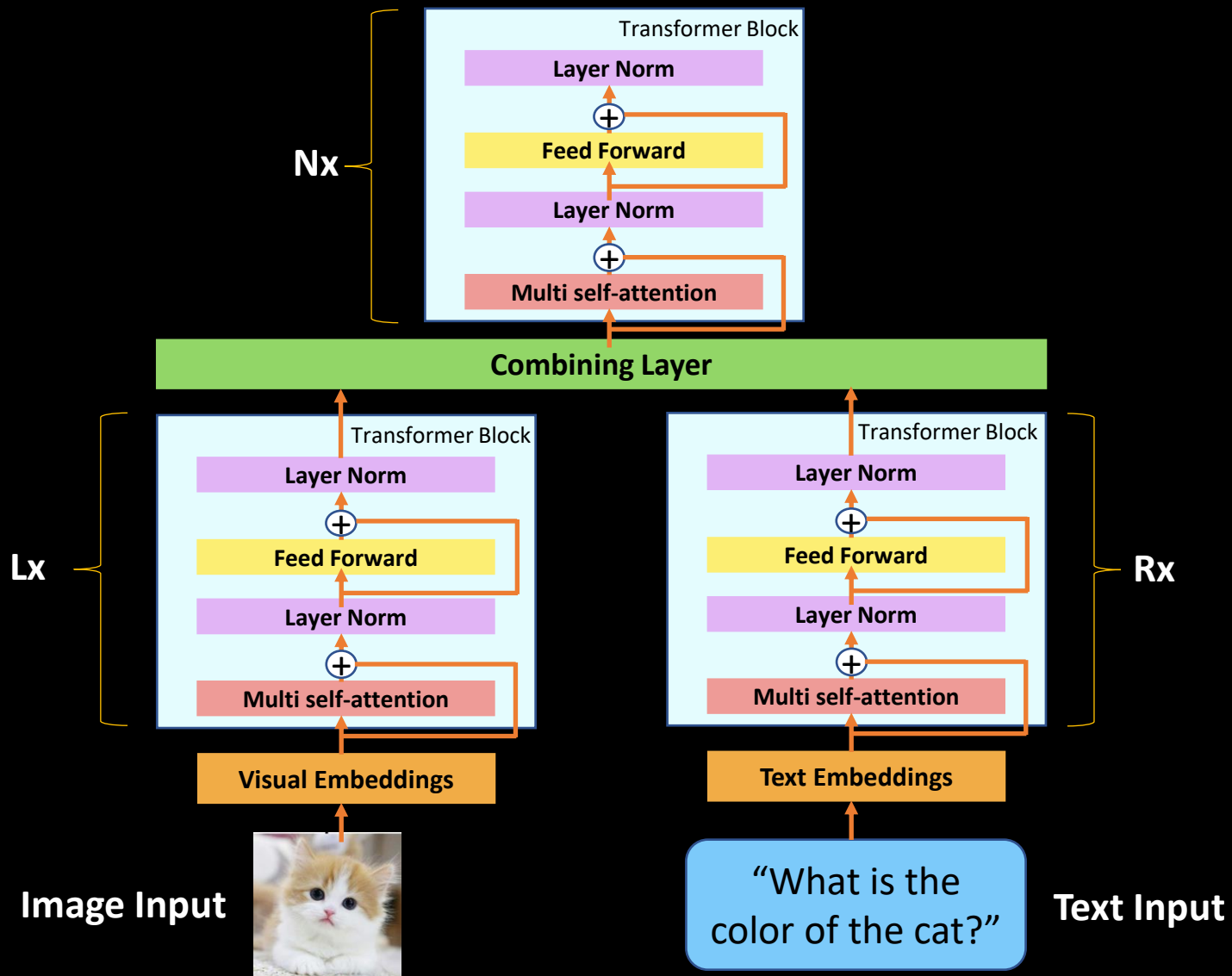
GPT-4 is **more reliable, creative**, and able to handle much more **nuanced instructions** than GPT-3.5.

Can do everything that GPT 3.5 can do

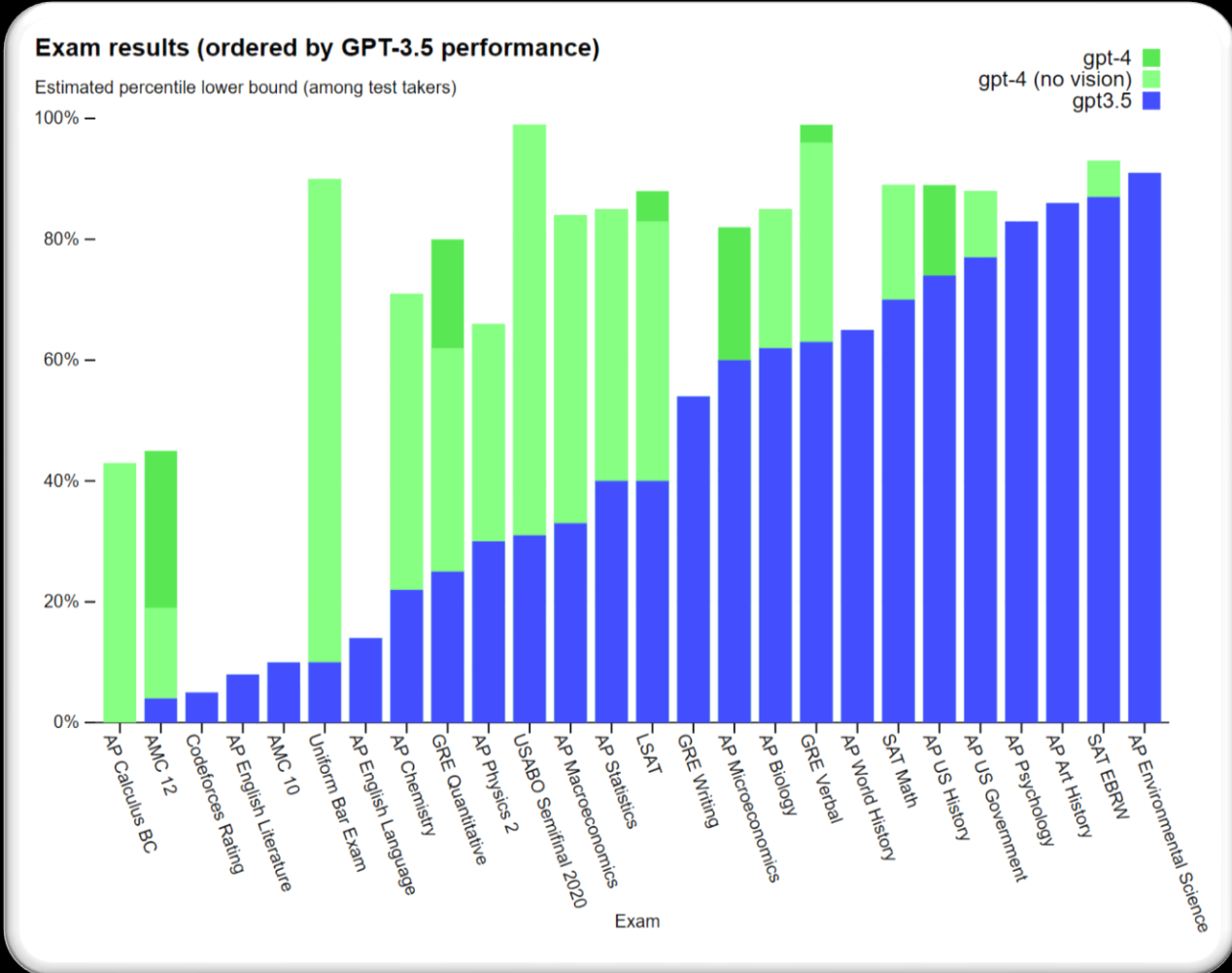
- + Visual Question and Answering (VQA)
- + Steerability
- + context length of 8,192 tokens

Apply here for GPT-4 access: [applying for access to GPT-4 today](#)

# Multi-Modal Transformer Architecture



# GPT-4 Benchmarks

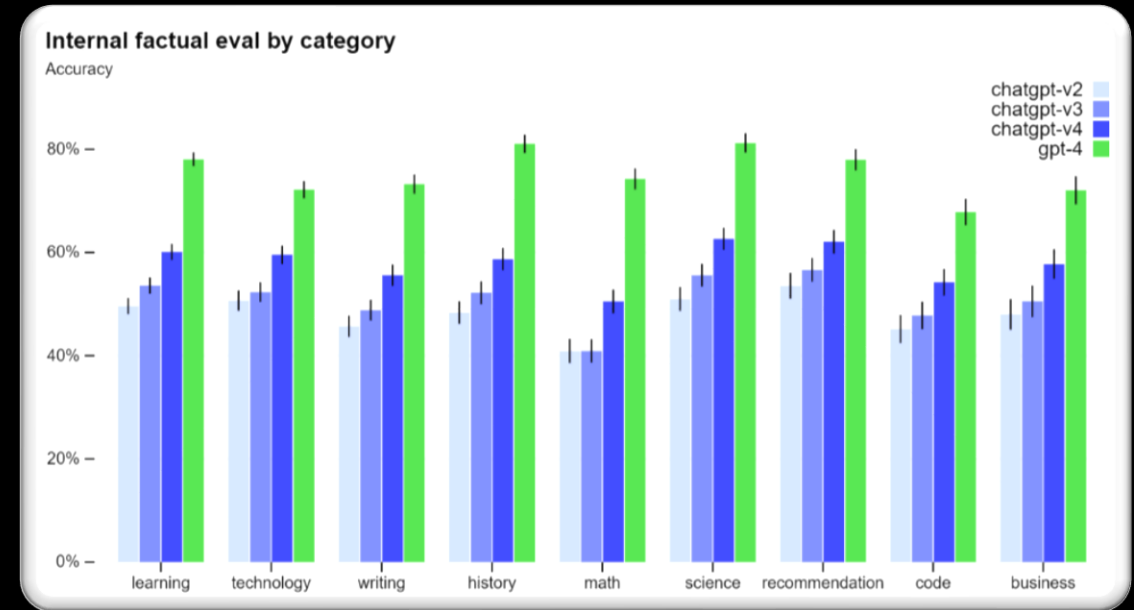


# GPT-4 Limitations

GPT-4 has similar limitations as earlier GPT models.

- not fully reliable
- Hallucinates
- makes reasoning errors
- biases in its outputs
- lacks knowledge of events that have occurred after the vast majority of its data cuts off (September 2021)
- Does not learn from its experience

GPT-4 significantly reduces hallucinations relative to previous models (which have themselves been improving with each iteration). GPT-4 scores 40% higher than the latest GPT-3.5 in OpenAI's internal adversarial factuality evaluations



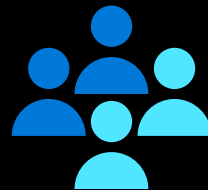
# DALL·E 2



Generate 50/images per minute  
(default rate, can be increased)  
with simple text prompts



Accelerate designs or inspire  
creative decision



Build capability into enterprise  
applications through APIs and SDKs



## Protection of species on our doorstep

[GEOlino](#)

Episode 14 • 04.08.2021

25:48 min left

▶ Continue

We humans are causing a loss of species with our way of life and in this episode, we show you ways in which you can help. Whether by building nesting spaces or counting hamster burrows, get ready to roll up our sleeves and support the endangered species!







A DTM race car like a hot rod

Let's start with your content


< Generate an image using a description ⓘ

Ombre cake decorated with flowers and fall foliage 



Tip  
Explore different camera angles  
"An aerial still of a seascape, Brazil"

Color ▾ Size ▾



Continue editing Share




# The ChatGPT model


Unlike previous GPT-3 models, the ChatGPT model is specifically designed to be a conversational interface.

The conversational nature of the model makes it easier to interact with and to take advantage of the full power of its capabilities.

The prompts used with the ChatGPT model are also different than previous models because of the conversation nature and a new Chat API is available for using with the model.

## ChatGPT

 Examples	 Capabilities	 Limitations
"Explain quantum computing in simple terms" →	Has awareness of what the user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Send a message... 

# Working with the ChatGPT model

There are two different APIs available for working with the ChatGPT model.

## Chat Completion API (Recommended)

A new API designed for working with Chat models (gpt-3.5-turbo and gpt-4). We recommend using this API moving forward.

In this API, you pass in a series of messages rather than the prompt as a raw string making it easier to manage the conversation.

---

```
messages = [  
  {  
    "role": "system",  
    "content": "Assistant is a large language model trained by OpenAI."  
  },  
  {  
    "role": "user",  
    "content": "What is a garbanzo bean?"  
  }  
]
```

## Completion API

The ChatGPT model also works with the existing completions API. In this API, you'll format your prompt as a string and need to use the special tokens from [Chat Markup Language](#).

We generally recommend using the Chat API instead because the special tokens may get updated in future version of gpt-3.5-turbo meaning you would need to update your prompts before updating to newer versions of the model.

---

```
<|im_start|>system  
Assistant is a large language model trained by OpenAI.  
<|im_end|>  
<|im_start|>user  
What is a garbanzo bean?  
<|im_end|>  
<|im_start|>assistant
```

# Working with the ChatGPT model

## Previous GPT-3 models

Previous models were text-in and text-out

(i.e., they accepted a prompt string and returned a completion to append to the prompt).

---

Answer questions from the context below.

Context:

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.

Q: What is a neutron star?

A:

## The ChatGPT model

The ChatGPT model is conversation-in and message-out.

(i.e., it expects a prompt string that is formatted in a specific chat-like transcript format and returns a completion that represents a model-written message in the chat)

---

[System Message]

Assistant is an AI Chatbot designed to answer questions from the context provided below.

Context:

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.

[User]

What is a neutron star?

[Assistant]

# Understanding the ChatGPT prompt format

## The 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" }
```

# ChatGPT limitations



## Hallucinations

While the ChatGPT model has proven to have extensive knowledge, it can still be wrong at times. It's important to understand this limitation and apply mitigations for your scenario.

## No custom-tuning

We do not currently offer custom-tuning for the ChatGPT model. Instead, we recommend customizing the model by tailoring the system message to your use case or using few shot examples of messages between the user and the assistant.

# ChatGPT benefits



## Conversational

The conversational nature of the model makes it easier to interact with so you can more easily get the most out of the model.

## Multi-turn

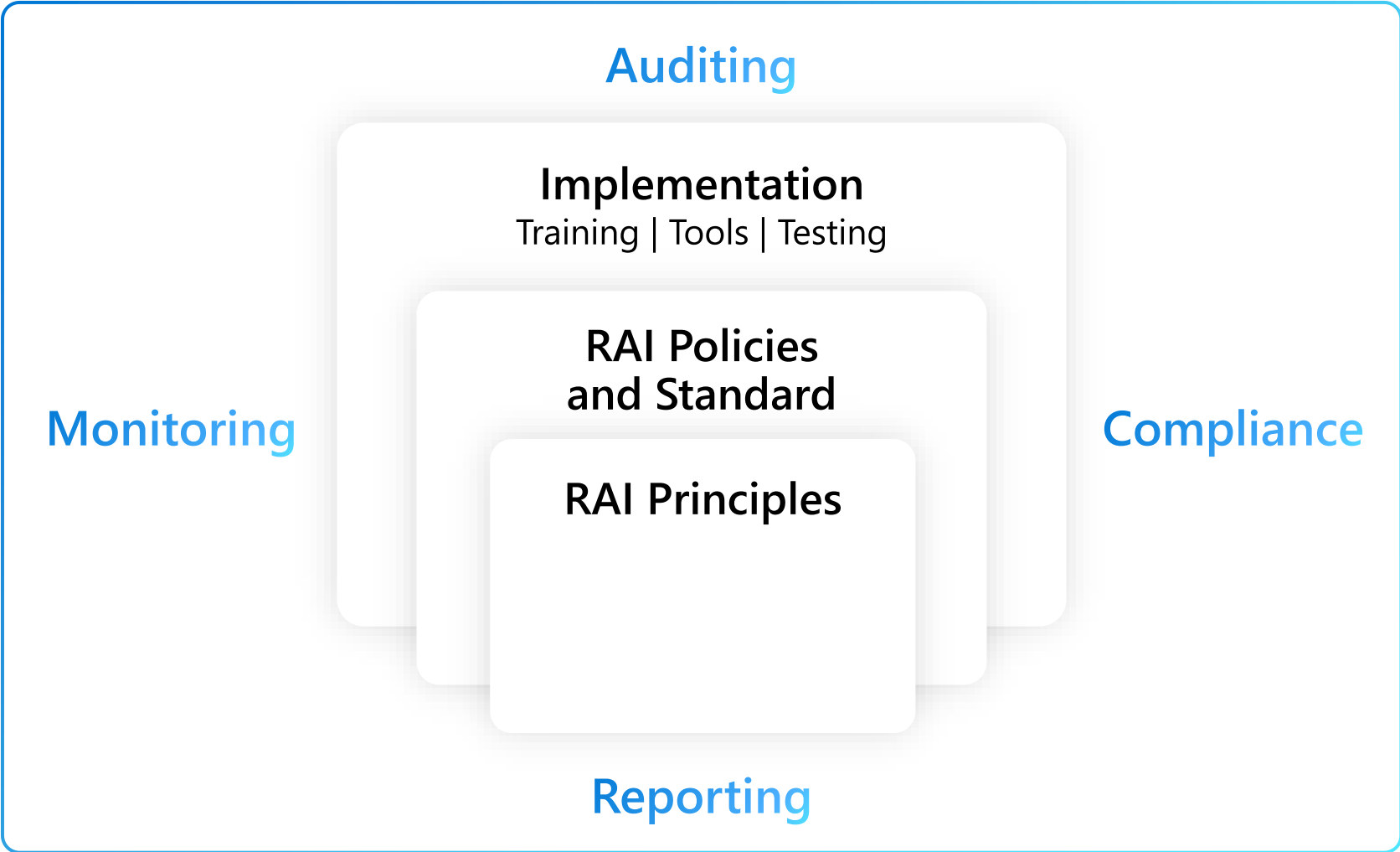
The conversational nature of ChatGPT makes it easy to follow up on the model's response. This gives users an easy mechanism to ask suggest edits, ask for clarification, etc.

## Creative

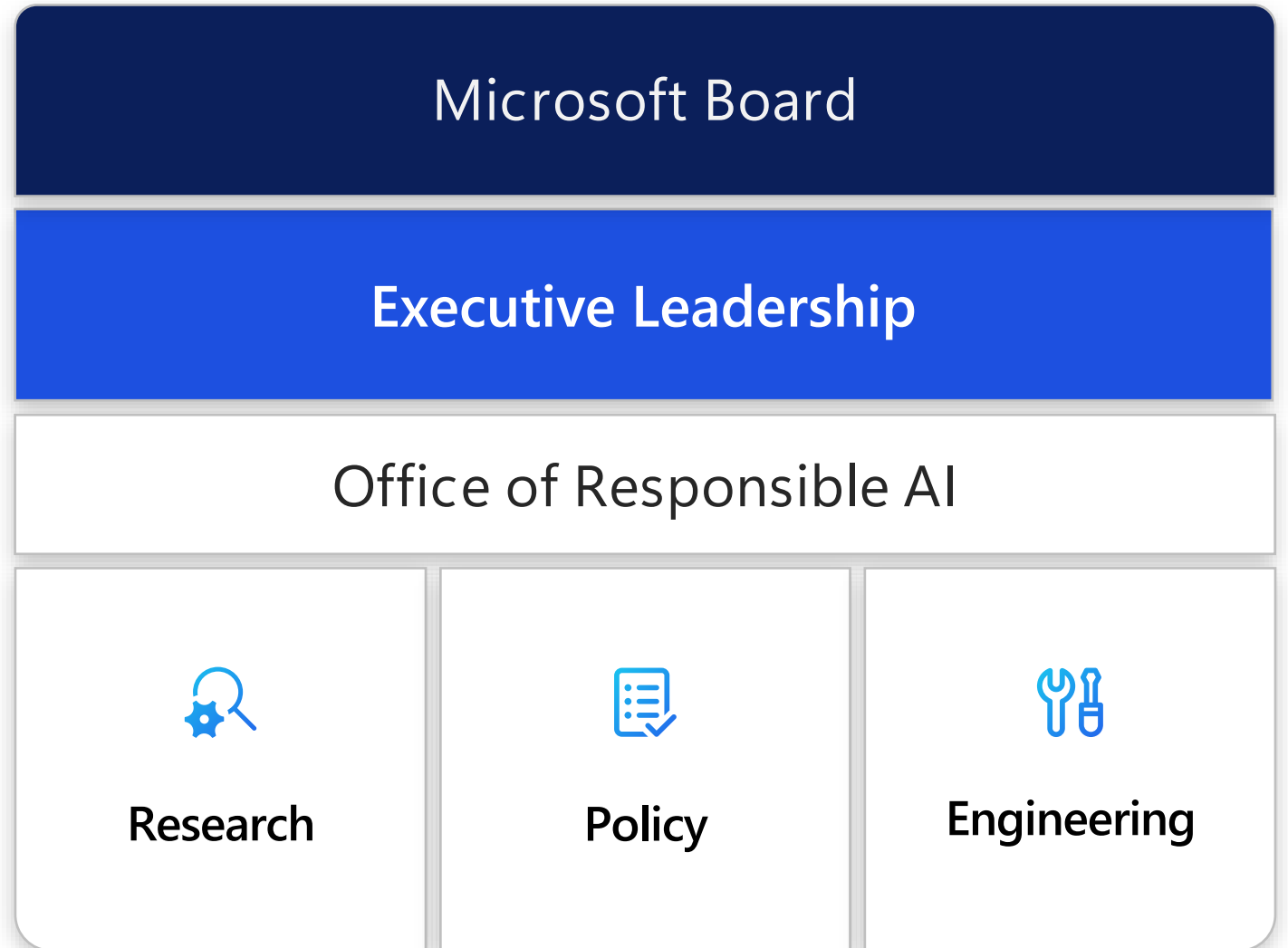
The ChatGPT model excels at creative tasks like content writing and storytelling.



# Responsible AI Governance Framework



# Our ecosystem



# The Anatomy of the Responsible AI Standard

## Principles

> Which **enduring values** guide our responsible AI work?

## Goals

> What are the **outcomes** that we need to secure?

## Requirements

> What are the **steps we must take** to secure the Goals?

## Tools and Practices

> Which **aids** can help us meet the Requirements?

# The Standard's Goals at a Glance

## Accountability

- A1: Impact Assessment
- A2: Oversight of significant adverse impacts
- A3: Fit for purpose
- A4: Data governance and management
- A5: Human oversight and control

## Transparency

- T1: System intelligibility for decision making
- T2: Communication to stakeholders
- T3: Disclosure of AI interaction

## Fairness

- F1: Quality of service
- F2: Allocation of resources and opportunities
- F3: Minimization of stereotyping, demeaning, and erasing outputs

## Reliability & Safety

- RS1: Reliability and safety guidance
- RS2: Failures and remediations
- RS3: Ongoing monitoring, feedback, and evaluation

## Privacy & Security

- PS1: Privacy Standard compliance
- PS2: Security Policy compliance

## Inclusiveness

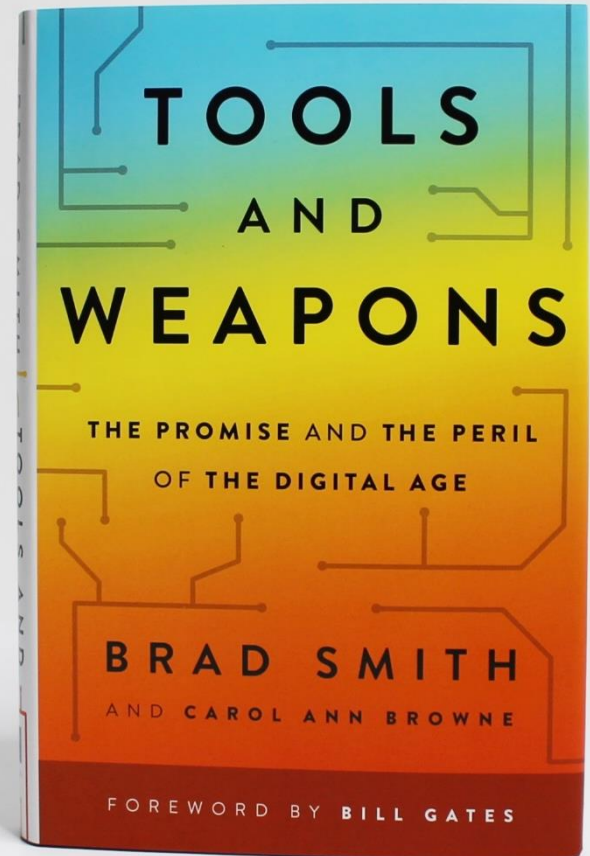
- I1: Accessibility Standards compliance

## Why responsible AI?

// When your technology changes the world, you bear a responsibility to help address the world that you have helped create. //

**Brad Smith**

President and Chief Legal Officer, Microsoft

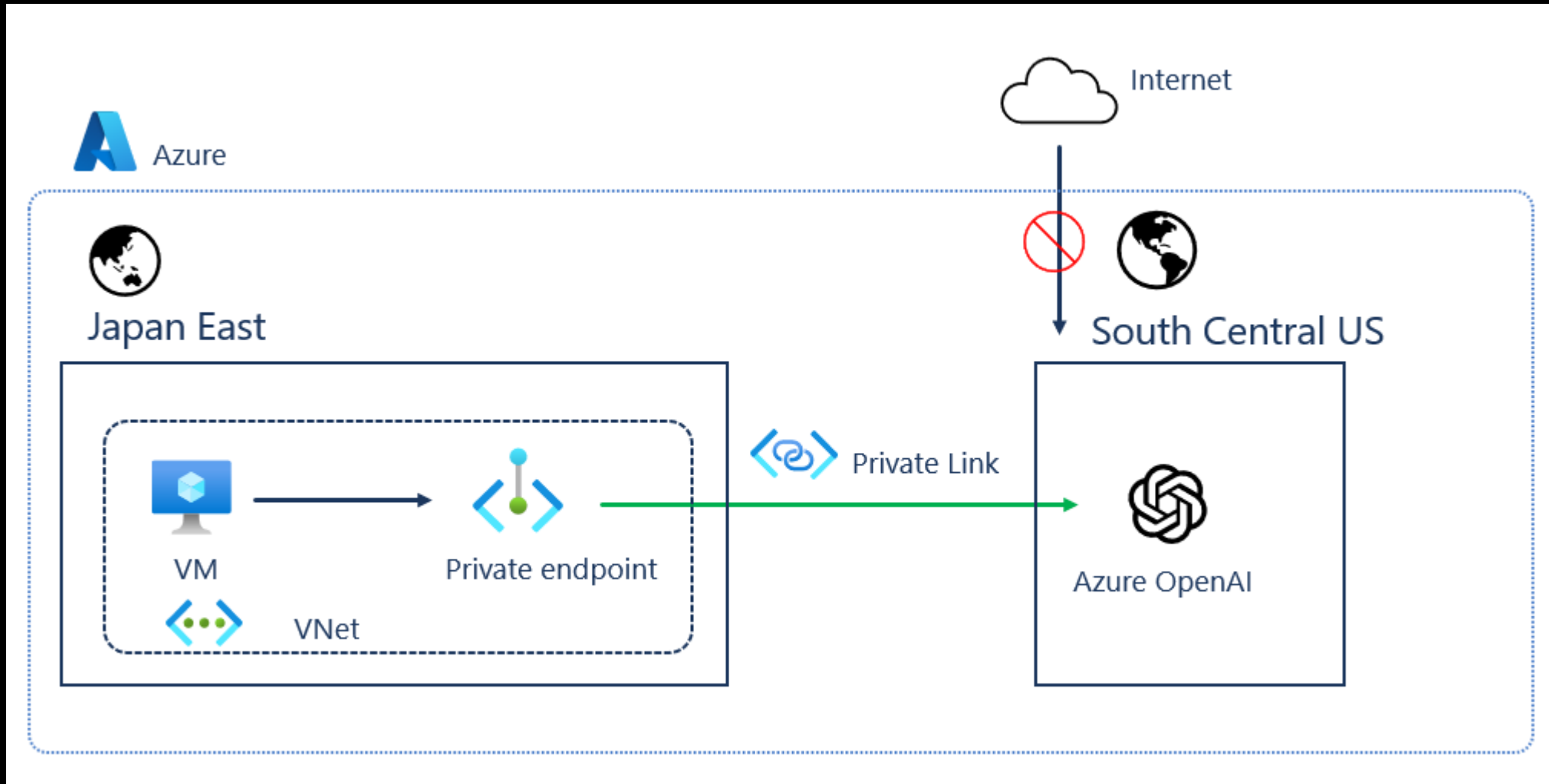


[Responsible AI principles from Microsoft](#)

# Why Azure OpenAI

- Security – AAD RBAC
- Security – secure access via Microsoft private network
- Latency - DC location
- Privacy/compliance – opt out from data being stored
- Privacy/compliance – Azure compliance certifications
- Monitoring – integrates with Azure Monitor
- Harmful content filtering

# Why Azure OpenAI

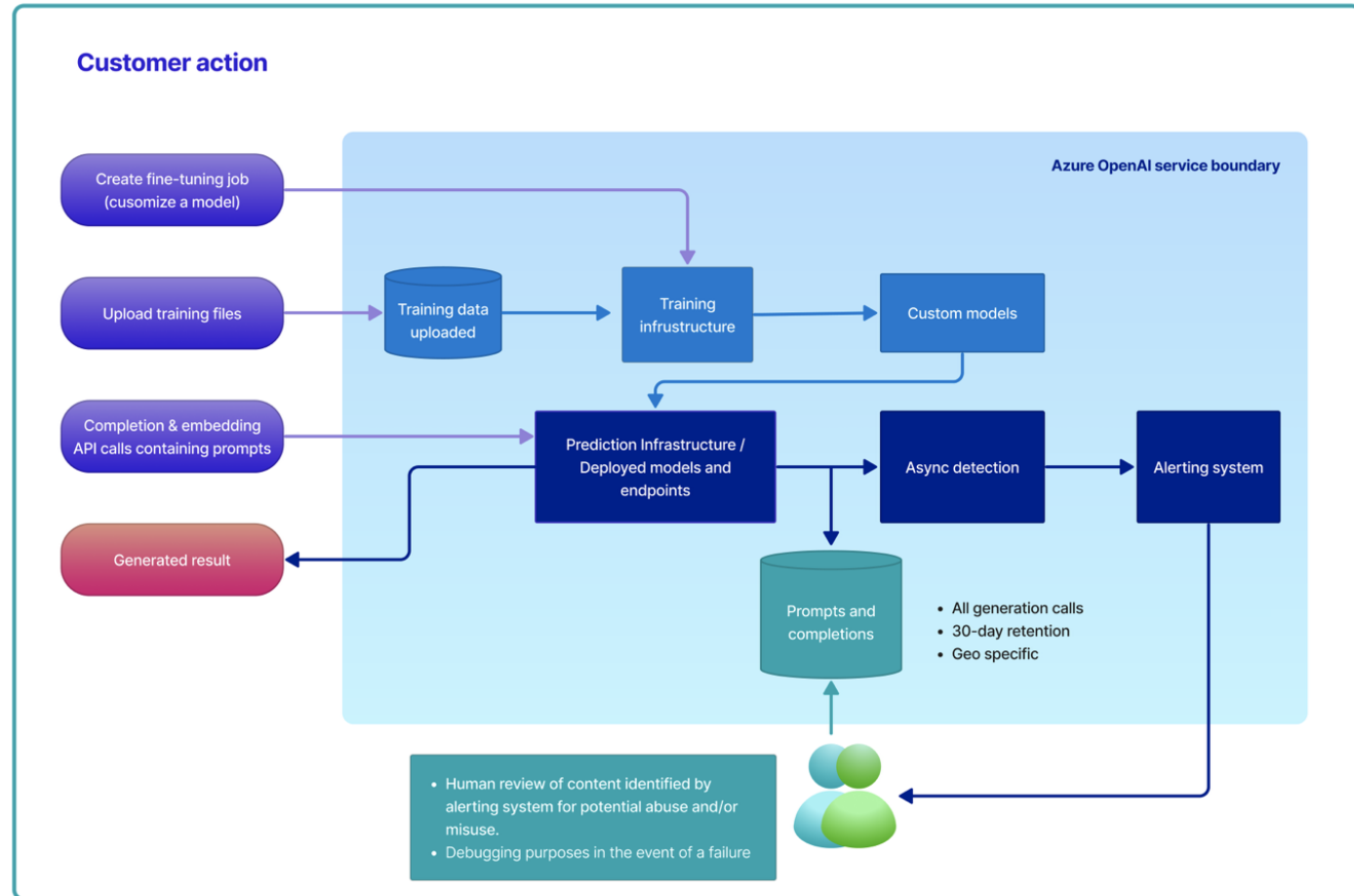


# Data privacy

---



# How Azure OpenAI processes data



# Azure OpenAI Data Processing

## CAN A CUSTOMER OPT OUT OF HAVING PROMPTS-AND-COMPLETIONS LOGGED FOR CONTENT FILTERING & ABUSE MONITORING?

- Customers can apply to have abuse monitoring and/or content filtering configured off by completing this [Limited Access form](#).
- Once approved, a customer can determine if they want Microsoft to configure off content filtering, abuse monitoring, or both.
  - If abuse monitoring is configured off, no prompt and completion data is stored, and no human review is possible. This configuration happens at the subscription level, and the [product documentation](#) explains how a customer can [verify that logging is off](#).
- Because the [content filtering](#) system does not store prompt and completion data, content filtering can remain active and no prompt and completion data will be stored.

# Uses and Capabilities

---

# Unlock new use cases



## Content generation

Call Center Analytics: Automatically generate responses to customer inquiries



## Code generation

Aircraft company using to convert natural language to SQL for aircraft telemetry data

Consulting service using Azure OpenAI Service to convert natural language to query propriety data models



## Semantic search

Financial services firm using Azure OpenAI Service to improve search capabilities and the conversational quality of a customer's Bot experience.

Insurance companies extract information from volumes of unstructured data to automate claim handling processes



## Summarization

International insurance company using Azure OpenAI Service to provide summaries of call center customer support conversation logs

Global bank using Azure OpenAI Service to summarize financial reporting and analyst articles

Government agency using Azure OpenAI Service to extract and summarize key information from their extensive library of rural development reports

Financial services using Azure OpenAI Service to summarize financial reporting for peer risk analysis and customer conversation summarization

# Examples of multiple model use cases

## End-to-end call center analytics

- Classification
- Sentiment
- Entity extraction summarization
- Email generation

## Customer 360

- Hyper-personalisation using timely Summarization of customer queries & trends
- Search
- Content generation

## Business process automation

- Search through structured & unstructured documentation
- Generate Code to query data models
- Content Generation



# Azure OpenAI Service

## Use cases

### Codex

- Natural Language to Code
- Natural Language to SQL
- Code to Natural Language
- Code documentation
- Refactoring

### DALL·E 2

- Creative ideation
- Podcast and music playlist images
- Content syndication
- Marketing campaign personalization
- Hyper-personalization



# Azure OpenAI Service

## Use cases

### Language

**Reason over structured and unstructured data:  
Classification, Sentiment, Entity Extraction, Search**

- Product feedback sentiment
- Customer and employee feedback classification
- Claims and risk analyses
- Support emails and call transcripts
- Social media trends

### Writing assistance

- Marketing copy/email taglines
- Long format text
- Paragraphs from bullets



# Azure OpenAI Service

## Use cases

### Language

#### Summarization

- Call center call transcripts
- Subject Matter Expert Documents
  - Competitive analysis
  - Peer Analysis
  - Technical reports
- Product and service feedback
- Social media trends

#### Conversational AI

- Smart assists for call centers
- Tech support chat bots
- Virtual assistants






# Azure OpenAI Service

## Use cases

### Language

#### Use Cases that use multiple model capabilities

- Contact Centers
  - Classification—route mails to appropriate team
  - Sentiment—prioritize angry customers
  - Entity extraction and search—analyze liability and risk
  - Mail and call transcript summarization
  - Customer response email generation
- Rapid response marketing campaigns: classification, sentiment, summarization, content generation, image generation



# Additional considerations

This is only a REST API. It needs to be connected to applications and/or business processes to be useful.

These models use input from instructions and examples in the prompt to identify the task.

The model completes the task by predicting the most probable next text.

In the next section(s), we will cover three main approaches for in-context learning.

There are some additional considerations when choosing a use case. [Learn more](#)



# Azure OpenAI – based Architectures Deep Dive

Andrey Vykhodtsev  
Sr. AI Architect – CEMA SE



# Agenda

- Azure OpenAI use cases overview
- “Retrieval-Augmented Generation” conceptual overview
- Key components to build an OpenAI solution
- Scenario deep dive: AI Powered Enterprise Q&A
- Scenario deep dive: Contact center analytics
- Scenario deep dive: Document Process Automation
- Scenario deep dive: Application CoPilot

# Azure OpenAI

## Top Capabilities and Use Cases



### Content generation

**Call center analytics:** automatically generate responses to customer inquiries

Generate personalized UI for your website



### Summarization

**Call center analytics:** summary of customer support conversation logs

**Subject matter expert document:** summarization (e.g. Financial reporting, analyst articles)

Social media trends summarization



### Code generation

Convert natural language to SQL (or vice versa) for telemetry data

Convert natural language to query proprietary data models

Code documentation



### Semantic search

Search reviews for a specific product/service

Information discovery and knowledge mining

## Examples of multiple model use cases

**End to end call center analytics:** classification, sentiment, entity extraction, summarization and email generation

**Customer 360:** hyper-personalisation using timely summarization of customer queries & trends, search, and content generation

**Business process automation:** search through structured & unstructured documentation, generate code to query data models, content generation

# Azure OpenAI | When to use it



I need a general-purpose model that can handle multiple tasks.  
e.g., translation+entity recognition+sentiment analysis



I need to generate high-quality content, while preserving data privacy and security  
e.g., abstractive summarization, content writing, paraphrasing, code



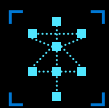
I need rapid prototyping and quick time to market for many use cases



I could use a model with little or no additional training



I want to explore solutions / use cases that have been described previously



Vision



Speech

Azure OpenAI Service



Language



Decision

Azure Cognitive Services



**Patient Support**  
 Enable presenting a summary of customer conversations, removing the need to scroll through the long conversation history to understand customer needs.

**Pattern Recognition**  
 Topic Analysis will help us classify customer conversations to identify emerging patterns in customer complaints and asks.

**Hyper-personalisation**  
 Topic Analysis will help us classify customer conversations to suggest alternatives to customers.

**News and Trends**  
 Topic Analysis will enable mining the web content for latest news and trends to know what are the latest innovation topics in the industry news or to understand what is leading to a drug recall, etc.

**Sentiment Analysis**  
 Advanced Clinical Language Understanding can help predict and even diagnose diseases by analyzing clinical notes and patient conversations and summarizing them to understand if the patient's tone is indicative of a deeper illness or depression / anxiety?

**Claims Processing**  
 Automate the claims processing process by answering questions and providing information to patients and providers about coverage, deductibles, and reimbursement through Semantic Search and ChatGPT.

**Cost Estimation**  
 Search through cost models and alternative treatments to provide patients with an estimation of cost for healthcare services.

**Ambient clinical intelligence**  
 Automatically generate responses to customers to field basic questions based on knowledge base, so as to prioritise more complex care.

**Clinical documentation**  
 Reduce manual work in going through extensive documentation to derive an answer for a patient or to identify key patterns that can be done through summarization.

**Alternative Treatments**  
 Advanced language understanding that can help find connections between different treatments, understanding available alternatives and their side effects through summarizing internal and external data.

**HEALTHCARE INDUSTRY BENEFITS**

Personalised Customer Care

— Cost Savings through Optimised Engagement

— Innovation w/ Better / Alternative Treatment

### Contact Center Analytics

Call summarization to extract key information from call logs in order to identify trends in customer complaints and involve human agents who have succinct and relevant information to make faster decisions / responses. Use Q&A features to automatically handle calls through bot agents.

### Customer Q&A

Generate human-like responses for live chat/voice bot applications.

### Virtual health

Automatically generate personalized correspondence to land the best messaging based on the customer needs and patterns

### Regulatory Compliance

Help agents comply to regulatory compliance in internal research and due diligence through assisted Q&A policy and underwriting rules.

### Sentiment Analysis

Agent training mechanism to simulate customer interaction and evaluating the end 'customer' satisfaction via sentiment analysis using GPT models.

### Contact Center Analytics

Call summarization to extract key information from call logs in order to identify trends in customer complaints and involve human agents who have succinct and relevant information to make faster decisions / responses. Use Q&A features to automatically handle calls through bot agents.

### Document Process Automation

Digitise documents, extract key information through Semantic or Cognitive Search and summarise them (e.g. legal document summarization through internal and external information).

### Risk Management

Recognise current patterns and trends to identify risks through content (including social media and blogs) search and summarization.

### Cross-Business Unit Information Gathering

NER, Sentiment, Classification and Summarization of news sources across multiple business units

### Contact Center Analytics

Call summarization to extract key information from call logs in order to identify trends in customer complaints and involve human agents who have succinct and relevant information to make faster decisions / responses. Use Q&A features to automatically handle calls through bot agents.

### Market Analysis

Automatically generate top-of-mind / editorial opinions based on a day's market dynamics, produce a summary view and a few key-points analysis



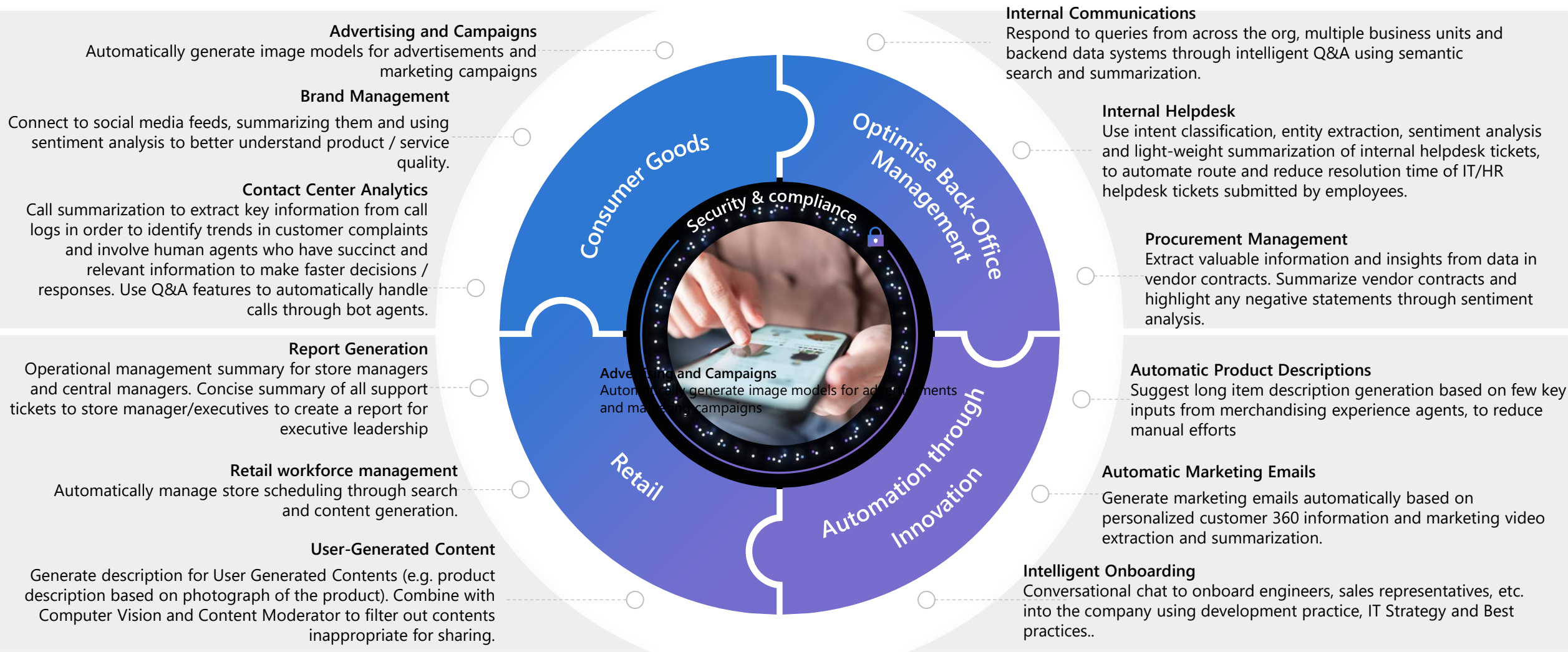
## FSI INDUSTRY BENEFITS

Personalised Customer Care

— Cost Savings through Optimised Engagement

— Intelligent Compliance & Risk Management





**RETAIL & CPG INDUSTRY BENEFITS**



### Contact Center Analytics

Call summarization to extract key information from call logs in order to identify trends in customer complaints and involve human agents who have succinct and relevant information to make faster decisions / responses. Use Q&A features to automatically handle calls through bot agents (e.g. outage information, etc.)

### Pattern Recognition

Topic Analysis will help us classify customer conversations to identify emerging patterns in customer complaints and asks.

### Hyper-personalisation

Topic Analysis will help us classify customer conversations to suggest alternatives to customers.

### Accelerating pre-publication work

Accelerate and analyze their work before publication, including suggesting content edits to comply with editorial style guide, generating compelling informative summaries, generating SEO optimized headlines for articles, and assisting in article composition from wires or from fact lists, etc.

### Summarise audio transcripts

Perform entity extraction and generate summaries of audio transcripts extracted from their video content.

### Summarise and automatically generate new content to fans

Summarise and innovative the content provided to fans through company's website. Even generate new UI for website.

### Trends Analysis

Use Social Media trends and data to understand customer sentiments on products and services, in order to better service to the customers and personalize content based on new trends and patterns.

### Internal Communications

Respond to queries from across the org, multiple business units and backend data systems through intelligent Q&A using semantic search and summarization.

### Internal Helpdesk

Use intent classification, entity extraction, sentiment analysis and light-weight summarization of internal helpdesk tickets, to automate route and reduce resolution time of IT/HR helpdesk tickets submitted by employees.

### Co-Pilot Approach

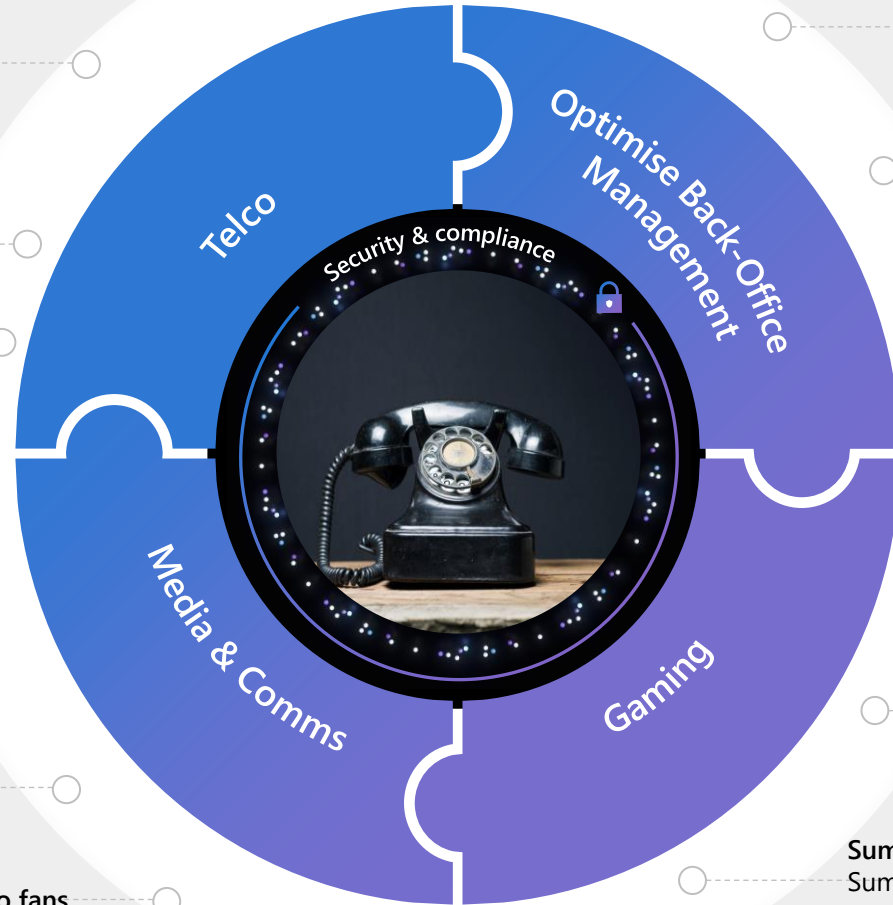
Embedding Codex into the game development platform to have a co-pilot approach to developing code on the platform with code completion service within the company's IDE.

### Automatic Marketing Emails

Generate marketing emails automatically based on personalized customer 360 information and marketing video extraction and summarization.

### Summarise and automatically generate new content to gamers

Summarise and innovative the content provided to gamers on the gaming platform.



### CMT INDUSTRY BENEFITS

Hyper-personalisation

Intelligent Contact Center

Innovative Content Creation

### Contact Center Analytics

Call summarization to extract key information from call logs in order to identify trends in customer complaints and involve human agents who have succinct and relevant information to make faster decisions / responses. Use Q&A features to automatically handle calls through bot agents.

### Pattern Recognition

Topic Analysis will help us classify customer conversations to identify emerging patterns in customer complaints and asks, as well as in most effective consumption patterns..

### Hyper-personalisation

Topic Analysis will help us classify customer conversations to suggest alternatives to customers.

### Remote Worker Assistance

Personal assistance/knowledge provider to business users as well as operations personnel at remote locations.

### Automatically generate email responses

Understand customer sentiment and automatically generate email responses to customers.

### Equipment Reporting

Text summarization and pattern recognition to automatically generate reports on equipment failures, etc.

### Trends Analysis

Use Social Media trends and data to understand customer sentiments on products and services, in order to better service to the customers and personalize content based on new trends and patterns.

### Internal Communications

Respond to queries from across the org, multiple business units and backend data systems through intelligent Q&A using semantic search and summarization (e.g. information on prevention guidelines, etc.)

### Internal Helpdesk

Use intent classification, entity extraction, sentiment analysis and light-weight summarization of internal helpdesk tickets, to automate route and reduce resolution time of IT/HR helpdesk tickets submitted by employees.

### Staff Onboarding

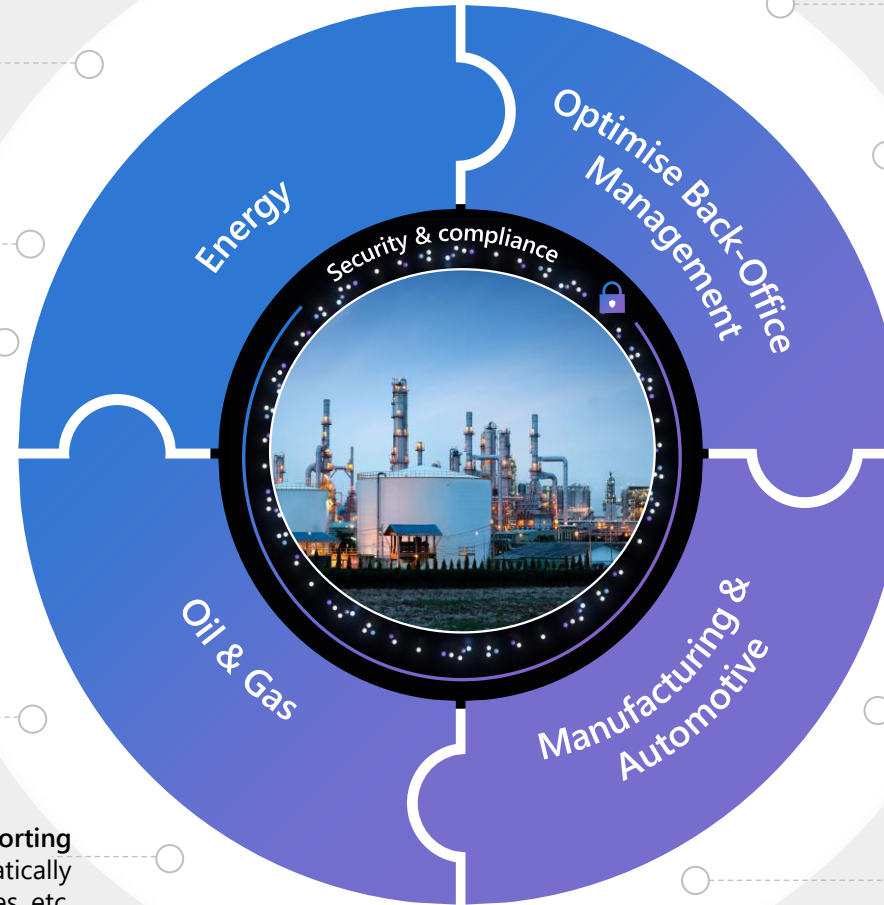
Use Generative Models capabilities to build training material for new personnel (e.g. customer service agents, factory workers, etc.) onboarding, based on recorded conversation history.

### Automatic Marketing Emails

Generate marketing emails automatically based on personalized customer 360 information and marketing video extraction and summarization.

### Summarise and automatically generate new content to customers

Automatically summarise and generate new and relevant content to customers to reduce the amount of manual effort required in responding to customers.



## MANUFACTURING / ENERGY / OIL&GAS INDUSTRY BENEFITS

Worker Assistance and Onboarding

Intelligent Contact Center

Innovative Content Creation

# Aviation Example

An example Passenger Name Record (PNR) data below:

```
###  
MU99XZ  
1EZRRK1/22FEB2E2/SGN  
DL/KE/AF/SQ  
LAXLHR/AF/LHRJNB/DL/JNB JFK/SQ  
22FEB22/28FEB22  
1/2  
SMITH/JOHNM  
987654321  
DL 23 Y 22FEB 2 LAXLHR HK1 1025A 545P 1A/E  
AF 168 Y 22FEB 2 LHRJNB HK1 930P 955A 1A/E  
DL 54 Y 25FEB 5 JNB JFK HK1 1240A 610A 1A/E  
SQ 238 Y 26FEB 6 JFKSIN HK1 1050A 525P 1A/E  
SQ 186 Y 26FEB 6 SGNBKK HK1 820P 1000P 1A/E  
KE 652 Y 28FEB 1 BKKICN HK1 1215A 720P 1A/E  
KE 601 Y 28FEB 1 ICNLAX HK1 1000P 540P 1A/E  
E-TKT 0123456789ABCDEF  
###
```

Summarize the above in an easy to understand language in one paragraph, in layman words and no airport acronyms. Describe the city or origin and date of departure of every leg.



John M. Smith is going on a journey starting on February 22nd, 2022. He'll be traveling from Los Angeles to London, then from London to Johannesburg on the same day. On February 25th, he'll fly from Johannesburg to New York, and on February 26th, he'll travel from New York to Singapore and then from Singapore to Bangkok. Lastly, on February 28th, he will visit Seoul before returning back to Los Angeles. He has an electronic ticket with a unique number to help him throughout his travels.

# Finance Example

An example SWIFT message from Banking Industry as below:

```
###
{1:F01KNABNL2HAXXX0000000000}
{2:1940KNABNL2HXXXN3020}
{4: :20:B4G30MS9D00A003D :25:123456789 :28C:999/1 :60F:C140729EUR3058,
98 :61:1407290729D2760,NTRFNONREF//B4G29PGCVO7Q06US /OCMT/EUR10,
00/ :86:FACTUUR 201403110, 201403113 REK: NL65INGB0123456789/NAAM:
PICQER :61:1407290729C500NTRF29-07-2014 10:05//B4G29PGDCK1QFV3E
:86:12160475 0050001631430920 ORDERID: 264267 MEDIA MARKT ONLINE NE REK:
NL84INGB0234561789/NAAM: MMS ONLINE NEDERLAND B.V. :62F:C140730EUR798,98
-}
###
```

Tasks:

1. Summarize the above SWIFT message.
2. Determine the risk level in [low, medium, high] in terms of Fraud Detection or Money Laundering, and briefly give reasons.

Answer above questions one at a time, in an easy to understand language in one paragraph, in layman words and no airport acronyms.



1. This is a message about a bank account's transactions. There is a payment of 2,760 Euros going out for an invoice with some numbers and the name "PICQER". There is also a payment of 500 Euros coming in with a date and time, and it is related to an order from a company called "MEDIA MARKT ONLINE" with the name "MMS ONLINE NEDERLAND B.V.". After these transactions, the account will have 798.98 Euros left.
2. The risk of fraud or money laundering in this message seems low. The payment amounts are not very large, and the descriptions have clear details about what they are for. However, to be sure, more information about the people or companies involved and their transaction history would be needed.

# Biomedical Example

"The ERX-41 compound did not kill healthy cells, but it wiped out tumor cells regardless of whether the cancer cells had estrogen receptors," Ahn said. "In fact, it killed the triple-negative breast cancer cells better than it killed the ER-positive cells. "This was puzzling to us at the time. We knew it must be targeting something other than estrogen receptors in the TNBC cells, but we didn't know what that was." To investigate the ERX-41 molecule, Ahn worked with collaborators, including co-corresponding authors Dr. Ganesh Raj, professor of urology and pharmacology at the Harold C. Simmons Comprehensive Cancer Center at UT Southwestern Medical Center, as well as Dr. Ratna Vadlamudi, professor of obstetrics and gynecology at UT Health San Antonio. Dr. Tae-Kyung Lee, a former UTD research scientist in Ahn's Bio-Organic/Medicinal Chemistry Lab, was involved in synthesizing the compound. The researchers discovered that ERX-41 binds to a cellular protein called lysosomal acid lipase A (LIPA). LIPA is found in a cell structure called the endoplasmic reticulum, an organelle that processes and folds proteins. "For a tumor cell to grow quickly, it has to produce a lot of proteins, and this creates stress on the endoplasmic reticulum," Ahn said. "Cancer cells significantly overproduce LIPA, much more so than healthy cells. By binding to LIPA, ERX-41 jams the protein processing in the endoplasmic reticulum, which becomes bloated, leading to cell death." "Triple-negative breast cancer is particularly insidious — it targets women at younger ages; it's aggressive; and it's treatment resistant. I'm really glad we've discovered something that has the potential to make a significant difference for these patients." Dr. Jung-Mo Ahn, associate professor of chemistry and biochemistry in the School of Natural Sciences and Mathematics The research team also tested the compound in healthy mice and observed no adverse effects. "It took us several years to chase down exactly which protein was being affected by ERX-41." Ahn said. "That was the hard part. We chased many dead ends, but we did not give up. "Triple-negative breast cancer is particularly insidious — it targets women at younger ages; it's aggressive; and it's treatment resistant. I'm really glad we've discovered something that has the potential to make a significant difference for these patients." The researchers fed the compound to mice with human forms of cancerous tumors, and the tumors got smaller. The molecule also proved effective at killing cancer cells in human tissue gathered from patients who had their tumors removed. They also found that ERX-41 is effective against other cancer types with elevated endoplasmic reticulum stress, including hard-to-treat pancreatic and ovarian cancers and glioblastoma, the most aggressive and lethal primary brain cancer. "As a chemist, I am somewhat isolated from patients, so this success is an opportunity for me to feel like what I do can be useful to society," Ahn said. Ahn is a joint holder of patents issued and pending on ERX-41 and related compounds, which have been licensed to the Dallas-based startup EtiraRX, a company co-founded in 2018 by Ahn, Raj and Vadlamudi. The company recently announced that it plans to begin clinical trials of ERX-41 as early as the first quarter of 2023. Ahn's research on this project is supported by the National Cancer Institute, part of the National Institutes of Health (1R01CA223828); the Cancer Prevention and Research Institute of Texas; and The Welch Foundation. In addition to researchers from UT Southwestern and UT Health San Antonio, other study authors from Howard Hughes Medical Institute's Janelia Research Campus, Northwest A&F University in China and the Medical College of Xiamen University in China contributed. Media Contact: Amanda Siegfried, UT Dallas, 972-883-4335, amanda.siegfried@utdallas.edu, or the Office of Media Relations, UT Dallas, (972) 883-2155, newscenter@utdallas.edu. Tagged: chemistry Dr. Jung-Mo Ahn New Dimensions: The Campaign for UT Dallas NSM research

&nbsp;

Detect the new drug name above, and list any adverse events related to the use of this drug.



The new drug name mentioned in the text is ERX-41. According to the information provided, the researchers tested the compound in healthy mice and observed no adverse effects.

# The Standard's Goals at a Glance

## Accountability

- A1:** Impact Assessment
- A2:** Oversight of significant adverse impacts
- A3:** Fit for purpose
- A4:** Data governance and management
- A5:** Human oversight and control

## Transparency

- T1:** System intelligibility for decision making
- T2:** Communication to stakeholders
- T3:** Disclosure of AI interaction

## Fairness

- F1:** Quality of service
- F2:** Allocation of resources and opportunities
- F3:** Minimization of stereotyping, demeaning, and erasing outputs

## Reliability & Safety

- RS1:** Reliability and safety guidance
- RS2:** Failures and remediations
- RS3:** Ongoing monitoring, feedback, and evaluation

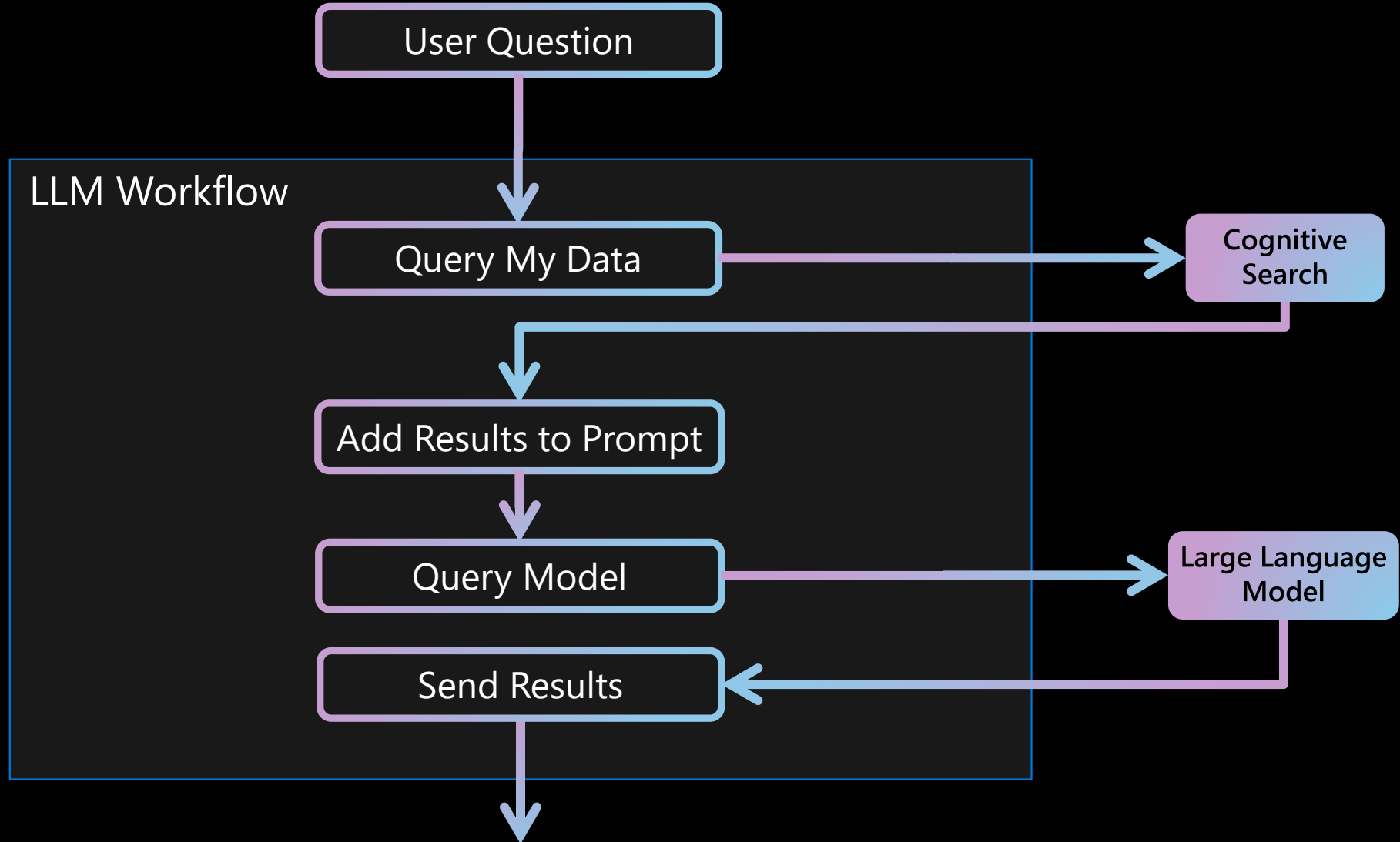
## Privacy & Security

- PS1:** Privacy Standard compliance
- PS2:** Security Policy compliance

## Inclusiveness

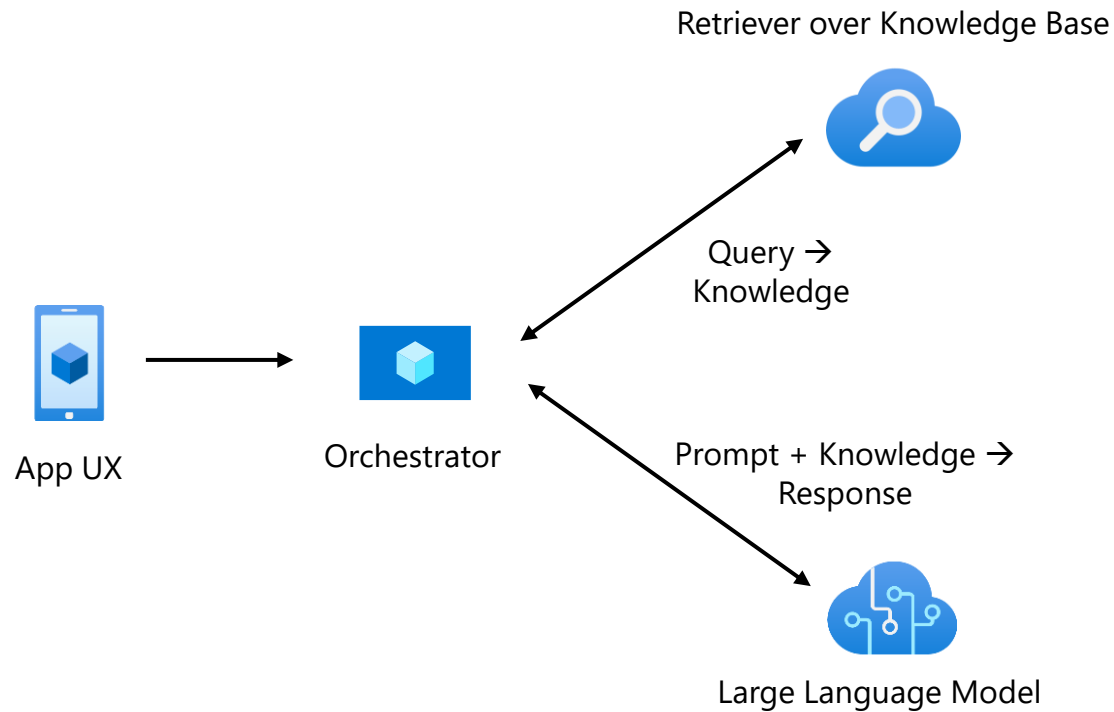
- I1:** Accessibility Standards compliance

# Retrieval Augmented Generation





# 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.

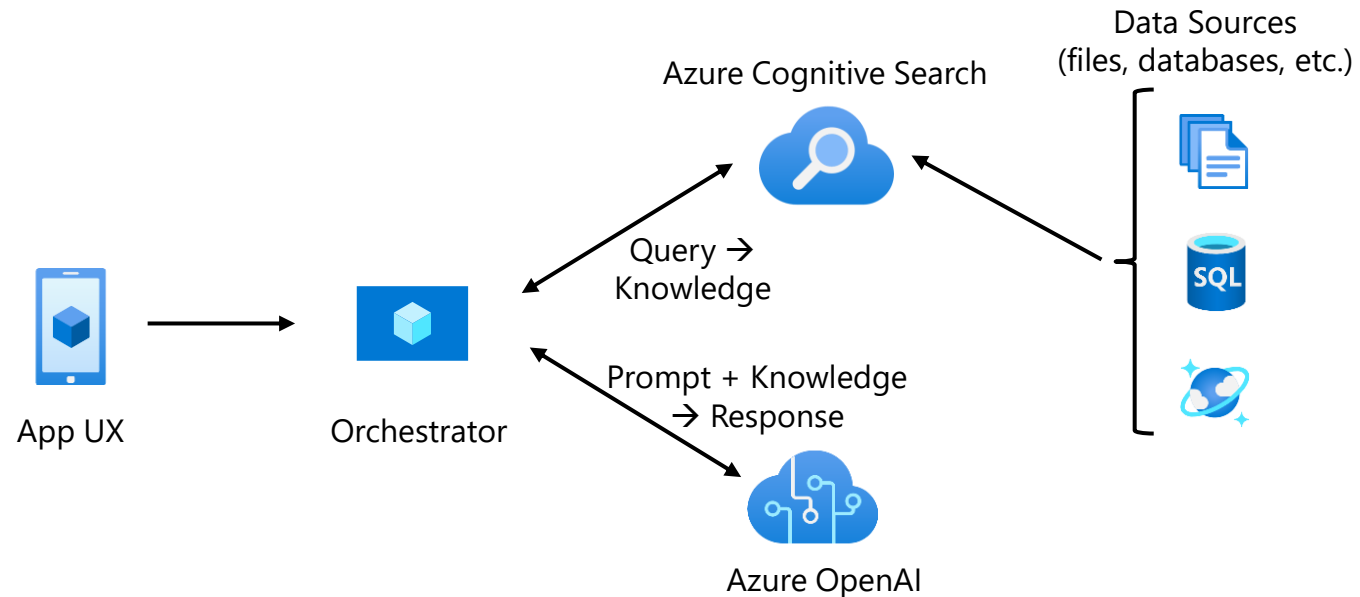
e.g., plugins for OpenAI ChatGPT

# Retrievers: Externalizing Knowledge

“Find the most relevant snippets in a large data collection, using unstructured input as query”  
== search engine

## Azure Cognitive Search

- Azure’s complete retrieval solution
- Data ingestion, enterprise-grade security, partitioning and replication for scaling, support for 50+ written languages, and more



# RAG vs. FineTuning

·Prompt:

·Extract job titles from the following sentences.

·Sentence: John Doe has been working for Microsoft for 20 years as a Linux Engineer.

·Job title: Linux Engineer

·###

·Sentence: John Doe has been working for Microsoft for 20 years and he loved it.

·Job title: none

·###

·Sentence: Marc Simoncini | Director | Meetic

·Job title: Director

·###

·Sentence: Damien is the CTO of Platform.sh, he was previously the CTO of Commerce Guys, a leading ecommerce provider.

·Completion:

·Job title: CTO

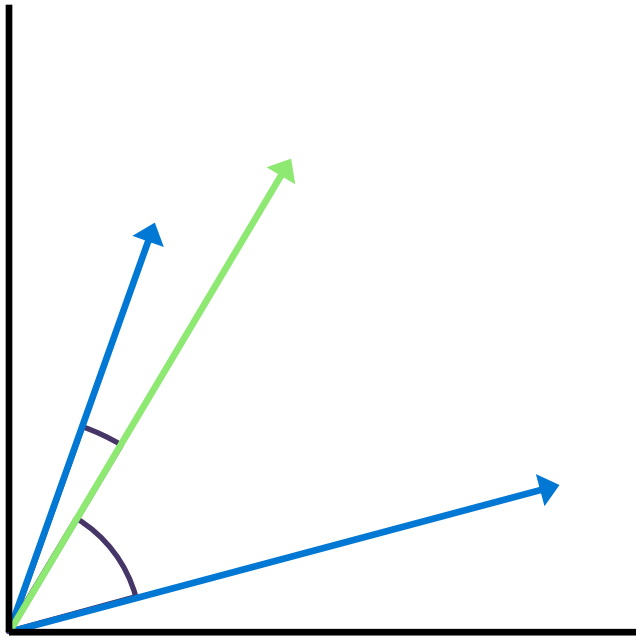
· Finetuning:

- Need to provide labeled data
- Need to deploy finetuned model (\$\$)
- Model "internalizes" knowledge

· RAG

- No need for labelled data
- Externalized knowledge

# Retrieving Using Semantic Similarity



## Vector representations (or embeddings)

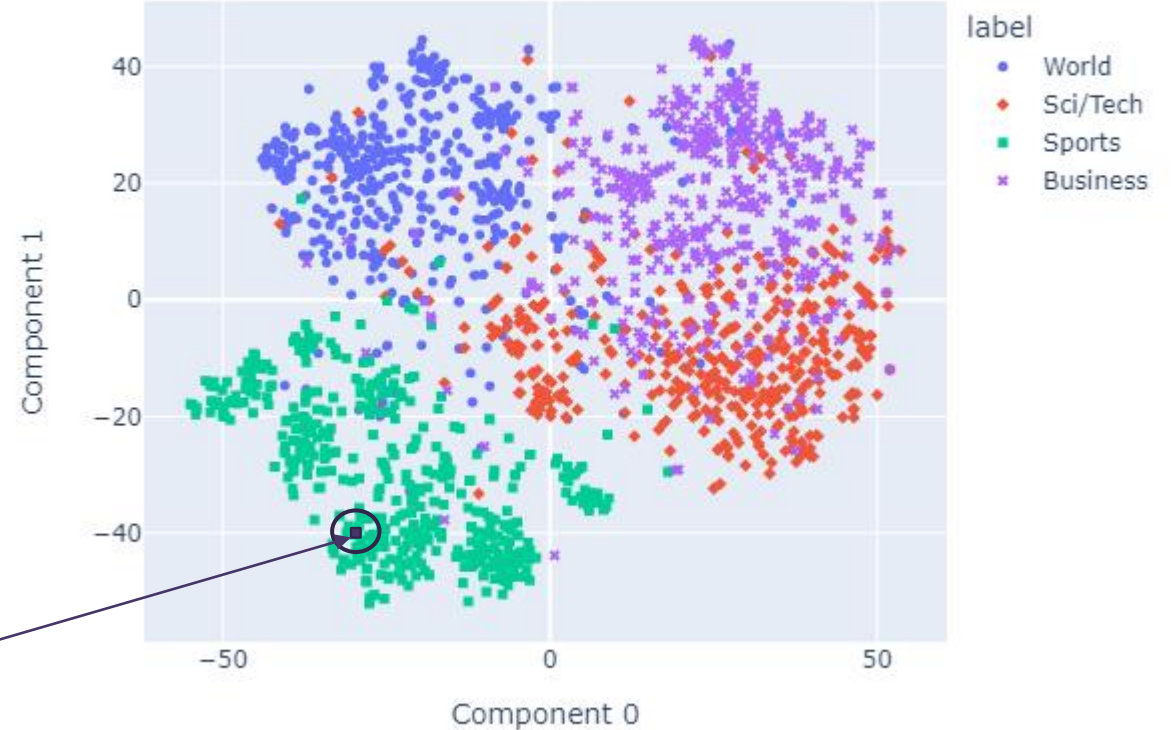
- Learned such that “close” vectors represent items with similar meaning
- May encode words, sentences, images, audio, etc.
  - Some map multiple media types into the same space
- Azure OpenAI embeddings API, OSS embeddings (e.g., SBERT, CLIP)

# GPT: Embeddings - Recommendations

- Recommendations are widespread across the web:
  - 'Bought that item? Try these similar items.'
  - 'Enjoy that book? Try these similar titles.'
  - 'Not the help page you were looking for? Try these similar pages.'
- With embeddings, we can answer the question: given an article, what other articles are most similar to it?

If current Article is the center dot, then recommend all closest articles falling inside the circle, as being "Recommended Readings"

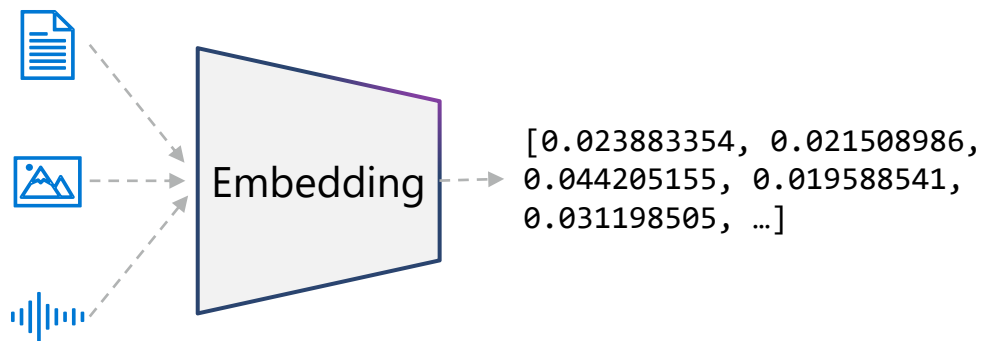
t-SNE components of article descriptions



# Vector-based Retrieval

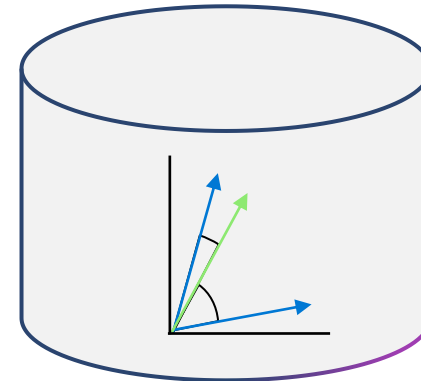
## Encoding (vectorizing)

- 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



# Vector Search in Azure Cognitive Search

## New vector type for index fields

- Users indicate vector size, distance function, algorithm and algo-specific parameters

## Pure Vector Search & Hybrid Search

- Filters, faceting, etc. all works with vectors
- Integrates with existing search indexes
- Existing data ingestion and augmentation machinery entirely applicable

## Combines well with L2 re-ranker powered by Bing's models

- Enables improved ranking for hybrid search scenarios
- L1: keywords + vector retrieval
- L2: Bing's ranker refreshed with GPT-enhanced work

## Enterprise-grade

- Scalability (partitioning, replication)
- Security: network isolation, managed identities, RBAC, etc.

Private preview

# Azure Cognitive Search – Vector search

## Revolutionizing Indexing and Retrieval for LLM-powered Apps

Power your retrieval-augmented generation applications



Images



Audio



Video



Graphs



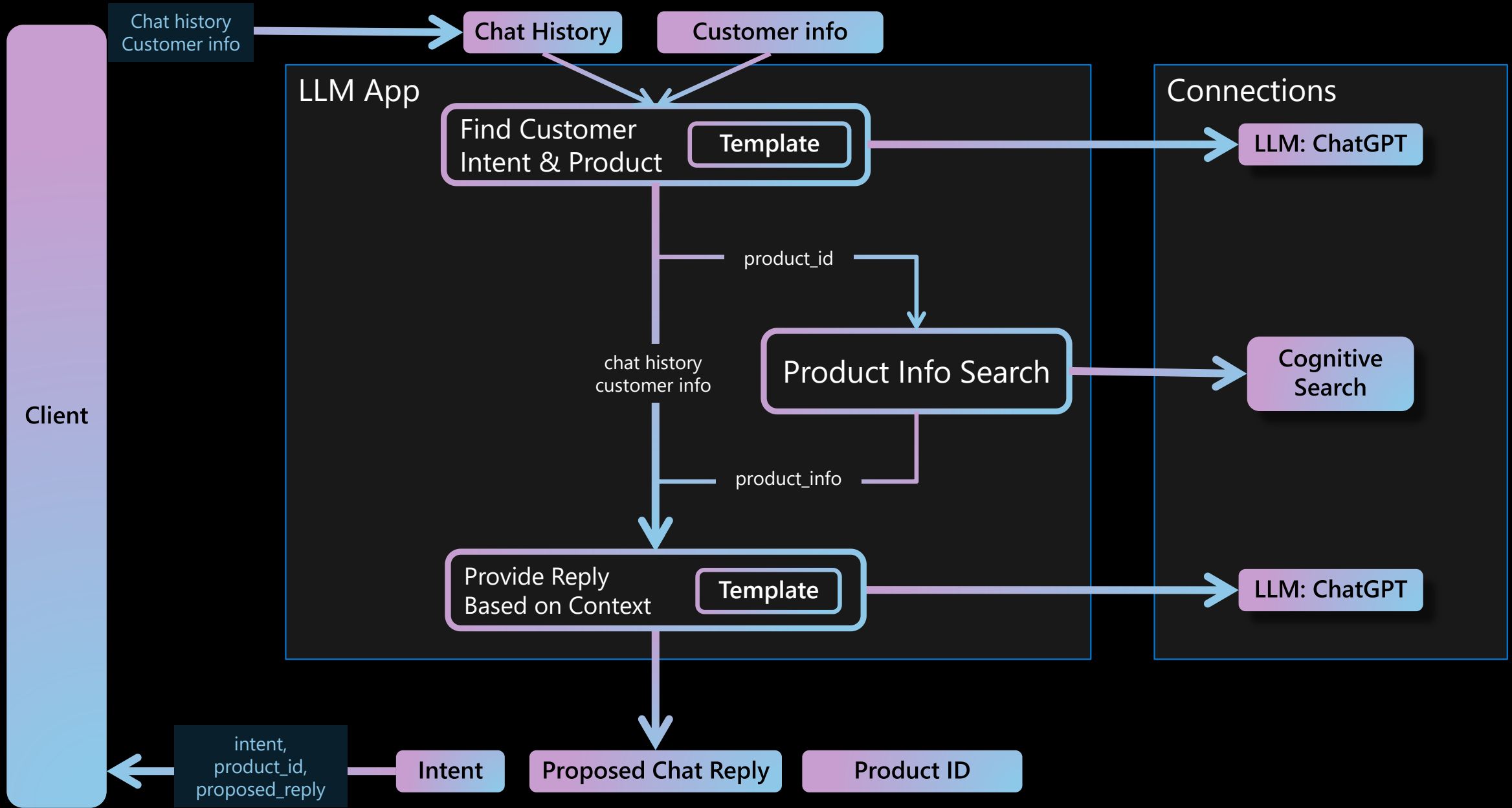
Documents

- Use vector or hybrid search
- Use Azure OpenAI embeddings or bring your own
- Deeply integrate with Azure
- Scale with replication and partitioning
- Build generative AI apps and retrieval plugins

Sign up today

<https://aka.ms/VectorSearchSignUp>





# Azure AI

## Applications



Partner Solutions

## Application Platform

AI Builder



Power BI



Power Apps



Power Automate



Power Virtual Agents

## 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

Azure OpenAI Service

## ML Platform



Azure Machine Learning

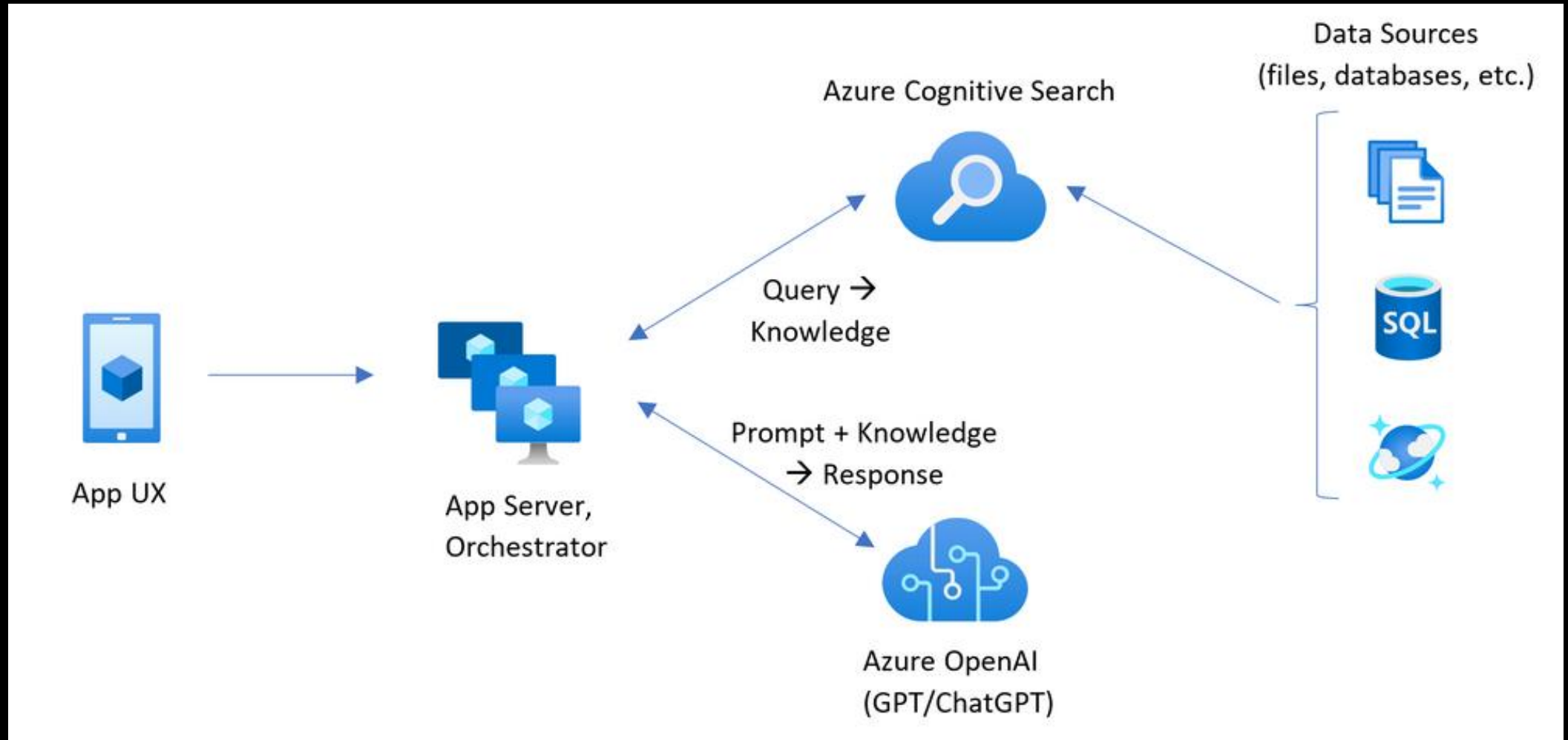


Business Users



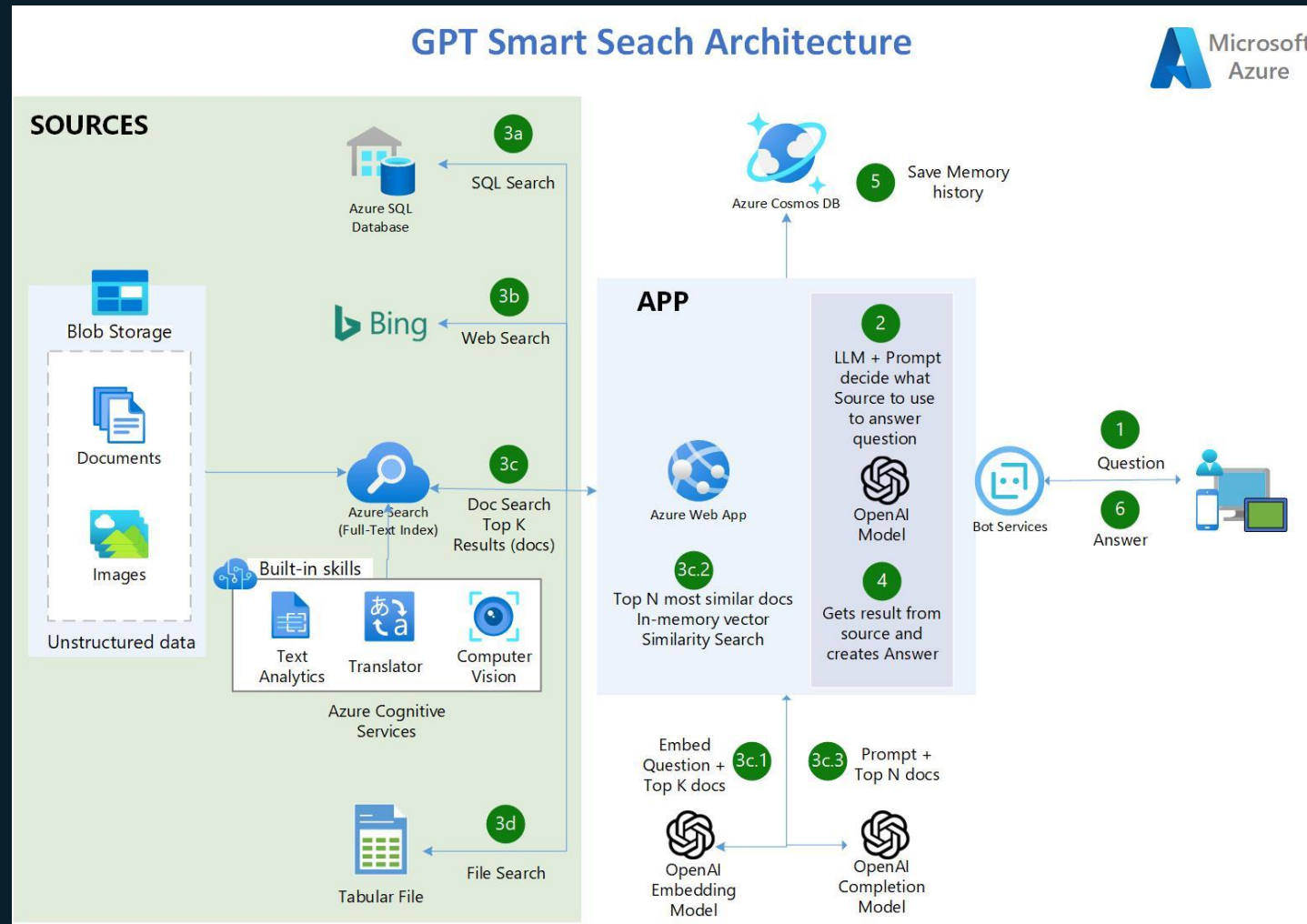
Developers & Data Scientists

# Q&A Search App



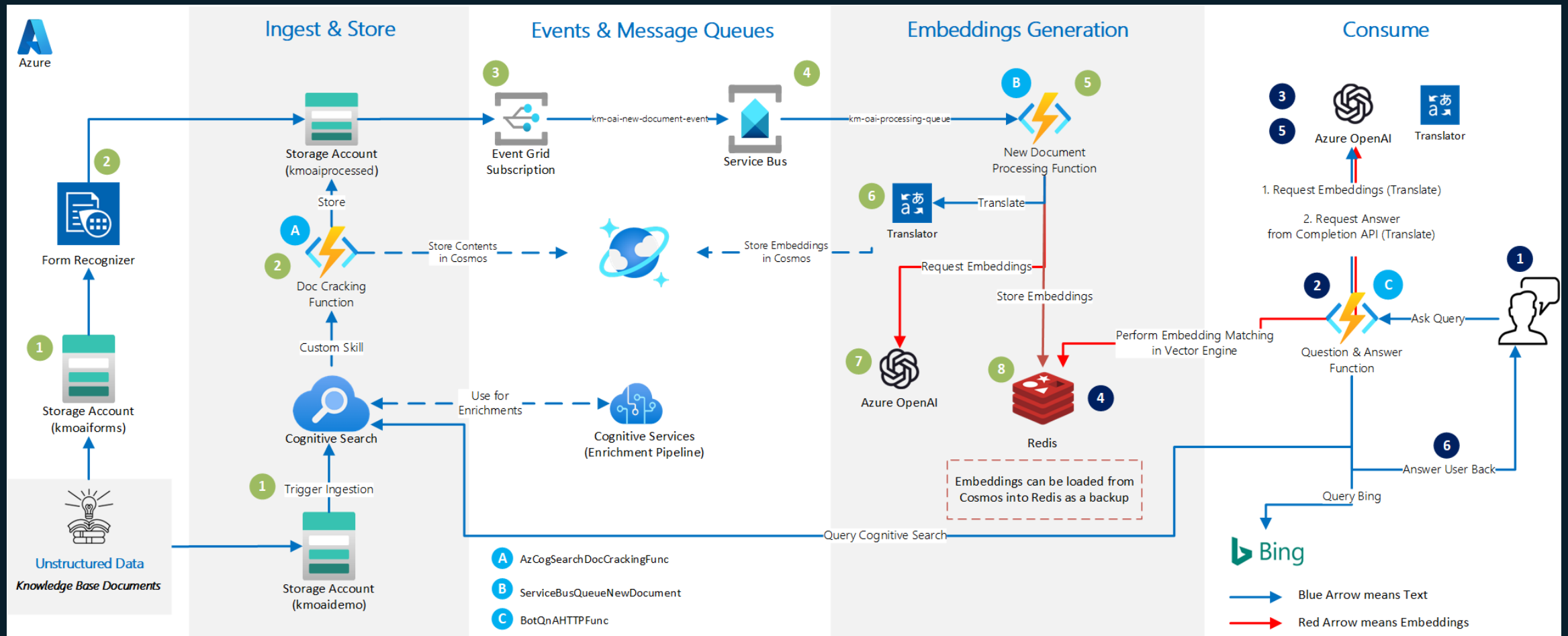
<https://github.com/Azure-Samples/azure-search-openai-demo>

# Q&A Chatbot architecture



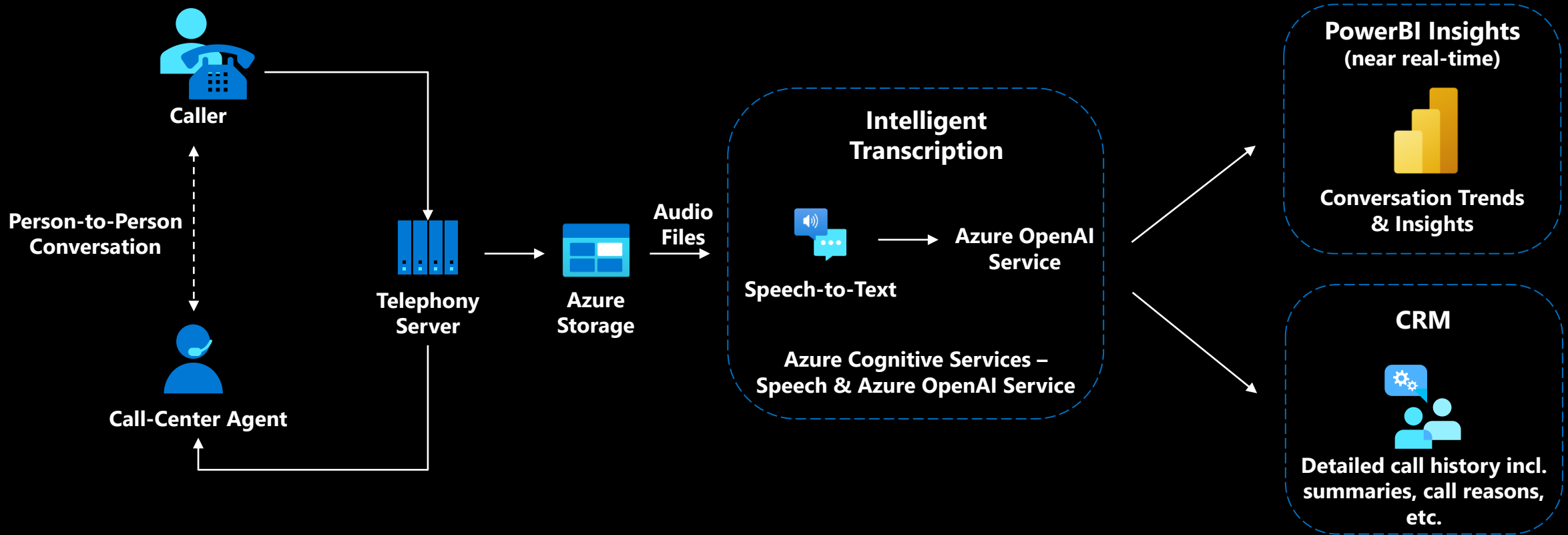
<https://github.com/pablomarin/GPT-Azure-Search-Engine>

# Q&A PoC architecture



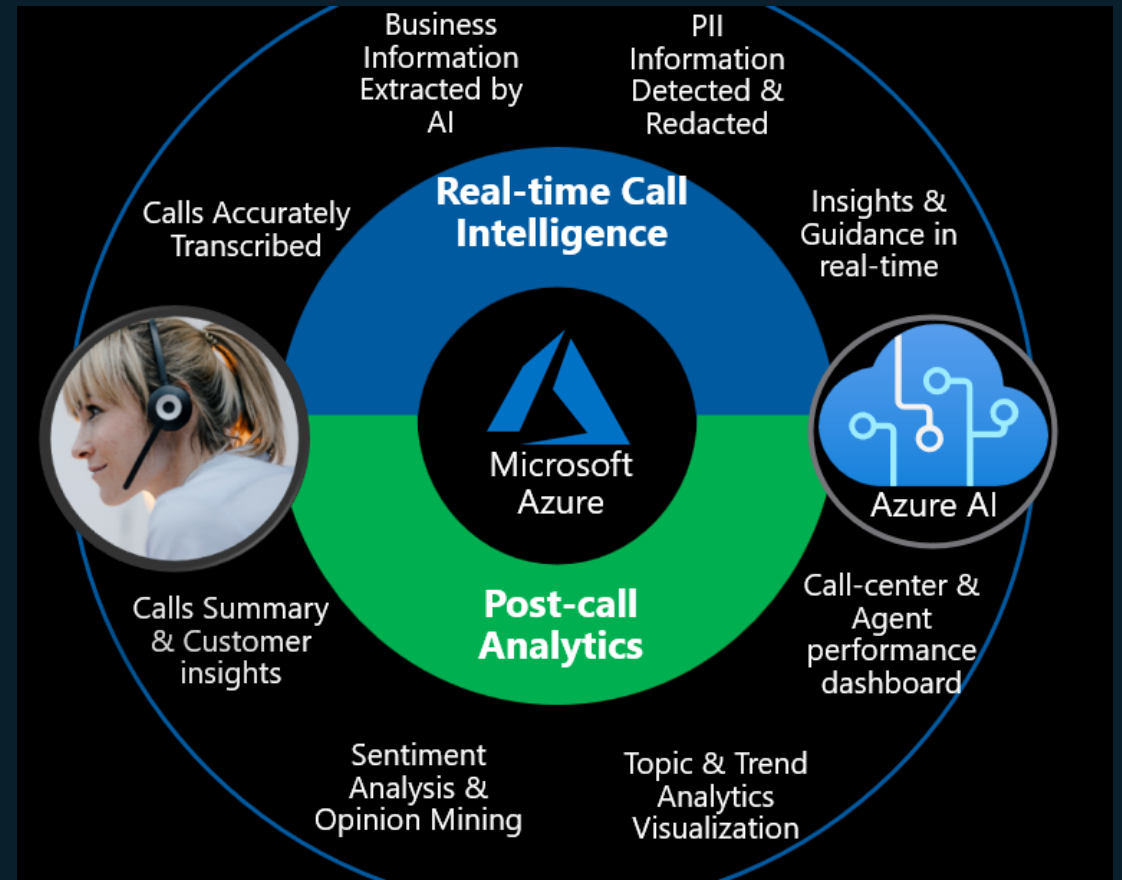
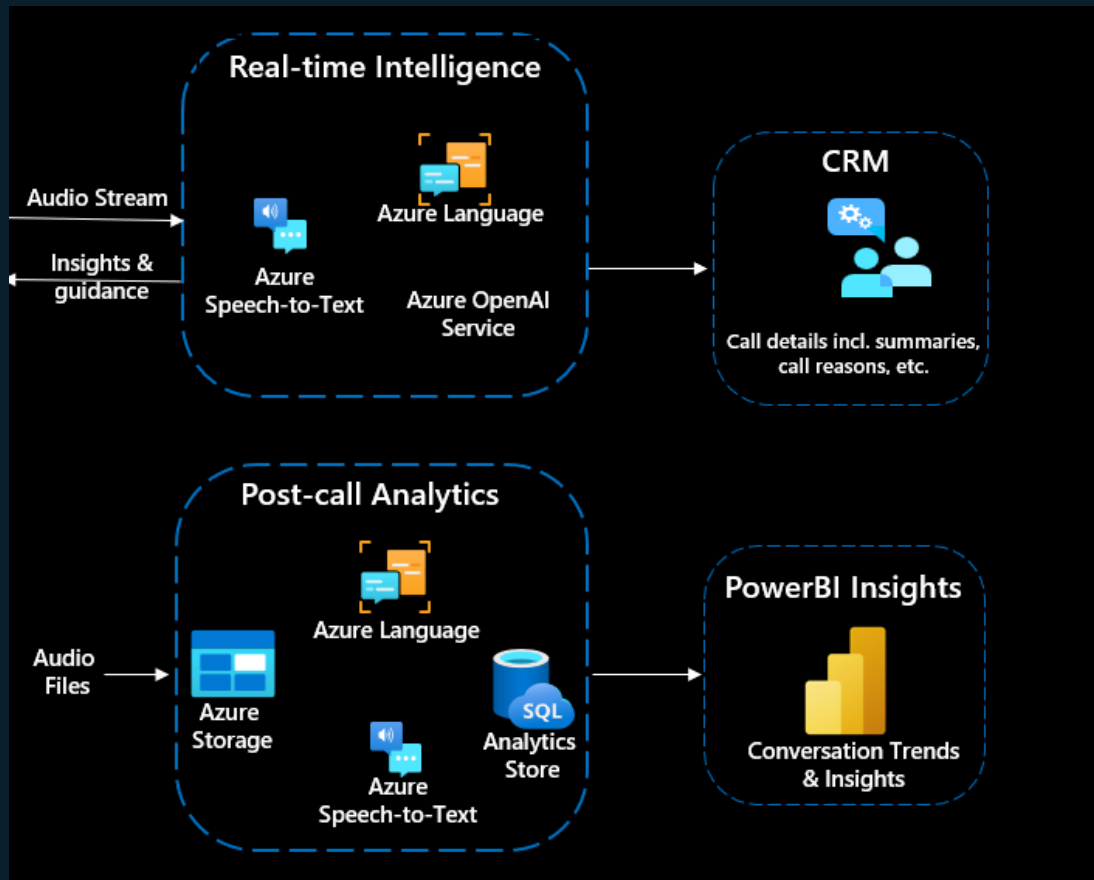
# Contact Center Analytics using Speech API & Azure OpenAI Service

Extract rich insights from call transcripts



# Call Center intelligence

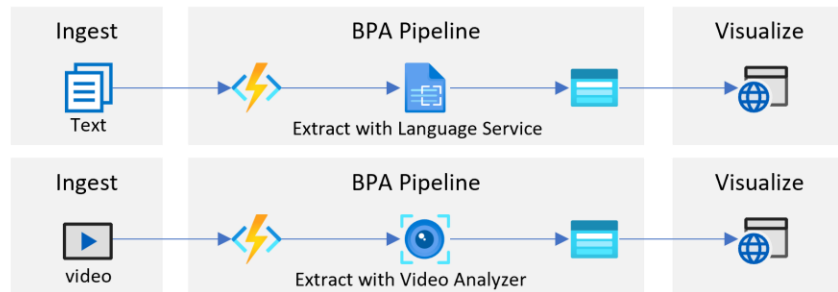
<https://github.com/amulchapla/AI-Powered-Call-Center-Intelligence>



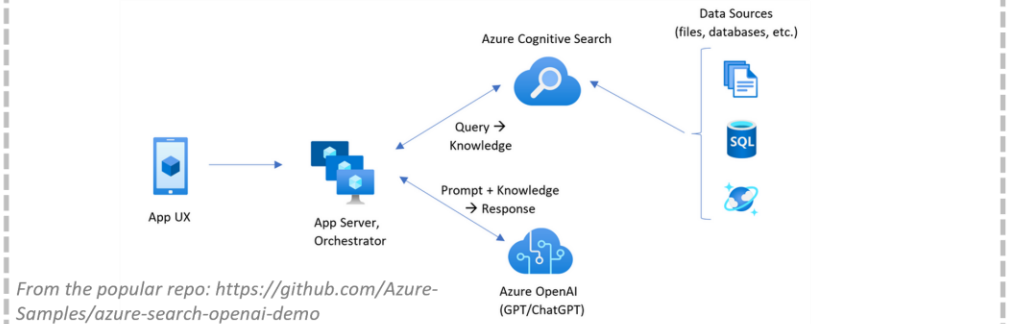
# Business Process Automation accelerator

## Sample Pipelines

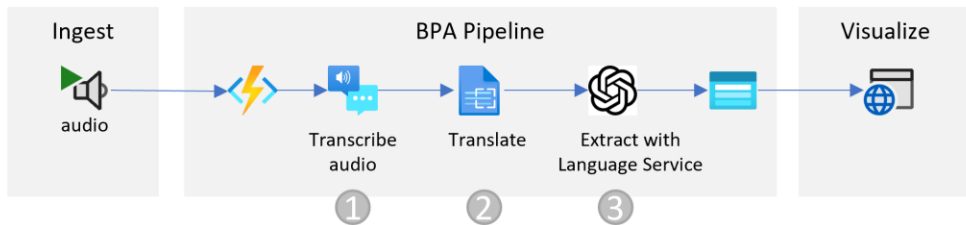
### Single Service Pipeline Examples



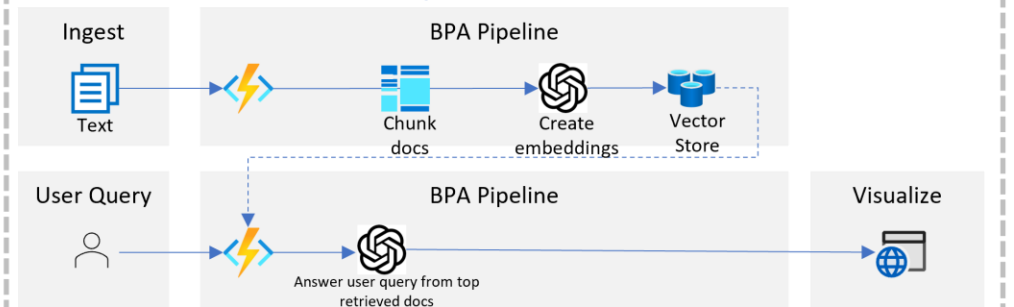
### ChatGPT + Enterprise data with Azure OpenAI & Cognitive Search



### Multiple Services with Multiple Domains (e.g. speech + text)

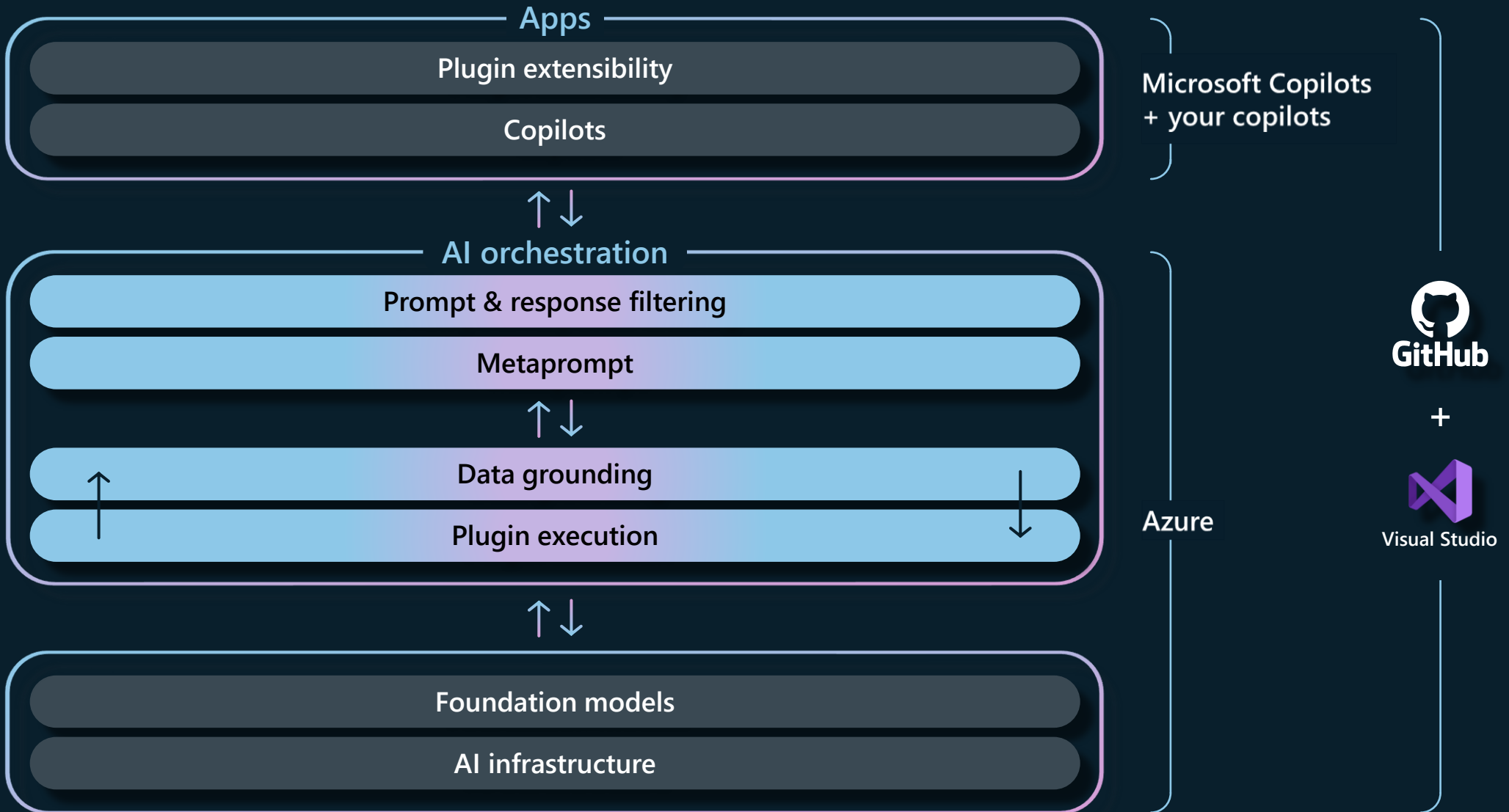


### Vector Store Pipeline (Conversational UI)





# Copilot stack



# Generative AI Applications

# Azure Machine Learning

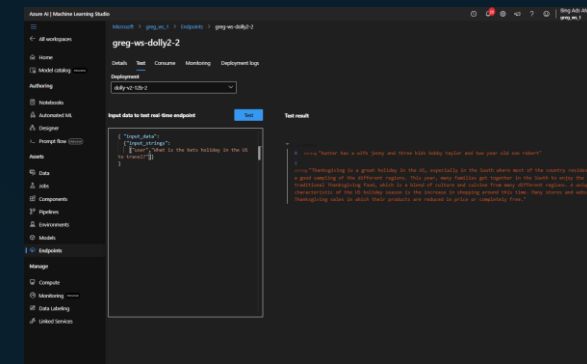
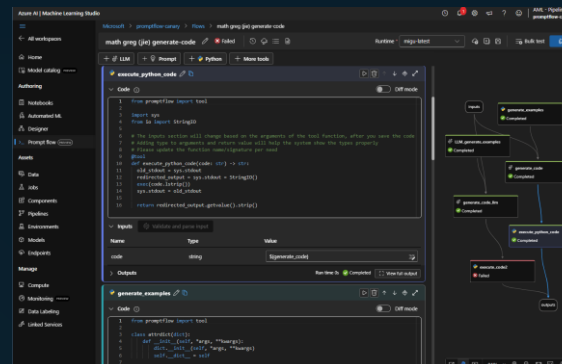
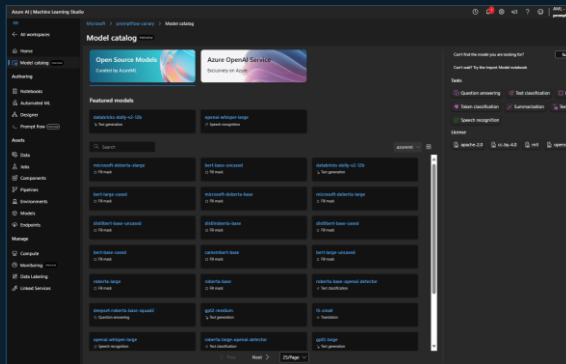
Native OSS  
Model Catalog

Prompt Eng/Eval  
Prompt flow

Responsible AI  
Azure Content Safety

High Scale Gen AI  
App Deployment

Gen AI Model  
Monitoring



# Announcing

# Azure Machine Learning Prompt flow

## Customer Benefits

- Create AI workflows that consume various language models and data sources using the frameworks and APIs of your choice
- One platform to quickly iterate through build, tune, & evaluate for your GenAI workflow
- Evaluate the quality of AI workflows with pre-built and custom metrics
- Easy historical tracking and team collaboration
- Easy deployment and monitoring

The screenshot displays the Microsoft Azure Machine Learning Studio interface for a Prompt Flow named 'Q&A with GPT3.5'. The flow is currently in a 'Completed' state. The main workspace shows the configuration for the 'extract\_query\_from\_question' step, which uses the 'azure\_open\_ai\_connection' and 'Api completion' model. The prompt is configured with a 'question' input and a system message: 'I want you to act as a web browser browsing an imaginary internet. Given an input question, infer user r...'. The prompt also includes two example questions and answers. The flow is composed of several steps: 'extract\_query\_from\_question', 'search\_on\_bing', 'process\_search\_result', 'extract\_answer', and 'detect\_harmful\_content'. The 'search\_on\_bing' step is currently selected, showing its configuration and a preview of the search results. The interface includes a left-hand navigation pane, a top toolbar with 'Run', 'Clone', and 'Save' buttons, and a right-hand pane showing the flow's execution history and logs.

# Start building plugins today

<https://aka.ms/PluginRepo>

The screenshot shows a GitHub repository page for 'Azure-Samples / openai-plugin-fastapi'. The repository is private and has 10 watchers, 2 forks, and 0 stars. The main branch is 'main' with 3 branches and 0 tags. A recent merge pull request #4 from binderjoe/pa... is shown, merged 1 hour ago with 16 commits. The file list includes folders like .devcontainer, .github/workflows, .vscode, .well-known, data, infra, routers, scripts, and files like .gitattributes, .gitignore, and Dockerfile. The right sidebar shows 'About' with no description, 'Releases' with no releases published, and 'Packages' with no packages published.

File/Folder	Commit Message	Last Commit
.devcontainer	Init	yesterday
.github/workflows	Remove extra branch	20 hours ago
.vscode	Init	yesterday
.well-known	openapi yaml out of the box (#2)	2 hours ago
data	Init	yesterday
infra	Init	yesterday
routers	openapi yaml out of the box (#2)	2 hours ago
scripts	Init	yesterday
.gitattributes	Init	yesterday
.gitignore	Init	yesterday
Dockerfile	Init	yesterday

# Project Florence – Large foundation model

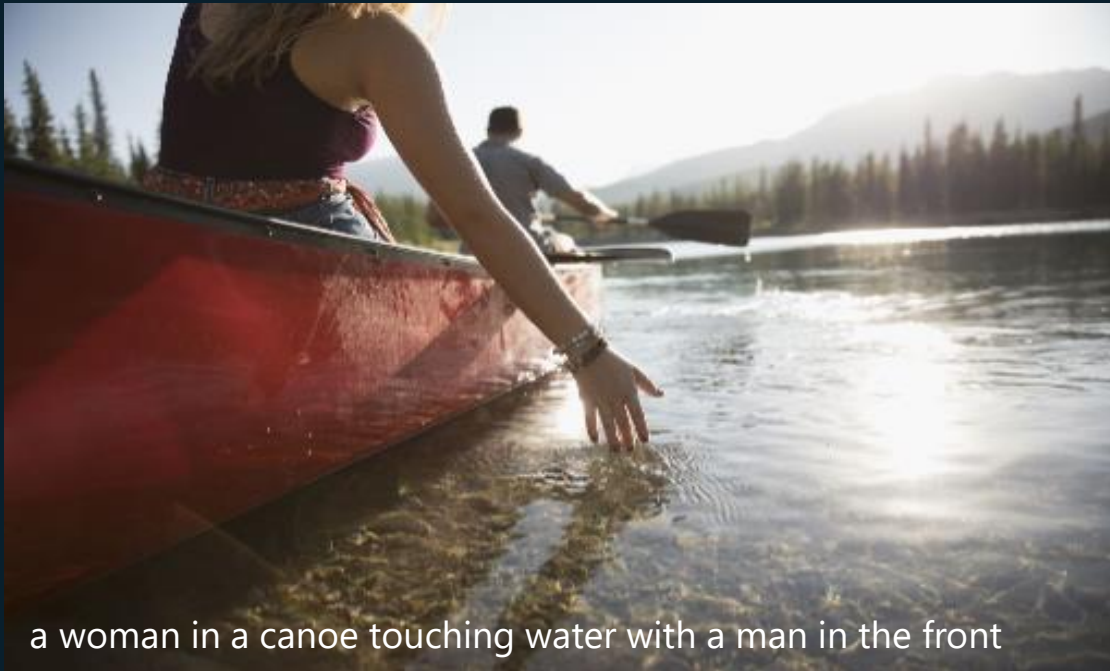


# Image Analysis APIs

## REST API

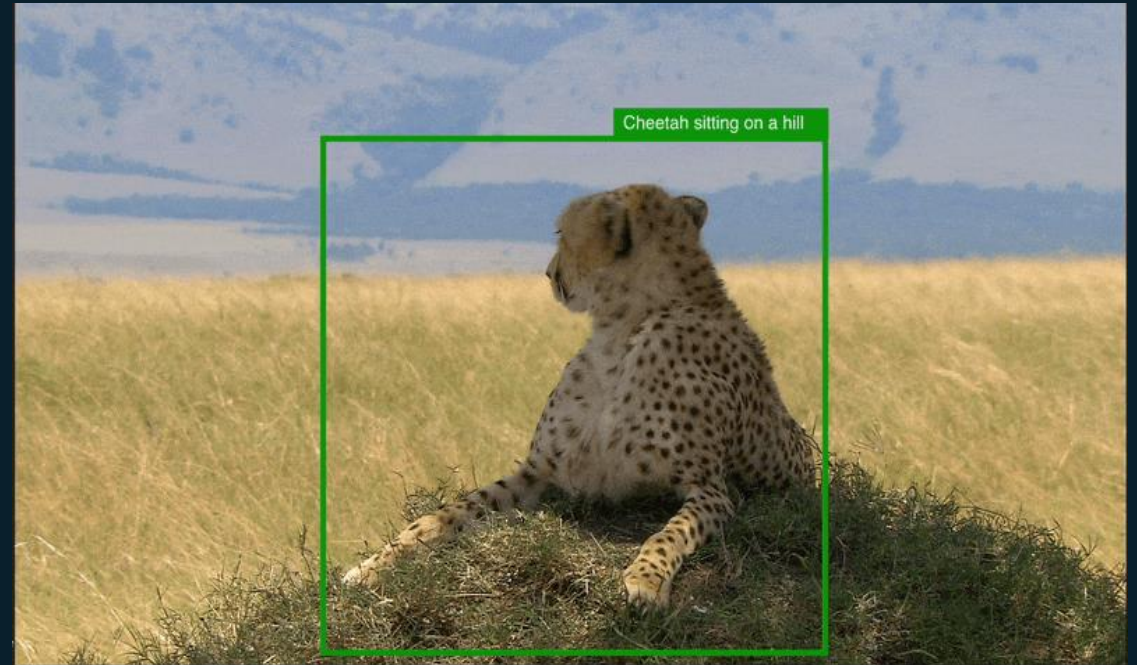
<https://{endpoint}/computervision/imageanalysis:analyze?features=caption>

<https://{endpoint}/computervision/imageanalysis:analyze?features=densecaptions>



a woman in a canoe touching water with a man in the front

Image captions



Dense captions

# Image retrieval APIs

Vision Studio > Search photos with natural language

## Search photos with natural language PREVIEW



Retrieve specific moments within your photo album. For example, you can query: a wedding you attended last summer, your pet, your favorite city. Search for images based on the content of the image itself, rather than relying solely on manually assigned keywords or tags

Platforms

☁ Cloud

[View documentation](#) [View SDK reference](#) [Use the REST API](#) [View samples on Github](#)

### Try it out

To try out this feature choose from a sample below. To try searching your own images, [sign in with Azure](#)

Sample image sets Try with your own images



Nature

No.of photos: 260



Manufacturing

No.of photos: 245



Education

No.of photos: 264



Retail

No.of photos: 265

### Select a retrieval query or create your own

Enter a custom query

Search

[Reset search](#)

### Query results

Query results vary from most relevant in the dataset to least relevant. Utilize the slider below to view more or less images based on their relevance to the retrieval query.

Most relevant   Least relevant



# Image retrieval APIs cont.

## REST API

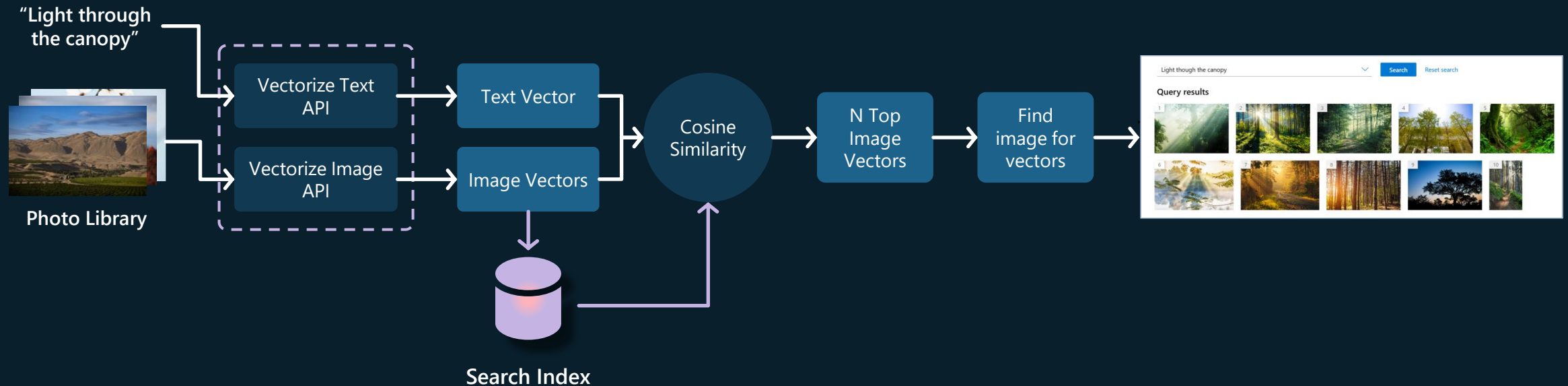
`https://<endpoint>/computervision/retrieval:vectorizeImage`

`https://<endpoint>/computervision/retrieval:vectorizeText`

### 1. Vectorize

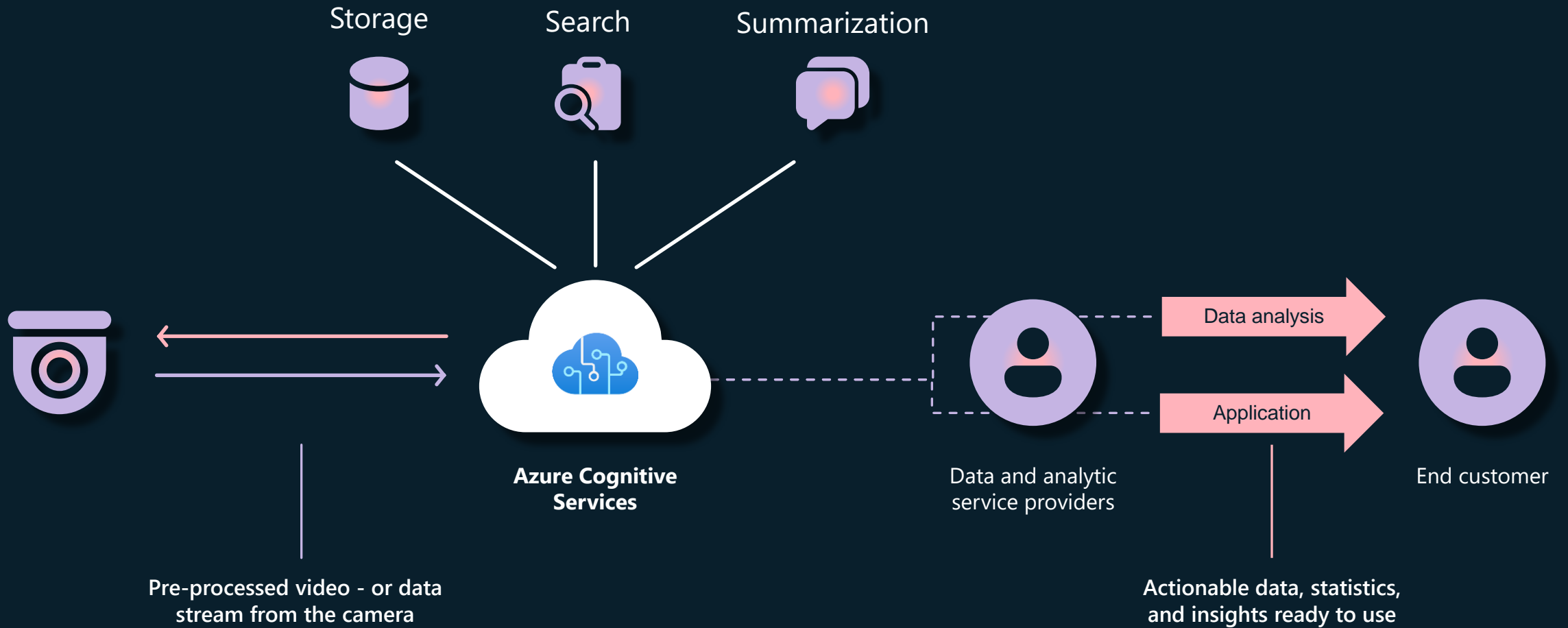
### 2. Measure

### 3. Retrieve

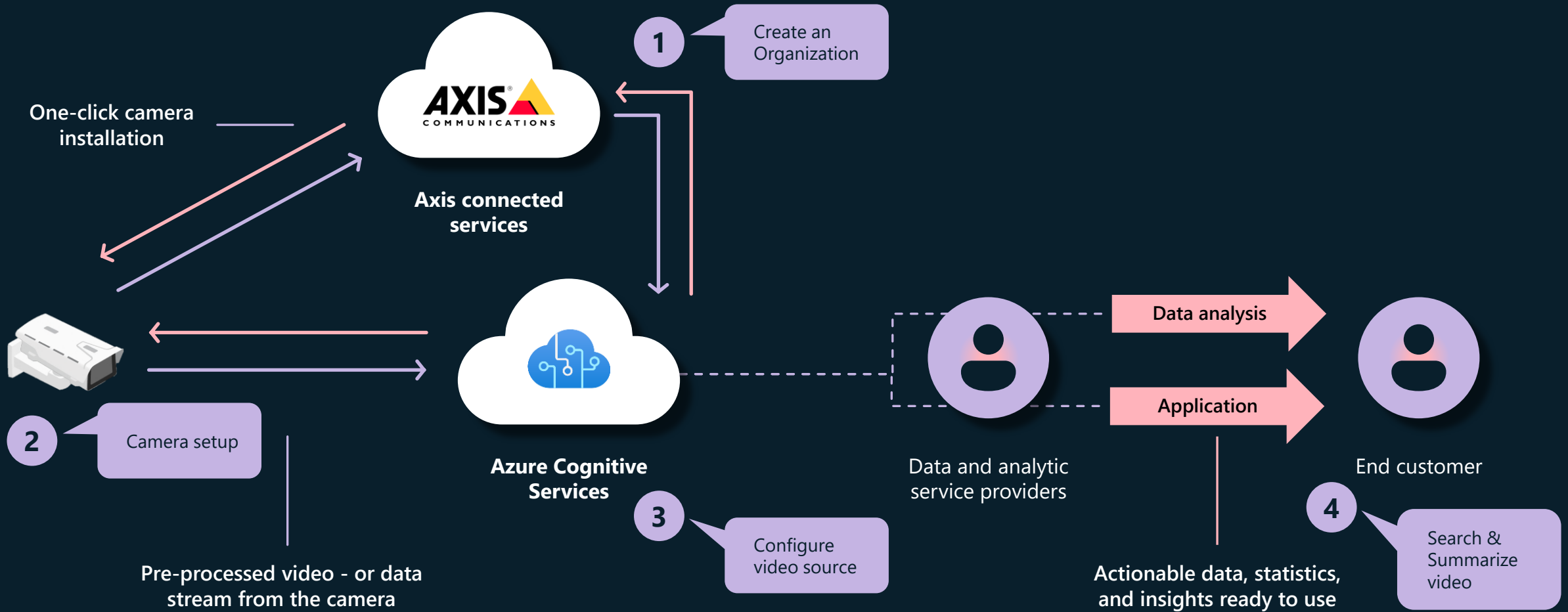




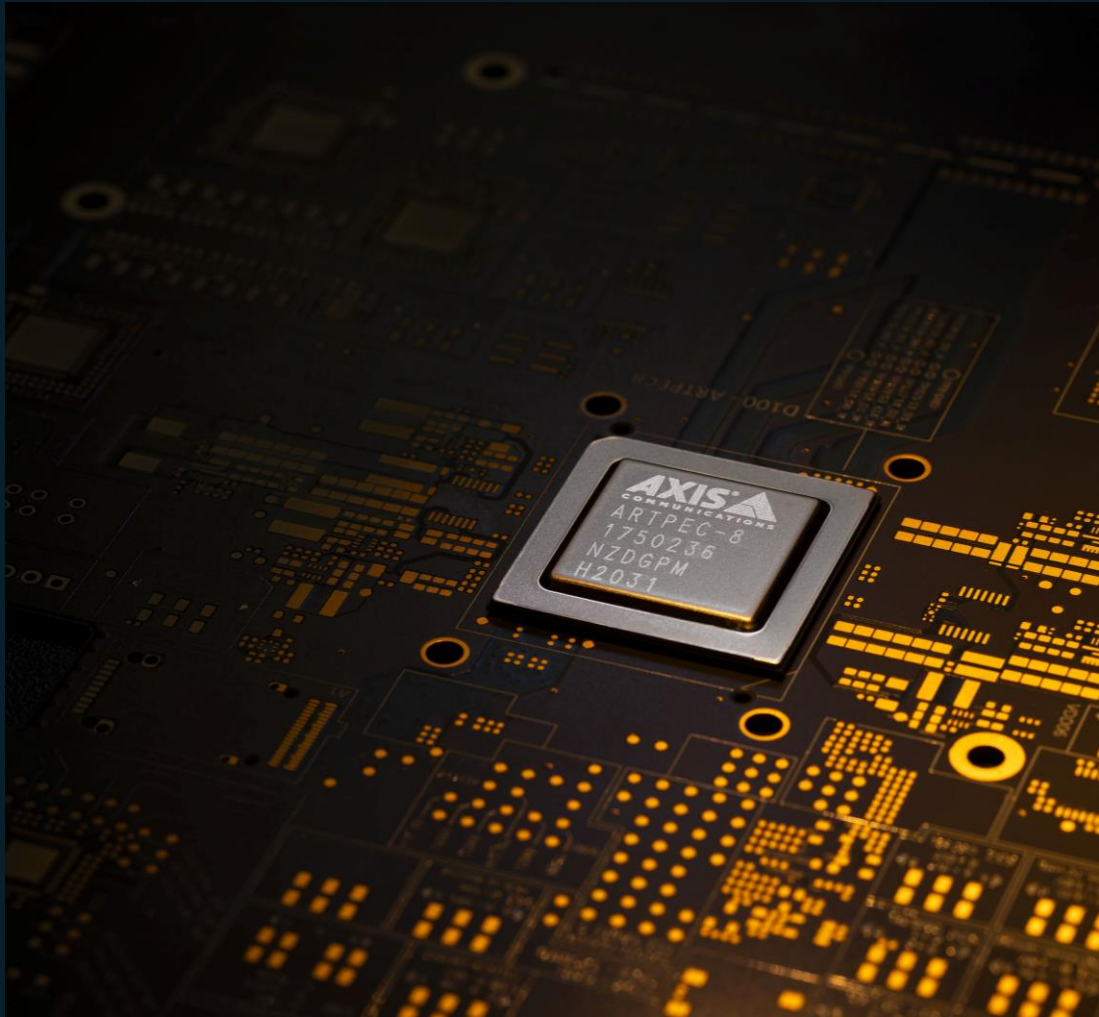
# Video Frame Locator & Summarization



# Axis Cameras Connected to Azure Cognitive Services cont.



# Axis supported portfolio



The Azure Cognitive Services for Vision integration is supported on camera models with ARTPEC-8 and CV25 System-on-chip (SoC).

Find a compatible camera at:

<https://www.axis.com/support/tools/product-selector>

[AxisCommunications/acap-integration-examples-azure: Integration between Axis devices and Microsoft Azure \(github.com\)](https://github.com/AxisCommunications/acap-integration-examples-azure)

# Microsoft Build 2023 Book of News

## Announcements

Azure Speech, Multistyle Auto Predict

Language Summarization

Azure Form Recognizer, Query Fields

Azure OpenAI, On your Data

Azure OpenAI, Plugins

Promptflow

Azure AI Content Safety

Azure Cognitive Search, Vector Search

# Links to assets

[microsoft/Customer-Service-Conversational-Insights-with-Azure-OpenAI-Services \(github.com\)](#)

[openai/End to end Solutions at main · Azure-Samples/openai \(github.com\)](#)

[azure-openai-samples/use cases at main · Azure/azure-openai-samples \(github.com\)](#)

[https://github.com/Azure/ai-solution-accelerators-list/tree/main/OpenAIDemos](#)

[pablomarin/GPT-Azure-Search-Engine: Azure Cognitive Search + Azure OpenAI Accelerator \(github.com\)](#)



# Prompt Engineering Guide and Techniques



Mahmoud Moussa & Jurgens de Bruin  
Cloud Solution Architect

# Text classification

## Named Entity Recognition (Zero-Shot)

Prompt:

Extract the name and mailing address from this email:

Dear Kelly,

It was great to talk to you at the seminar. I thought Jane's talk was quite good.

Thank you for the book. Here's my address 2111 Ash Lane, Crestview CA 92002

Best,

Maya

**Name:** Maya

**Mailing Address:** 2111 Ash Lane, Crestview CA 92002

# Text classification

## Custom Named Entities (Few-Shot)

**Prompt:**

**Extract job titles from the following sentences.**

**Sentence: John Doe has been working for Microsoft for 20 years as a Linux Engineer.**

**Job title: Linux Engineer**

**###**

**Sentence: John Doe has been working for Microsoft for 20 years and he loved it.**

**Job title: none**

**###**

**Sentence: Marc Simoncini | Director | Meetic**

**Job title: Director**

**###**

**Sentence: Damien is the CTO of Platform.sh, he was previously the CTO of Commerce Guys, a leading ecommerce provider.**

**Completion:**

**Job title: CTO**



# Text classification

## Multi-Label Text Classification (Few-Shot)

**Prompt:**

The following is a list of movies and the categories they fall into:

**Topgun:** As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.

**Category:** Drama, Action

**Hustle:** A basketball scout discovers a phenomenal street ball player while in Spain and sees the prospect as his opportunity to get back into the NBA.

**Category:** Comedy, Drama, Sport

**The Northman:** From visionary director Robert Eggers comes The Northman, an action-filled epic that follows a young Viking prince on his quest to avenge his father's murder.

**Category:**

**Completion :**

Action, Adventure, Drama

# Text classification

## Paraphrasing (Few-Shot)

Prompt:

Article: Searching a specific search tree for a binary key can be programmed recursively or iteratively.

**Paraphrase:** Searching a specific search tree according to a binary key can be recursively or iteratively programmed.

Article: It was first released as a knapweed biocontrol in the 1980s in Oregon , and it is currently established in the Pacific Northwest.

**Paraphrase:** It was first released as Knopweed Biocontrol in Oregon in the 1980s , and is currently established in the Pacific Northwest.

Article: 4-OHT binds to ER , the ER / tamoxifen complex recruits other proteins known as co-repressors and then binds to DNA to modulate gene expression.

**Paraphrase:** The ER / Tamoxifen complex binds other proteins known as co-repressors and then binds to DNA to modulate gene expression

Article: Microsoft Corporation is a multinational technology company based in Redmond, Washington. It is best known for its Windows operating system, Microsoft Office suite, and Xbox video game consoles. The company was founded in 1975 by Bill Gates and Paul Allen and became dominant in the personal computer market with MS-DOS and Windows.

Completion:

**Paraphrase:** Microsoft Corporation is a technology company with a global presence, headquartered in Redmond, Washington. Its most recognizable products include the Windows operating system, Microsoft Office suite, and Xbox gaming consoles. The company was established in 1975 by Bill Gates and Paul Allen, and it achieved market dominance in personal computers with the MS-DOS and Windows operating systems.

# Reason over unstructured text

## Question and answering

Prompt:

On Monday, Microsoft launched an Azure OpenAI Service as part of its Azure cloud platform, offering businesses and start-ups the ability to incorporate models like ChatGPT into their own systems. The company has already been building AI tools into many of its consumer products, such as a DALL-E 2 feature in its Bing search engine that can create images based on a text prompt, and the Information reported recently that it's working to bring more of them to Microsoft Office as well.

What will Microsoft incorporate in its own systems?

Microsoft plans to incorporate AI tools into its own systems, such as a DALL-E 2 feature in its Bing search engine that can create images based on a text prompt, and the Information reported recently that it's working to bring more of them to Microsoft Office as well

# Text classification

## Keyword Extraction (Few-Shot)

Prompt:

Information Retrieval (IR) is the process of obtaining resources relevant to the information need. For instance, a search query on a web search engine can be an information need. The search engine can return web pages that represent relevant resources.

**Keywords:** searching, missing, desert

###

I believe that using a document about a topic that the readers know quite a bit about helps you understand if the resulting keyphrases are of quality.

**Keywords:** document, understand, keyphrases

###

Since transformer models have a token limit, you might run into some errors when inputting large documents. In that case, you could consider splitting up your document into paragraphs and mean pooling (taking the average of) the resulting vectors.

Completion:

**Keywords:** transformer, models, token

# Summarization

## Extractive vs Abstractive

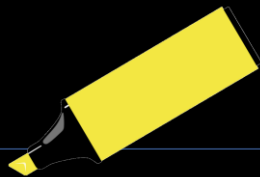
### Extractive

Select parts of the original text to form a summary

- Easier
- Restrictive (no paraphrasing)

Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and was rushed to the hospital.

Peter and Elizabeth attend party city. Elizabeth rushed hospital



### Extractive (Semantic)

Generates new text using natural language generation techniques

- More difficult
- More flexible (more human)

Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and was rushed to the hospital.

Elizabeth was hospitalized after attending a party with Peter.



# Abstractive summarization

## Contact Center Summarization

Prompt:

Summarize this for a call center agent:

Agent: Thank you for calling ADM. Who am I am speaking to?

Customer: Hello, my name is Peter Smith. I own a small business and have some questions regarding payroll processing.

Agent: Good morning, Peter, before we get started may I ask you a few questions so that I better answer your questions today?

Customer: Thank you that is quite helpful. Are there specific regulations that I need to follow?

Agent: Certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL)

Completion:

The agent is speaking to a customer named Peter Smith. Peter is asking for tips on manual payroll processing. The agent also informs Peter that certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL).

# Prompt Insert

Only DaVinci-002+

## Before

These are agenda topics for a customer presentation.

1. Introduction to GPT3

[insert]

10. Conclusion

## After

These are agenda topics for a customer presentation.

1. Introduction to GPT3

2. Overview of GPT3 Features

3. Benefits of GPT3

4. GPT3 Use Cases

5. GPT3 Pricing

6. GPT3 Security

7. GPT3 Support

8. GPT3 Integration

9. Q&A Session

10. Conclusion

# Insert classification

listen to WestBam album allergic on google music: PlayMusic

give me a list of movie times for films in the area: SearchScreeningEvent

show me the picture creatures of light and darkness: SearchCreativeWork

I would like to go to the popular bistro in oh: BookRestaurant

what is the weather like in the city of Frewen in the country of Venezuela: GetWeather

I want to book a flight to Delhi: bookFlight



# Chatbot with personality

This is a discussion between a [human] and a [robot].  
The [robot] is very nice and empathetic.

[human]: Hello nice to meet you.

[robot]: Nice to meet you too.

###

[human]: How is it going today?

[robot]: Not so bad, thank you! How about you?

###

[human]: I am ok, but I am a bit sad...

[robot]: Oh? Why that?

###

[human]: I broke up with my girlfriend...

Oh? How did that happen?

# Analyze customer feedback

## **Customer review:**

The red mountain bike I recently purchased is an excellent choice for anyone who wants a reliable, high-quality bike for off-road adventures. The bike is built with sturdy materials and the red color gives it a bold and stylish look. The bike also has many great features that make it a joy to ride. One of my favorite things about this bike is how customizable it is. The seat can be easily adjusted to fit different heights, and the handlebars can be adjusted for a comfortable riding position. The bike also has multiple gears, allowing you to adjust the resistance and speed as needed.

## **Analyze the customer review and provide feedback and sentiment**

### Feedback:

Overall, the customer feedback is positive and indicates satisfaction with the red mountain bike purchased. The customer highlights the bike's reliability and high-quality construction, as well as its stylish appearance. The customer also praises the bike's many features that make it enjoyable to ride, including its customization options and multiple gears.

# Topic classification for Bot routing

Message: When the spaceship landed on Mars, the whole humanity was excited

Topic: space

###

Message: I love playing tennis and golf. I'm practicing twice a week.

Topic: sport

###

Message: Managing a team of sales people is a tough but rewarding job.

Topic: business

###

Message: I am trying to cook chicken with tomatoes

Topic: Cooking

# Summarize conversational transcripts

Summarize this conversation transcript:

**Agent:** Thank you for calling ADM. Who am I am speaking to? **Customer:** Hello, my name is Peter Smith. I own a small business and have some questions regarding payroll processing. **Agent:** Good morning, Peter, before we get started may I ask you a few questions so that I better answer your questions today? **Customer:** Yes, of course. **Agent:** In case we get disconnected, can you please share your phone number and email address so that we may contact and share additional information? **Customer:** Yes, my phone number is (514) 777-5232 and my email address is petersmith@gmail.com **Agent:** May I ask where is your business located so I can answer your questions based on your location **Customer:** Yes, my business is located in Miami, Florida. **Agent:** Thank you, what can I help you with today? **Customer:** I am fairly new to payroll processing and do everything manually today. Do you have any advice or tips on manual payroll processing? **Agent:** If you're a small business with only a few employees and choose to process payroll manually, you will need to keep precise records of hours worked, wages paid and worker classifications, among other details. You must also ensure your calculations are correct and remember to file all the necessary taxes and paperwork with government authorities on time. **Customer:** Thank you that is quite helpful. Are there specific regulations that I need to follow? **Agent:** Certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL). Some of the laws you must comply with include: Fair Labor Standards Act (FLSA), Federal Insurance Contributions Act (FICA), Federal Unemployment Tax Act (FUTA) **Customer:** This sounds very complicated. I don't really feel like managing all of this. **Agent:** If you would like I can transfer you to one of our specialized agents who can talk to you about payroll service providers that would help you manage all of that. **Customer:** That would be great, thank you very much for al your help.

The agent is speaking to a customer named Peter Smith. Peter is asking for tips on manual payroll processing. The agent advises Peter to keep precise records and to file all necessary taxes and paperwork on time. The agent also informs Peter that certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL). Finally, the agent offers to transfer Peter to a specialized agent who can help him find a payroll service provider.

# Product description and ad generation

Generate a product description out of keywords.

Keywords: shoes, women, \$59

Sentence: Beautiful shoes for women at the price of \$59.

###

Keywords: trousers, men, \$69

Sentence: Modern trousers for men, for \$69 only.

###

Keywords: gloves, winter, \$19

Sentence: Amazingly hot gloves for cold winters, at \$19.

###

Keywords: t-shirt, men, \$39

Sentence:

Extraordinary t-shirt for men, for \$39 only.

# Product recommendation

Generate product recommendations:

Product: table, chair, plate

Recommendation: utensils

###

Product: skis, skateboard, bike

Recommendation: golf clubs

###

Product: shoes, shirt, hat

Recommendation: Pants

# Naming Convention

Azure OpenAI's model names typically correspond to the following standard naming convention:

{family} - {capability} [-{input-type}] - {identifier}

Element	Description
{family}	The model family of the model. For example, <u>GPT-3 models</u> uses text, while <u>Codex models</u> use code.
{capability}	The relative capability of the model. For example, GPT-3 models include ada, babbage, curie, and davinci.
{input-type}	( <u>Embeddings models</u> only) The input type of the embedding supported by the model. For example, text search embedding models support doc and query.
{identifier}	The version identifier of the model.

text – davinci – 003

{family} - {capability}[ {identifier}]

This model is a GPT-3 text model, the most powerful (davinci), and of the latest version (003)

*\*<https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/models>*

**“Insert quote text here. This layout  
accommodates longer quotes very well.”**

**Name attribute**

Job title or another attribute



# Tokens

You can think of tokens as pieces of words used for natural language processing. For English text, 1 token is approximately 4 characters or 0.75 words.

---

As a point of reference, the collected works of Shakespeare are about 900,000 words or 1.2M tokens.

# 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:

IF TEMPERATURE IS 0

Horses are my favorite animal

Horses are my favorite animal

Horses are my favorite animal

Horses are my favorite animal

IF TEMPERATURE IS 1

Horses are my favorite animal

Horses are my favorite animals

Horses are my favorite !

Horses are my favorite animal

## 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)

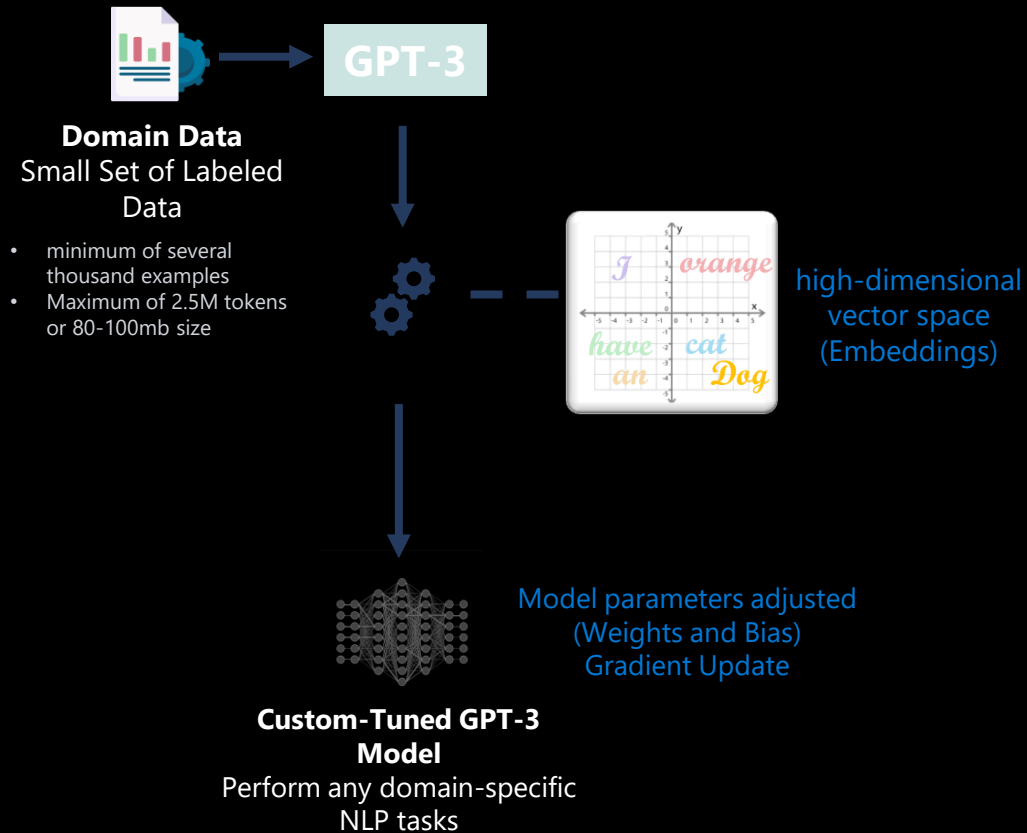
Blaze the Miracle Mare, Pegasus the Winged Warrior, Secretariat the Superhorse

## Completion Temperature 1 (often different)

Blaze of Glory, Sterling Silver, Thunderbolt

# Model Adaptation with specific domain data

## Custom-Tuning



Custom-tuning results is a new model being generated with updated weights and biases.

This is in contrast to **few-shot learning** in which model weights and biases are not updated.

# When Custom-Tuning is needed

If model is making untrue statements (“hallucinations”), then mitigate the hallucinations

Accuracy of results of the model does not meet customer requirements

Custom-tuning lets you get more out of the models available through the API by providing:

- Higher quality results than prompt design
- Ability to train on more examples than can fit in a prompt
- Lower latency requests

Custom-tuning improves over few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks.

# Best practices of Custom-Tuning

Custom-tuning data set must be in JSON format

A set of training examples that each consist of a single input ("prompt") and its associated output ("completion")

For classification task, the prompt is the problem statement, completion is the target class

For text generation task, the prompt is the instruction/question/request, and completion is the text ground truth

# Best practices of Custom-Tuning

Custom-tuning data size: Advanced model (Davinci) performs better with limited amount of data; with enough data, all models do well.

Custom-tuning performs better with more high-quality examples.

To custom-tune a model that performs better than using a high-quality prompt with base models, you should provide at least a few hundred high-quality examples, ideally vetted by human experts.

From there, performance tends to linearly increase with every doubling of the number of examples. Increasing the number of examples is usually the best and most reliable way of improving accuracy.

# Best practices of Custom-Tuning

Make sure to completely remove wrong labels in custom-tuning dataset. If you are custom-tuning on a pre-existing dataset rather than writing prompts from scratch, be sure to manually review your data for offensive or inaccurate content if possible, or review as many random samples of the dataset as possible if it is large.

# Custom-Tuning data formatting

To custom-tune a model, you'll need a set of training examples that each consist of a single input ("prompt") and its associated output ("completion").

This is notably different from using the base models, where you might input detailed instructions or multiple examples in a single prompt.

Each prompt should end with a fixed separator to inform the model when the prompt ends and the completion begins.

A simple separator which generally works well is `\n\n###\n\n`.

The separator should not appear elsewhere in any prompt.



# Custom-Tuning data formatting

Each completion should start with a whitespace due to tokenization, which tokenizes most words with a preceding whitespace.

Each completion should end with a fixed stop sequence to inform the model when the completion ends.

A stop sequence could be `\n`, `###`, or any other token that does not appear in any completion.

For inference, you should format your prompts in the same way as you did when creating the training dataset, including the same separator.

Also specify the same stop sequence to properly truncate the completion.

# Hyperparameters specific to Custom-Tuning

Parameter	Description	Recommendation
<b>n_epochs</b> controls how many times each example is trained on	The number of epochs to train the model for. An epoch refers to one full cycle through the training dataset.	Start from 4 and small datasets may need more epochs and large datasets may need fewer epochs. If you see low training accuracy (underfitting), try increasing n_epochs. If you see high training accuracy but low validation accuracy (overfitting), try lowering n_epochs.
<b>batch_size</b> controls the number of training examples used in a single training pass	The batch size to use for training. The batch size is the number of training examples used to train a single forward and backward pass.	We've seen good performance in the range of 0.01% to 4% of training set size. In general, we've found that larger batch sizes tend to work better for larger datasets.
<b>learning_rate_multiplier</b> controls rate at which the model weights are updated	The learning rate multiplier to use for training. The fine-tuning learning rate is the original learning rate used for pre-training, multiplied by this value.	We recommend experimenting with values in the range 0.02 to 0.2 to see what produces the best results. Empirically, we've found that larger learning rates often perform better with larger batch sizes. Empirically, we found learning_rate_multiplier has minor impact compared to n_epochs and batch_size.
<b>prompt_loss_weight</b> controls how much the model learns from prompt tokens vs completion tokens	The weight to use for loss on the prompt tokens. This value controls how much the model tries to learn to generate the prompt (as compared to the completion, which always has a weight of 1.0.) Increasing this value can add a stabilizing effect to training when completions are short.	When a model is fine-tuned, it learns to produce text it sees in both the prompt and the completion. In fact, from the point of view of the model being fine-tuned, the distinction between prompt and completion is mostly arbitrary. The only difference between prompt text and completion text is that the model learns less from each prompt token than it does from each completion token. This ratio is controlled by the prompt_loss_weight, which by default is 0.1. If prompts are extremely long (relative to completions), it may make sense to reduce this weight to avoid over-prioritizing learning the prompt. Empirically, we found prompt_loss_weight has minor impact compared to n_epochs and batch_size.

# Guides and examples of Custom-Tuning

[How to customize a model with Azure OpenAI - Azure OpenAI | Microsoft Learn](#)

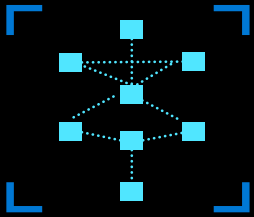
## [Custom-tuning—OpenAI API](#)

<https://platform.openai.com/docs/guides/fine-tuning/general-best-practices>

[\[PUBLIC\] Best practices for fine-tuning GPT-3 to classify text—Google Docs](#)

[Fine-tuning a Classifier to Improve Truthfulness | OpenAI Help Center](#)

# Embeddings



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

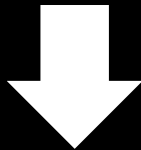
The embedding is an information dense representation of the semantic meaning of a piece of text.

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.

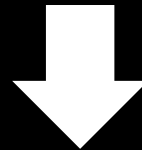
# Embeddings make it possible to map content to a “semantic space”

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



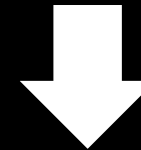
[ 15 34 24 13 ...]

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



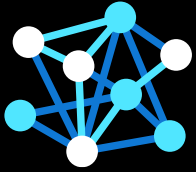
[16 22 89 26 ...]

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



[ 20 13 31 89 ...]

# Embedding models



Different Azure OpenAI Service embedding models are specifically created to be good at a particular task.

- **Similarity embeddings** are good at capturing semantic similarity between two or more pieces of text.
- **Text search embeddings** help measure long documents relevant to a short query.
- **Code search embeddings** are useful for embedding code snippets and embedding natural language search queries.

Embeddings make it easier to do machine learning on large inputs representing words by capturing the semantic similarities in a vector space.

Therefore, we can use embeddings to determine if two text chunks are semantically related or similar, and provide a score to assess similarity.

# Similarity Search with embeddings

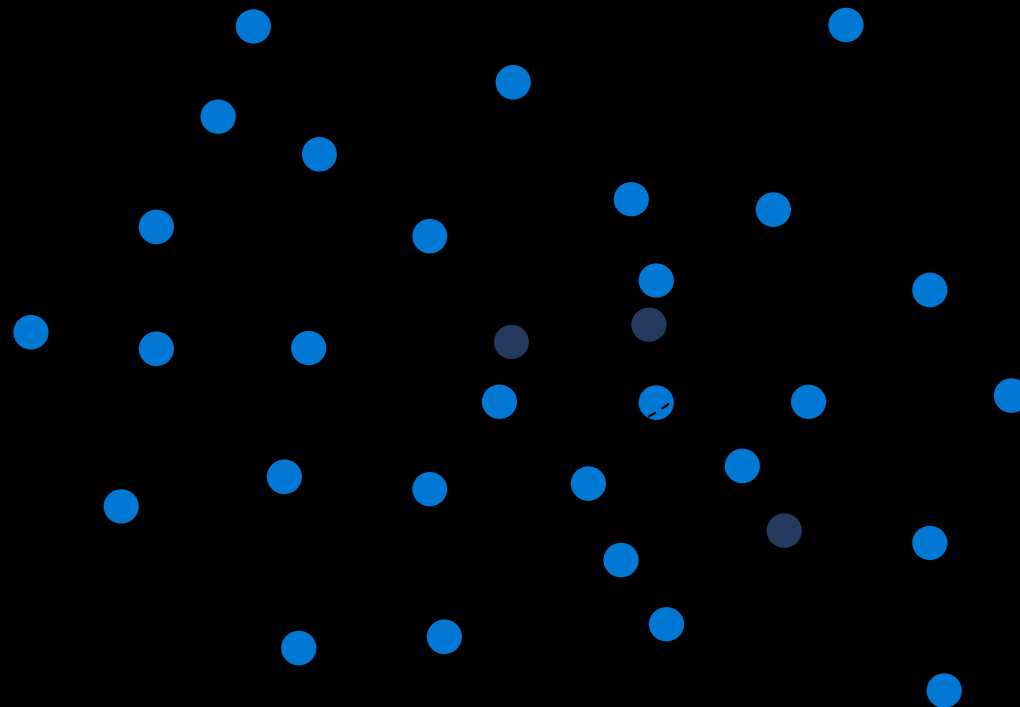
Once you encode your content as embeddings, you can then get an embedding from the user input and use that to find the most semantically similar content.

*user  
input*

"What is a neutron star?"

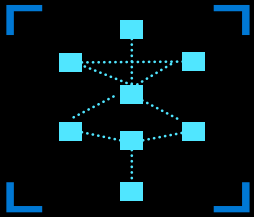
*embedding*

[ 13 33 34 13 ... ]



*result set*

# Embeddings



We strongly recommend using text-embedding-ada-002 (Version 2). This model/version provides parity with OpenAI's text-embedding-ada-002.

To learn more about the improvements offered by this model, please refer to this [blog post](#).

Even if you are currently using Version 1, you should migrate to Version 2 to take advantage of the latest weights/updated token limit. Version 1 and Version 2 are not interchangeable, so document embedding and document search must be done using the same version of the model.



# 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.

Prompt typically includes problem descriptions, instructions on how to solve the problem, and examples of correct problem and solution pairs.

# Prompt Engineering Guide

---

1. Give clearer instructions

2. Split complex tasks into simpler subtasks

3. Structure the instruction to keep the model on task

4. Prompt the model to explain before answering

5. Ask for justifications of many possible answers, and then synthesize

6. Generate many outputs, and then use the model to pick the best one

7. Custom-tune custom models to maximize performance

# An example of Prompt Engineering– Topic Classifier

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Instructions

High level and detailed

Customer question: Hi there, do you know how to choose flood insurance?

Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.

Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?

Classified topic: 3

Examples

Order of examples matter

Customer question: {insert new question here}

Classified topic:

Task and Prompting  
answer

# Prompt Chaining

Please **extract entities** from the following news article: 'The new iPhone model is set to be released next month. It has been highly anticipated by Apple fans and is expected to feature a larger screen and improved camera

```
{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}
```

Please **summarize** the information about the product

```
[{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}]
```

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

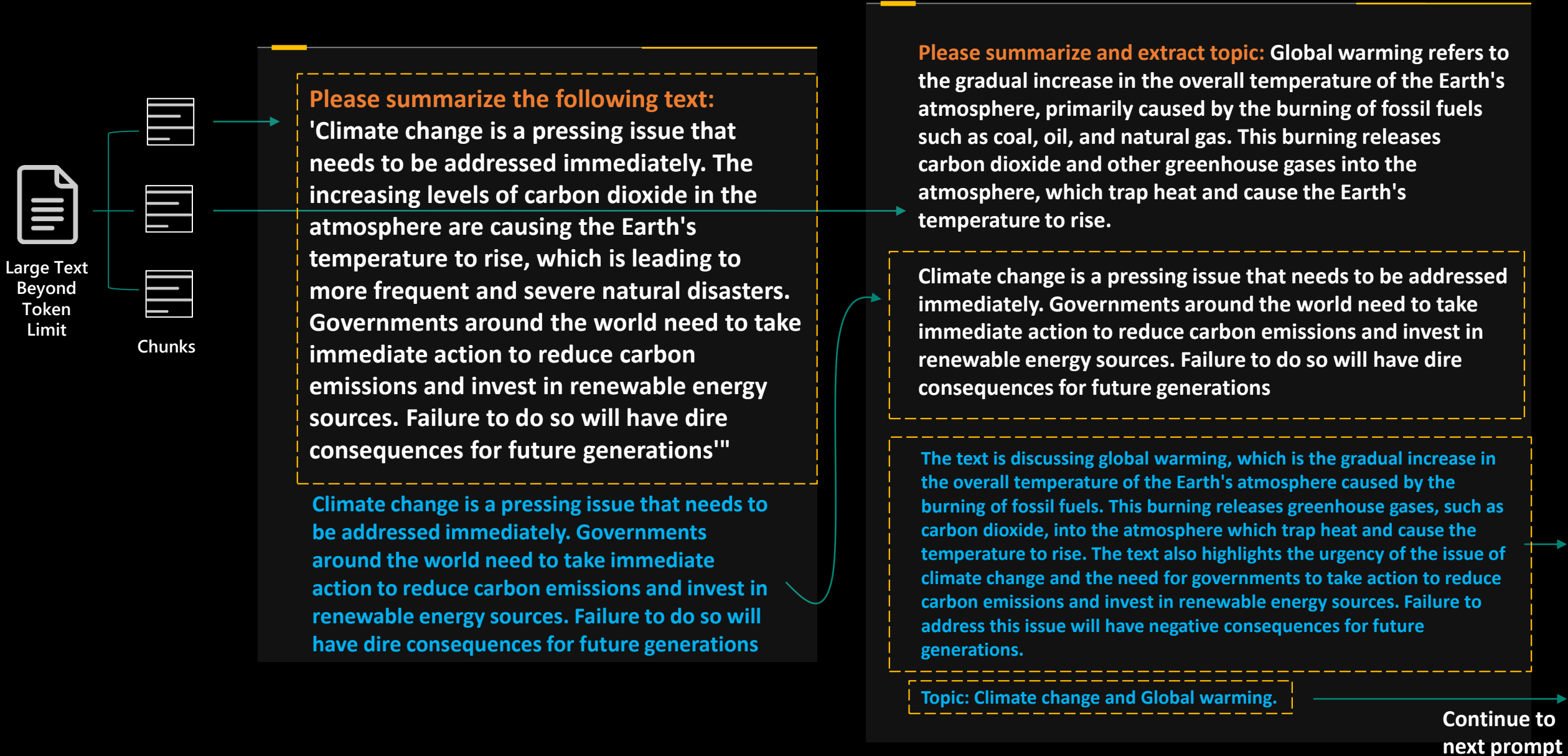
Please provide a **sentiment** for the following text:

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

**Positive**

# Prompt Chunking (token limit)

Then all summaries are passed to GPT again to summarize, i.e. second round of summarization. The sum of the summaries has to fit within the token limit.



# Few-Shot Reasoning (Human version)

Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

**The answer is 27**



Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5+6 = 11$ . The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

**The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23-20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9.**



# Two Types of Prompt Engineering

## Zero-shot

The model predicts the answer given only a natural language description of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: {insert new question here}

Classified topic:

## Few-shot

In addition to the task description, the model sees a few examples of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: Hi there, do you know how to choose flood insurance?  
Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.  
Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?  
Classified topic: 3

Customer question: {insert new question here}

Classified topic:

# Model Reasoning—Chain-of-Thought

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

The annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day is 8,000,000 liters per year.

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

**Let's think step by step and explain the calculation step by step.**

**Step 1:** Calculate the total number of days in a year:  
There are 365 days in a year.

**Step 2:** Calculate the total number of days the family is at home:  
The family is at home an average of 200 days per year.

**Step 3:** Calculate the total water demand of the family:  
The family uses an average of 100 liters of water per day. Therefore, the total water demand of the family is  $100 \text{ liters} \times 200 \text{ days} = 20,000 \text{ liters per year}$ .

**Step 4:** Calculate the annual water demand of the single-family household:  
The annual water demand of the single-family household is 20,000 liters per year.



# Using Chain-of-Thought for Reasoning

On average Joe throws 25 punches per minute.  
A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

**Let's think step by step and explain the calculation step by step.**

In one minute, Joe throws 25 punches.  
In three minutes, Joe throws  $3 * 25 = 75$  punches.  
In five rounds, Joe throws  $5 * 75 = 375$  punches.

One average Joe throws 25 punches per minute.  
A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

In one minute, Joe throws 25 punches.  
In three minutes, Joe throws  $3 * 25 = 75$  punches.  
In five rounds, Joe throws  $5 * 75 = 375$  punches.

Therefore, the answer (Arabic numerals) is

# Few-Shot– Chain-of-Thought--Examples

## Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ .

The answer is 11.

---

## StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about  $0.6 \text{ g/cm}^3$ , which is less than water. Thus, a pear would float

So the answer is no.

---

## SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

## Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500.  $9 + 90(2) + 401(3) = 1392$ .

The answer is (b).

---

## Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943.

So the answer is 05/23/1943.

---

## Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya".

So the answer is ya

## CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do.

So the answer is (b).

---

## Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer.

So the answer is no.

---

## Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up.

So the answer is no.

# Selection-inference prompting

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

14 times



It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

To solve "How many times can she slide before it closes" we need to first solve: **Chain-of-Thought**

How much time does it take for Amy to climb and slide?

Decomposes problem into a smaller task (Reasoning)

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

How much time does it take for Amy to climb and slide?

It takes Amy 5 minutes to climb and slide

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

It takes Amy 5 minutes to climb and slide.

She can slide 3 times before the water slides close.



# Reduce Hallucination using Prompt Engineering

## Include

Include instructions of requesting the model not to make up stuff but stay with facts.

## Restrict

Restrict the output (e.g., choose from a confined list instead of generating free form strings)

## Add

Add Chain of Thought style of instruction, "Solve the problem step by step."

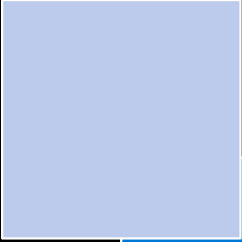
## Repeat

Repeat most important instructions in the prompt a couple of times.

## Position

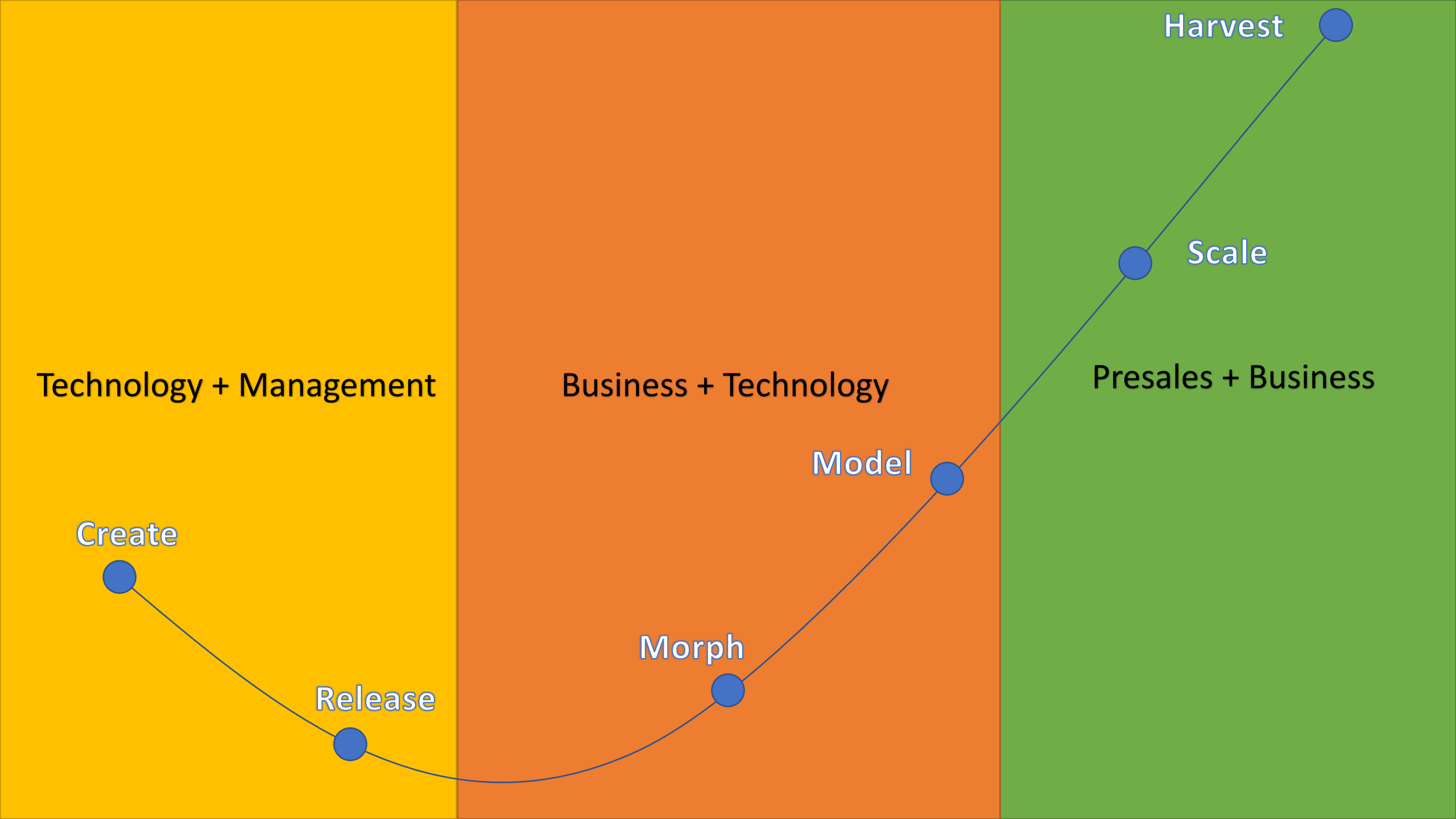
Position most important instructions in the last making use of latency effect.

# LLM Solution Pieces

- 
- Open AI
    - Prompts
    - Embeddings
    - Models
    - Hyper Parameters
    - APIs

- 
- Customization
    - Vector Databases
    - LangChain
      - Agents
      - Connectors
    - Data Engineering
    - Embedding

- 
- Roles
    - Developer
    - Prompt Engineering
    - Data Engineer
    - Solution Architect
    - Analyst
    - Presales



Technology + Management

Business + Technology

Presales + Business

Create

Release

Morph

Model

Harvest

Scale



Q&A

Azure GPT

| Ask the team anything



# Azure OpenAI Service partner Resources

---

## Partner access

- [Request Access to Azure OpenAI Service \(microsoft.com\)](#)

## Product pages

- [Azure OpenAI Service | Microsoft Azure](#)
- [Azure OpenAI Service – Advanced Language Models | Microsoft Azure](#)

## Partner-facing sales assets

- [Pricing \(Azure.com\)](#)
- [Azure OpenAI Briefing Deck \(GearUp\)](#)
- [Azure AI EBC \(GearUp\)](#)
- [Document analysis demo using Azure Form Recognizer and Azure OpenAI \(GearUp/Stream\)](#)

## Public blogs/announcements

- [Advancing human-centered AI: Updates on responsible AI research - Microsoft Research Blog \(2023 01 12\)](#)
- [Azure OpenAI Service Announcement Blog \(2023 01 16\)](#)
- [Revolutionize your Enterprise Data with ChatGPT: Next-gen Apps w/ Azure OpenAI and Cognitive Search | Microsoft Community Hub \(2023 03 09\)](#)
- [Azure OpenAI Service Chat GPT Announcement Blog \(2023 03 09\)](#)
- [Azure OpenAI Service Chat GPT Announcement Partner Blog \(2023 03 09\)](#)

## Partner skilling

- [Azure AI Bootcamp Registration Page](#)
- [Introduction to Azure OpenAI Service - Training \(Microsoft Learn\)](#)
- [Learning Plan Resources for Azure OpenAI Service \(GitHub\)](#)
- [Microsoft Mechanics ChatGPT & OpenAI powering your apps/Open AI Studio in Microsoft Azure \(YouTube\)](#)
- [Financial Services with OpenAI Exploring Use Cases and Opportunities \(YouTube\)](#)
- [Azure OpenAI Service Playlist \(YouTube\)](#)
- [Library of L200/300 technical slides \(GearUp\)](#)
- [Azure Open AI Service Workshop \(GitHub\)](#)
- [Azure OpenAI Service technical documentation \(Microsoft Learn\)](#)
- [What's new in Azure OpenAI Service \(Microsoft Learn\)](#)