



Deep Drive Architecture for Azure OpenAI



Agenda

- **Deep Dive Architecture for Azure OpenAI**
 - Different Azure Open AI Architecture
 - Introduction to AOA Models
 - Overview of ChatGPT (Large Conversational Foundation Model)
 - 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)
 - Contact Centre Analytics using Speech API & Azure OpenAI Service
 - AI-Powered Q&A over Enterprise Data Sources
- **Available Azure OpenAI Accelerators**

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

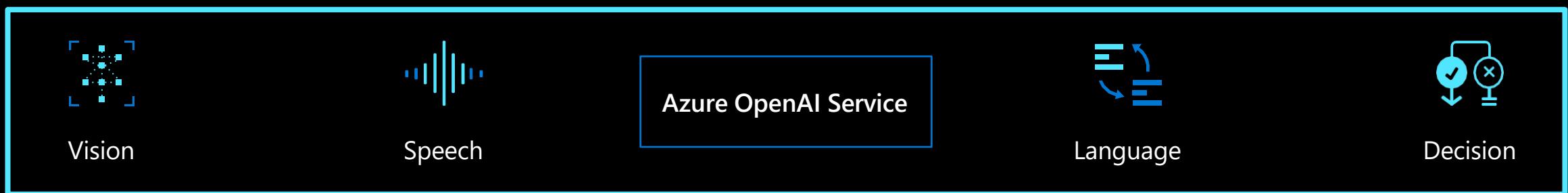
End to end call center analytics: classification, sentiment, entity extraction, summarization and email generation

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

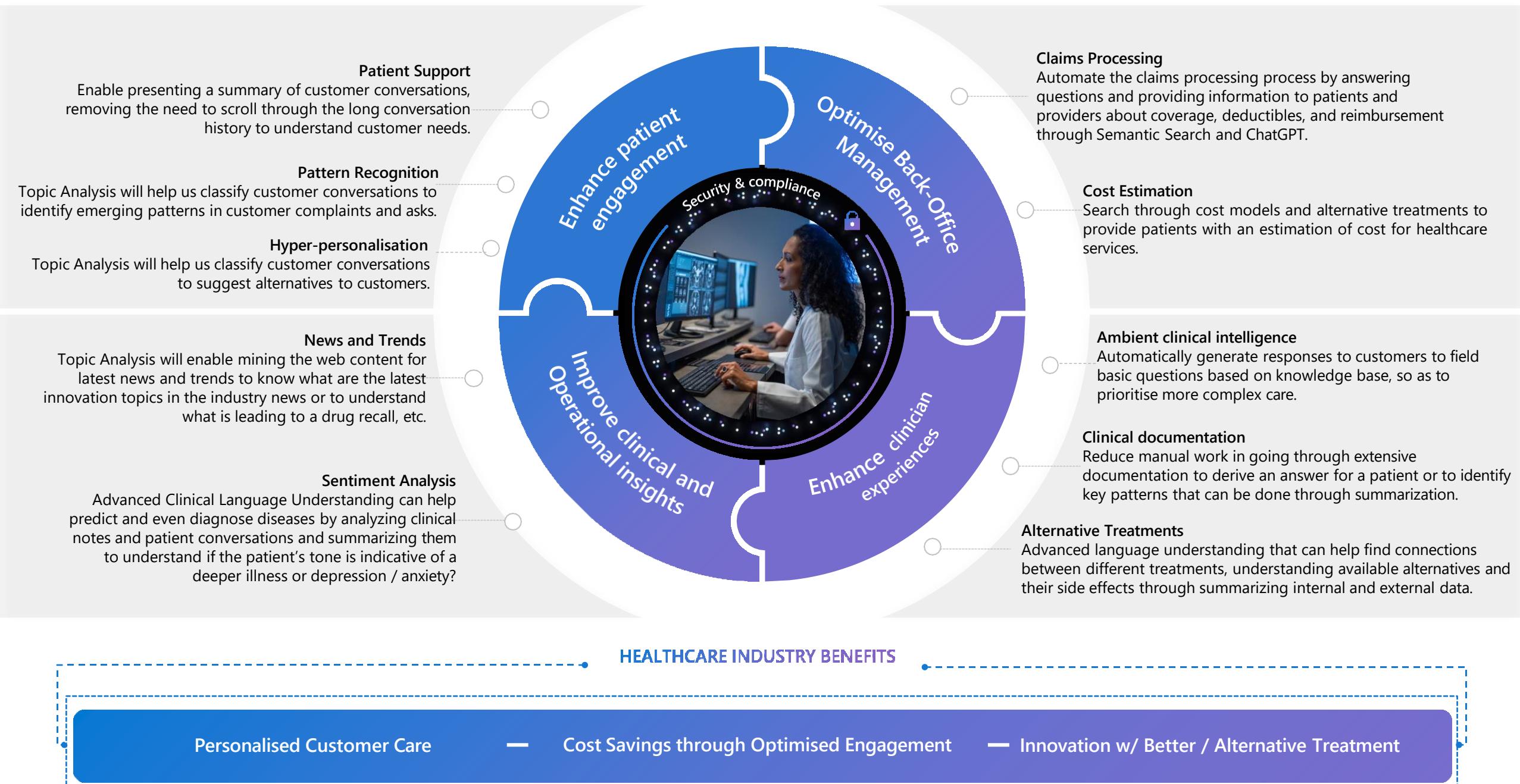
Business process automation: search through structured & unstructured documentation, generate code to query data models, content generation

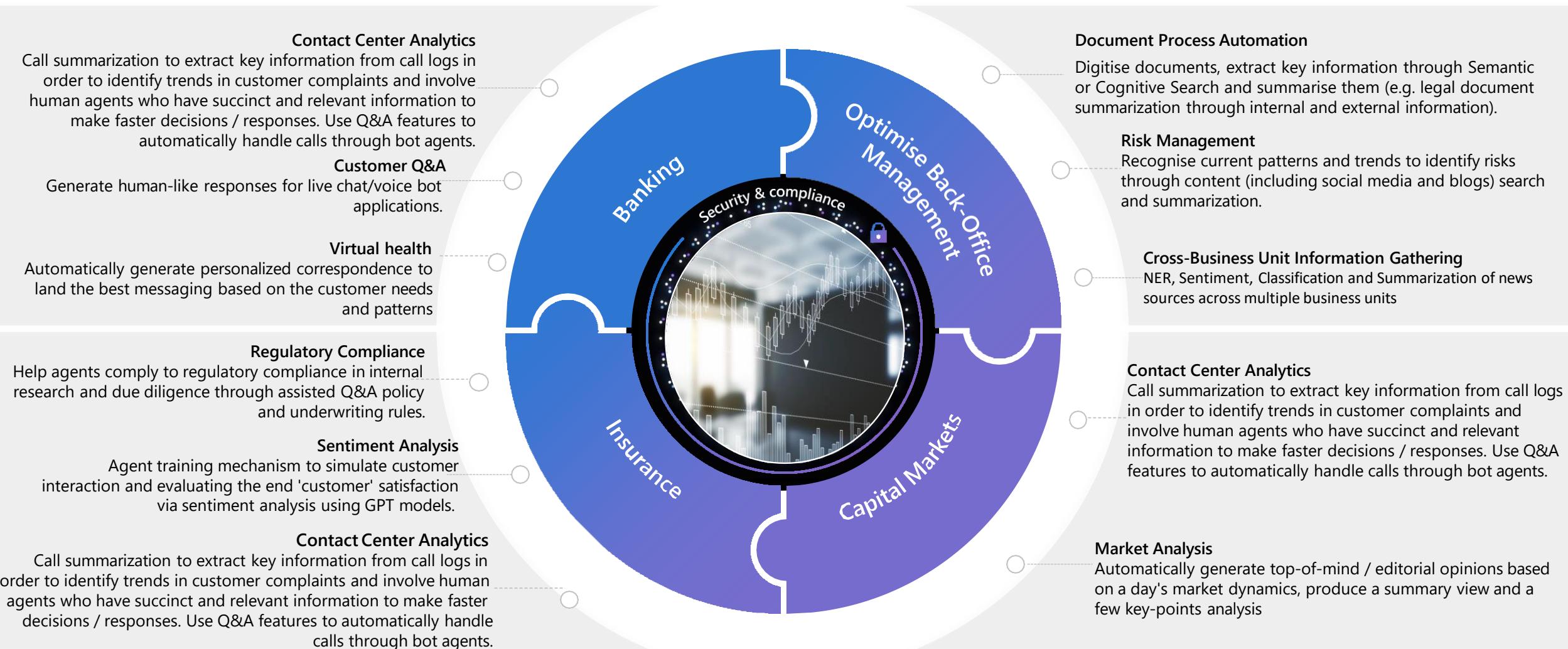
Azure OpenAI | When to use it

-  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



Azure Cognitive Services

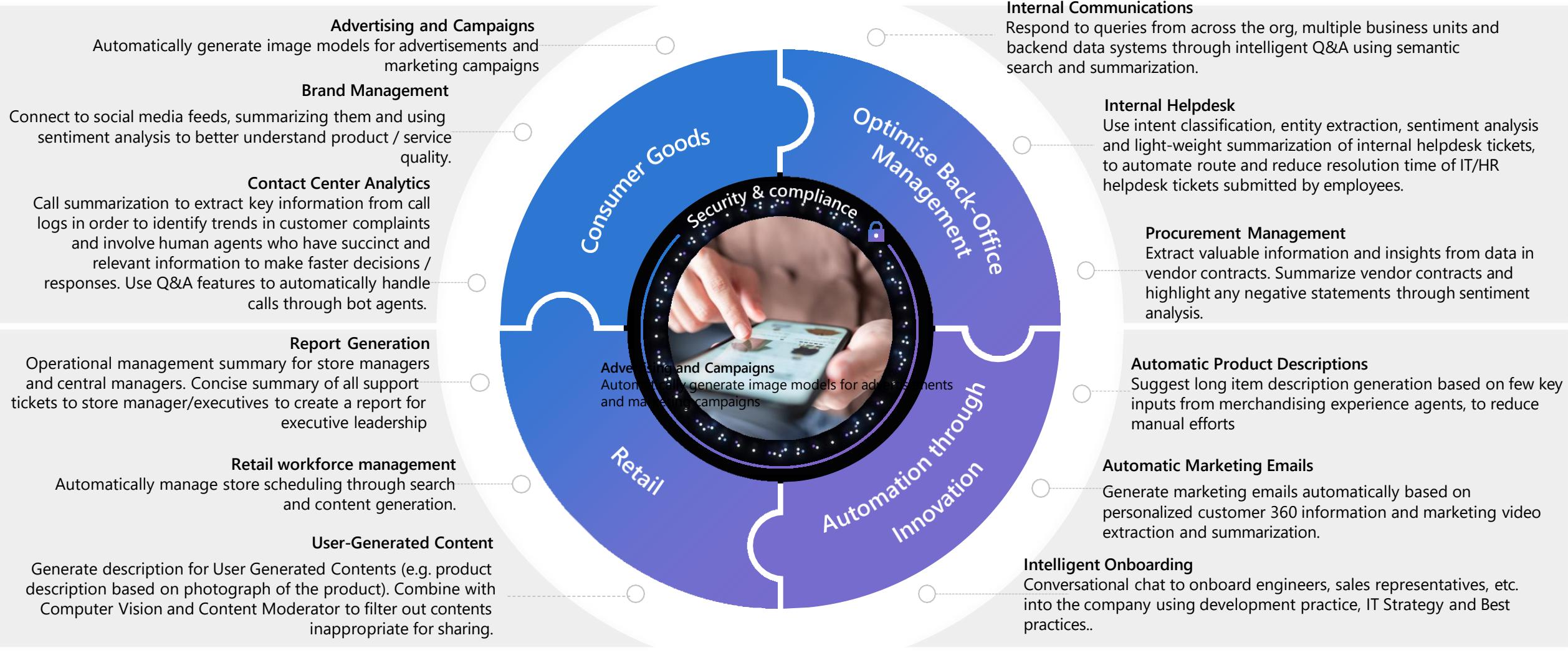




Personalised Customer Care

— Cost Savings through Optimised Engagement

— Intelligent Compliance & Risk Management

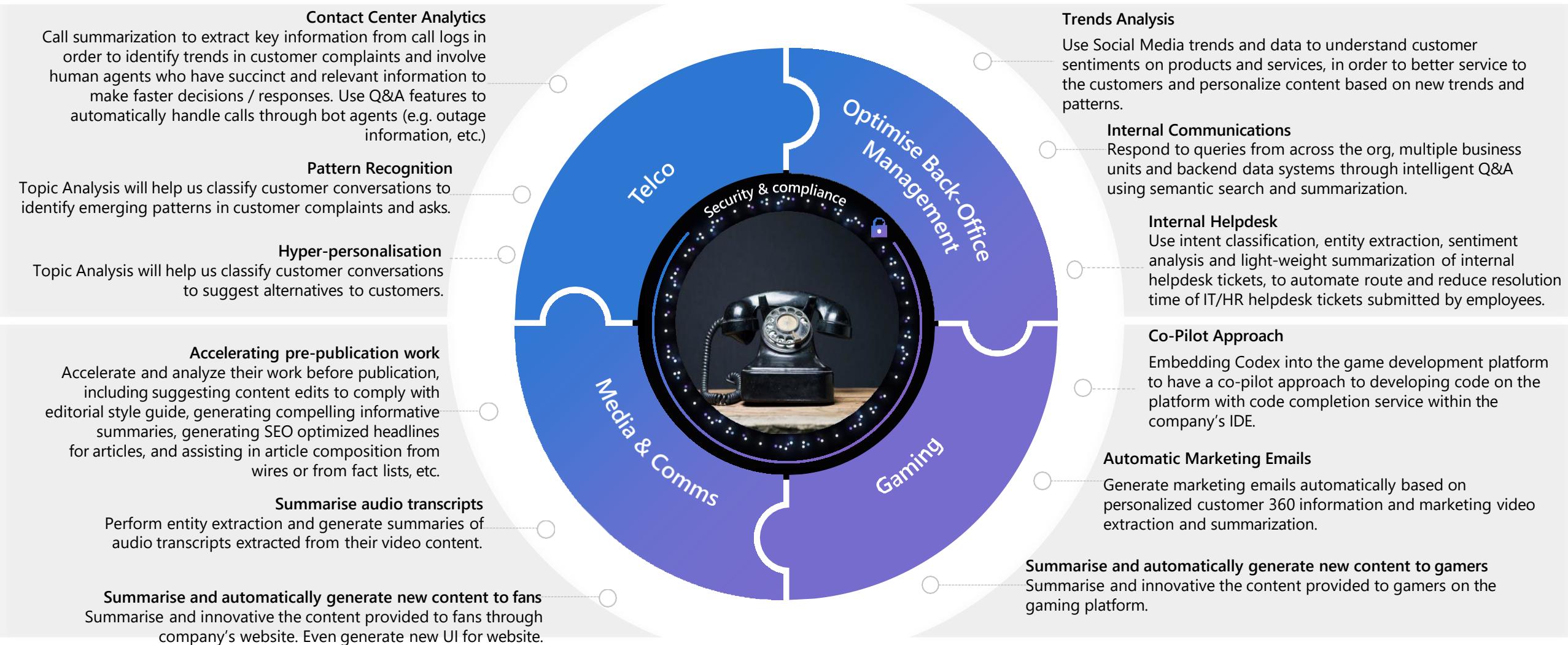


RETAIL & CPG INDUSTRY BENEFITS

Hyper-personalisation

Innovative Marketing and Campaigning

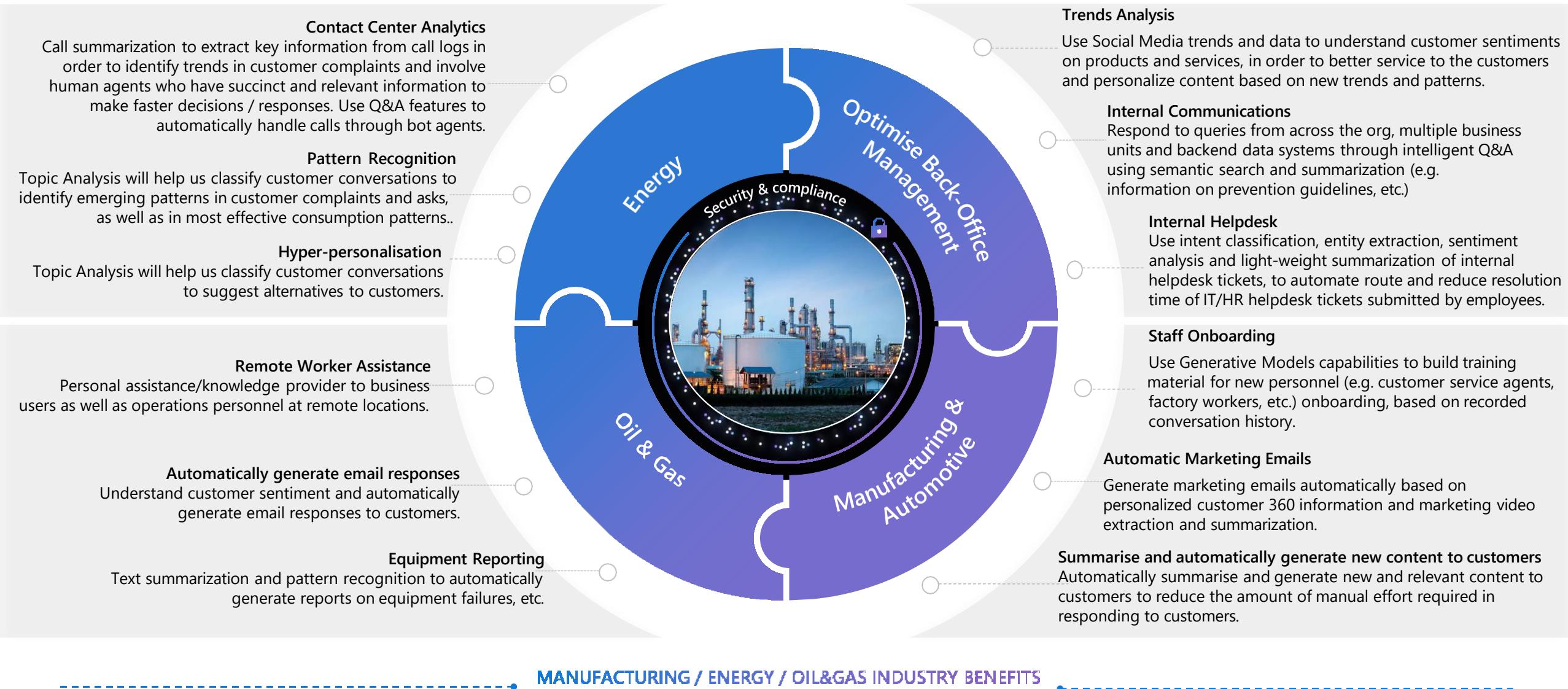
Intelligent Contact Center



Hyper-personalisation

Intelligent Contact Center

Innovative Content Creation



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/JNBJFK/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 JNBJFK 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}  
{2:1940KNABNL2HXXXXN3020}  
{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

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.

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

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



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.



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





NUWA | Infinity

A cloudy morning
on the beach with
the tide coming in

Text to image



Image to video



Mountains
with clouds



Azure OpenAI | GPT-3 Family of Models

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

Azure OpenAI | Comparing the GPT-3.5 models

Model	Description
text-davinci-002	<ul style="list-style-type: none">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	<ul style="list-style-type: none">An improvement over the text-davinci-002 model. The model is similar to its predecessor but generally more capable across all tasks.
ChatGPT model	<ul style="list-style-type: none">A model custom-tuned for working with dialogue. ChatGPT is a great model to use for conversational tasks.
(gpt-3.5-turbo)	

Model Use Cases

GPT-3

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



ChatGPT (preview)

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



GPT-4 (preview)

Built-in responsible AI to detect and mitigate harmful use



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

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

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:

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:

1. Turner is so important defensively to
2. Griner pulled way out, Hartley with

Sample response

Optimize accuracy and
latency to validate proof
of concept fast

Prompt and completion examples



Azure OpenAI
Service

Azure OpenAI GPT-4 Overview

new slide

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.

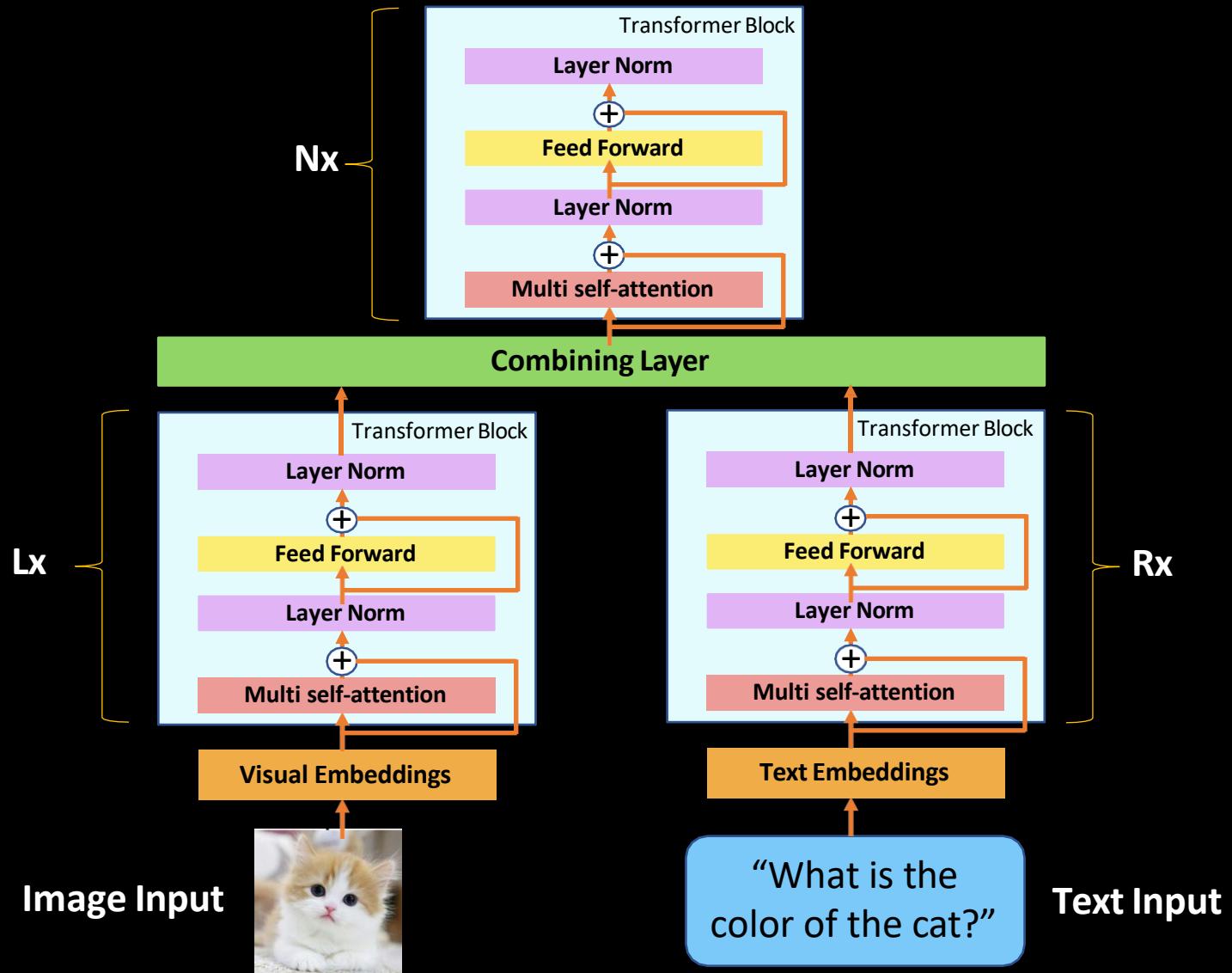
Can do everything that GPT 3.5 can do

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

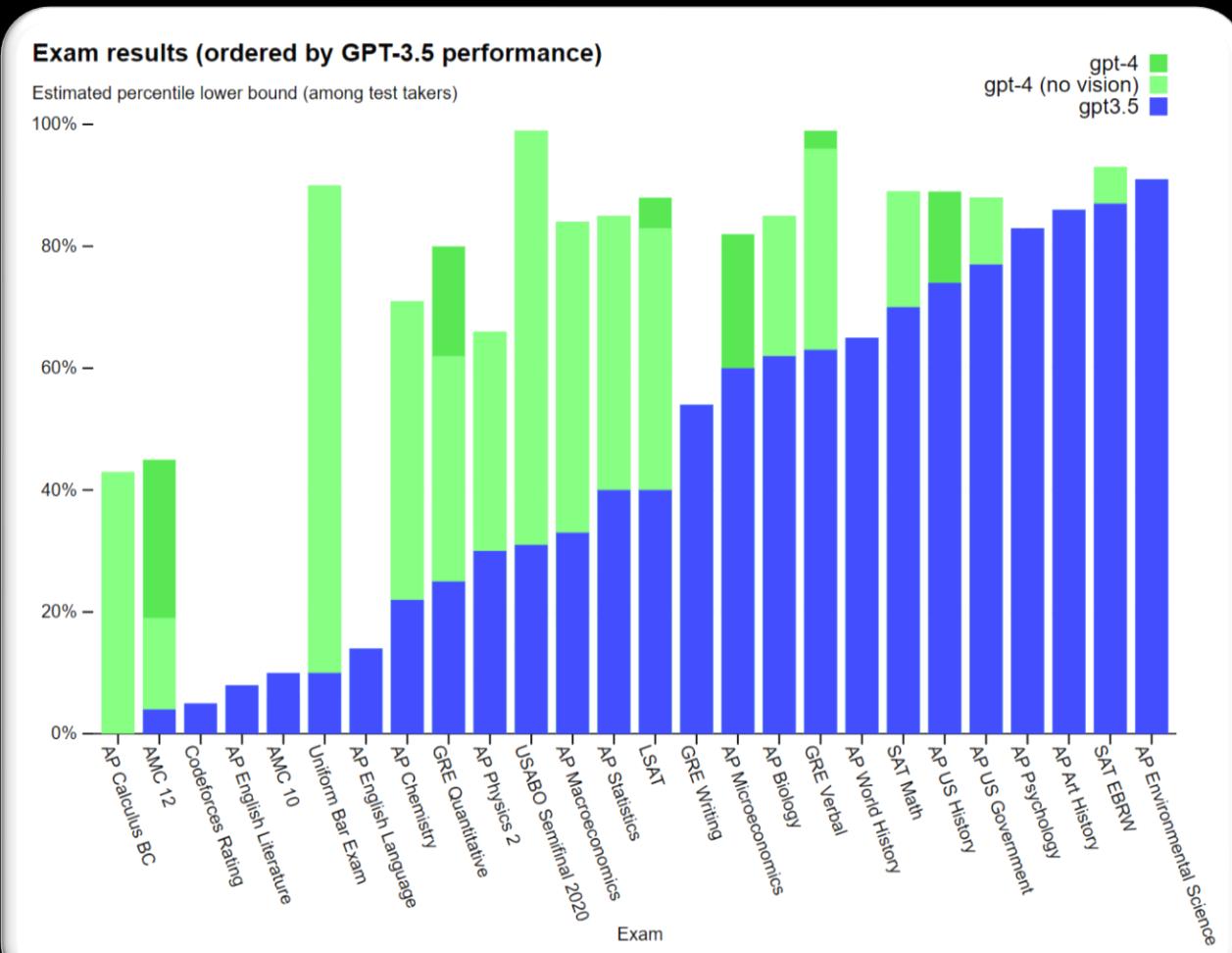
Apply here for GPT-4 access: [applying for access to GPT-4 today](#)

Multi-Modal Transformer Architecture

new slide



GPT-4 Benchmarks



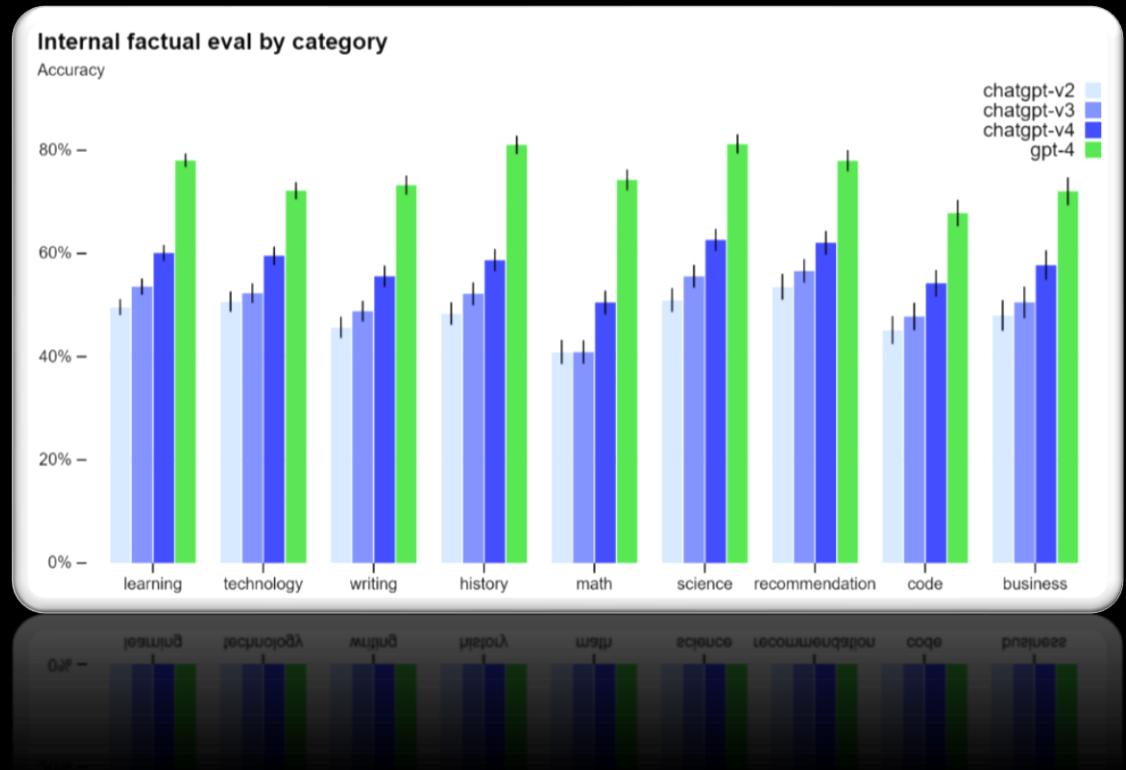
GPT-4 Limitations

OpenAI

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



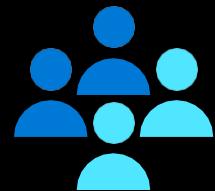
DALL•E 2



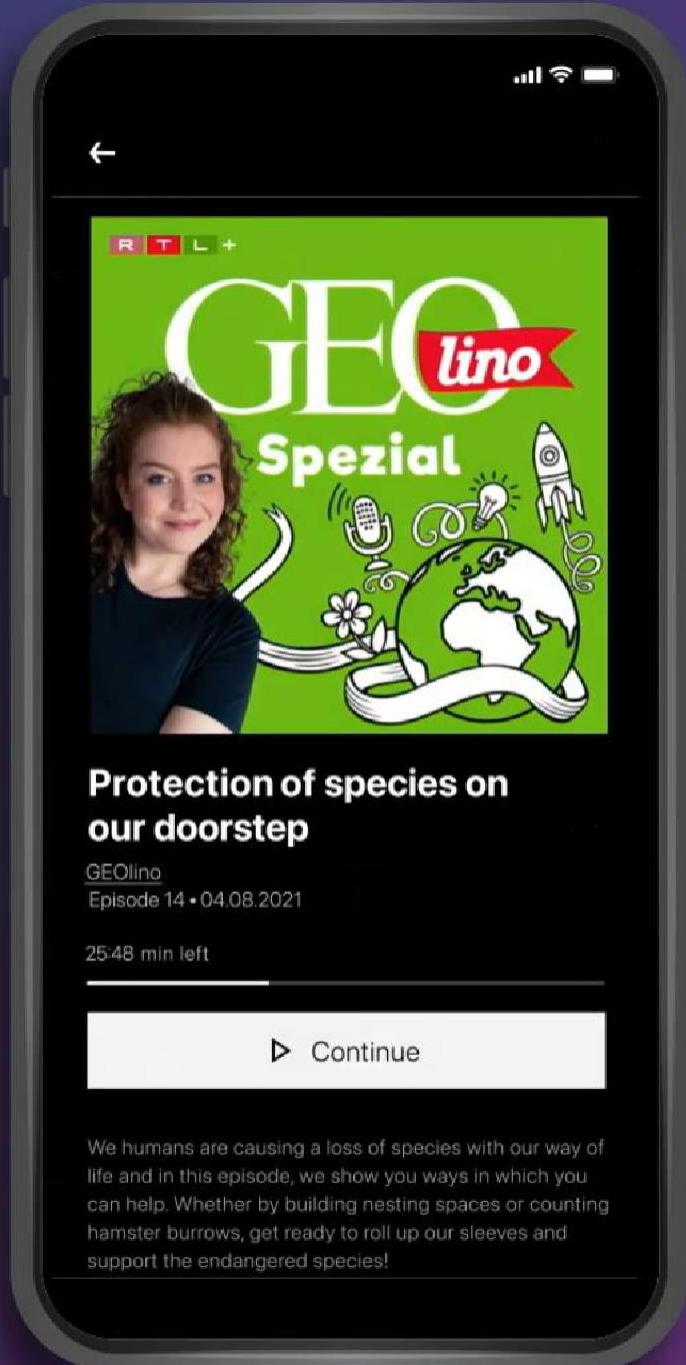
Generate 50/images per minute
(default rate, can be increased)
with simple text prompts



Accelerate designs or inspire
creative decision



Build capability into enterprise
applications through APIs and SDKs





A DTM race car like a hot rod

Microsoft Designer x + - □ ×

← → ⌂ https://designer.microsoft.com

Designer + New design My design 100% ⌂ ⌂ Share ⌂ ⌂ DS

Templates
My media
Visuals
Text
Styles

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

Cakes by Davi

CAKES BY DAVI

Cakes by Davi

CAKES BY DAVI

Continue editing Share

Design Ideas

The ChatGPT model

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

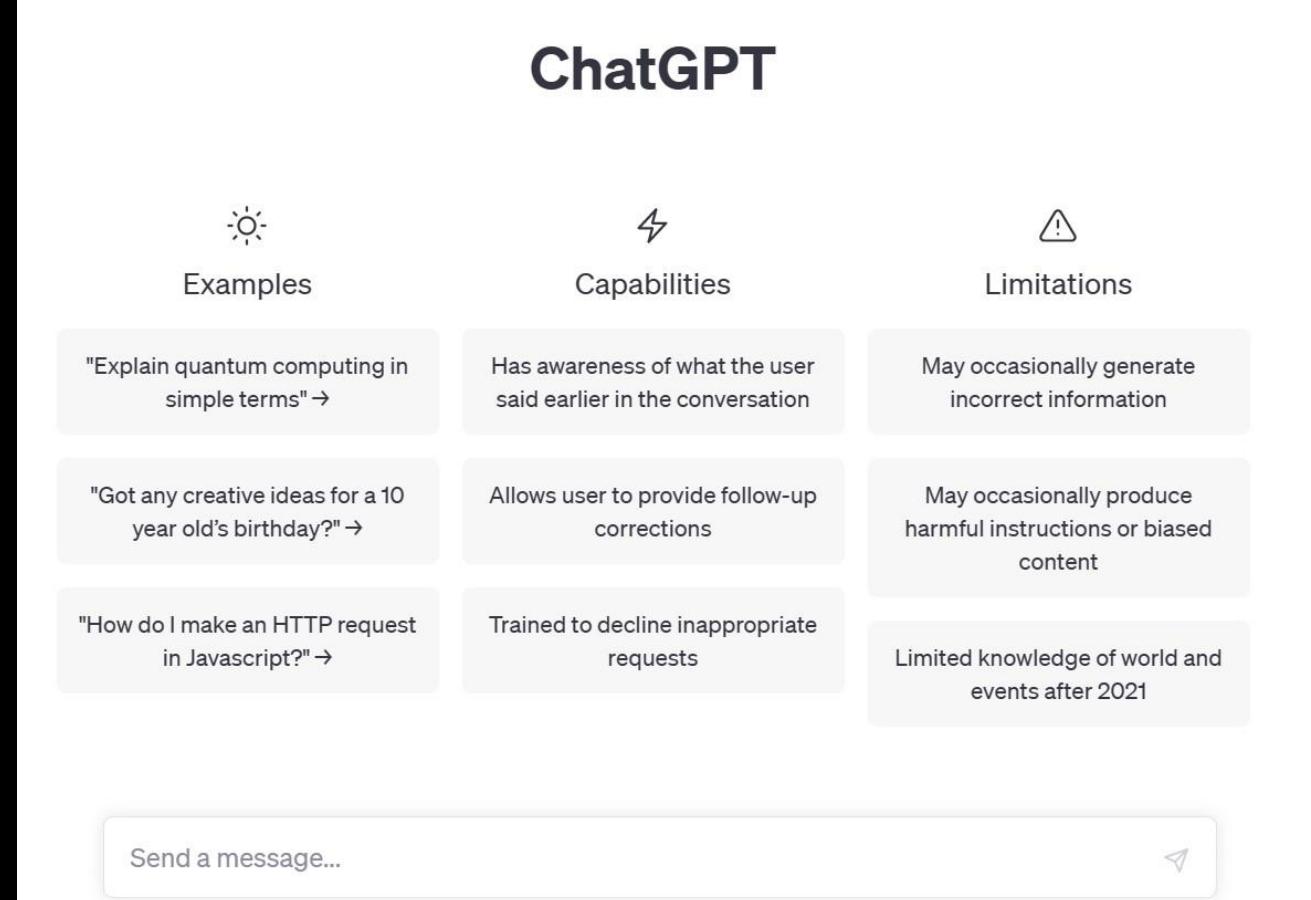
The conversational nature of the model makes it easier to interact with and to take advantage of the full power of its capabilities.

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-35-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-35-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

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

Category	Completion	ChatCompletion
Dialog Form	Single-Turn Dialogue (System generates a response based on a single input prompt.)	Multi-Turn Dialogue (System records dialogue history and generates a response.)
Model	GPT-3.5 Turbo Instruct (previously text-davinci-003)	GPT-3.5 Turbo GPT-4
Purpose	More suitable for single-prompt tasks.	Suitable for interactive dialogue applications.
API Setup	<code>openai.Completion.create</code>	<code>openai.ChatCompletion.create</code>
Advanced Settings	Prompt design and engineering	System message and System message template design

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.

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

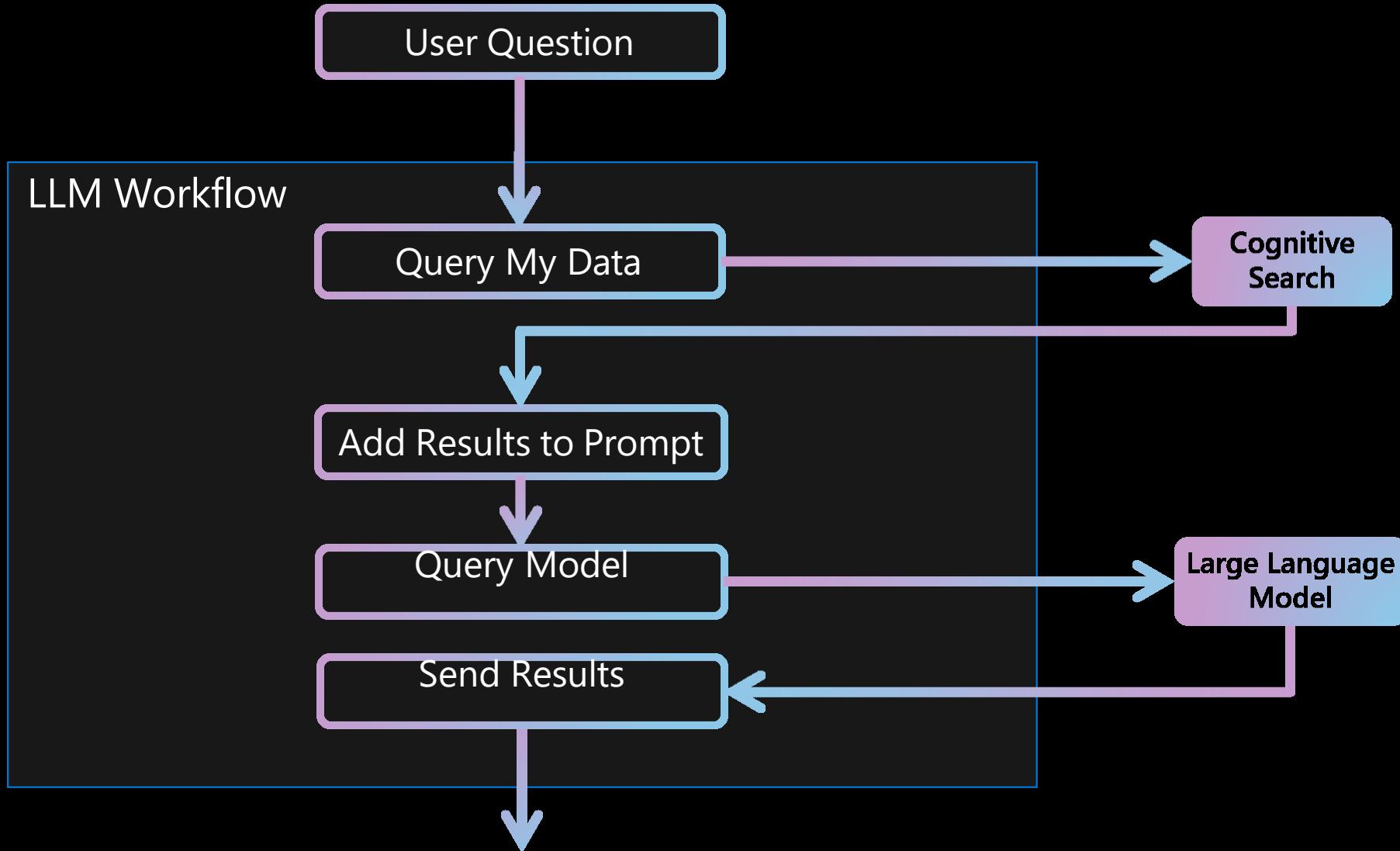
Privacy & Security

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

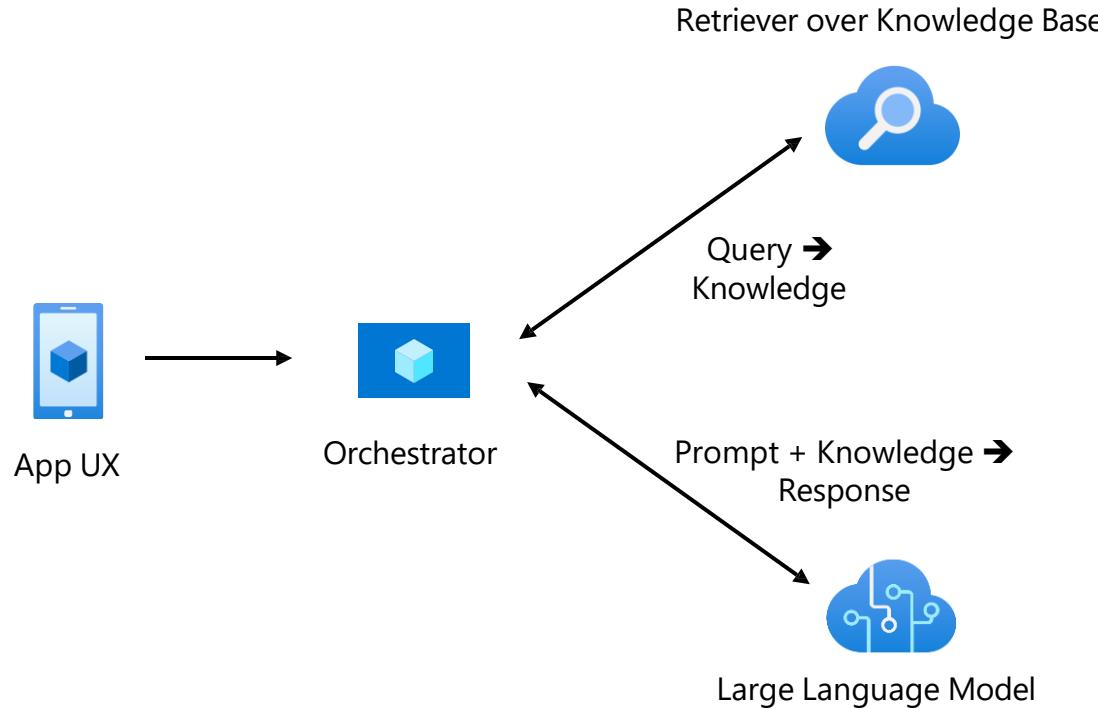
Inclusiveness

- 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

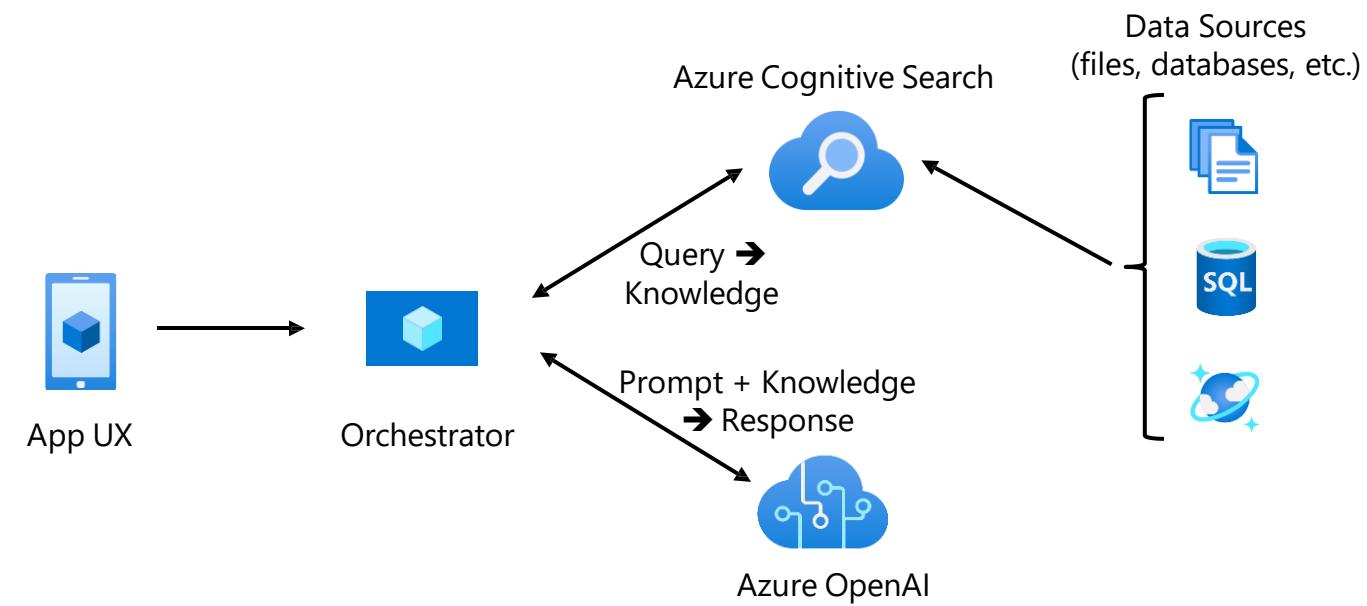
Plugins for retrieval, symbolic math,
app integration, etc.
e.g., plugins for OpenAI ChatGPT

Retrievers: Externalizing Knowledge

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

Azure Cognitive Search

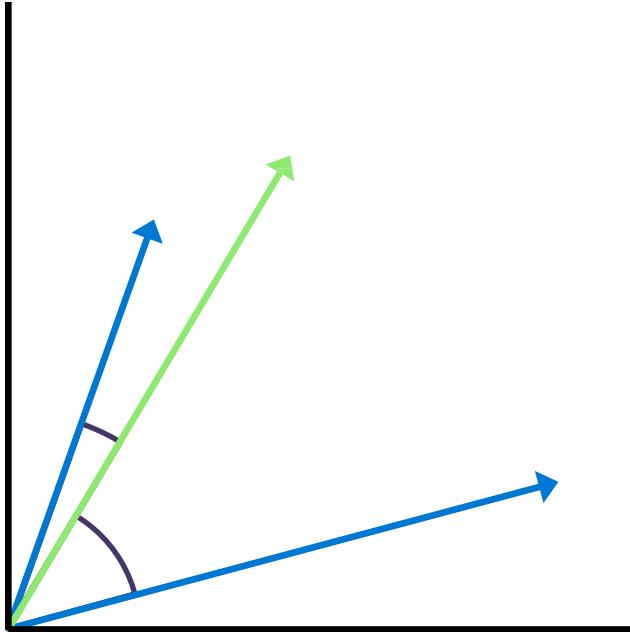
- Azure's complete retrieval solution
- Data ingestion, enterprise-grade security, partitioning and replication for scaling, support for 50+ written languages, and more



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)

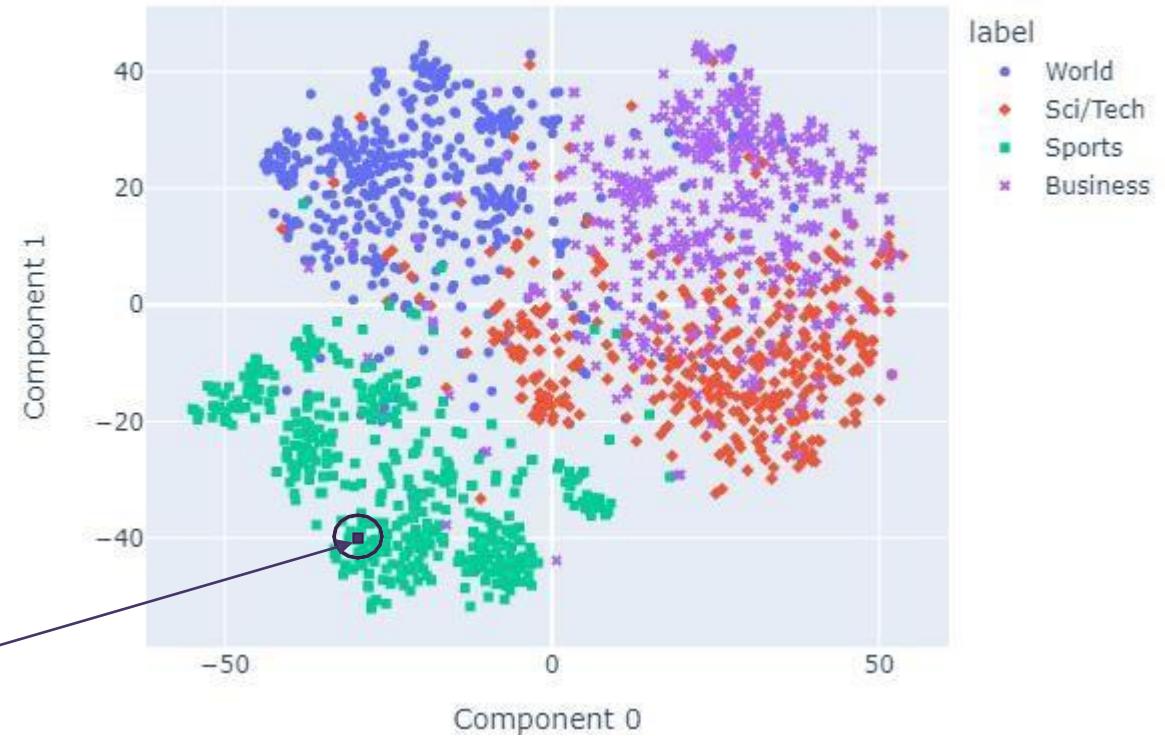
- Learned such that “close” vectors represent items with similar meaning
- May encode words, sentences, images, audio, etc.
 - Some map multiple media types into the same space
- Azure OpenAI embeddings API, OSS embeddings (e.g., SBERT, CLIP)

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"

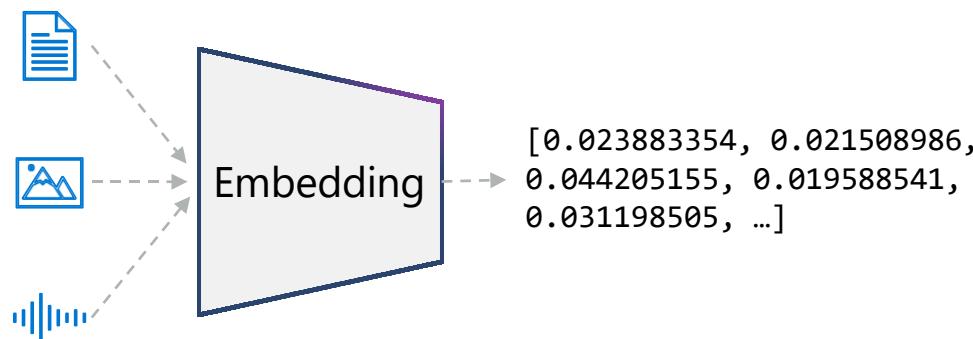
t-SNE components of article descriptions



Vector-based Retrieval

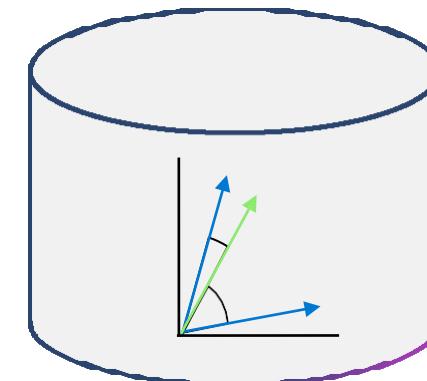
Encoding (vectorizing)

- 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

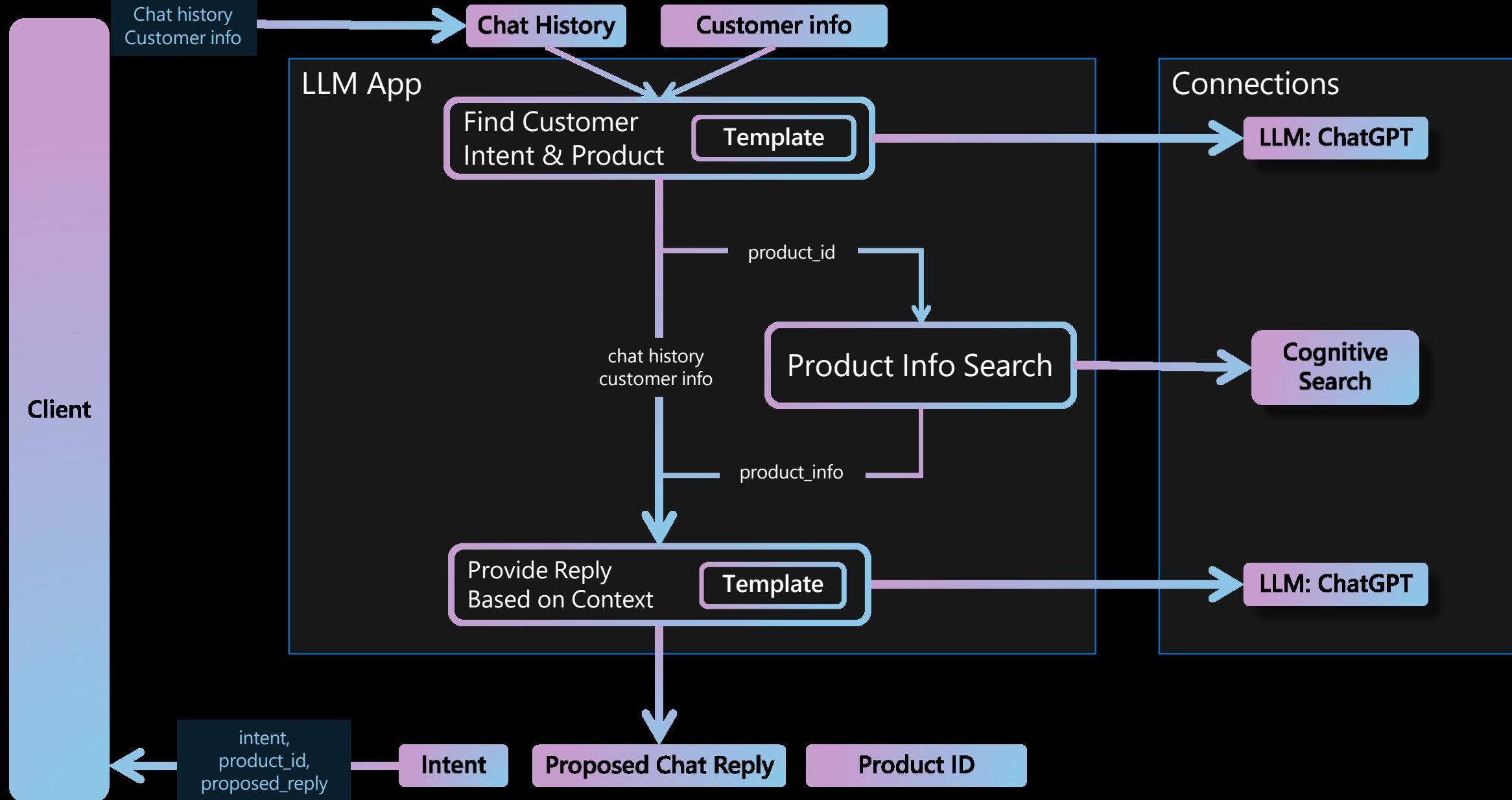


Documents

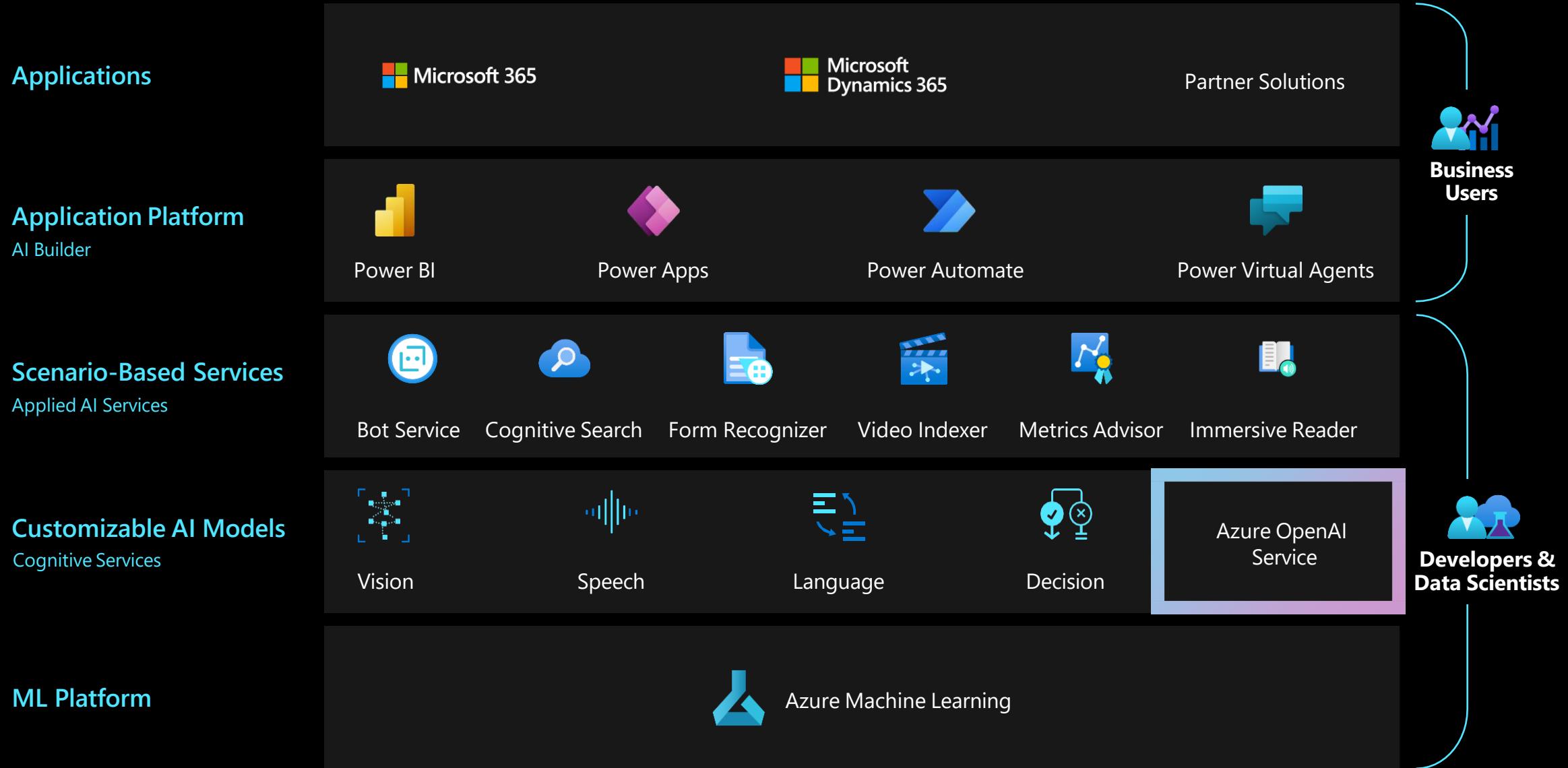
- Use vector or hybrid search
- Use Azure OpenAI embeddings or bring your own
- Deeply integrate with Azure
- Scale with replication and partitioning
- Build generative AI apps and retrieval plugins

Sign up today

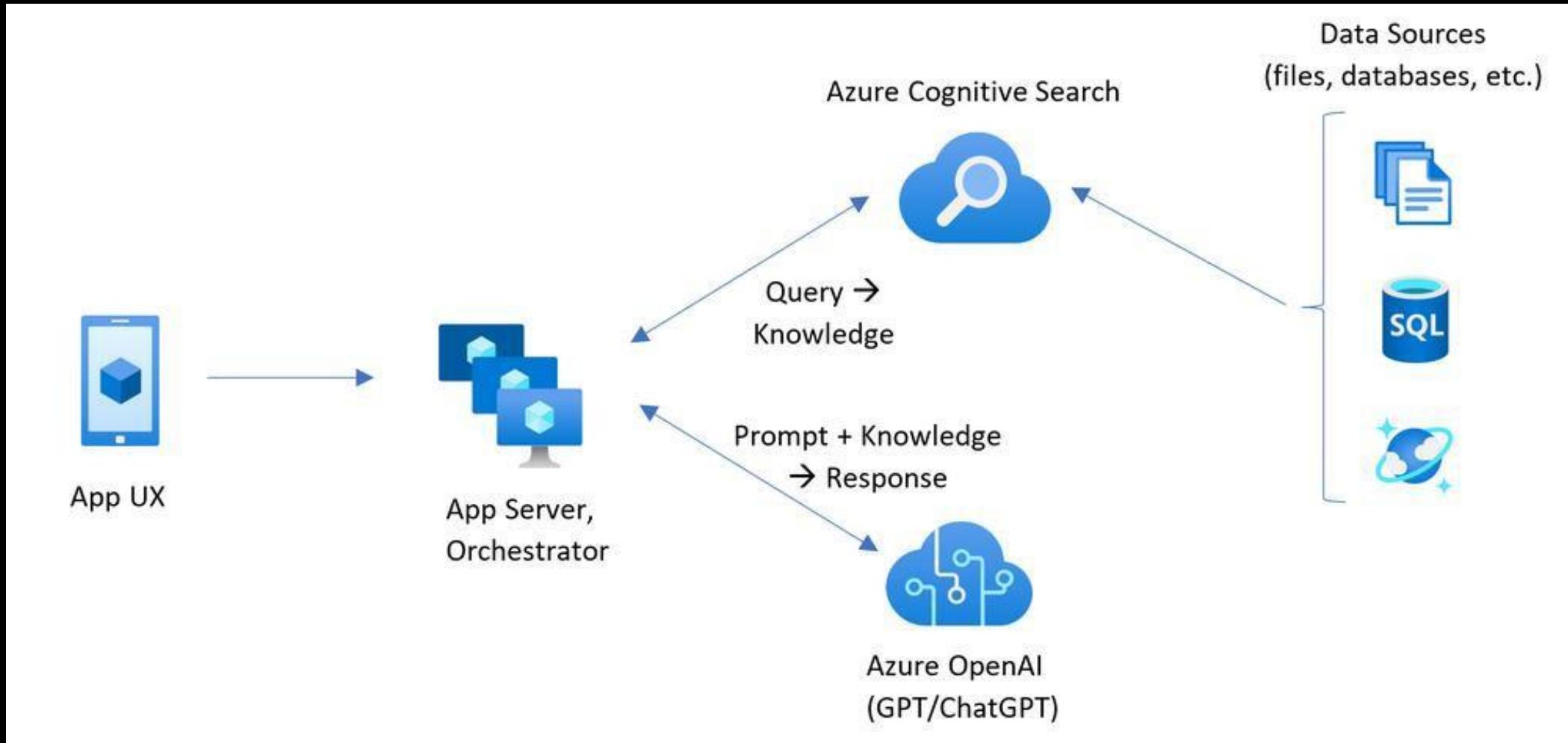
<https://aka.ms/VectorSearchSignUp>



Azure AI

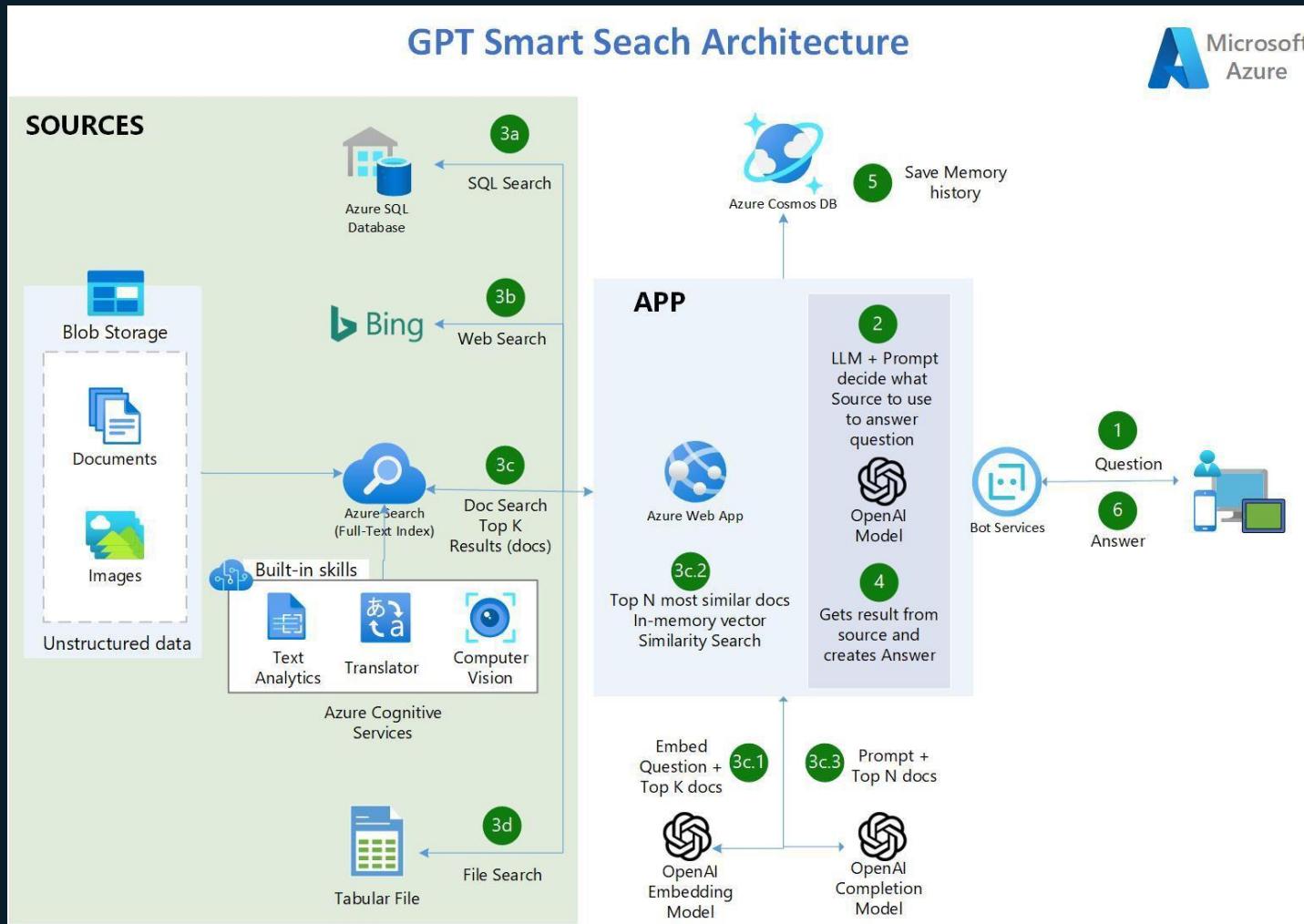


Q&A Search App



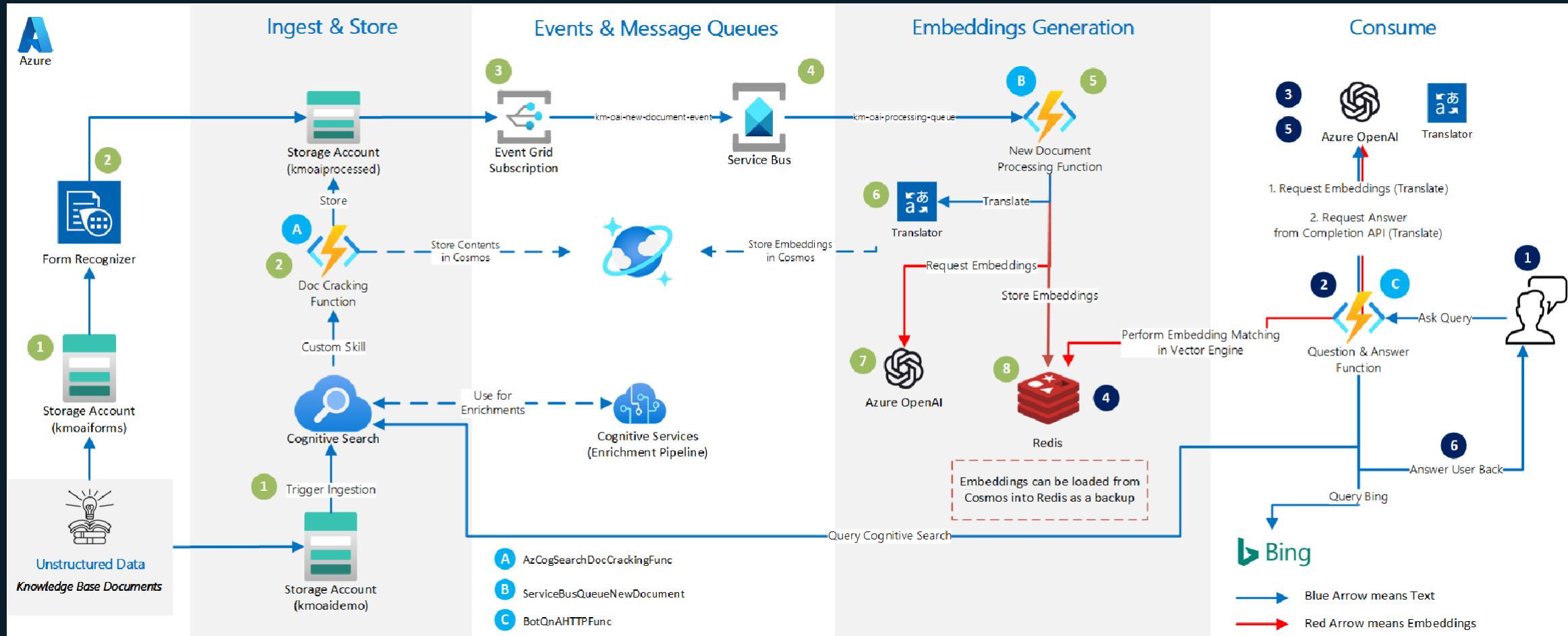
<https://github.com/Azure-Samples/azure-search-openai-demo>

Q&A Chatbot architecture



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

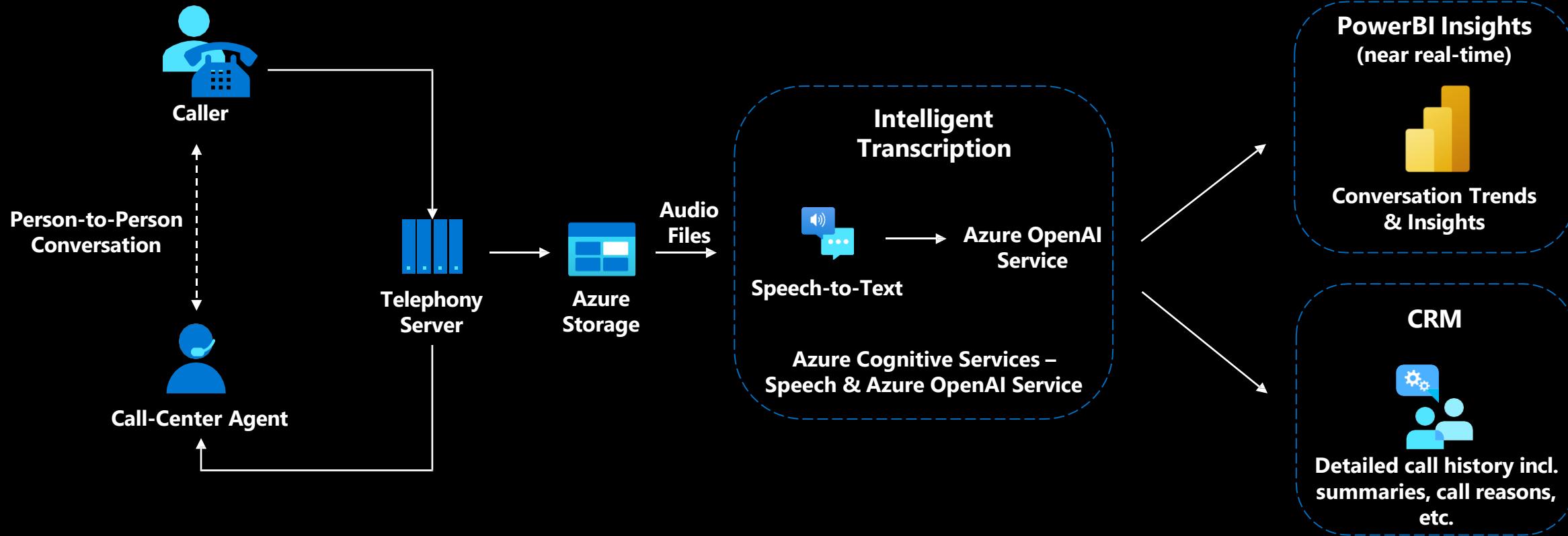
Q&A PoC architecture



<https://github.com/samelhousseini/km-openai/>

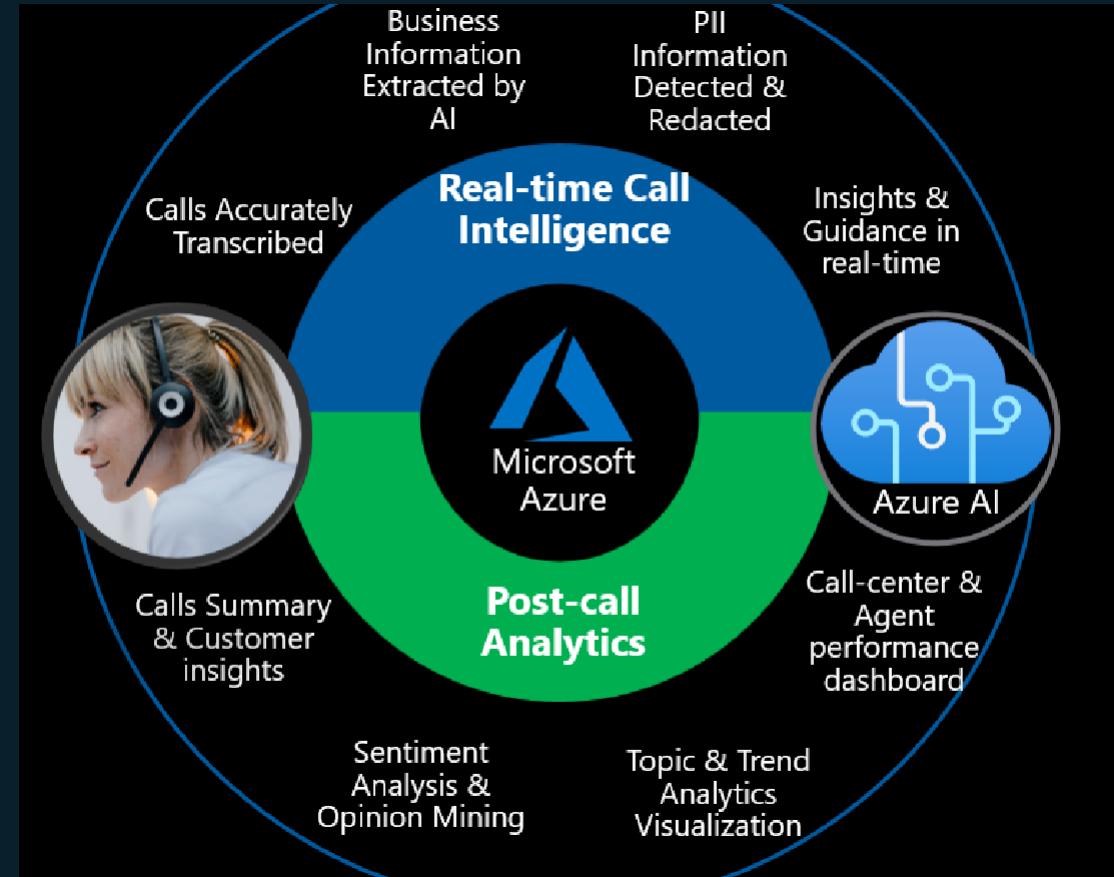
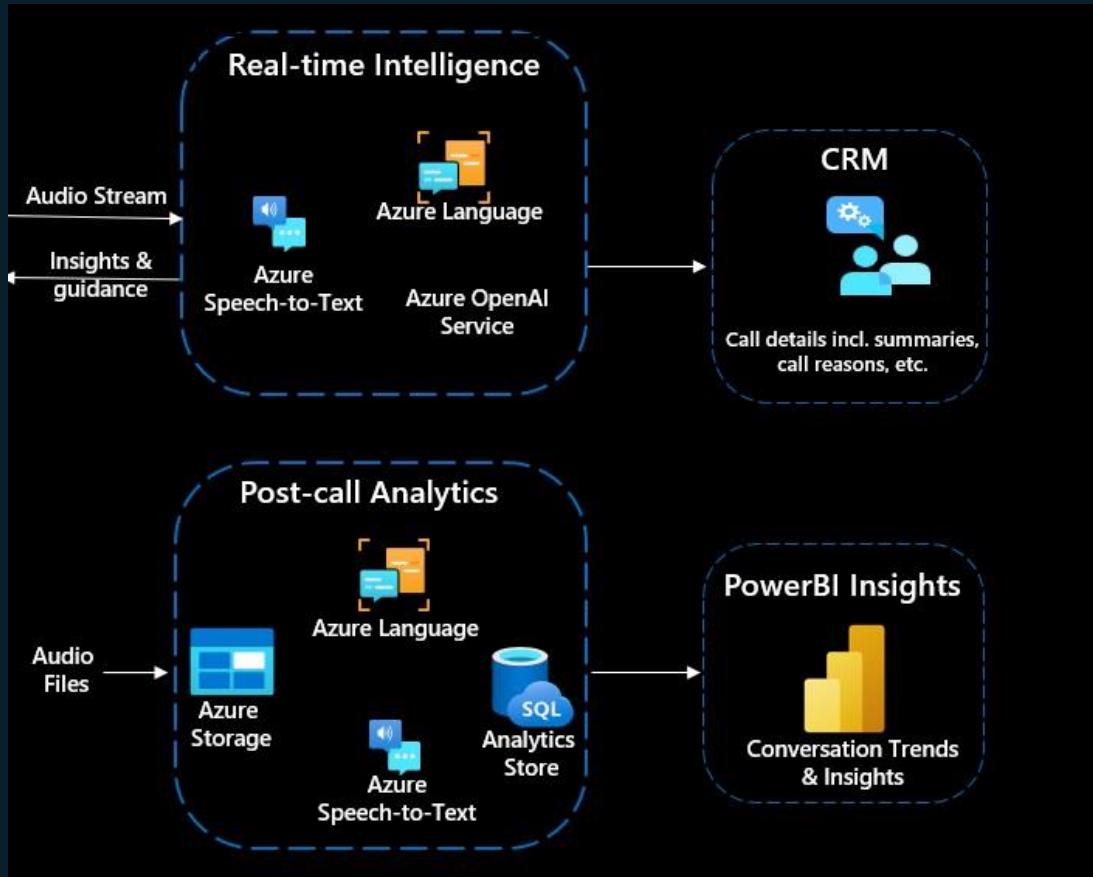
Contact Center Analytics using Speech API & Azure OpenAI Service

Extract rich insights from call transcripts



Call Center intelligence

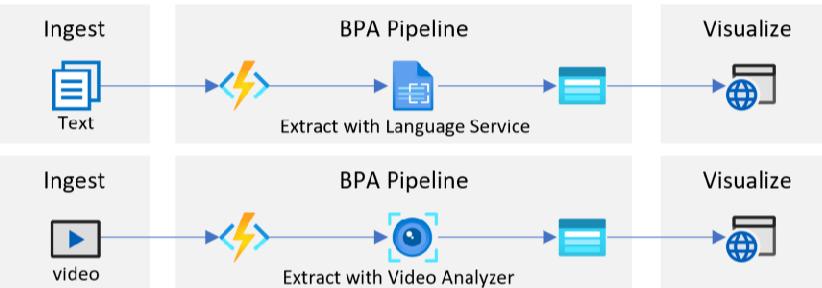
<https://github.com/amulchapla/AI-Powered-Call-Center-Intelligence>



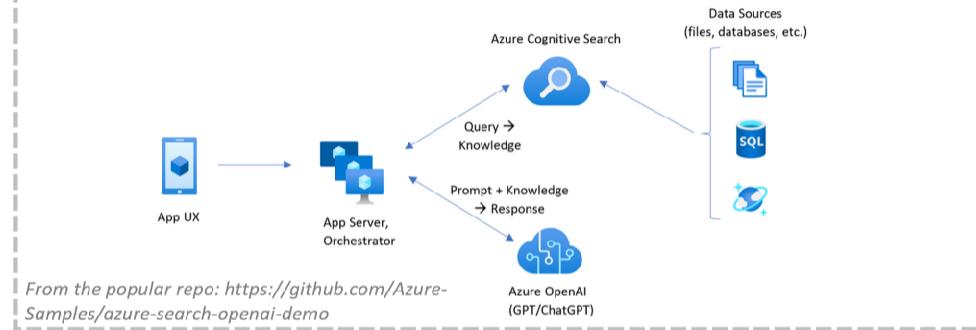
Business Process Automation accelerator

Sample Pipelines

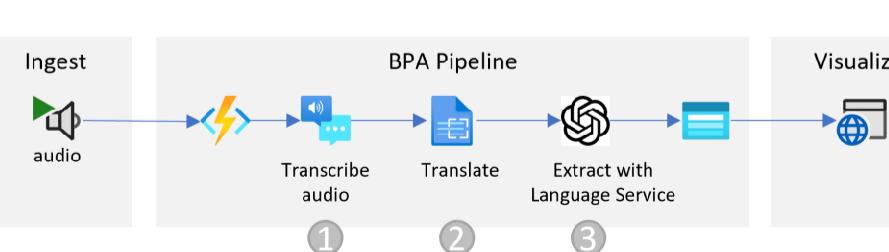
Single Service Pipeline Examples



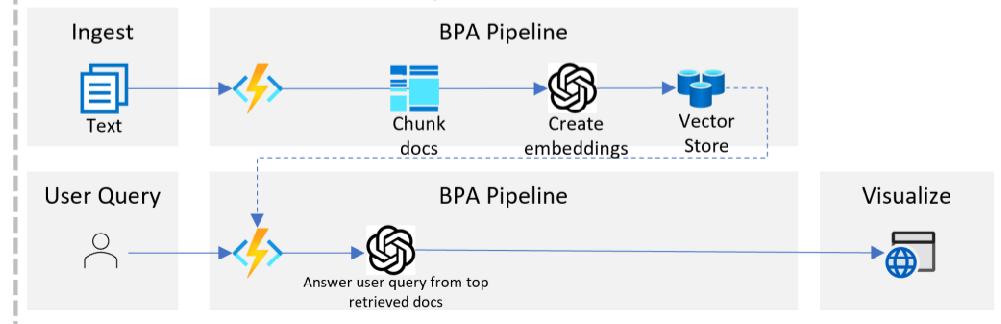
ChatGPT + Enterprise data with Azure OpenAI & Cognitive Search



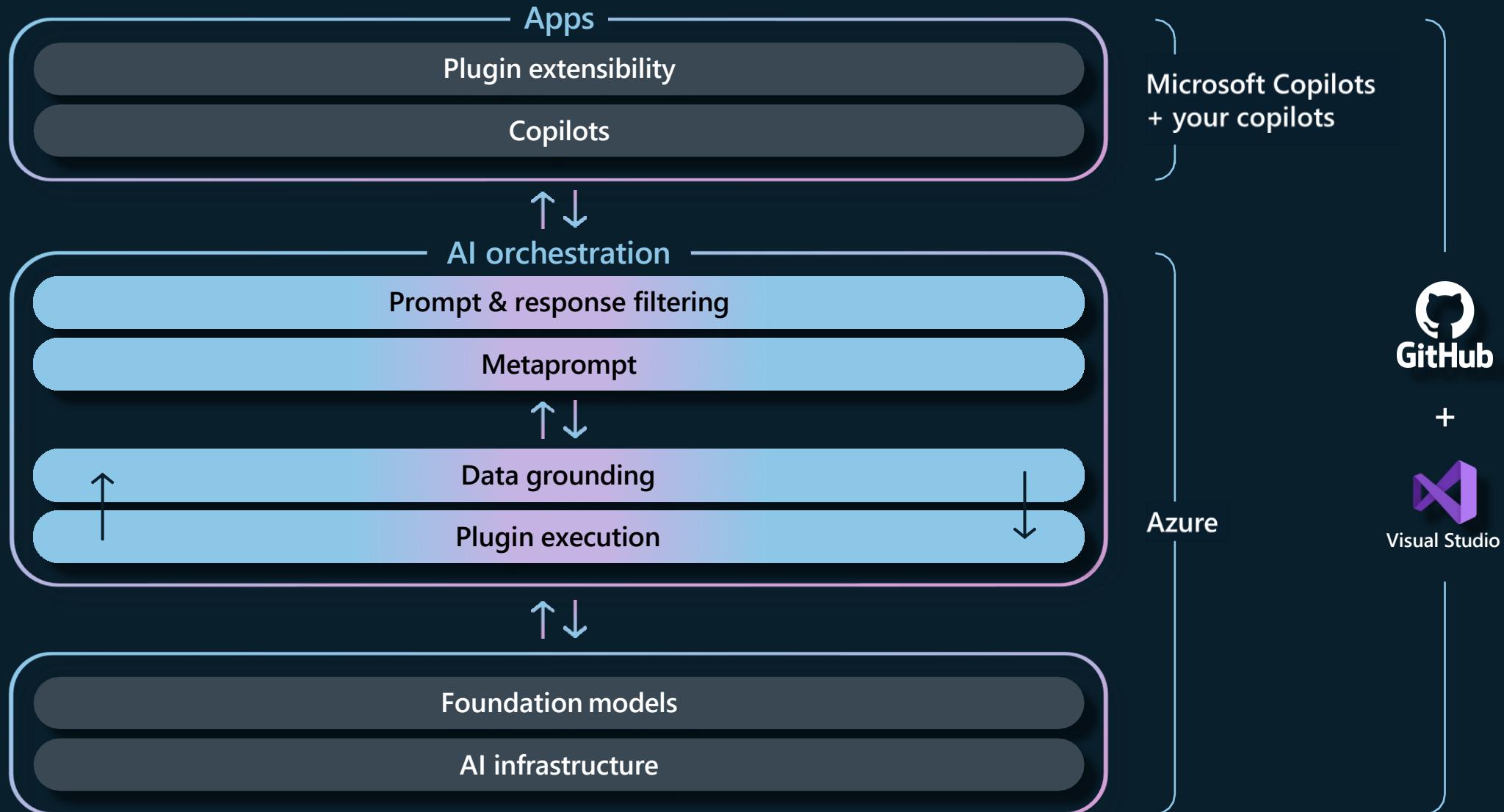
Multiple Services with Multiple Domains (e.g. speech + text)



Vector Store Pipeline (Conversational UI)



Copilot stack



Generative AI Applications

Azure Machine Learning

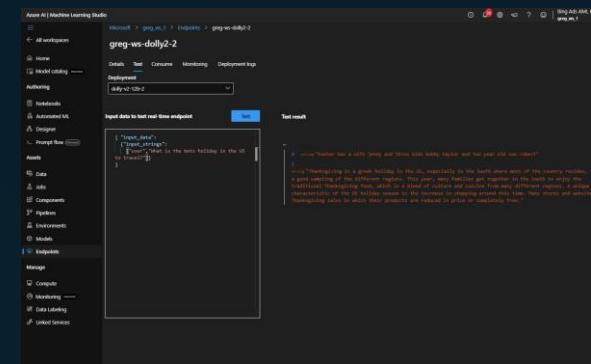
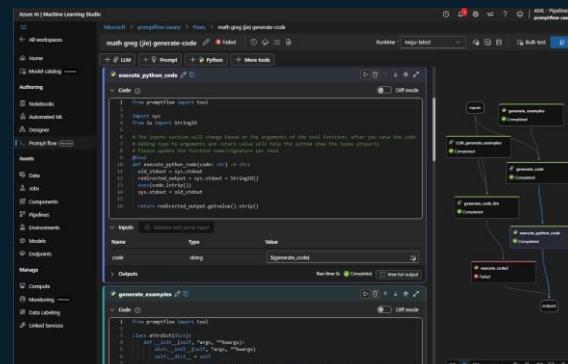
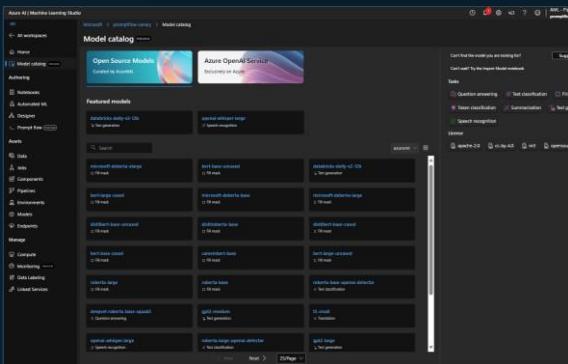
Native OSS
Model Catalog

Prompt Eng/Eval
Prompt flow

Responsible AI
Azure Content Safety

High Scale Gen AI
App Deployment

Gen AI Model
Monitoring

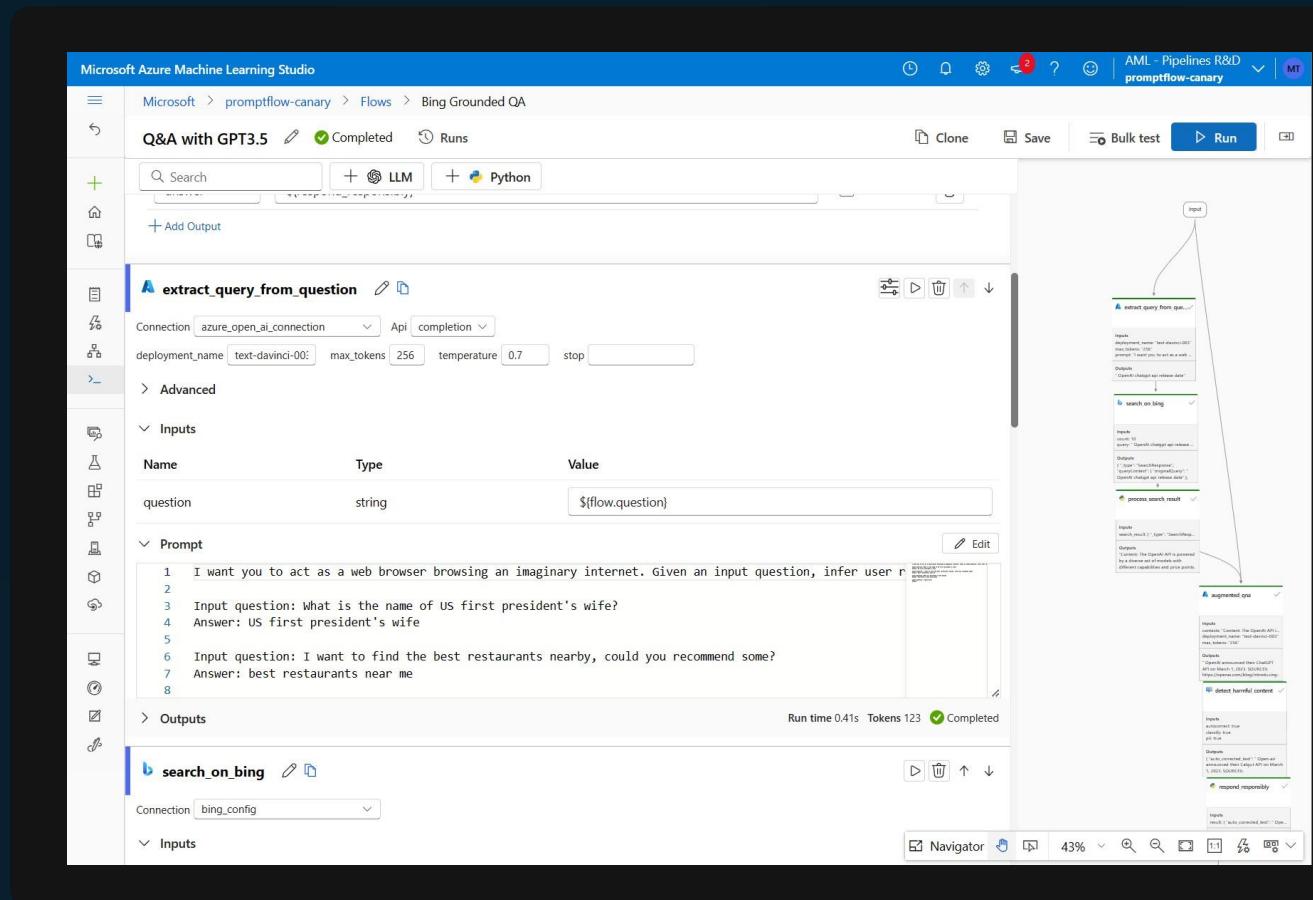


Announcing

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



Start building plugins today

[https://
github.com/Azure-Samples/openai-plugin-fastapi](https://github.com/Azure-Samples/openai-plugin-fastapi)

The screenshot shows a GitHub repository page with the following details:

- Repository Name:** Azure-Samples / openai-plugin-fastapi (Private)
- Code tab:** Selected. Shows 3 branches and 0 tags.
- Commits:** A list of 16 commits by **digitalrajd**, merged from **binderjoe/pa...**. The commits are:
 - .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)
- About:** No description, website, or topics provided.
- Readme:** Readme
- Licenses:** MIT license
- Code of conduct:** Code of conduct
- Stars:** 0 stars
- Watching:** 10 watching
- Forks:** 2 forks
- Releases:** No releases published
- Packages:** No packages published

Project Florence – Large foundation model



Image Analysis APIs

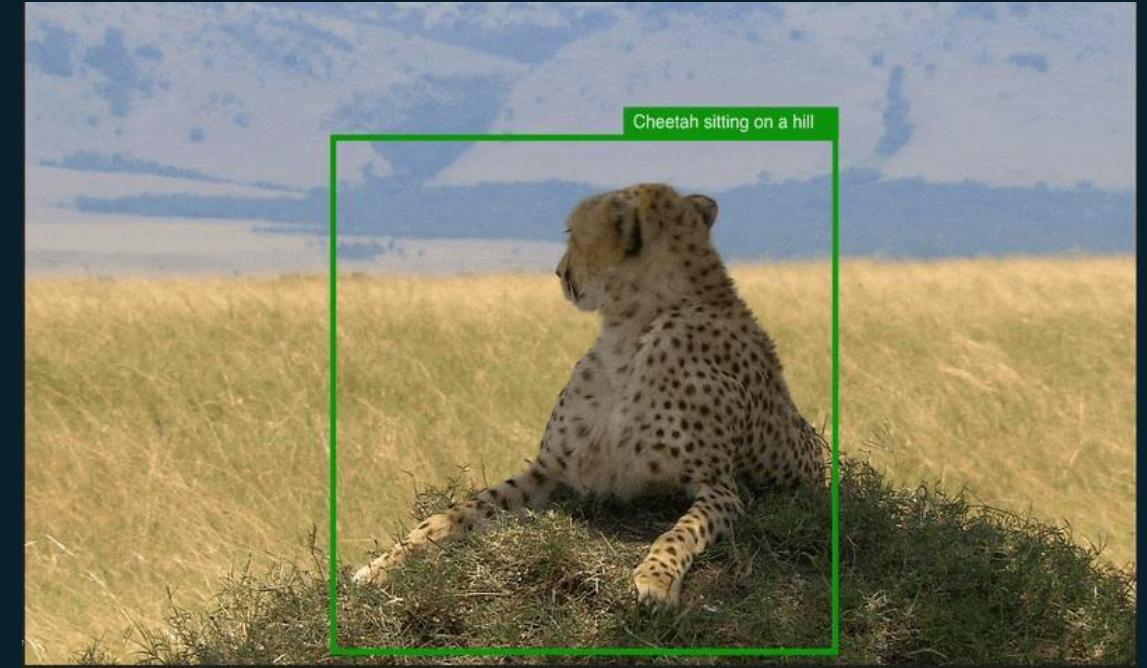
REST API

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

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



Image captions



Dense captions

Image retrieval APIs

Vision Studio > Search photos with natural language

Search photos with natural language PREVIEW

X

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

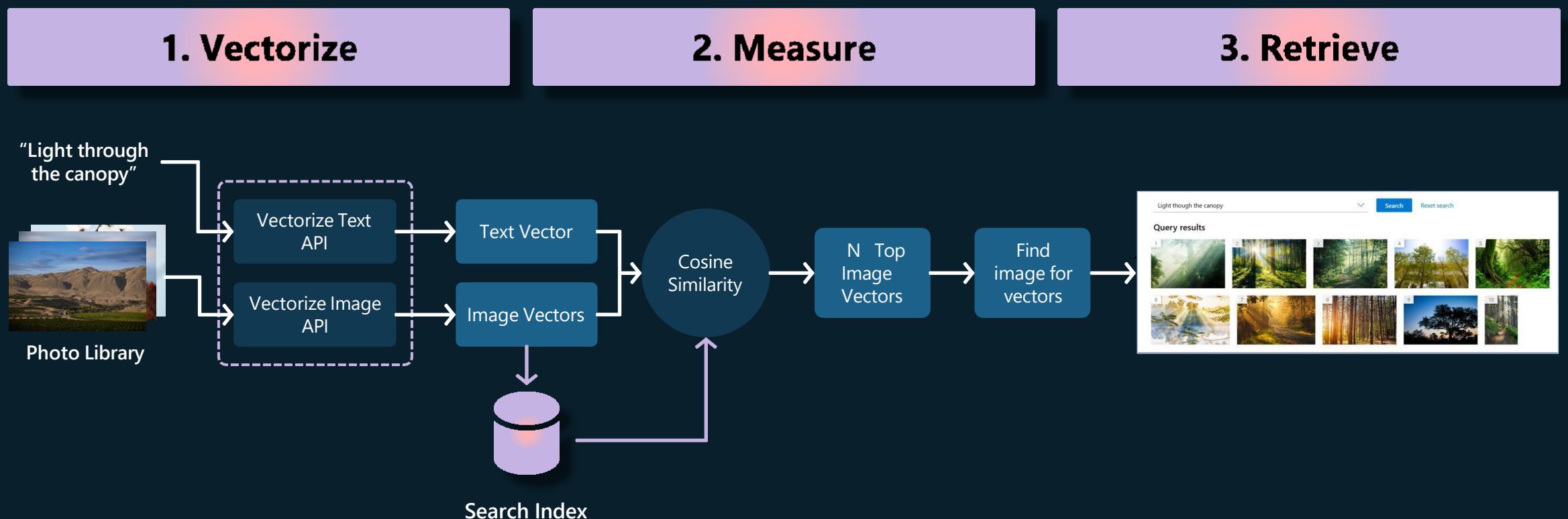


Image retrieval APIs cont.

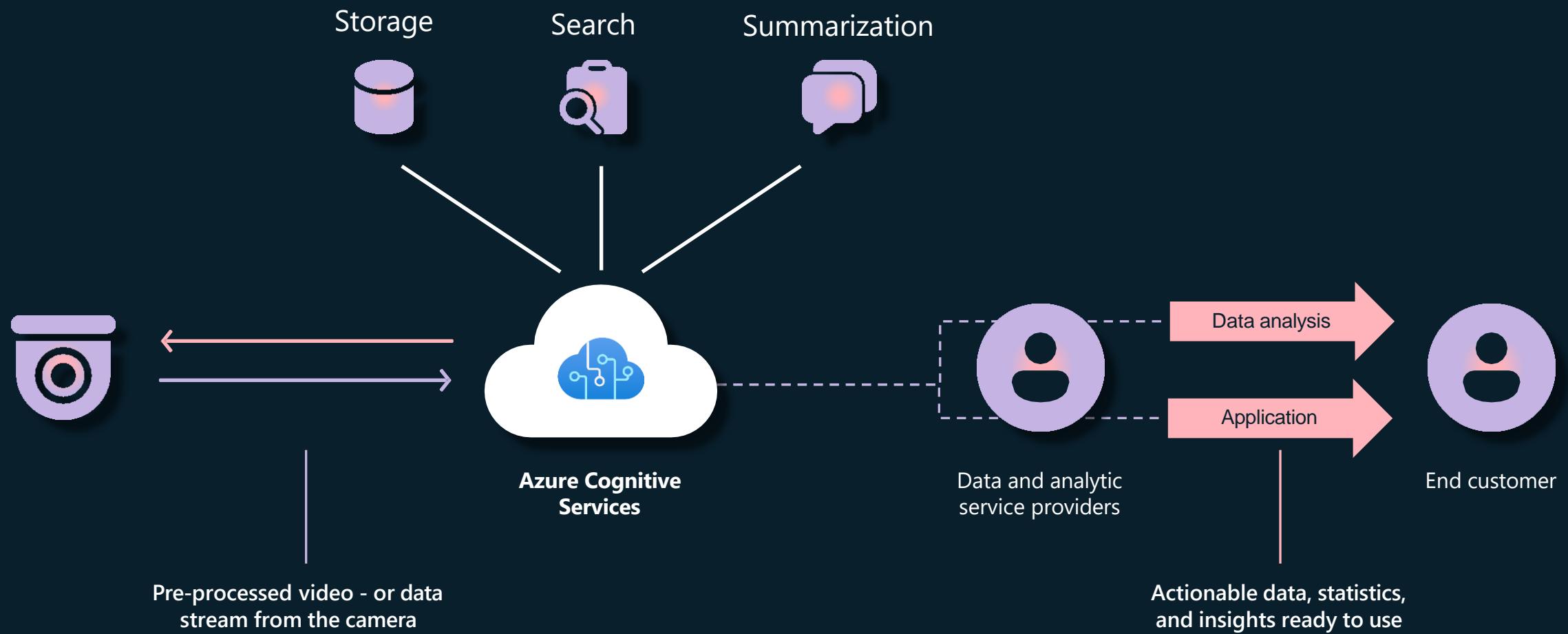
REST API

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

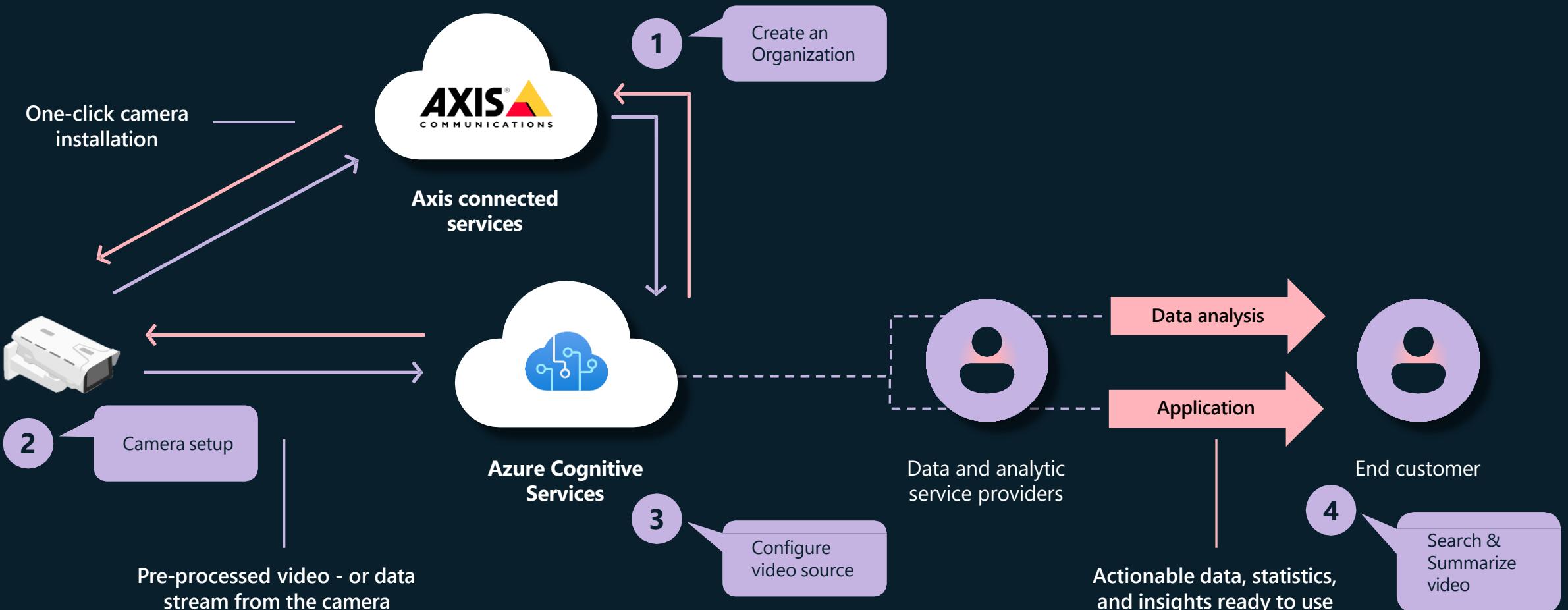
<https://<endpoint>/computervision/retrieval:vectorizeText>



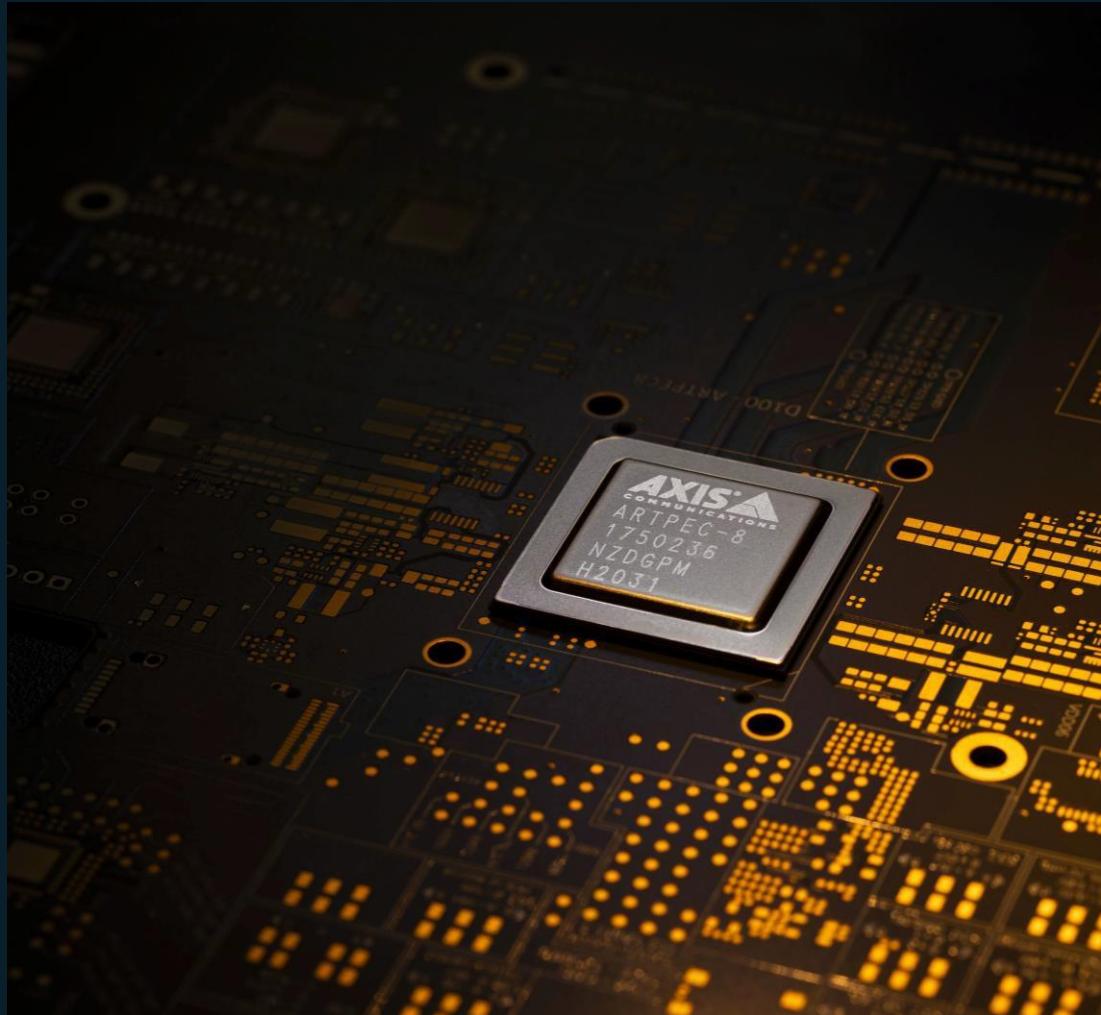
Video Frame Locator & Summarization



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

Q&A

Azure GPT

| Ask the team anything

