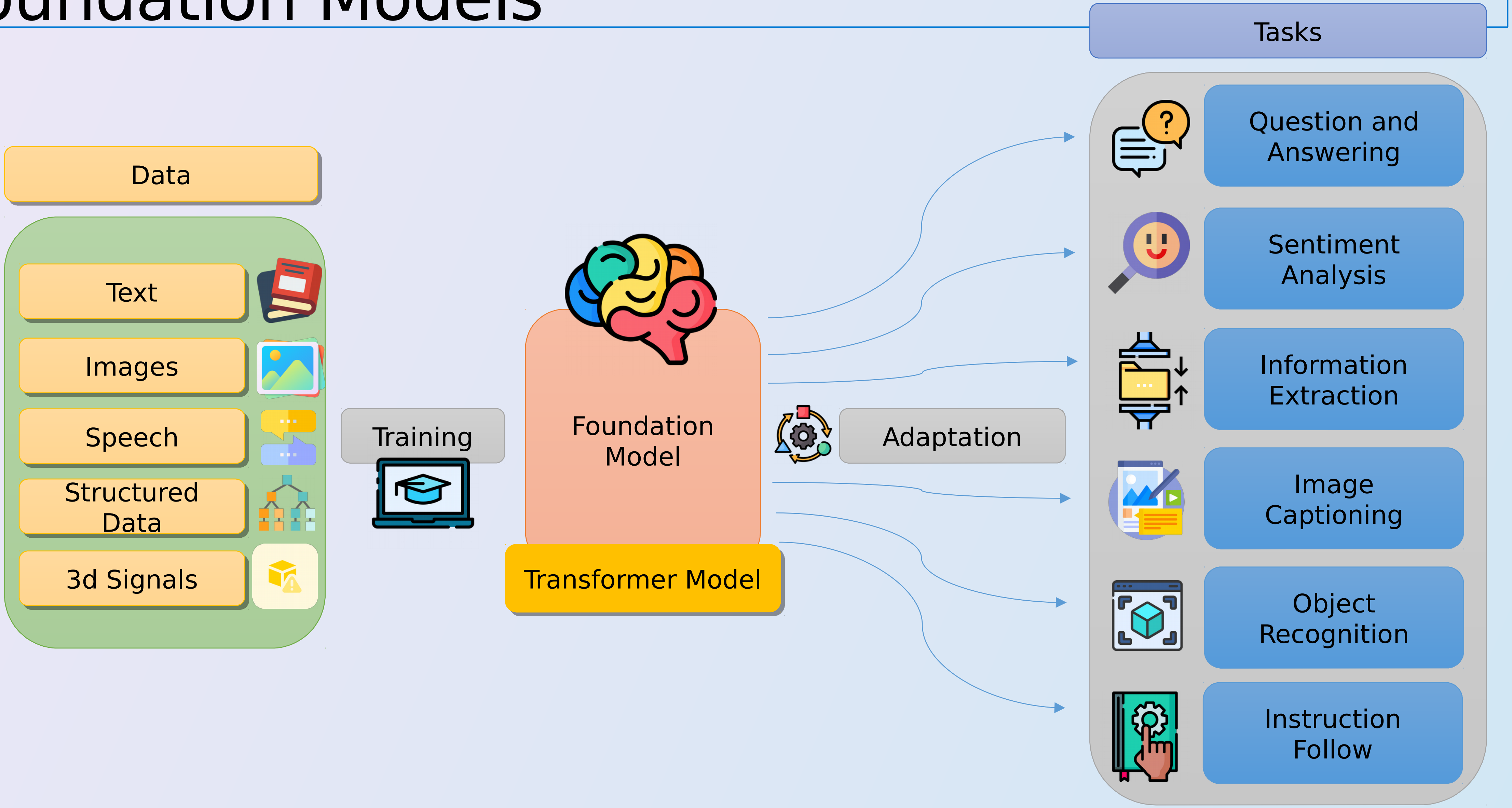
#GlobalAzure
#GlobalAzureMilano

Parla con i tuoi video grazie ad Azure AI e a GPT-4 Turbo with Vision!

Roberta Bruno

MICROSOFT

Foundation Models



Data

Text

Images

Speech

Structured Data

3d Signals

Training

Foundation Model

Transformer Model

Adaptation

Tasks

Question and Answering

Sentiment Analysis

Information Extraction

Image Captioning

Object Recognition

Instruction Follow

Microsoft and OpenAI partnership

OpenAI

Ensure that artificial general intelligence (AGI) benefits humanity

Microsoft

Empower every person and organization on the planet to achieve more

Azure OpenAI Service

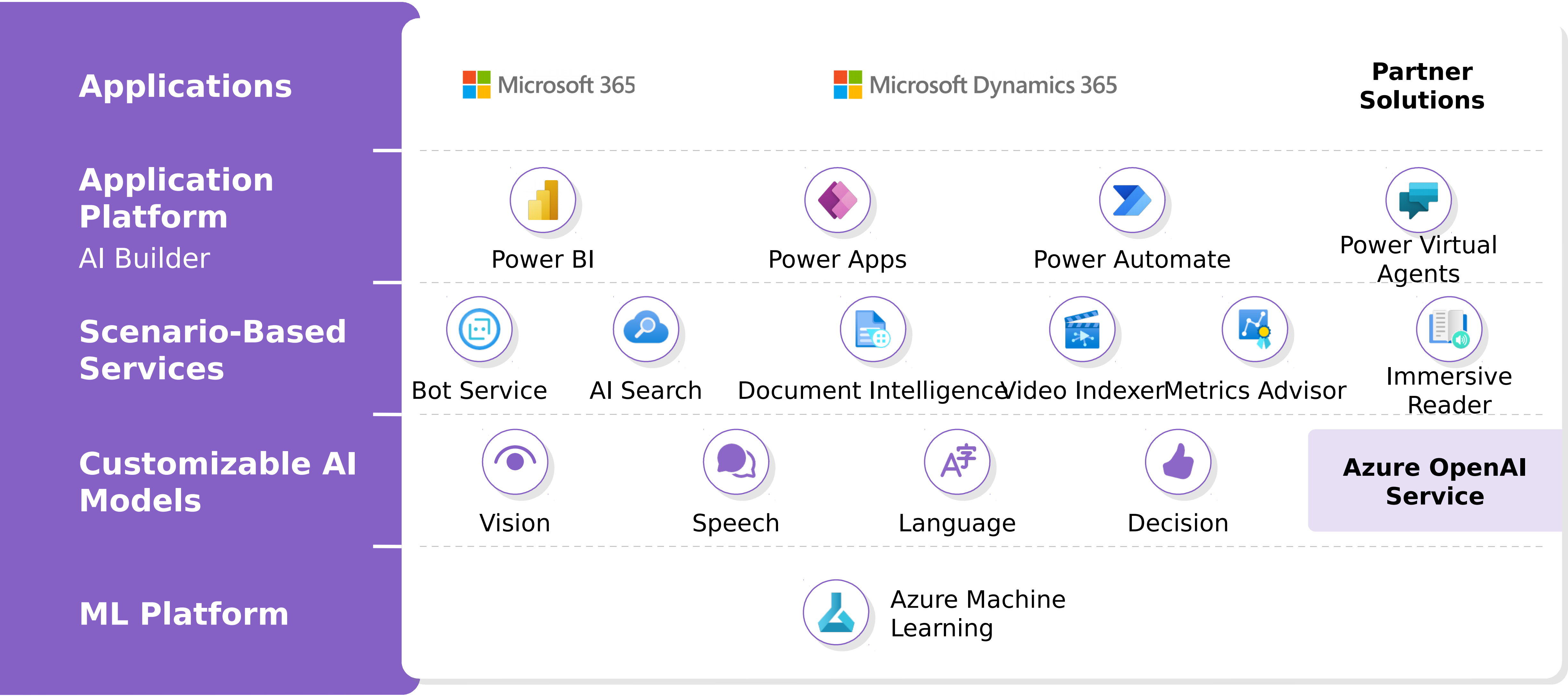
GPT-4, 4-Turbo and 3.5-Turbo	GPT-4 Vision	Babbage and Davinci	DALL·E 3	Whisper
Language	Multi-Modal	Fine Tuning	Images	Transcription & Translation

On Your Data

Azure AI Studio

Assistants

Microsoft is powered by Azure AI



Applications

Microsoft 365

Microsoft Dynamics 365

Partner Solutions

Application Platform

AI Builder

Power BI

Power Apps

Power Automate

Power Virtual Agents

Scenario-Based Services

Bot Service

AI Search

Document Intelligence

Video Indexer

Metrics Advisor

Immersive Reader

Customizable AI Models

Vision

Speech

Language

Decision

Azure OpenAI Service

ML Platform

Azure Machine Learning

Microsoft Azure Cloud

Runs on trust

Your data is your data

- Data is stored encrypted in your Azure subscription

Your data is not used to train underlying foundation models in the model catalog, without your permission

- Azure OpenAI Service provisioned in your Azure subscription
- Model fine tuning stays in your Azure subscription

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

- Encrypted with Customer Managed Keys
- Private Virtual Networks, Role Based Access Control
- Soc2, ISO, HIPPA, CSA STAR Compliant

Azure OpenAI Service

GPT-3.5-Turbo

GPT-4

GPT-4 Turbo

GPT-4 Turbo with Vision

Generative Text Models, with varying capabilities and uses

Whisper

Transcription and Translation

DALL·E 3

Generative Image Model

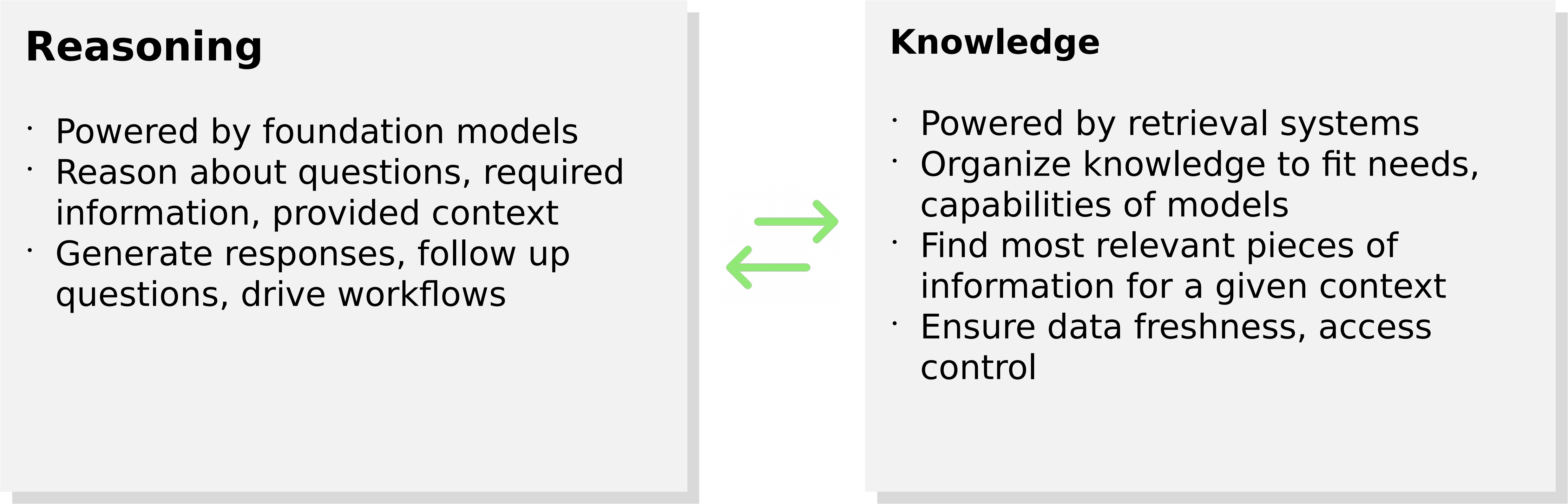
Deploy on your own data

Provisioned throughput units (PTUs)

Assistants, Functions and Plugins

RAG

Reasoning + Knowledge



Reasoning

- Powered by foundation models
- Reason about questions, required information, provided context
- Generate responses, follow up questions, drive workflows

Knowledge

- Powered by retrieval systems
- Organize knowledge to fit needs, capabilities of models
- Find most relevant pieces of information for a given context
- Ensure data freshness, access control

Bringing domain knowledge to LLMs

Prompt engineering

In-context learning

Fine tuning

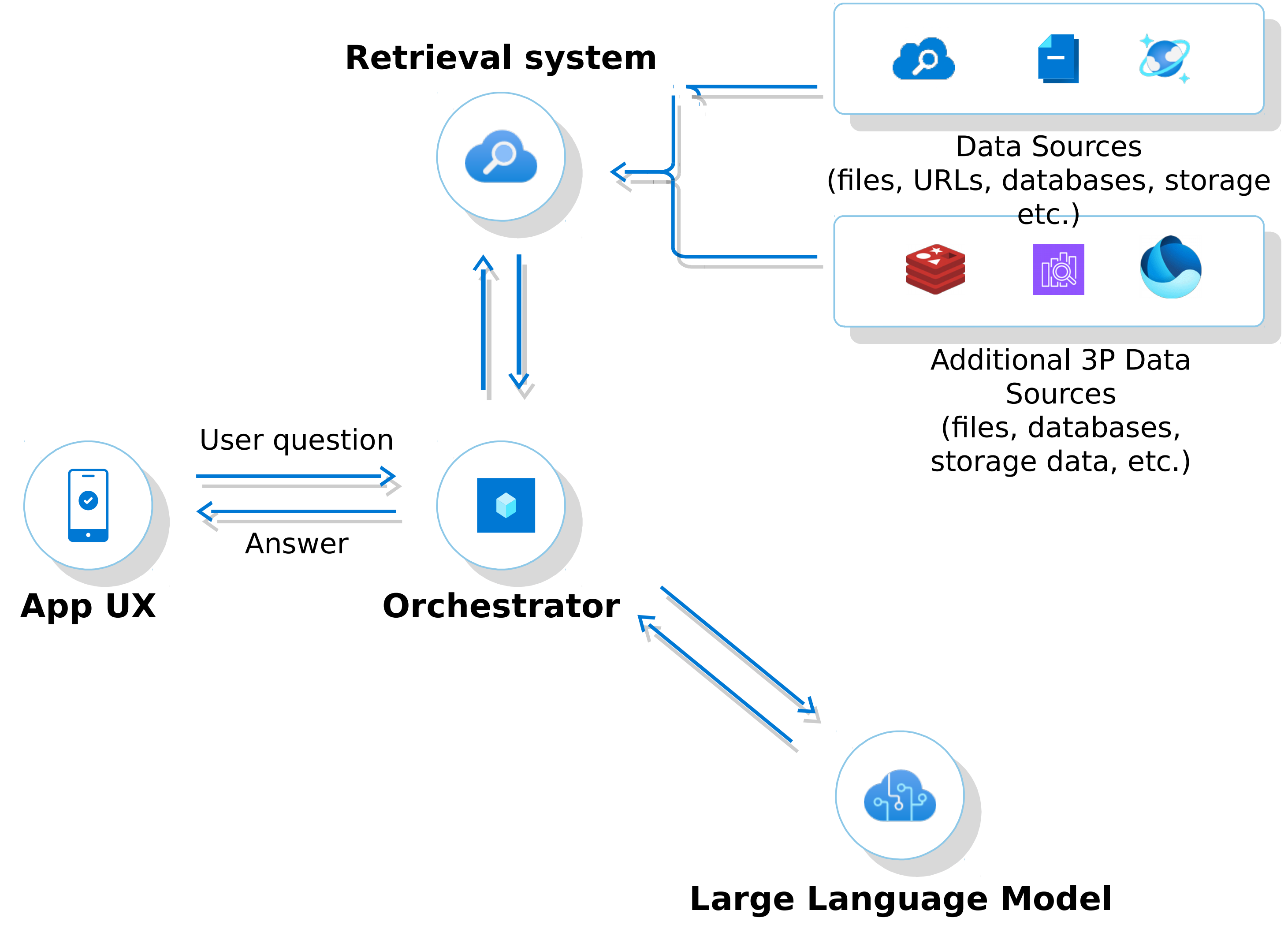
Learn new skills

Retrieval augmentation

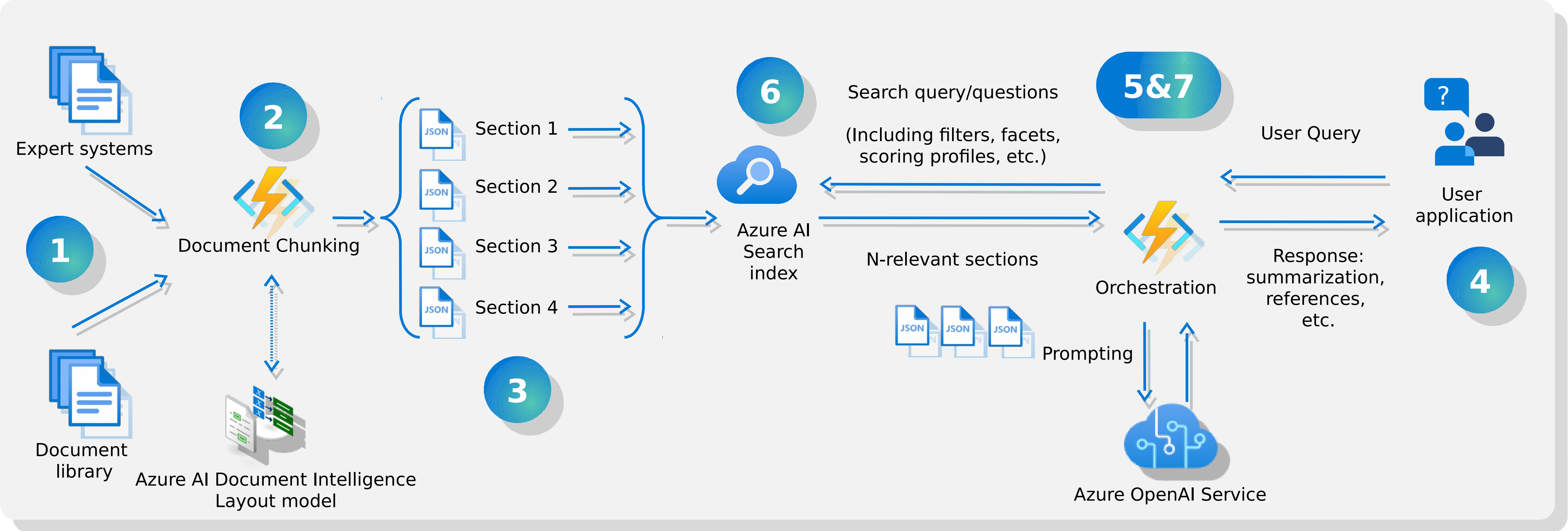
Learn new facts

Retrieval-augmented generation

Anatomy of the workflow


Retrieval system
Data Sources
(files, URLs, databases, storage etc.)
Additional 3P Data Sources
(files, databases, storage data, etc.)
User question
Answer
App UX
Orchestrator
Large Language Model

Anatomy of RAG



1. Data ingestion

Different data formats and system of records

2. Chunking

What is the best Chunking strategy?

3. Indexing

Shall I use vector embeddings data transformation, mappings?

4. User interface

Chatbot for Q&A surfaced to end users

5. Orchestration

Communication coordination and prompting—Prompt to get retriever query

6. Data retrieving

Shall I use vector, semantic,

keyword or hybrid approach?

7. Orchestration

Communication coordination: create user response based on retrieve data and send to User app

Azure AI Search

Feature-rich vector database

Optimized vector storage

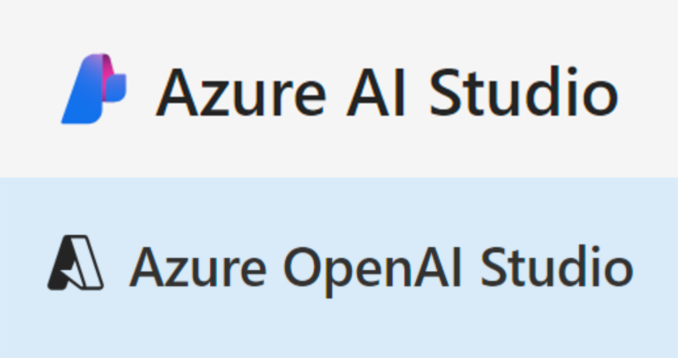
Seamless data & platform integrations

State-of-the-art search technology

Enterprise-ready foundation

Expanded storage and vector index size

Tools for ingesting data into AI Search for RAG



Azure AI Studio

Azure OpenAI Studio

	AI Studio	Azure OpenAI "On your data"	Azure AI Search built-in indexer
Incremental indexing	Using versioning only	Yes	Yes
Multiple data source support	Yes	Yes	Yes
Different data sources going to the same index	No	No	Yes (one indexer per data source, multiple indexers pointing to the same index)
Configurable deletion policy	No	No	Yes
Chunking	Yes	Yes	Yes (through split skill/custom skill)
Vectorization	Yes	Yes	Yes (through embedding skill/ custom skill)
Using an existing AI Search Chunked index	No	Yes	Yes
AI enrichment	Options to transform data can be added	Some transformations can be done using plugins	Yes

Ingestion options provided by AI Search

1

Data source support from AI Search directly through built-in pull indexers:
 Data sources gallery—Azure AI Search | Microsoft Learn
 Integrated Vectorization

2

Data source supported by Microsoft Partners:
 Data sources gallery—Azure AI Search | Microsoft Learn

3

Push API/SDK for any data source not supported with pull method:
 Data import and data ingestion—Azure AI Search | Microsoft Learn
 Push SDK in RAG

Use data from all over Azure

Supported data sources include

- Azure Storage
  - Blob
  - Data Lake Storage Gen2
  - Table
  - Files
- Azure Cosmos DB
  - NoSQL
  - Gremlin
  - MongoDB
- Azure SQL
- Azure Database for MySQL
- A variety of partner-supported data sources

The technology behind Azure AI Search

Retrieval modes

Keyword-based retrieval

- Traditional full-text search method
- Content is broken into terms; uses the BM25 probabilistic model for scoring

Vector-based retrieval

- Text is converted into vector representations
- Uses embedding models, e.g., Azure Open AI text-embedding-ada-002

Hybrid retrieval

- Combines strengths of Keyword and Vector
- Fusion step selects the best results from both methods, using Reciprocal Rank Fusion (RRF)

Semantic ranking

What is Semantic ranking?

- Bing technology that uses transformer models with cross-attention to simultaneously processes query and document text

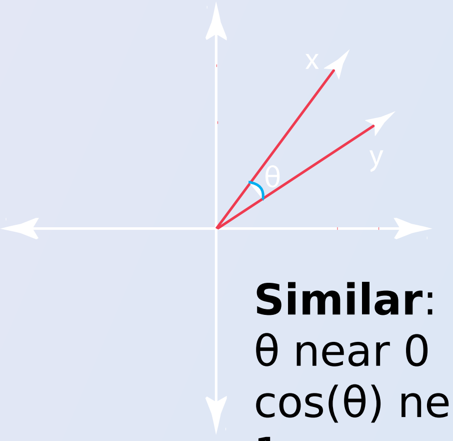
What does it do?

- Prioritizes the most important results
- Normalized relevance score filters out low-quality results
- Score Range: 0 (irrelevant) to 4 (highly relevant)

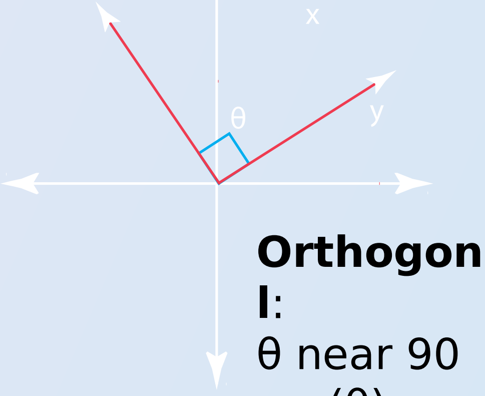
Vector similarity

We compute embeddings so that we can calculate similarity between inputs.
The most common distance measurement is **cosine similarity**.

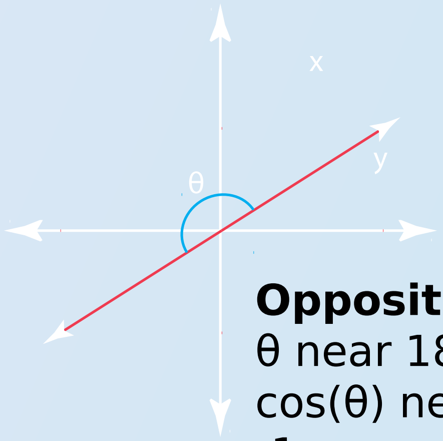
```
def cosine_sim(a, b):
 return dot(a, b) /
  (mag(a) * mag(b))
```



Similar:
θ near 0
cos(θ) near 1



Orthogonal:
θ near 90
cos(θ) near 0



Opposite:
θ near 180
cos(θ) near -1

*For ada-002, cos(θ) values range from 0.7-1

 Demo: Compare vectors with cosine similarity (aka.ms/aitour/vectors)

 Demo: Vector Embeddings Comparison (aka.ms/aitour/vector-similarity)

Orchestration with AI Search

Azure AI Studio

- Explore, build, test, and deploy cutting-edge LLM-powered genAI solutions responsibly
- Evaluate LLM responses and pinpoint fine-tuning opportunities
- Scale PoCs with a paved path to full production

Copilot Studio

- Build your own copilot using intuitive building experiences
- Customize Microsoft Copilots with your own enterprise scenarios
- Leverage a connected, integrated platform

Open Source

- Semantic Kernel
- Langchain
- LlamaIndex

GPT-4 Turbo with Vision

Public preview

Announcing GPT-4V with Azure AI Vision

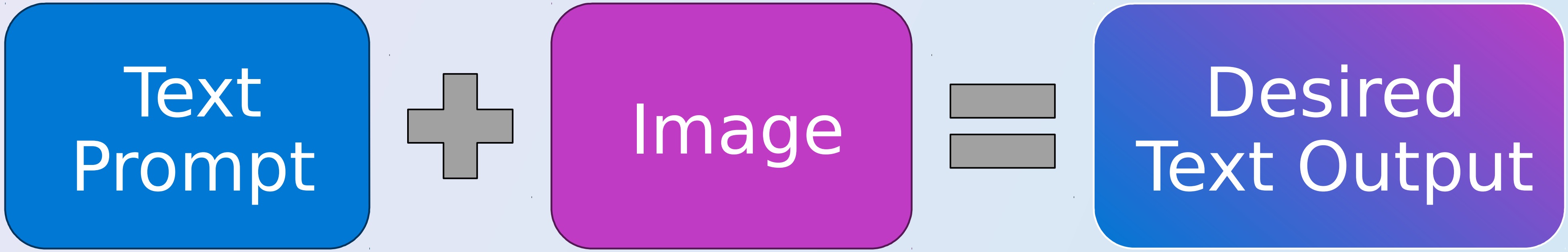
Unlock new scenarios with GPT-4V, Azure Open AI Service and Azure AI Vision integration

Add images to retrieval augment generation (RAG) patterns

Prompt with video, images, and text

What GPT4 Turbo with Vision Offers

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



Text Prompt

Image

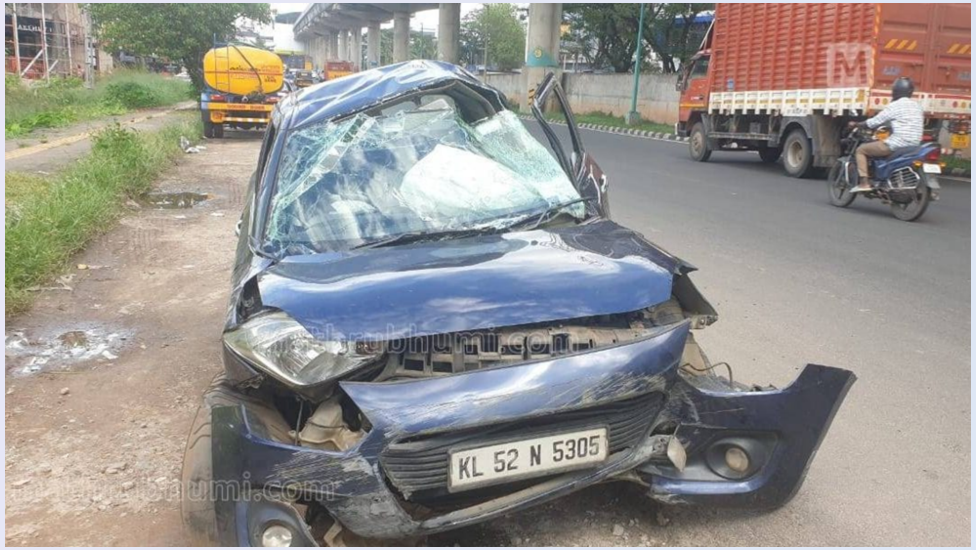
Desired Text Output

Note: GPT-4 Turbo w/Vision doesn’t generate image outputs

Vision AI Emerging Application Highlights - Auto Insurance (Insurance Report Generation)

User

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please fill the incident report for the accident shown in image below, with the following format in JSON (note xxx is placeholder, if you cannot determine that based on the image, put "N/A" instead).
{"**make": xxx, "model": xxx, "license plate": xxx, "damage description": xxx, "estimated cost of repair": xxx**}



Vision AI

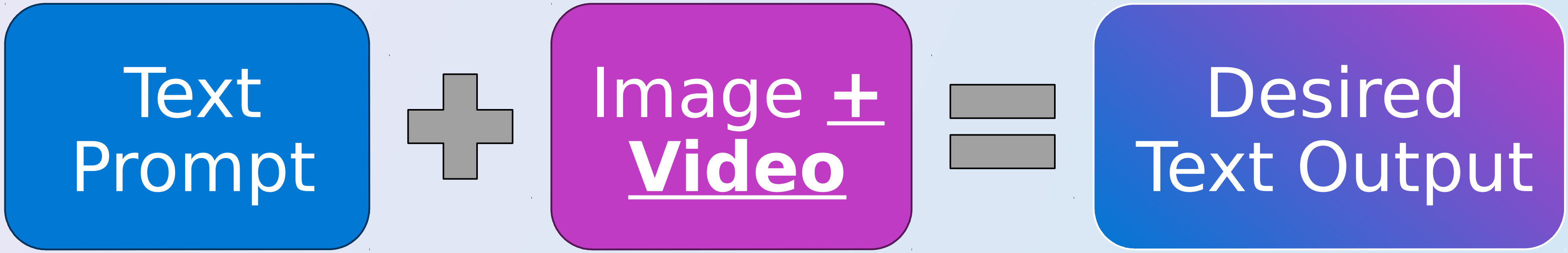
{"make": "Maruti Suzuki", "model": "Celerio", "license plate": "KL 52 N 5305", "damage description": "Front bumper, hood, and windshield damaged. Airbags deployed.", "estimated cost of repair": "N/A"}

GPT-4 Turbo with Vision +Azure AI Vision Enhancements

Use GPT4 Turbo w/Vision with Azure AI Vision Enhancements to make the output more robust.

What 'GPT-4 Turbo with Vision' **PLUS** Azure AI Vision Enhancements does

GPT-4 Turbo with Vision is able to take text and video inputs when used with the Azure AI Vision Enhancement feature.



Text Prompt

+

Image + **Video**

=

Desired Text Output

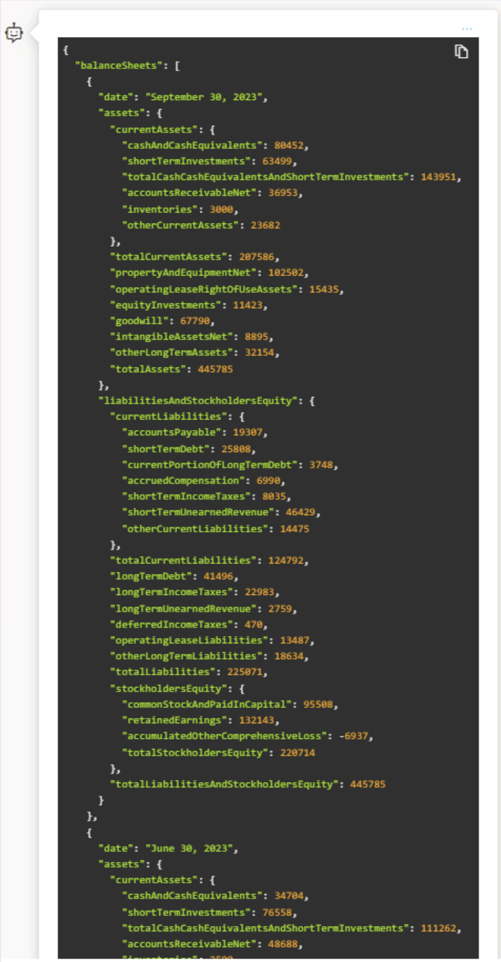
Note: GPT-4 Turbo w/Vision doesn't generate image outputs

Enhancement Type: OCR

Improve number accuracy in dense text

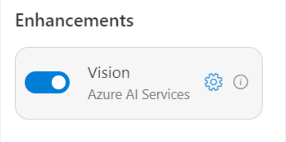
However, with the Azure AI Vision Enhancement turned on, we can minimize errors in output.

Extracted json



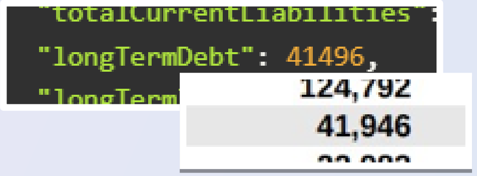
```
{
  "balanceSheets": [
    {
      "date": "September 30, 2023",
      "assets": {
        "currentAssets": {
          "cashAndCashEquivalents": 80452,
          "shortTermInvestments": 63499,
          "totalCashCashEquivalentsAndShortTermInvestments": 143951,
          "accountsReceivableNet": 36953,
          "inventories": 3000,
          "otherCurrentAssets": 23682
        },
        "totalCurrentAssets": 207586,
        "propertyAndEquipmentNet": 102502,
        "operatingLeaseRightOfUseAssets": 15435,
        "equityInvestments": 11423,
        "goodwill": 67790,
        "intangibleAssetsNet": 8895,
        "otherLongTermAssets": 32154,
        "totalAssets": 445785
      },
      "liabilitiesAndStockholdersEquity": {
        "currentLiabilities": {
          "accountsPayable": 19307,
          "shortTermDebt": 25808,
          "currentPortionOfLongTermDebt": 3748,
          "accruedCompensation": 6990,
          "shortTermIncomeTaxes": 8035,
          "shortTermUnearnedRevenue": 46429,
          "otherCurrentLiabilities": 14475
        },
        "totalCurrentLiabilities": 124792,
        "longTermDebt": 41496,
        "longTermIncomeTaxes": 22983,
        "longTermUnearnedRevenue": 2759,
        "deferredIncomeTaxes": 470,
        "operatingLeaseLiabilities": 13487,
        "otherLongTermLiabilities": 18634,
        "totalLiabilities": 225071,
        "stockholdersEquity": {
          "commonStockAndPaidInCapital": 95508,
          "retainedEarnings": 132143,
          "accumulatedOtherComprehensiveLoss": -6937,
          "totalStockholdersEquity": 220714
        },
        "totalLiabilitiesAndStockholdersEquity": 445785
      }
    },
    {
      "date": "June 30, 2023",
      "assets": {
        "currentAssets": {
          "cashAndCashEquivalents": 34704,
          "shortTermInvestments": 76558,
          "totalCashCashEquivalentsAndShortTermInvestments": 111262,
          "accountsReceivableNet": 48688,
          "inventories": 2500
```

GPT-4 Turbo with Vision:



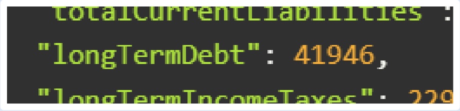
Enhancements

Vision
Azure AI Services

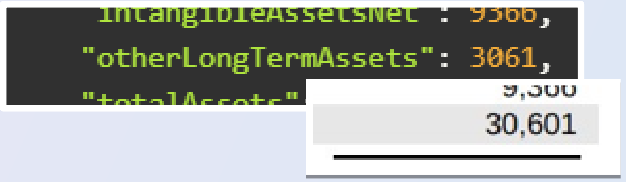


```
"totalCurrentLiabilities":
"longTermDebt": 41496,
"longTerm
```

124,792
41,946
22,983

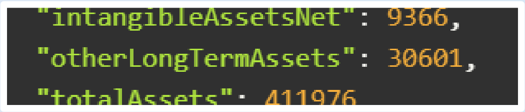


```
"totalCurrentLiabilities":
"longTermDebt": 41946,
"longTermIncomeTaxes": 229
```

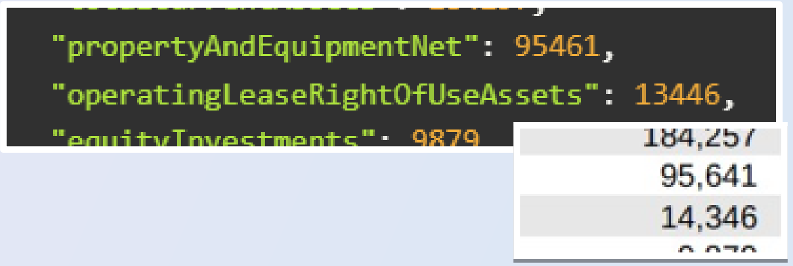


```
"intangibleAssetsNet": 9366,
"otherLongTermAssets": 3061,
"totalAssets"
```

9,366
30,601

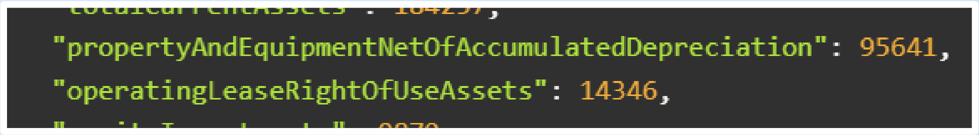


```
"intangibleAssetsNet": 9366,
"otherLongTermAssets": 30601,
"totalAssets": 411976
```



```
"totalCurrentAssets": 184257,
"propertyAndEquipmentNet": 95461,
"operatingLeaseRightOfUseAssets": 13446,
"equityInvestments": 9879
```

184,257
95,641
14,346
9,879

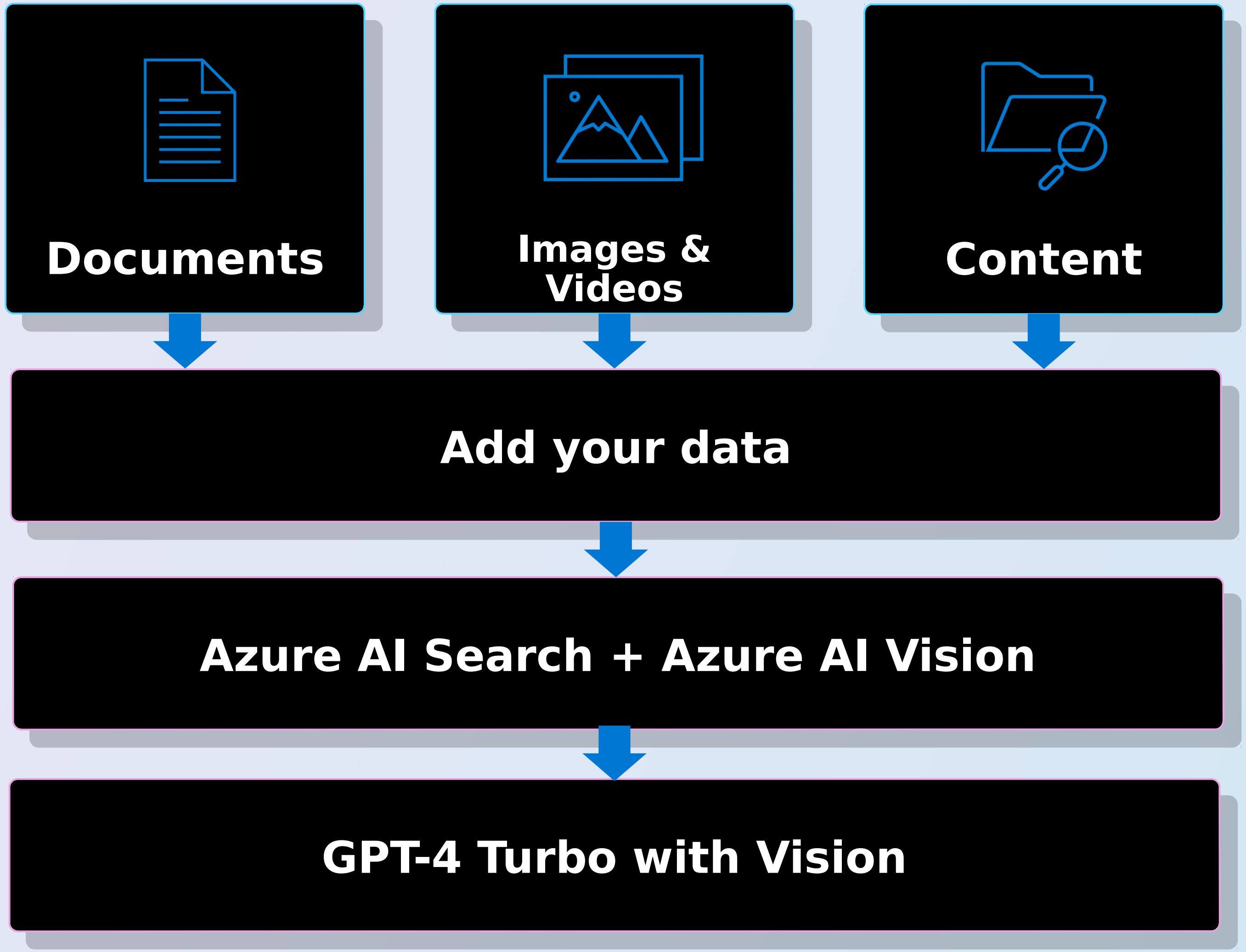


```
"totalCurrentAssets": 184257,
"propertyAndEquipmentNetOfAccumulatedDepreciation": 95641,
"operatingLeaseRightOfUseAssets": 14346,
"equityInvestments": 9879
```

Azure OpenAI Service on your data, with images

Ground the information provided on your company data

Retrieval Augmented Generation



Documents

Images & Videos

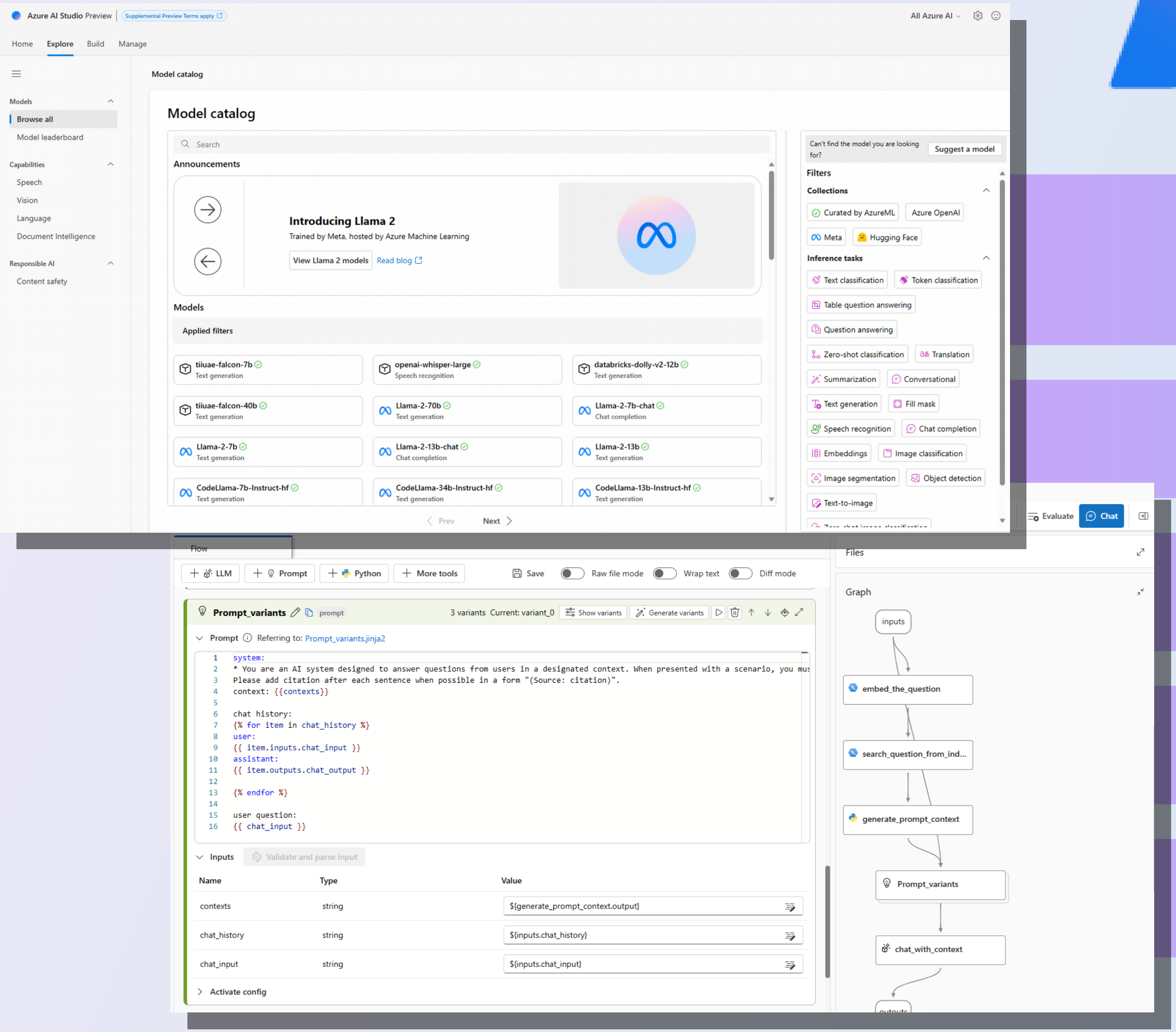
Content

Add your data

Azure AI Search + Azure AI Vision

GPT-4 Turbo with Vision

Azure AI is a platform for Generative AI



Model catalog

Introducing Llama 2

Trained by Meta, hosted by Azure Machine Learning

Can't find the model you are looking for?

tiiuae-falcon-7b Text generation	openai-whisper-large Speech recognition	databricks-dolly-v2-12b Text generation
tiiuae-falcon-40b Text generation	Llama-2-70b Text generation	Llama-2-7b-chat Chat completion
Llama-2-7b Text generation	Llama-2-13b-chat Chat completion	Llama-2-13b Text generation
CodeLlama-7b-Instruct-hf Text generation	CodeLlama-34b-Instruct-hf Text generation	CodeLlama-13b-Instruct-hf Text generation

Prompt_variants prompt 3 variants Current: variant_0

Prompt Referring to: Prompt_variants.jinja2

```
system:
* You are an AI system designed to answer questions from users in a designated context. When presented with a scenario, you mus
Please add citation after each sentence when possible in a form "(Source: citation)".
context: {{contexts}}

chat history:
{% for item in chat_history %}
user:
{{ item.inputs.chat_input }}
assistant:
{{ item.outputs.chat_output }}

{% endfor %}

user question:
{{ chat_input }}
```

Inputs

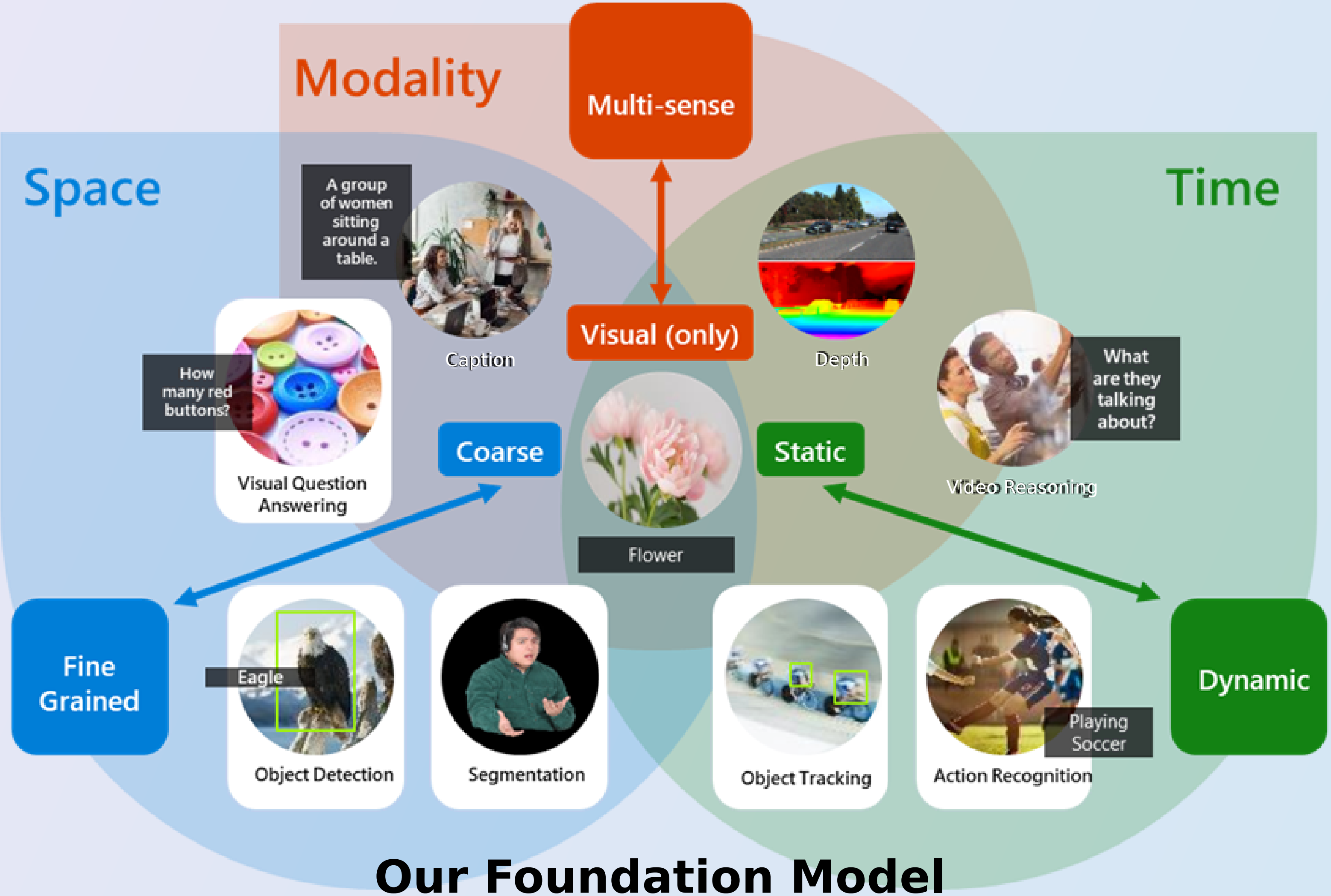
Name	Type	Value
contexts	string	${generate_prompt_context.output}
chat_history	string	${inputs.chat_history}
chat_input	string	${inputs.chat_input}

- Access to thousands of LLMs from OpenAI, Meta, Hugging Face
- Data grounding with RAG
- Prompt engineering/evaluation
- Built-in safety and responsible AI
- Continuous monitoring for LLMs

Florence Vision AI Capabilities

Azure AI Vision’s Florence Model is our in-house developed Large Vision Model

Project Florence – Large foundation model



Our Foundation Model

When to use the Florence Model?

- Shorter Descriptions of Images
- Image Retrieval or Image Search
- Image Segmentation / Background Removal
- Video Retrieval or Video Search
- Customised Product Recognition
- Shelf Analysis
- RGB Camera / 3D Body Checking in Manufacturing, Sports, Mining, etc. (example: think health and safety scenarios)

Resources

GPT-4_Turbo-with-Vision_Pricing

How to use the GPT-4 Turbo with Vision model

Use your image data with Azure OpenAI Service in Azure OpenAI studio

Video Retrieval API reference

how-to/gpt-with-vision.md at main · GitHub

Let’s connect!

#GlobalAzure
#GlobalAzureMilano

GRAZIE!!!

Le slide saranno disponibili sulla pagina
Global Azure 2024 del sito di Azure Meetup Milano