

# Azure AI Apps in a Day

Extend your app's capabilities with Azure's AI services

# Meet our team

## Cloud App and Integration (East Coast)



Insight

Ross Johnson  
Practice Lead  
[ross.johnson@insight.com](mailto:ross.johnson@insight.com)



Insight

Daniel Fang  
Senior Architect  
[daniel.fang@insight.com](mailto:daniel.fang@insight.com)



Insight

Rakesh Lakshminarayana  
Architect  
[rakesh.lakshminarayana@insight.com](mailto:rakesh.lakshminarayana@insight.com)

# Content for today

## Presentations:

- What are AI Apps?
- Building AI Apps & Techniques
- Azure's Broad Set of AI Services
- Azure's AI Supporting Capabilities
- RAG Pattern & Techniques
- Common Patterns and Architectures

## Hands-on Labs

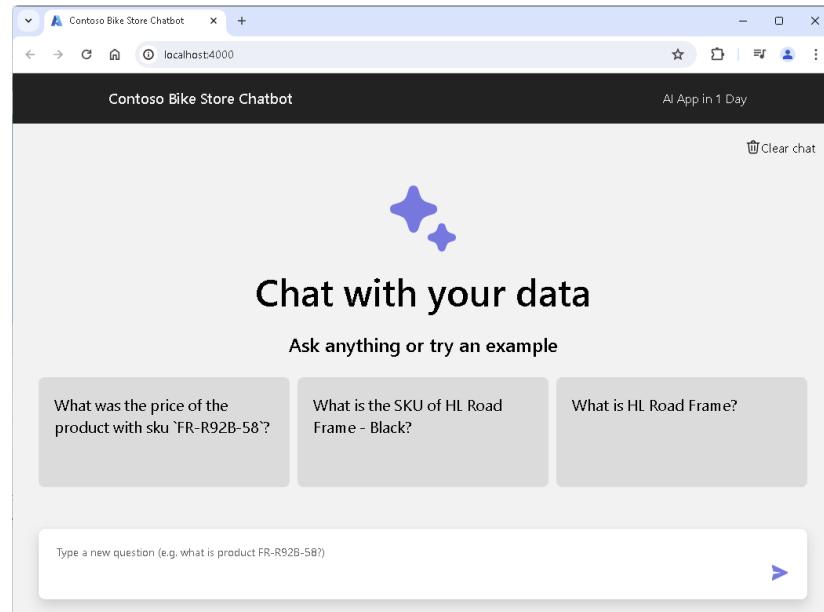
- Lab 1 - Interact with OpenAI Models
- Lab 2 - Building a Chatbot using RAG



# Your AI App for today

Get ready to build your very own Chatbot using RAG on Azure!

- Chatbot frontend (React)
- Restful API backend (node.js)
- Azure OpenAI GPT4o & Embedding
- Azure App Service
- Azure Cosmos DB for MongoDB
- Azure Deployment with Bicep
- Langchain (node.js)

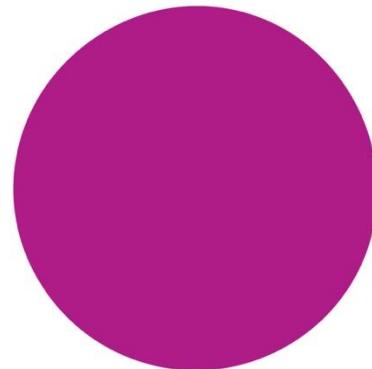


# Schedule

Time	Activity
08:30 (30m)	Presentation 1: AI Apps using Azure
09:00 (30m)	Tech check
09:30 (30m)	Morning break ☕
10:00 (2h)	Hands-on lab time (Lab 1)
12:00 (30m)	Lunch 🍕🍔🥗
12:30 (30m)	Presentation 2: RAG Pattern & Techniques
12:30 (3h)	Hands-on lab time (Lab 2)
	Afternoon break ☕ @ 14:00 (30m)
15:30 (15m)	Day review and wrap

You are here

# Be ambitious.



## AI Apps using Azure

# What are AI Apps ?

Software applications that leverage AI technologies to carry out tasks traditionally requiring human intelligence. These tasks include learning from data, identifying patterns, making decisions, and solving complex problems

## **Natural Language Processing (NLP):**

Enables apps to understand, interpret, and generate human language.

## **Machine Learning:**

Algorithms that allow apps to learn from and adapt to new data over time.

## **Predictive Analytics:**

Predictive analytics to forecast outcomes based on data

## **Computer Vision:**

Allows apps to interpret and process visual information from the world.

## **Automation:**

Automates repetitive tasks, improving efficiency and accuracy.

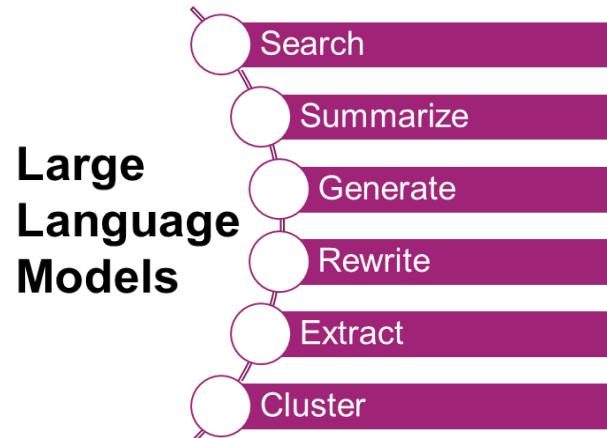


# Large Language Model (LLM)

Large Language Models are transforming the way applications engage with users, enabling more natural and intelligent interactions. With LLMs, AI apps can comprehend, generate, and communicate in human language with remarkable complexity and precision.

## Key Features:

- **Contextual Understanding:** Ability to grasp context and nuances in conversation, leading to more relevant and coherent responses.
- **Generative Abilities:** Create human-like text based on prompts, enabling applications like content creation and storytelling.
- **Adaptability:** Learn from diverse datasets, allowing for versatile applications across various domains.



# Think broader, Generative AI – text, image, video

Room of the House	Top AI-enabled Use Cases
Product	<ul style="list-style-type: none"><li>AI-enabled product experiences: Virtual assistant, document automation, content creation</li><li>Research at scale to inform product design</li><li>Rapid design and prototyping</li></ul>
Legal	<ul style="list-style-type: none"><li>Summarization of legal precedents or docs for litigation support</li><li>Automating contract creation, review and management</li><li>Real-time tracking of changing laws and regulations impacting policy</li></ul>
Customer Service and Contact Centers	<ul style="list-style-type: none"><li>Customer self-service assistant</li><li>Contact center agent assistant</li><li>Call summarization and analytics</li></ul>
Marketing	<ul style="list-style-type: none"><li>Shopping assistant / product recommendation and upsell</li><li>Auto-generate personalized marketing content</li><li>Customer micro-segmentation</li><li>Conversational / adaptive customer research surveys</li></ul>
IT/Software Development	<ul style="list-style-type: none"><li>IT and developer documentation, knowledge management</li><li>AI-enabled application development and modernization</li><li>Infrastructure-as-Code for service automation</li></ul>
HR	<ul style="list-style-type: none"><li>Recruitment process automation</li><li>Employee self-service and knowledge management</li><li>Personalized career planning</li></ul>

# Building AI Apps

Creating a successful AI app requires careful attention to several key components and considerations to ensure it is effective, reliable, and user-friendly

- **Objectives and Use Case:** Define the specific problem your AI app will solve
- **AI Models:** Choose or develop AI models that best fit the problem
- **Data Quality:** Ensure the data used for training and operation is clean, relevant, and representative
- **Integration:** Seamlessly integrate the AI component with existing systems and workflows
- **Infrastructure and Tools:** Utilize scalable infrastructure and development tools
- **Testing and Validation:** Rigorously test to ensure it performs as expected under different conditions and scenarios.
- **Ethics and Bias:** Incorporate mechanisms to detect, reduce, and address potential biases
- **Security:** Implement robust security measures to protect the AI app, data, and user information
- **Maintenance and Updates:** Regularly update the AI app to improve functionality, adapt to new data
- **Legal and Compliance:** Complies with relevant legal regulations, industry standards and data privacy policies

# Supercharge your AI models

Here are three techniques to use domain expertise and enhance the AI model behind your app



## Prompt engineering

In-context learning



## Retrieval Augmented Generation

Learn new facts  
(temporarily)



## Fine tuning

Learn new skills  
(permanently)

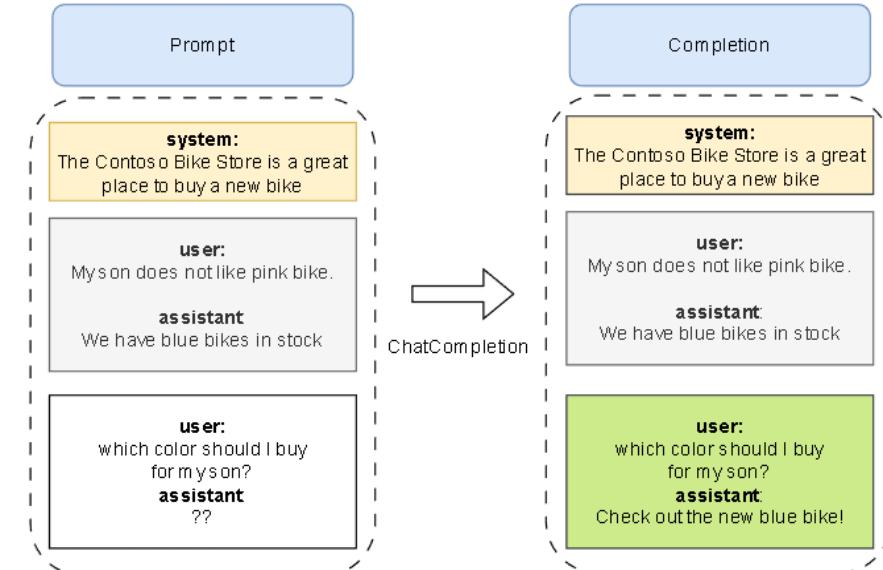
# Prompt Engineering

## Prompt Payload Structure:

- system prompt
- user prompt
- Assistant

## NLP techniques:

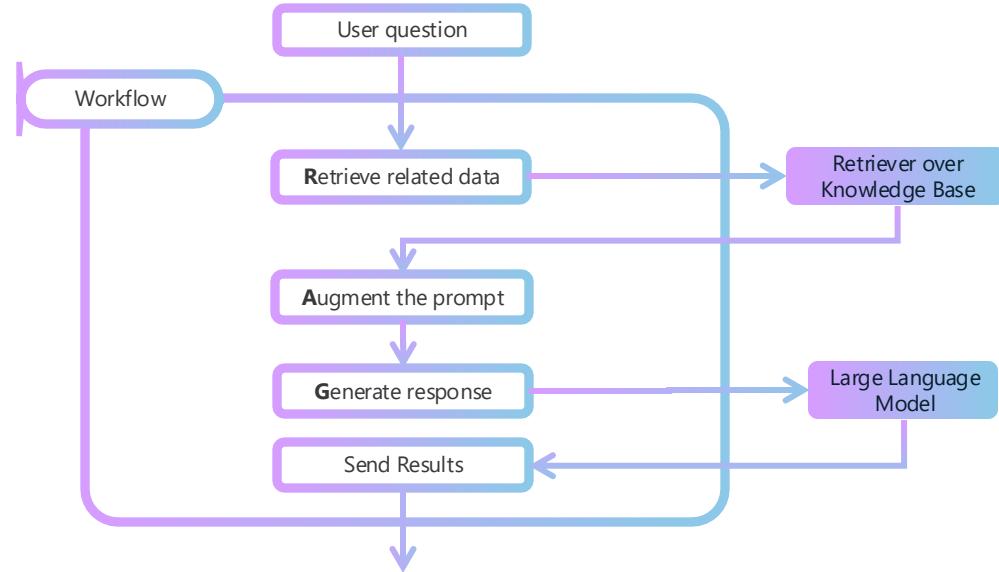
- Zero-shot learning: Provide prompts to an LLM without any prior examples.
- Few-shot learning: Provide prompts to an LLM with a few specific examples of the task.
- Knowledge base in system prompt: Incorporate a knowledge base directly into the system prompt.



# Retrieval Augmented Generation

RAG combines generative models with retrieval mechanisms to improve the quality and relevance of generated responses.

- **Retrieval Stage** : Relevant documents or pieces of information are retrieved from a knowledge base or database using a retriever model
- **Generation Stage** : The retrieved documents are used as context for the generative model. The generative model produces a response based on both the input query and the retrieved documents.



Azure AI Search has built-in data chunking and vectorization capabilities

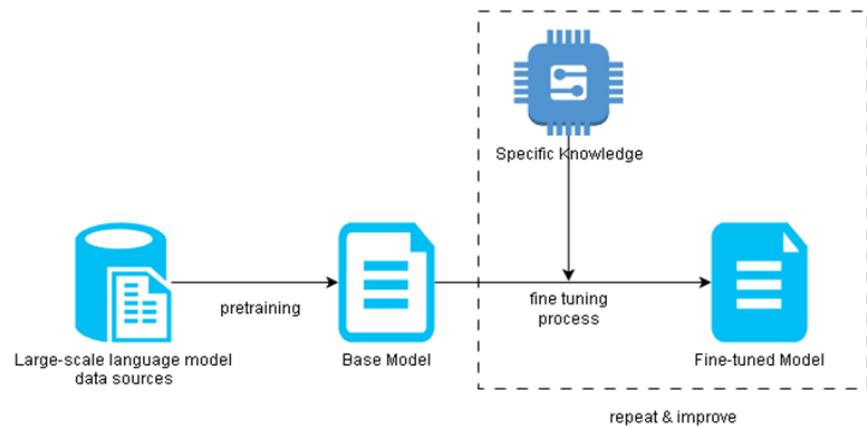
# Fine tuning & Custom Model

## What's LLM fine-tuning

- Take an open-source or proprietary model
- Re-train it on a variety of concrete examples
- And save the updated weights as a new model checkpoint
- So as to adapt the pre-trained model to specific tasks

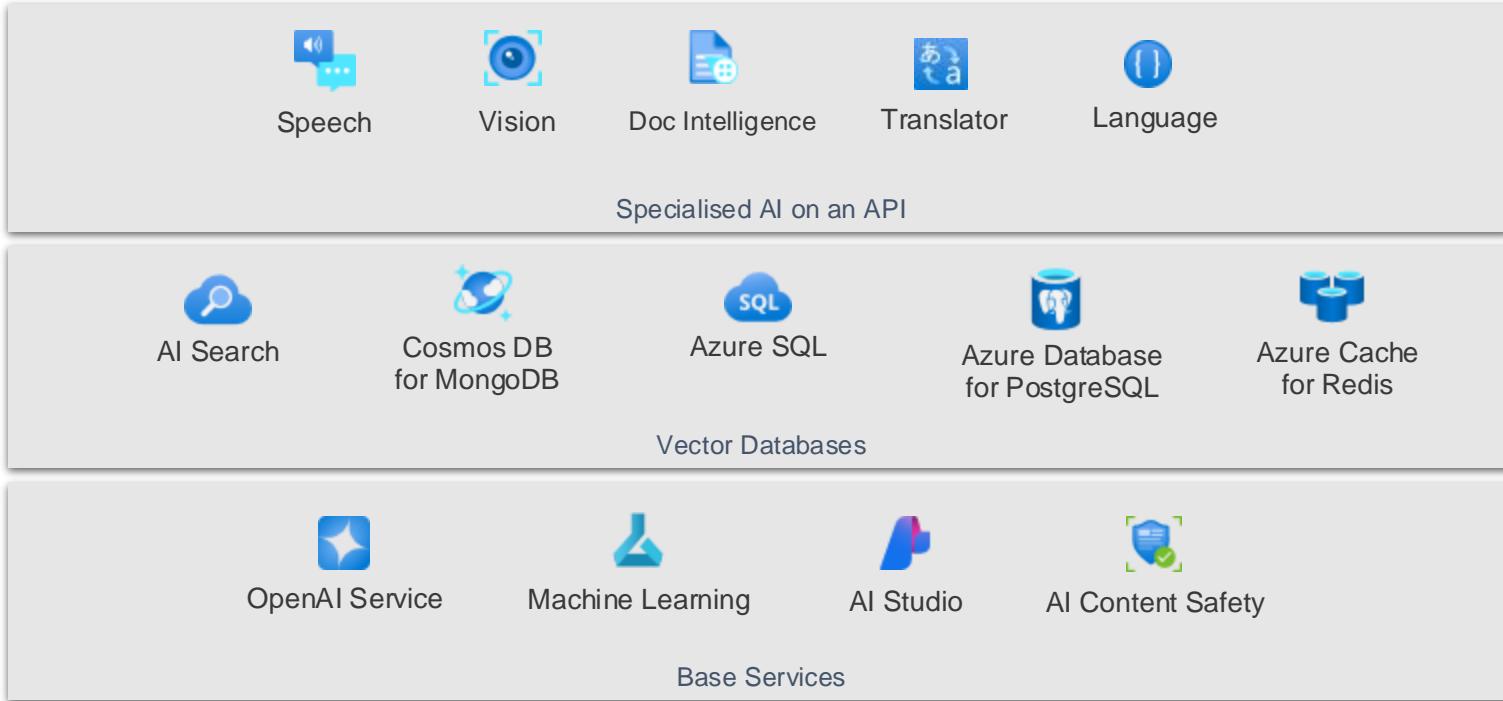
## Benefits from fine tuning:

- Higher quality results than prompt engineering
- Train on more examples than can fit into context limit
- Token savings due to shorter prompts
- Lower-latency requests



Azure OpenAI Service supports fine-tuning.

# Azure AI services - using AI for Intelligent Apps



# Azure OpenAI Service

Industry-Leading Coding and Language AI Models, plus Powerful and Customizable AI Solutions

## Many AI Models Available

- GPT-4o: Advanced language model for diverse tasks
- Turbo with Vision: Enhanced model with image understanding capabilities
- Embeddings: For semantic search and context-aware tasks

## Access Options

- REST API: Integrate with your existing applications
- Web-Based Interface: User-friendly access for quick tasks

## Use Cases

- Content Generation: Create high-quality, relevant content tailored to your needs
- Summarization: Condense information effectively
- Image Understanding: Analyze and interpret visual data
- Semantic Search: Improve search accuracy and relevance

# Model breakthroughs for innovation



## New Modes

Enrich interactions with **multimodal models** that understand diverse data types like videos, images and speech



## New Flexibility

Maximize efficiency for specific tasks with **small language models (SLMs)** that require less compute



## New Adaptability

Solve complex challenges with **multi-agent systems** that combine APIs and collaborate persistently



## New Control

Plan ahead with **provisioned throughput** or pay as you go with Models as a Service (MaaS)



# Find, customize, and operationalize models at scale

The Model Catalog is your starting point to explore and use the latest foundation models

The screenshot shows the Azure AI Studio Model Catalog interface. It features a sidebar with navigation links like Home, Explore, Build, and Manage, and a dropdown for 'Benchmarks'. The main area has sections for 'Explore' (with a chart titled 'Find the right model for your task' comparing models like gpt-4-32k, gpt-4, gpt-35-turbo, Llama-2-70b, and Llama-2-13b), 'Model catalog' (with a section for 'Find the right model to build your custom AI solution' featuring announcements about Mistral, Phi 3, and Meta Llama 3), and 'Evaluate' (with a 'Fine-tuning tasks' section listing models such as dalle-3, gpt-4, gpt-35-turbo-instruct, davinci-002, text-embedding-ada-002, gpt-35-turbo, bagpipe-002, gpt-4-32k, mistral-Mistral-7B-Instruct-v01, mistral-Mistral-8x7B-Instruct-v01, mistral-Mistral-8x7B-v01, mistral-Mistral-7B-v01, mistral-Mistral-7B-Instruct-v01, mistral-Mistral-8x22B-Instruct-v01, mistral-Mistral-8x22B-v01, and Llama-2-70b). A 'Deploy' section on the right shows a 'Pay-as-you-go deployment for Cohere-command-e-plus' with options to 'Overview', 'Pricing and terms', and 'Create a new project'.

## Discover

Find the best foundation model for your task and domain.

No more searching and downloading models from various sources.

## Fine-tune

Customize models to your specific needs.

Overcome the challenges of large-scale model training.

## Evaluate

Test the suitability and accuracy of models on your own data.

Eliminate the need to write code or set up infrastructure.

## Deploy

Operationalize models in your applications.

Don't worry about the scalability and reliability of inferencing endpoints.

# Azure AI Search (Azure Cognitive Search)

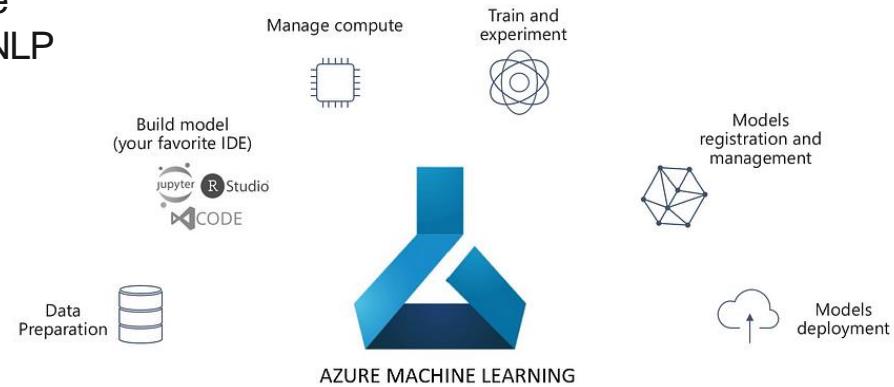
Provides secure information retrieval at scale over user-owned content in traditional and generative AI search applications.

- A search engine for vector search and full text and hybrid search over a search index
- Rich indexing with integrated data chunking and vectorization, lexical analysis for text, and optional applied AI for content extraction and transformation
- Rich query syntax for vector queries, text search, hybrid queries, fuzzy search, autocomplete, geo-search and others
- Relevance and query performance tuning with semantic ranking, scoring profiles, quantization for vector queries, and parameters for controlling query behaviors at runtime
- Azure integration at the data layer, machine learning layer, Azure AI services and Azure OpenAI

# Azure Machine Learning

Azure Machine Learning is a comprehensive machine learning platform that supports language model fine-tuning and deployment.

- Automated machine learning: Rapidly create accurate ML models for classification, regression, vision, and NLP
- Model catalog: Discover, fine-tune, and deploy foundation models using model catalog.
- Prompt flow: Design, construct, evaluate, and deploy language model workflows
- AI infrastructure: Purpose-built AI infrastructure



# More Azure AI Services

- **Azure Content Safety:**  
An AI service that detects unwanted contents (text and image)
- **Azure Document Intelligence:**  
Turn documents into intelligent data-driven solutions.
- **Azure Speech / Immersive Reader:**  
Speech to text, text to speech, translation, and speaker recognition
- **Azure Translator / Language:**  
translate more than 100 in-use, at-risk, and endangered languages and dialects.
- **Azure Vision & Custom Vision & Face:**  
Analyze content in images and videos, Detect and identify people and emotions in images
- **Azure Video Indexer:**  
Extract actionable insights from your videos
- **Azure Bot Service:**  
Create bots and connect them across channels.

# Azure AI Studio

A unified platform for developing and deploying generative AI apps responsibly



# Azure OpenAI Playground

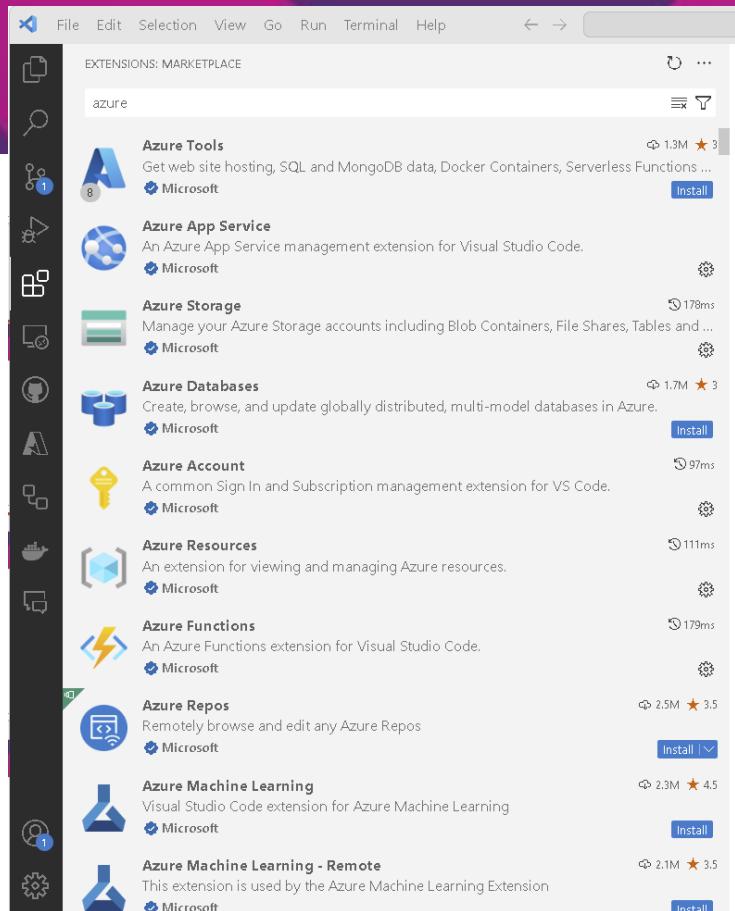
The screenshot shows the Azure OpenAI Playground interface. On the left, a sidebar lists navigation options: Azure OpenAI, Playground, Chat (highlighted with a red box), Completions, DALL-E (Preview), Management, Deployments, Models, Data files, Quotas, and Content filters (Preview). The main area is divided into several sections:

- System message**: Includes "Add your data (preview)", "Save changes", and a "Clear chat" button (highlighted with a red box).
- Specify how the chat should act**: A section for defining the AI's behavior, including a "Select a template" dropdown and a "System message" input field containing the placeholder "You are an AI assistant that helps people find information."
- Playground Settings**: Includes "View code", "Show raw JSON", and a volume icon.
- Deployment**: Shows "Deployment \* gpt35turboDply" (highlighted with a red box) and a dropdown menu.
- Session settings**: Includes "Past messages included" (set to 10), "Current token count" (11/4000), and an "Input tokens progress indicator".
- Start chatting**: A central area with a robot icon, instructions to "Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.", and a large input field for "Type user query here. (Shift + Enter for new line)". Below the input field are three small icons: a circular arrow, a downward arrow, and a right-pointing arrow.

# Visual Studio Code

Comprehensive and flexible environment that supports the entire lifecycle of AI app development, from coding and debugging to deployment and maintenance

- Lightweight IDE
- Extensions and Plugins
- Integration with AI and Cloud Services
- Jupyter Notebooks
- GitHub Co-pilot
- Version Control
- Code Snippets and Autocompletion
- Debugging Tools
- Data Visualization
- Customizable Workspace



# Azure OpenAI supported languages & libraries

Python

Libraries: openai, requests, http.client



JavaScript / Node.js

Libraries: openai, axios, node-fetch



C# (.NET)

Libraries: Azure.AI.OpenAI



Java

Libraries: azure-ai-openai, okhttp



Go

Libraries: openai-go, net/http



## LangChain

<https://python.langchain.com/v0.2/docs/integrations/platforms/microsoft/>



## Hugging Face

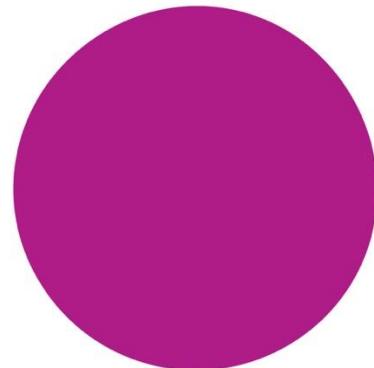
Open-source models from Hugging Face Hub to Azure Machine Learning

# What are we building today ?

Develop a powerful AI chatbot in one day, utilizing your own data, and leveraging Azure AI Services for scalable, enterprise-ready capabilities

- **Azure AI Services:** Leverage Azure Cognitive Services, including Azure OpenAI for natural language understanding and Cosmos DB (MongoDb) for vector search.
- **Bring Your Own Data:** Utilize your proprietary data sources to tailor the chatbot's responses, making it unique and relevant to your specific use case.
- **Retrieval-Augmented Generation (RAG):** Implement RAG to enhance the chatbot's responses by retrieving relevant information from your data before generating accurate, context-rich answers.
- **LangChain:** Seamlessly integrate with large language models (LLMs), enabling advanced conversational capabilities and chaining together complex tasks.

Be ambitious.



Tech check

# Tech check

## Development Environment

- Laptop / Internet / Wifi
- GitHub Account (Codespaces enabled)
- Visual Studio Code & Extensions
- Node.js / Azure CLI / Bicep

## Connections

- Azure OpenAI Subscription Key (shared)
- Cosmos DB connection string (shared)
- Azure Subscription (optional)

<https://github-insight-anz-lab.github.io/aiapp1day>



Docusaurus

# Share Idea and Experience



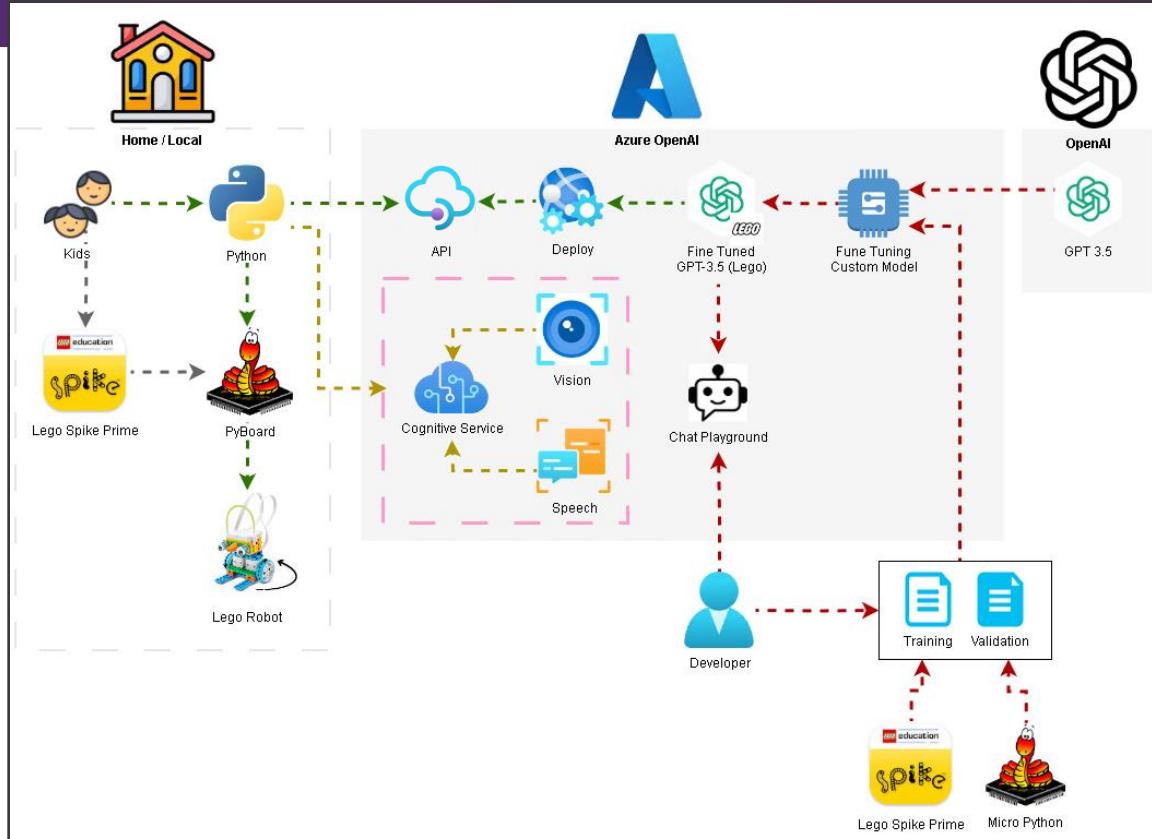
# Share Idea and Experience

## Lego Robot AI



**Lego + Robot + AI = Fun**

# Share Idea and Experience

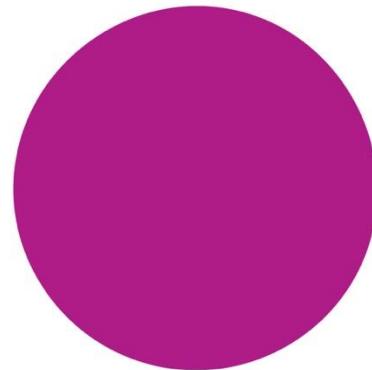


# Schedule

Time	Activity
08:30 (30m)	Presentation 1: AI Apps using Azure
09:00 (30m)	Tech check
09:30 (30m)	Morning break ☕
10:00 (2h)	Hands-on lab time (Lab 1)
12:00 (30m)	Lunch 🍕🍔🥗
12:30 (30m)	Presentation 2: RAG Pattern & Techniques
12:30 (3h)	Hands-on lab time (Lab 2)
	Afternoon break ☕ @ 14:00 (30m)
15:30 (15m)	Day review and wrap

You are here

Be ambitious.



Hands-on lab time - Lab 1

Morning

# Lab 1: Interact with Azure OpenAI models

**Time:** **2 hours**

**Topics :**

- Basic Prompting
- Prompt Engineering Techniques
- System Message
- Add External Knowledge
- Function Calling
- Create Images
- Using Azure OpenAI SDK
- Coding Challenge

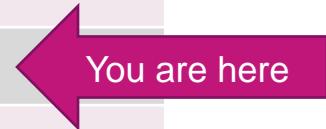
**YOU WILL NEED**

**OpenAI Key**

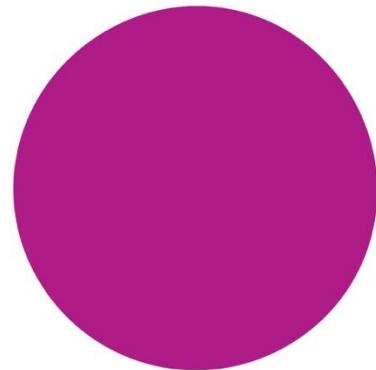


<https://github-insight-anz-lab.github.io/aiapp1day>

# Schedule

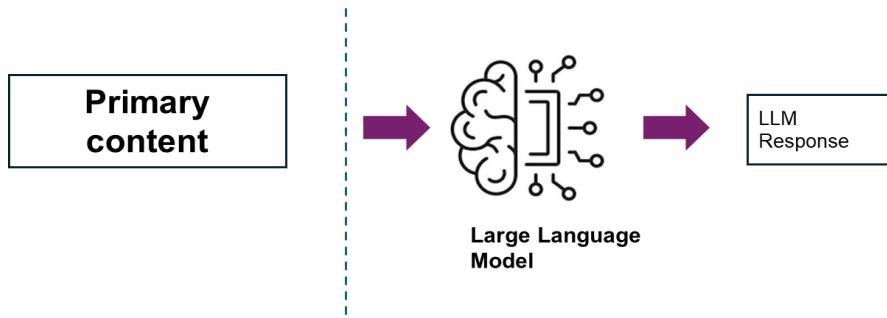
Time	Activity
08:30 (30m)	Presentation 1: AI Apps using Azure
09:00 (30m)	Tech check
09:30 (30m)	Morning break ☕
10:00 (2h)	Hands-on lab time (Lab 1)
12:00 (30m)	Lunch 🍕🍔🥗  You are here
12:30 (30m)	Presentation 2: RAG Pattern & Techniques
12:30 (3h)	Hands-on lab time (Lab 2)
	Afternoon break ☕ @ 14:00 (30m)
15:30 (15m)	Day review and wrap

Be ambitious.



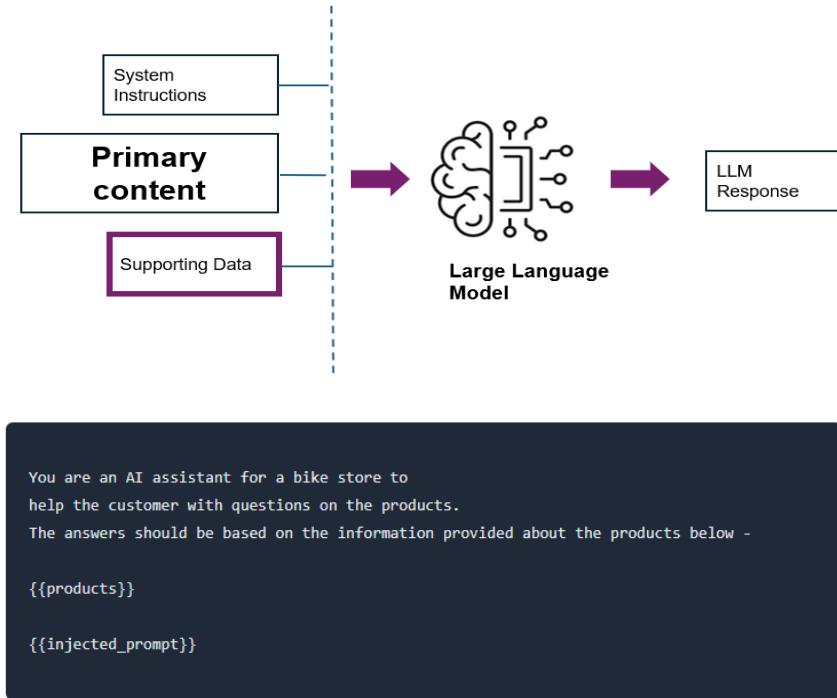
RAG Pattern & Techniques

# Limitations of LLM Models



- **Indeterministic** : they can generate different outputs for the same input. This can lead to inconsistency and unpredictability in the results.
- **Hallucination**: LLM models can generate texts that are not based on facts or logic, but on their own learned biases and assumptions. This can lead to false or misleading information in the outputs.
- **Context** : LLM models are trained on public data and are not updated with the latest data and events. They may also lack the specific context or domain knowledge that your application needs.

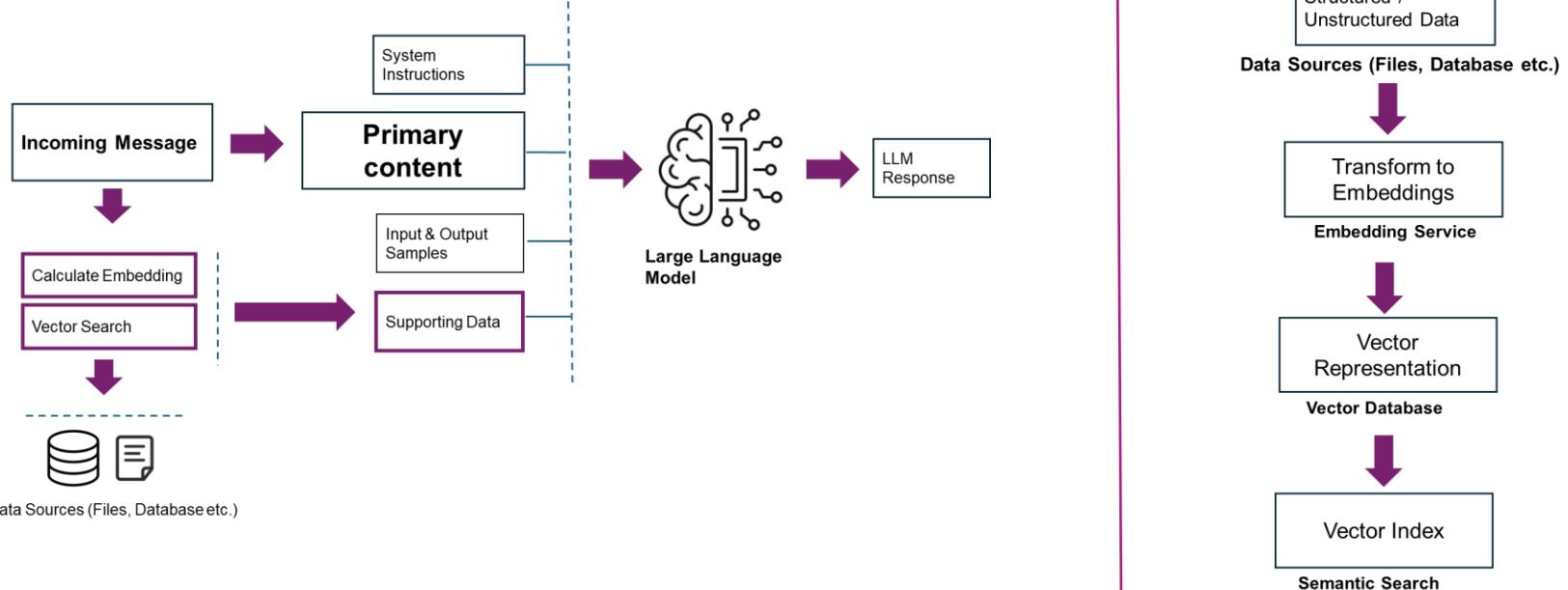
# Retrieval Augmented Generation – Deep Dive



RAG combines generative models with retrieval mechanisms to improve the quality and relevance of generated responses.

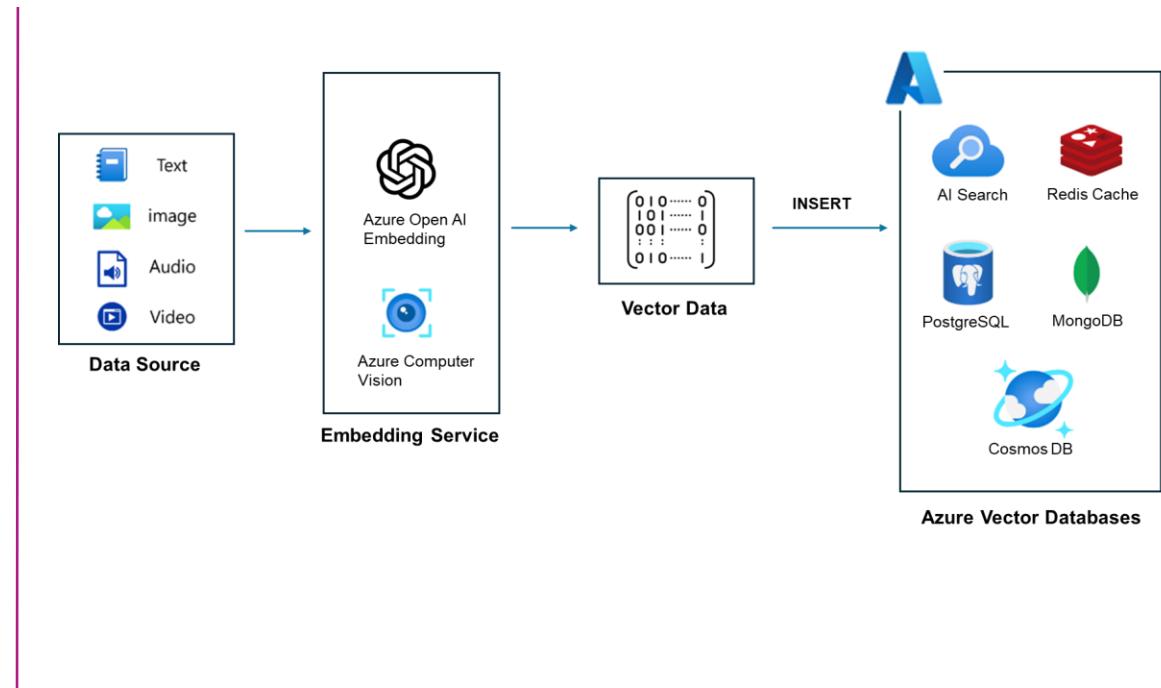
- **Retrieval Stage :** Relevant documents or pieces of information are retrieved from a knowledge base or database.
- **Generation Stage :** The retrieved documents are used as context for the generative model. The generative model produces a response based on both the input query and the retrieved documents.

# Retrieval Augmented Generation Contd.



# Vectorize text and images in Azure

- Azure offers support for multiple vector databases, including Azure Cache for Redis, Azure AI Search, PostgreSQL, Azure Cosmos, and MongoDB, providing efficient storage and querying capabilities.
- **Embedding Service** transforms text data into vector format.
- **Computer Vision Service** encodes the content and context of an image into vector format.



# Demo Azure AI Studio

The screenshot shows the Azure AI Studio interface. At the top, there are tabs for Work, Azure AI Studio, Home - Azure AI Studio, ExtractFromText - regulation-in, and MyCity Chatbot. The main navigation bar includes Home, Explore, Build, and Manage.

The Build section is active, showing a project named "mcasalaina-nous". The toolbar includes Manual evaluation, Deploy, Import, Export, View Code, Prompt samples, and Open in prompt flow.

The System message pane contains a message: "You are an AI assistant that helps people find information about Hyundai cars." Below it is a Variables section with an "Add variable" button.

The Chat pane displays a conversation:

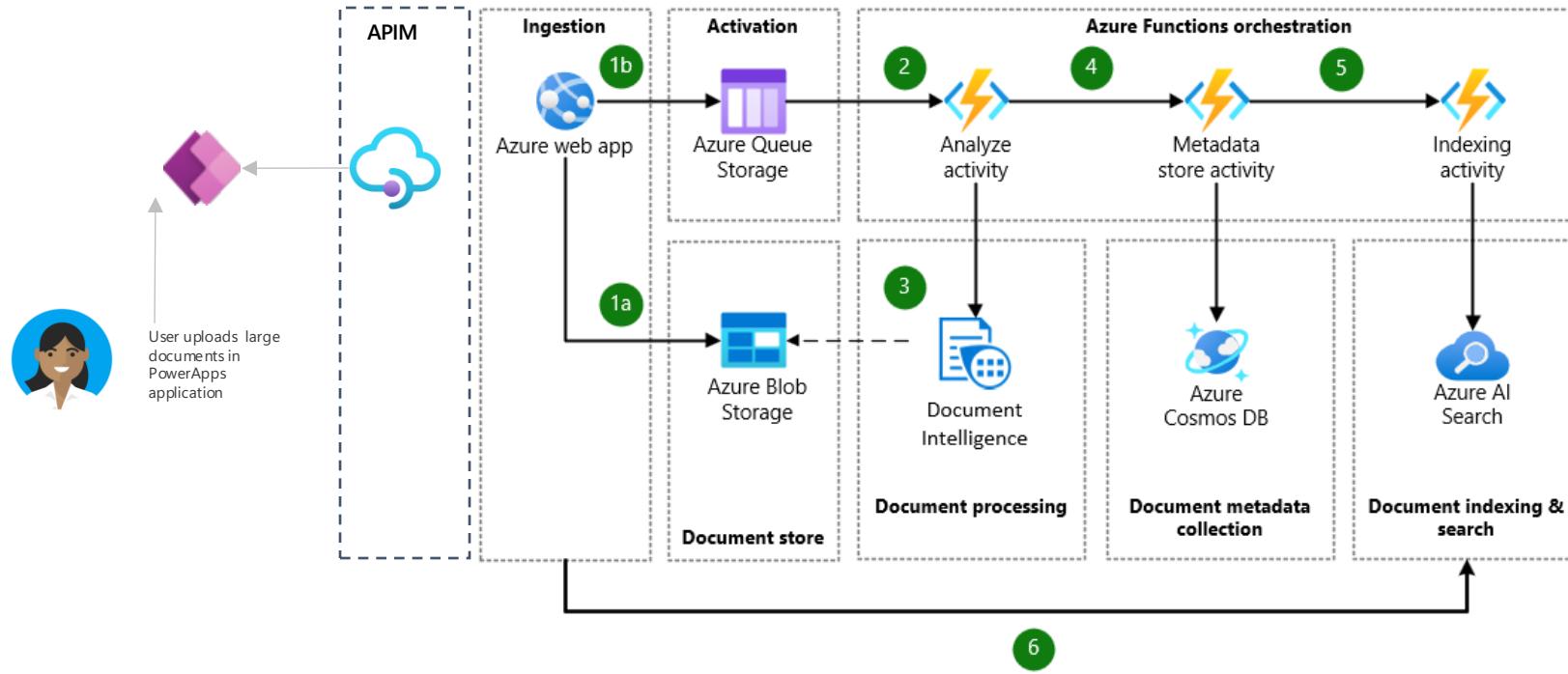
- User: How do I activate the heated seats in a Hyundai Ioniq 6?
- AI Assistant: I'm sorry, as of my last update in 2021, there's no specific model called Hyundai Ioniq 6. Hyundai Ioniq is available in models like Ioniq Hybrid, Ioniq Plug-In Hybrid, and Ioniq Electric. However, the process to activate heated seats in Hyundai vehicles is usually quite similar. Here are general steps:
  1. Locate the buttons for the seat heaters. These are typically located in the center console area, often near the controls for the air conditioning and heating.
  2. Press the button for the seat you want to heat. The driver's seat and front passenger seat usually have their own separate controls.
  3. The button usually has an indicator light that will illuminate to let you know the seat

The Deployment pane shows "Deployment" set to "GPT-4". The Session settings pane indicates "Past messages included" at 10.

# Document Processing

This technical pattern enables users to classify, extract, summarize, and gain deeper insights from documents using natural language prompts with high accuracy and limited-to-no machine learning (ML) training.

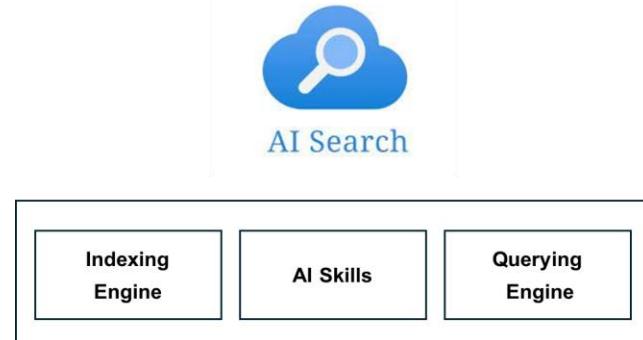
# Document identification and classification



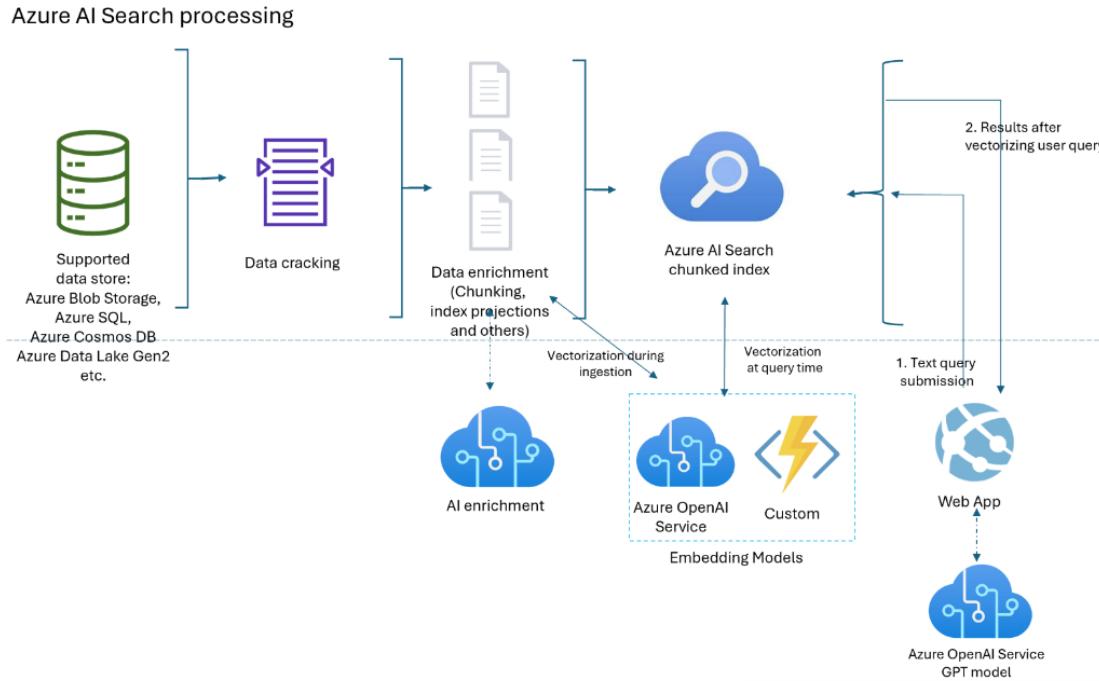
# Azure AI Search (Azure Cognitive Search)

**Azure AI Search** provides secure information retrieval at scale over user-owned content in traditional and generative AI search applications.

- Consolidate heterogeneous content into a search index composed of vectors and text.
- Supports full text, vector and hybrid search over a search index.
- Relevance and query performance tuning with semantic ranking, scoring profiles
- Integrate data chunking and vectorization for generative AI and RAG apps.
- Indexing through AI skills – Transform large unstructured text or image files into searchable chunks
- Apply granular access control at the document level.
- Enterprise-ready - scalability, security and compliance



# Azure AI Search – Integrated Vectorization



**Integrated vectorization** is an extension of the indexing and query pipelines in Azure AI Search.

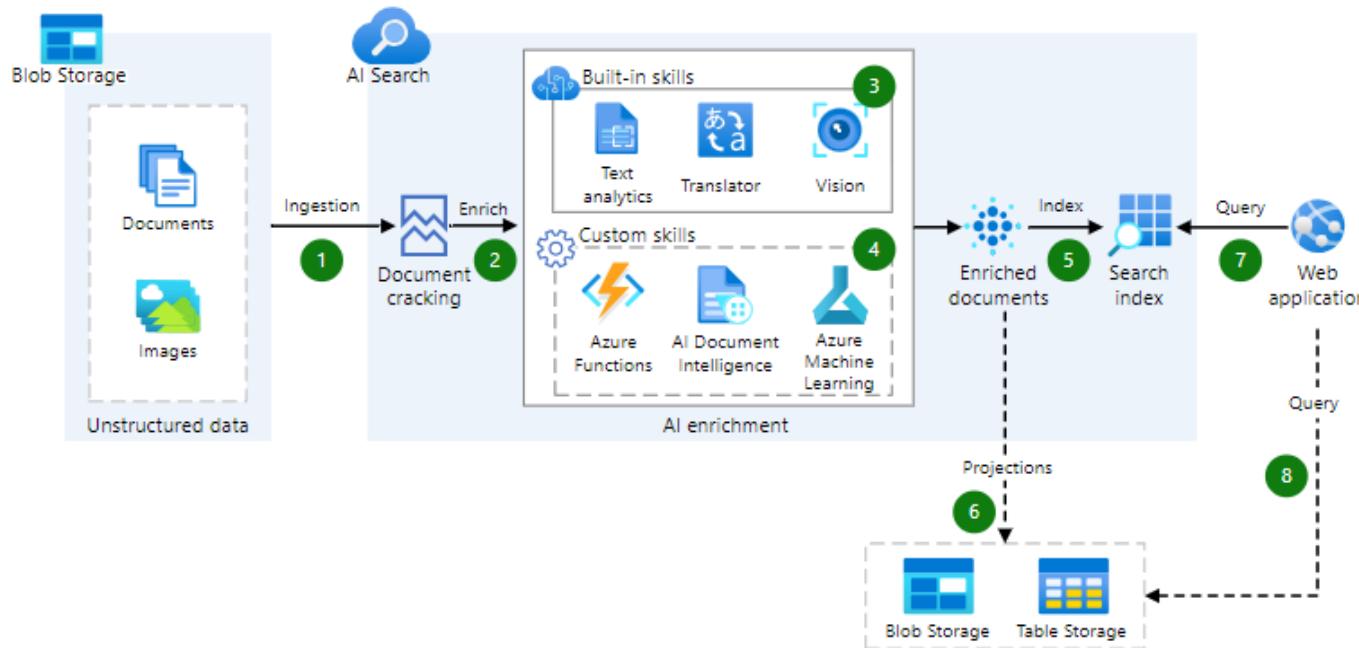
- Data chunking during indexing
- Text-to-vector conversion during indexing
- Text-to-vector conversion during queries

# Azure AI Search – AI Skills

A set of skills that prepare a document for indexing, calling either built-in AI search functions or custom code.

- **Entity Recognition:** Identifies and extracts entities such as names, dates, and locations from text.
- **Text Translation:** Translates text from one language to another
- **Text Analytics:** Analyzes text to extract key phrases, sentiments, and language information.
- **OCR (Optical Character Recognition):** Converts images of text into machine-readable text.
- **Image Analysis:** Analyzes images to extract information such as objects, faces, and text contained within them.
- **Sentiment Analysis:** Evaluates text to determine the sentiment or emotional tone, such as positive, neutral, or negative.
- **Language Detection:** Identifies the language of the provided text.
- **PII (Personally Identifiable Information) Detection:** Identifies and masks sensitive information like credit card numbers and social security numbers.
- **Form Recognizer:** Extracts text, key-value pairs, and tables from documents, such as forms and receipts.
- **Custom Skill:** Allows users to define and implement their own custom processing logic.

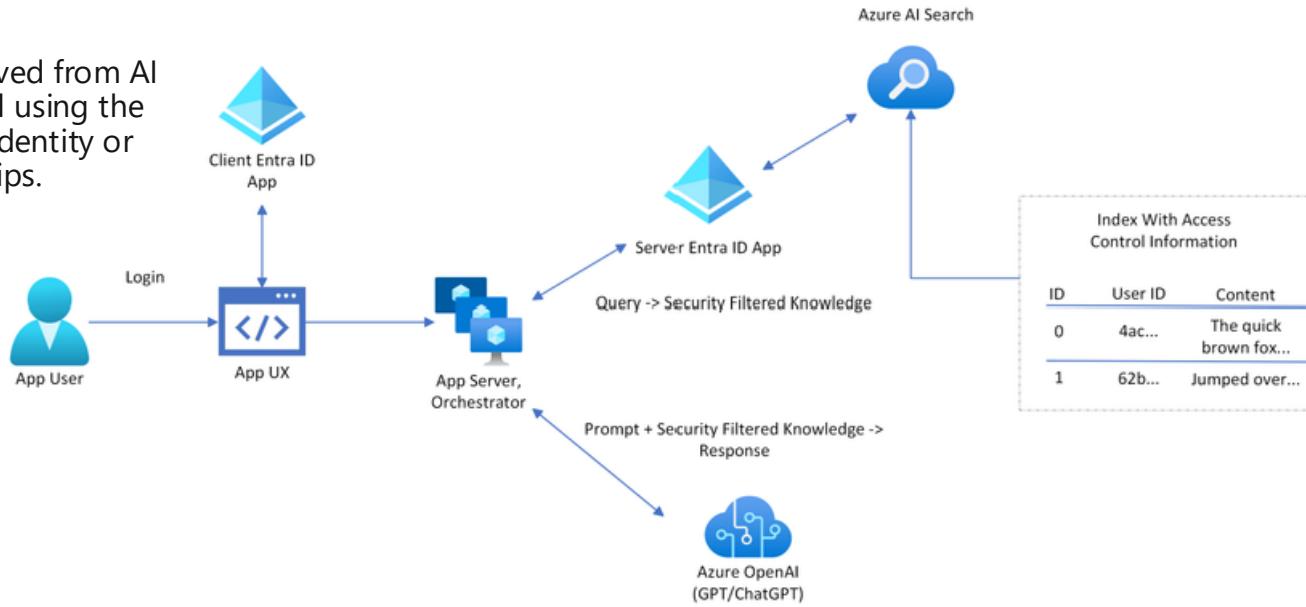
# Azure AI Search – Image & Text Processing



<https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/ai-search-skillsets>

# Azure AI Search – Data Security

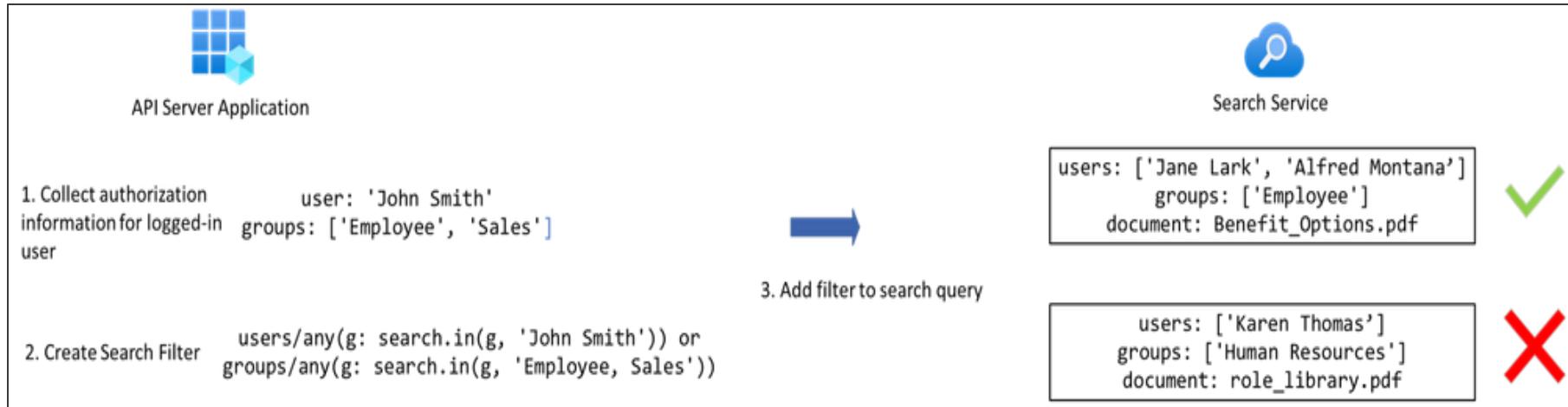
Documents retrieved from AI Search are filtered using the logged-in user's identity or group memberships.



<https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/access-control-in-generative-ai-applications-with-azure-ai/ba-p/3956408>

# Azure AI Search – Data Security Contd.

1. The user's identity is extracted from the token claims.
2. The API Server applies a filter containing the authorization information to the query sent to AI Search.



# Cosmos DB or Azure AI Search as Vector Database



**Azure Cosmos DB** is preferred when -

- you have structured or semi-structured data operational data (chat history, customer profile) in that database.
- simplified architecture with vector similarity search inline with database queries.



**Azure AI Search** is preferred when -

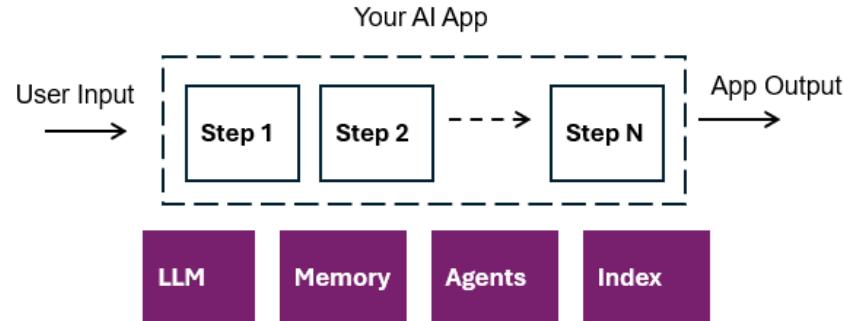
- you need to index structured/unstructured (e.g. images, docx, PDF) from a variety of internal and external data sources
- high quality search results with hybrid full-text / vector search semantic ranking etc.
- the workload requires multi-modal embeddings to perform OCR, image analysis and translation.
- you are building **Bing** like search experience in a custom application

# Workflow Orchestration in AI Apps



*I like to place an order for the Trek bicycle.  
Can you confirm the availability and the price?  
And what is the delivery time?*

- Find the product in product catalog that matches the user query.
- Check the availability and price of the product.
- Check the delivery time for the product based on the location of the user.



# Schedule

Time	Activity
08:30 (30m)	Presentation 1: AI Apps using Azure
09:00 (30m)	Tech check
09:30 (30m)	Morning break ☕
10:00 (2h)	Hands-on lab time (Lab 1)
12:00 (30m)	Lunch 🍕🍔🥗
12:30 (30m)	Presentation 2: RAG Pattern & Techniques
12:30 (3h)	Hands-on lab time (Lab 2)
	Afternoon break ☕ @ 14:00 (30m)
15:30 (15m)	Day review and wrap

You are here

# Be ambitious.

## Hands-on lab time – Lab 2

Afternoon

# Lab 2: Building a Chatbot using RAG

**Time:** 2 hours

**Topics :**

- Azure Deployment (optional)
- Load Product Catalog
- Vector search using text embeddings
- Workflow Orchestration using LangChain
- Build a Chatbot Backend
- Build a Chatbot Frontend
- App Deployment
- Coding Challenge

**YOU WILL NEED**

**OpenAI Key**  
**OpenAI Endpoint**  
**Cosmos DB Login**



<https://github-insight-anz-lab.github.io/aiapp1day>



**Docusaurus**

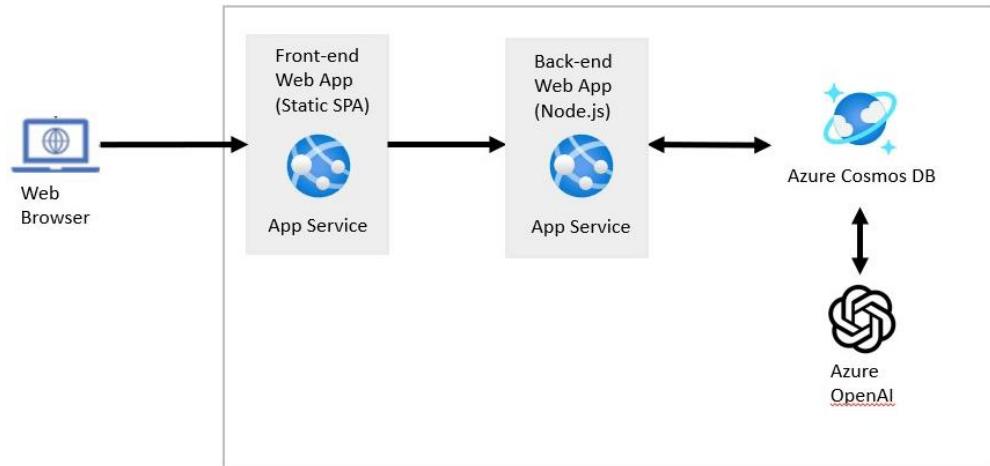
# Lab 2 : Azure Deployment (Optional)

## Shared Resources

Azure OpenAI Subscription Key

Cosmos DB connection string

- Cosmos DB instance is shared
- You will use your own database



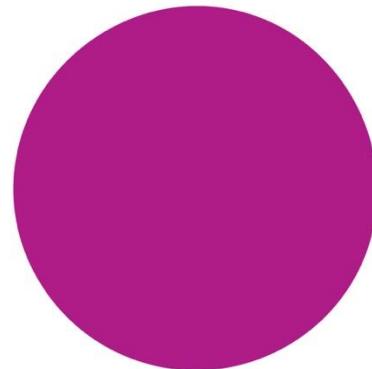
## Optional

Azure Subscription

- Bicep script is provided to deploy all services
- You can skip Bicep if use above shared resources

Source: <https://learn.microsoft.com/dotnet/ai/get-started-app-chat-template>

# Be ambitious.



## Wrap up

# AI App using Azure

## Create AI App power by Azure AI services

- Accelerate app creation with wide range of AI offerings
- Level up your app game and improve your process

## Deploy High-Quality AI Models as APIs

- Integrate industry-leading AI models into apps effortlessly
- Enhance user experiences with powerful Generative models

## Build and Train Models

- Rapid development with preferred tools and frameworks
- Wide range of models can be trained and fine-tuned easily

## Get Your AI App Live

- Leverage Azure's scalable and reliable services and infrastructure
- Automate processes with Azure services, and get time back in your day

# LLM Ops for AI App

## Idea Generation

- Define the problem
- Develop a prototype

## Technique Selection

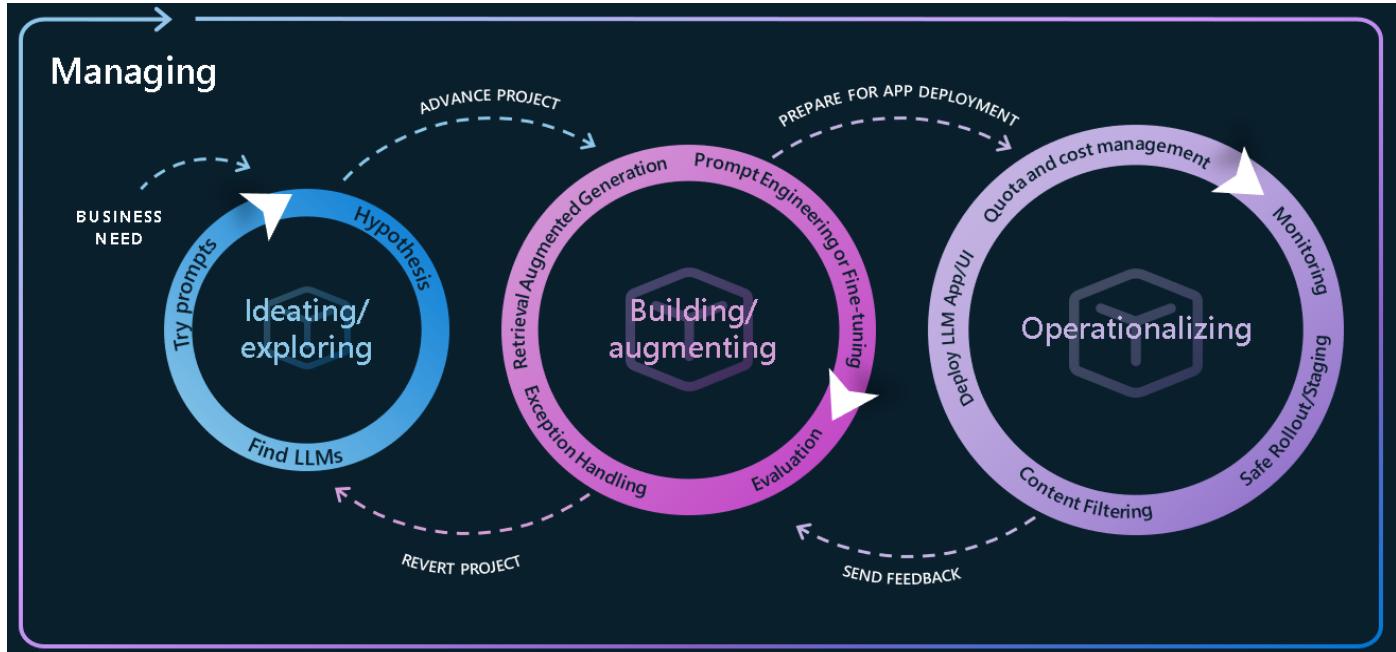
- Model selection
- Dataset Preparation
- Evaluation Metrics

## Operationalization

- Integration & Deployment
- Monitoring & Maintenance

## Continuous Improvement

- Feedback Loop
- Model Updates
- Adaptation & Optimization



# Empowering Developers

## AI enhances productivity

AI automates repetitive tasks, allowing us to focus on creative problem-solving and high-level design.

## Augmenting Human Efforts

AI enhances productivity by handling time-consuming tasks, allowing you to focus on more complex and innovative aspects of the work.

## Human expertise is irreplaceable

AI lacks the ability to understand business context, creativity, and nuanced decision-making.

## Collaboration between AI and developers:

Assisting with coding, debugging, and testing, but developers remain in control of the final product.

**AI + Developers = A Powerful Team**

# When devs are in the flow you get to market fast and scale smoothly

## The Azure AI difference:

- Code-first prompting tool
- Project starter templates
- Seamless integration with open-source tools
- Powerful tracing



# The threat landscape is more sophisticated than ever

## New protections:

- Identify new threats fast
- Safety system reinforcement
- Enhanced model security

Indirect prompt injection attacks

Jailbreak attacks

Protect Inputs

Compliant Outputs



# Responsible tooling for confidence



## Multimodal filtering

Scan videos, images and text to identify, block and monitor **harmful content**



## Customized systems

Create blocklists and **custom categories** to block entire topics, not just specific words



## Prompt shielding

Identify and mitigate prompts that could expose you to **prompt injection attacks**



## Protected materials

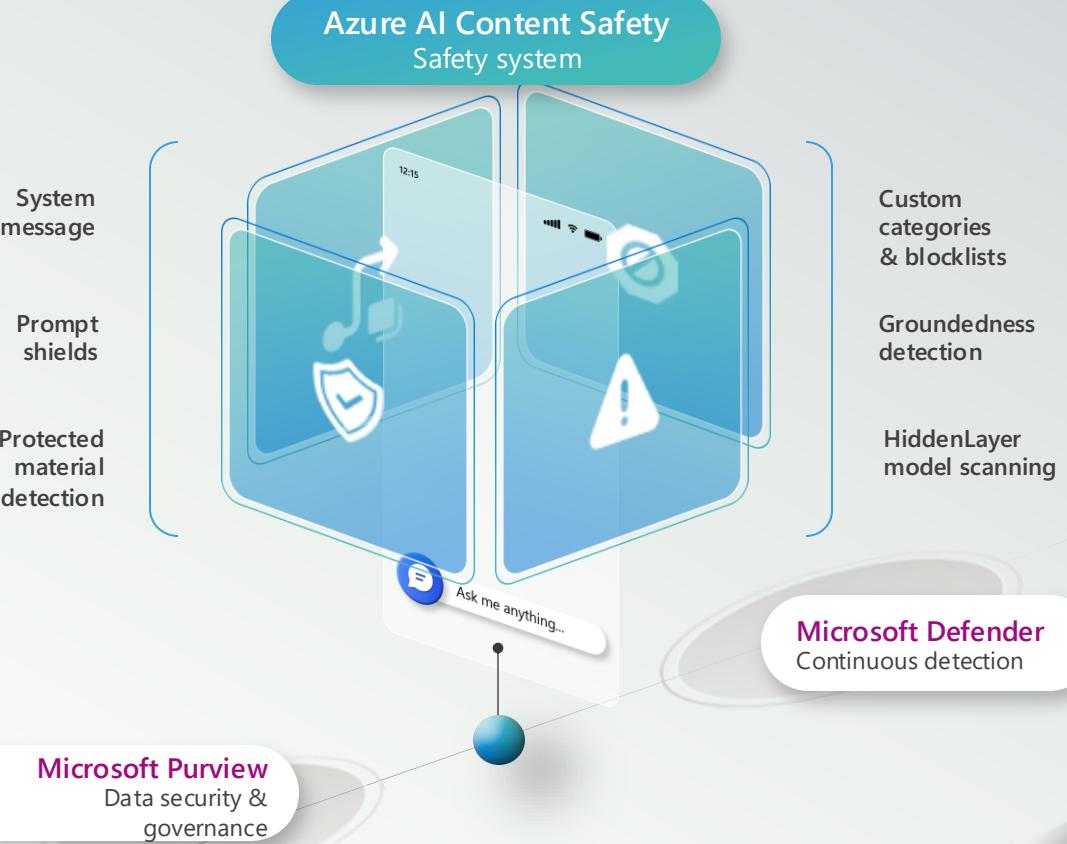
Avoid outputting **known or owned text** content with protected materials detection



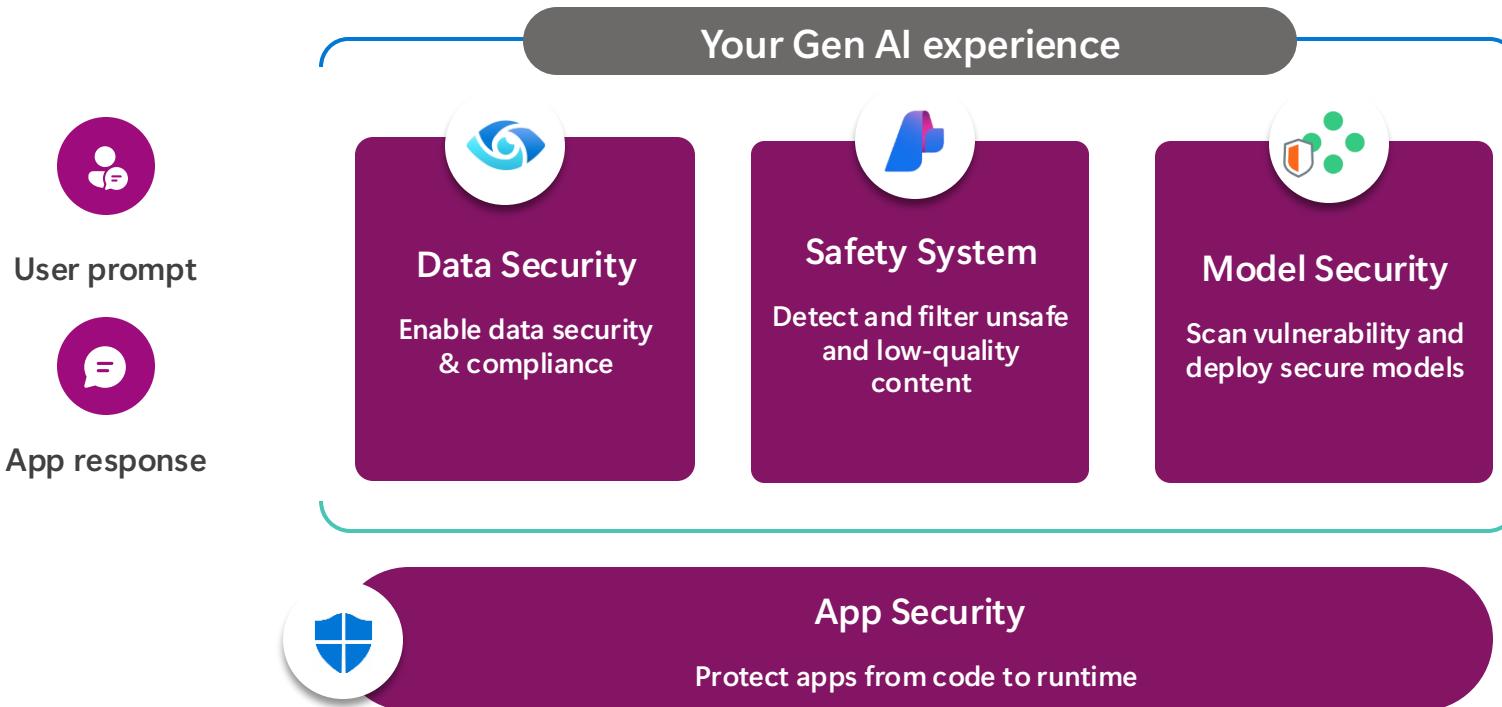
# Drive innovation with a built-in, comprehensive safety system

## The Azure AI difference:

- Security and governance
- HiddenLayer model scanning
- Copyright commitment



# Navigate the latest threat landscape with comprehensive **safety, trust and security**



# Responsible AI - Content Safety

The screenshot shows the Azure AI Studio Preview interface with the 'Content Safety' tab selected. The main content area displays information about Content Safety, including sections for 'Try out the capabilities' (Moderate text content, Moderate image content, Moderate multimodal content) and 'Explore safety solutions for Gen-AI' (Jailbreak risk detection, Protected material detection, Safety metaprompt). A sidebar on the left lists various AI categories, and a navigation bar at the top includes links for 'Explore', 'Build', and 'Manage'.

**Content Safety**

Azure AI Content Safety detects harmful user-generated and AI-generated content in applications and services. Content Safety includes text and image APIs that allow you to detect material that is harmful.

### Try out the capabilities

**Moderate text content**

Run moderation tests on text contents. Assess the test results with detected severities. Experiment with different threshold levels.

[Try it out](#)

**Moderate image content**

Run moderation tests on image contents. Assess the test results with detected severities. Experiment with different threshold levels.

[Try it out](#)

**Moderate multimodal content**

Run moderation tests on image and text combined contents. Assess the test results with detected severities. Experiment with different threshold levels.

[Try it out](#)

### Explore safety solutions for Gen-AI

**Jailbreak risk detection**

Use jailbreak risk detection to identify potential jailbreak attack from user in the Gen-AI scenario.

**Protected material detection**

Use protected material detection to detect and protect third-party text material in LLM output.

**Safety metaprompt**

Use the framework of metaprompt that helps you potentially mitigate different types of harm.

**ON THIS PAGE**

- Introduction
- Try out the capabilities
- Explore safety solutions for Gen-AI

**Quick links**

- Documentation
- Explore code samples
- Watch a video
- Microsoft Q&A

# Secure, Compliant, and Trustworthy AI

## Data Security on OpenAI Service with Azure

- **Comprehensive Data Protection:** Azure AI services prioritize data security, ensuring that your information remains confidential and protected.
- **Controlled Access:** Employs a range of security measures, such as encryption in transit and at rest, to safeguard data against threats.
- **Enterprise-Grade Security:** provides industry-leading security features ensuring that your data remains safe throughout processing and storage.
- **Data Privacy & Control:** Does not train models on your data, meaning your business information stays private.
- **Compliance with Global Standards**
- **Identity and Access Management**

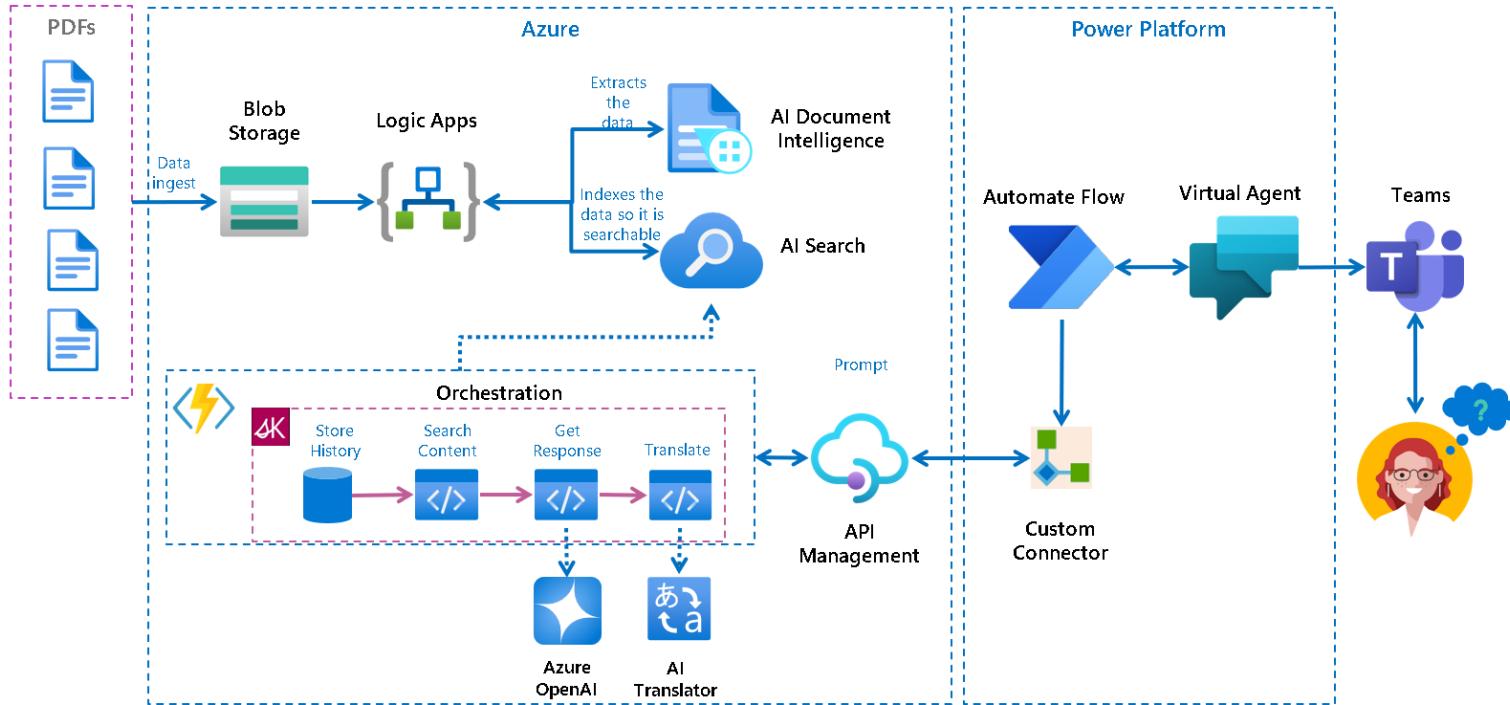
### Important

Your prompts (inputs) and completions (outputs), your embeddings, and your training data:

- are NOT available to other customers.
- are NOT available to OpenAI.
- are NOT used to improve OpenAI models.
- are NOT used to train, retrain, or improve Azure OpenAI Service foundation models.
- are NOT used to improve any Microsoft or 3rd party products or services without your permission or instruction.
- Your fine-tuned Azure OpenAI models are available exclusively for your use.

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

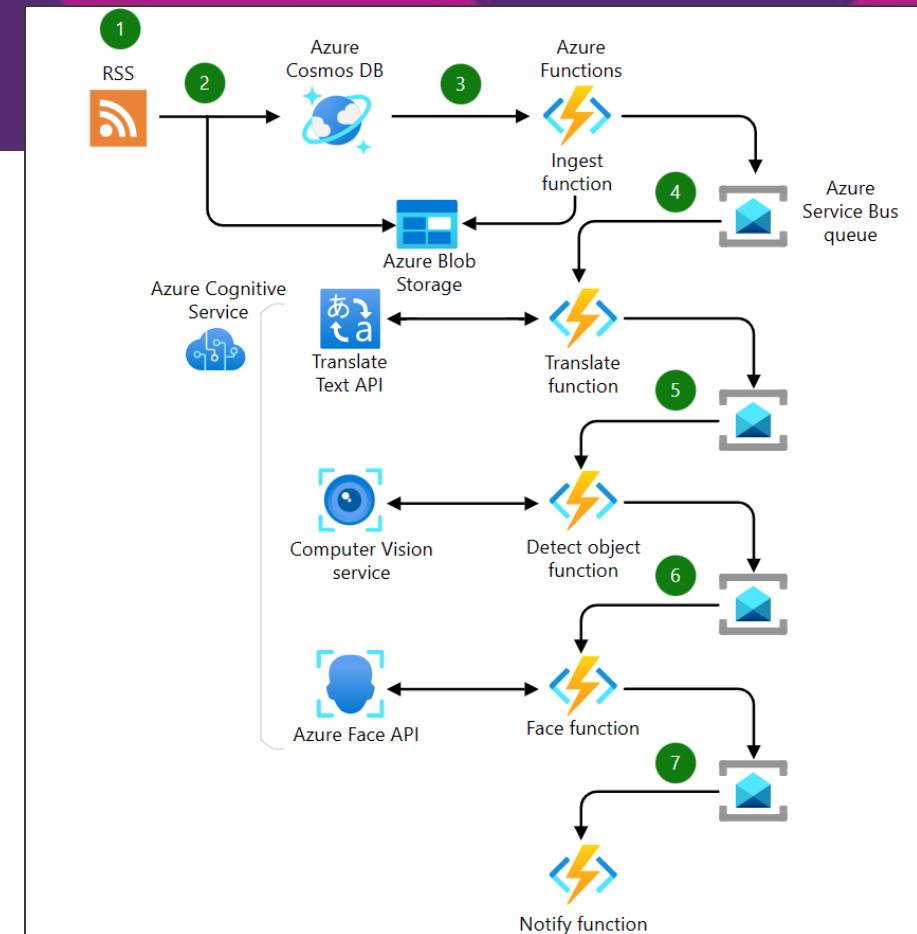
# Intelligent Q&A Chatbots



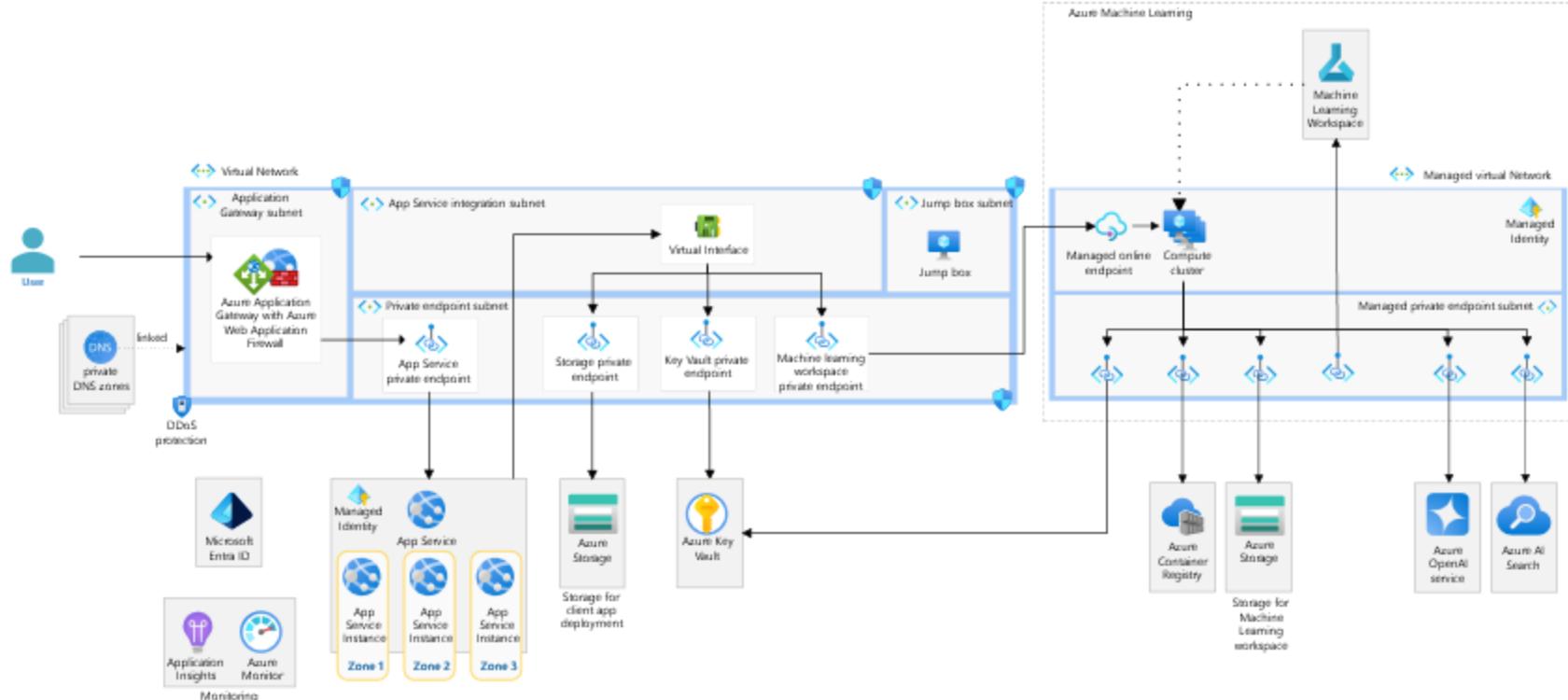
# Multiple Azure AI Services

Analyze news feeds with near real-time analytics using image and natural language processing

Source: <https://learn.microsoft.com/en-us/azure/architecture/example-scenario/ai/news-feed-ingestion-and-near-real-time-analysis>



# Enterprise chat using your data



Source: <https://learn.microsoft.com/azure/architecture/ai-ml/architecture/baseline-openai-e2e-chat>

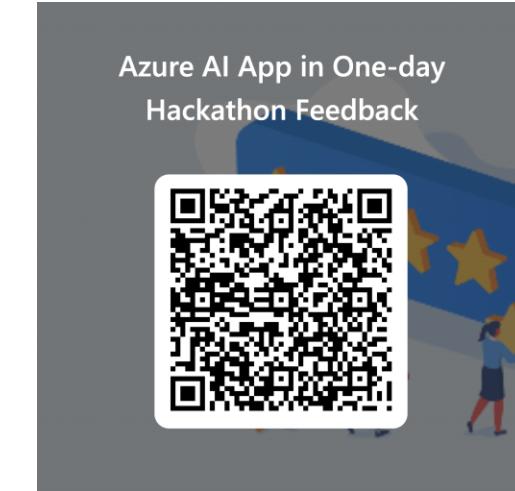
# Hackathon in review

## Presentations:

- AI App on Azure & Techniques
- Azure's broad set of AI services
- Supporting capabilities and tools
- RAG Pattern & Techniques
- AI App solutions design

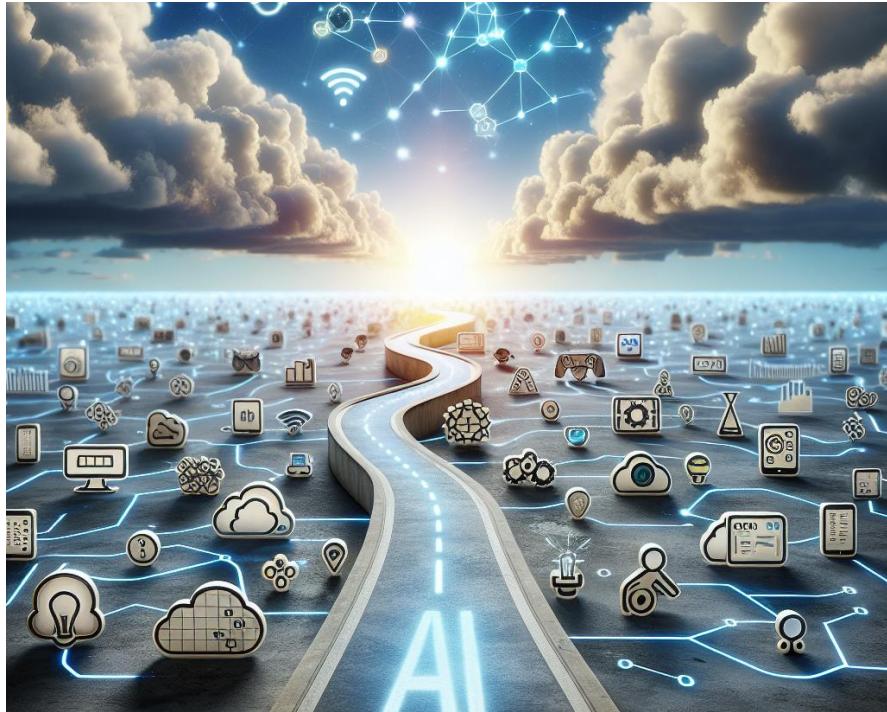
## Hands-on Labs:

- Lab 1: Interact with OpenAI Models
- Lab 2: Building a Chatbot using RAG



<https://forms.office.com/r/CefZaKwv6f>

# Next steps



## Sign-up Azure Subscription

- Utilize Free Azure Credits

## Microsoft Learn

- Engage in AI learning paths and modules
- Build foundational knowledge in AI and Azure services

## Applied Skills Certifications

- Search (Applied Skill & AI) in Microsoft Learn ([link](#))

## Microsoft Certification

- Microsoft Certified: Azure AI Fundamentals (AI-900)
- Microsoft Certified: Azure AI Engineer Associate (AI-102)

## Hands-on Exercise

- Sample repos on GitHub
- Build POC APPs

# References

The lab & tutorials are modified based on below Microsoft repos:

- Azure OpenAI Cosmos DB hackathon
  - <https://github.com/AzureCosmosDB/Azure-OpenAI-Node.js-Developer-Guide>
  - <https://github.com/AzureCosmosDB/Azure-OpenAI-Developer-Guide-Front-End>
- Azure OpenAI Proxy
  - <https://github.com/microsoft/Workshop-Interact-with-OpenAI-models/>

Other references:

- <https://azure.github.io/responsible-ai-hub/docs/content-safety-overview>
- <https://learn.microsoft.com/en-us/azure/architecture/ai-ml/architecture/baseline-openai-e2e-chat>



# Be ambitious.



## Thank you