Arizona Cloud Innovation Summit



RAG & Prompt Engineering in Azure OpenAI

Priyanshi Singh
Chris Baudo

04.22.2025

Microsoft

Speakers



Priyanshi Singh

AI Technical Specialist, Microsoft
US State and Local Government
priyanshi.singh@microsoft.com

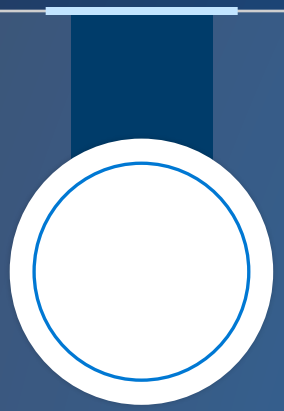


Chris Baudo

AI Technical Specialist, Microsoft
US State and Local Government
chris.baudo@microsoft.com

Agenda

- AI Foundry
- Retrieval Augmented Generation (RAG)
- Prompt Engineering
- Q&A



AI Foundry




Microsoft
Azure

Microsoft AI Options

Out-of-the-box AI

Just activate
Unlock productivity
Persona-based Copilots

M365 Copilots

M365 Copilot

M365 Copilot Chat

Users Business Units SOC IT

Other Copilots

Dynamics 365 Copilots

GitHub Copilot

Security Copilot

Composable AI

Deploy in days
No code to Pro code
Democratize innovation

Makers

Copilot Studio

Power Platform Copilot
Fabric Copilot

AI Builder

Dataverse

Developers

Semantic Kernel SDK

Azure Open AI service

GitHub Copilot

Azure Application Platform

Model-as-a-Service

Cognitive Services

Pretrained Models

Build-your-own AI

Tailored models
Develop monetizable IP
Maximize data science skills

Data Scientists

Azure Machine Learning

Azure DataBricks

Azure Prompt Flow

Microsoft Fabric

Model Catalog

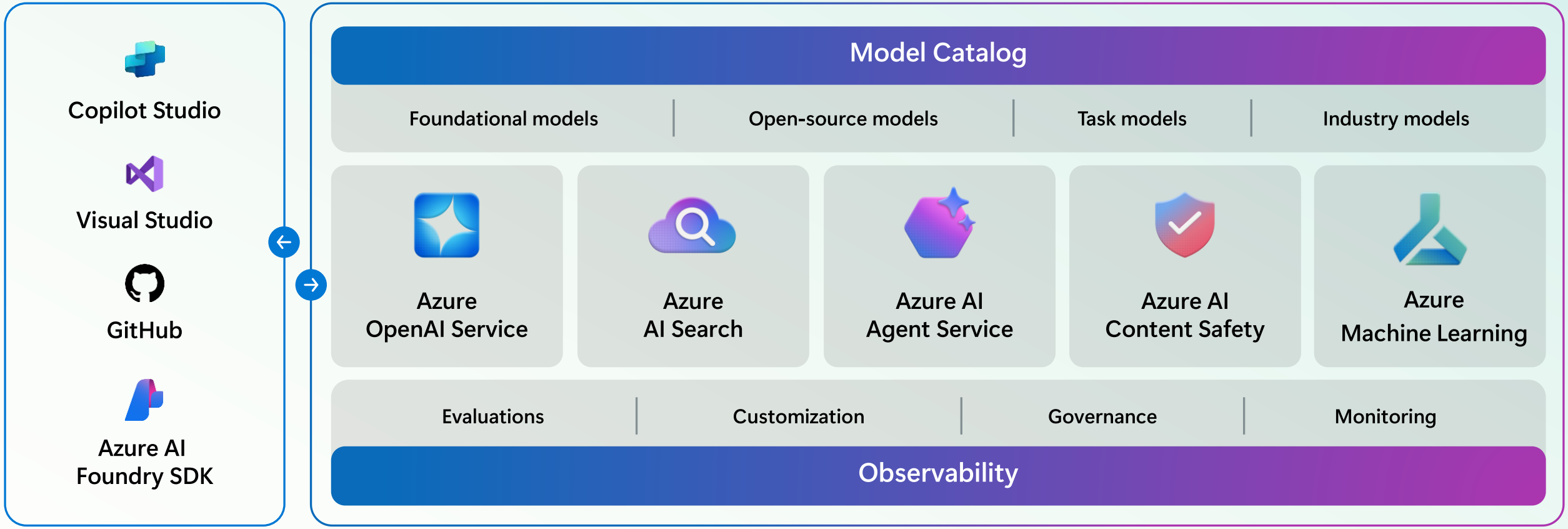
Fine Tuning

Azure AI Infrastructure

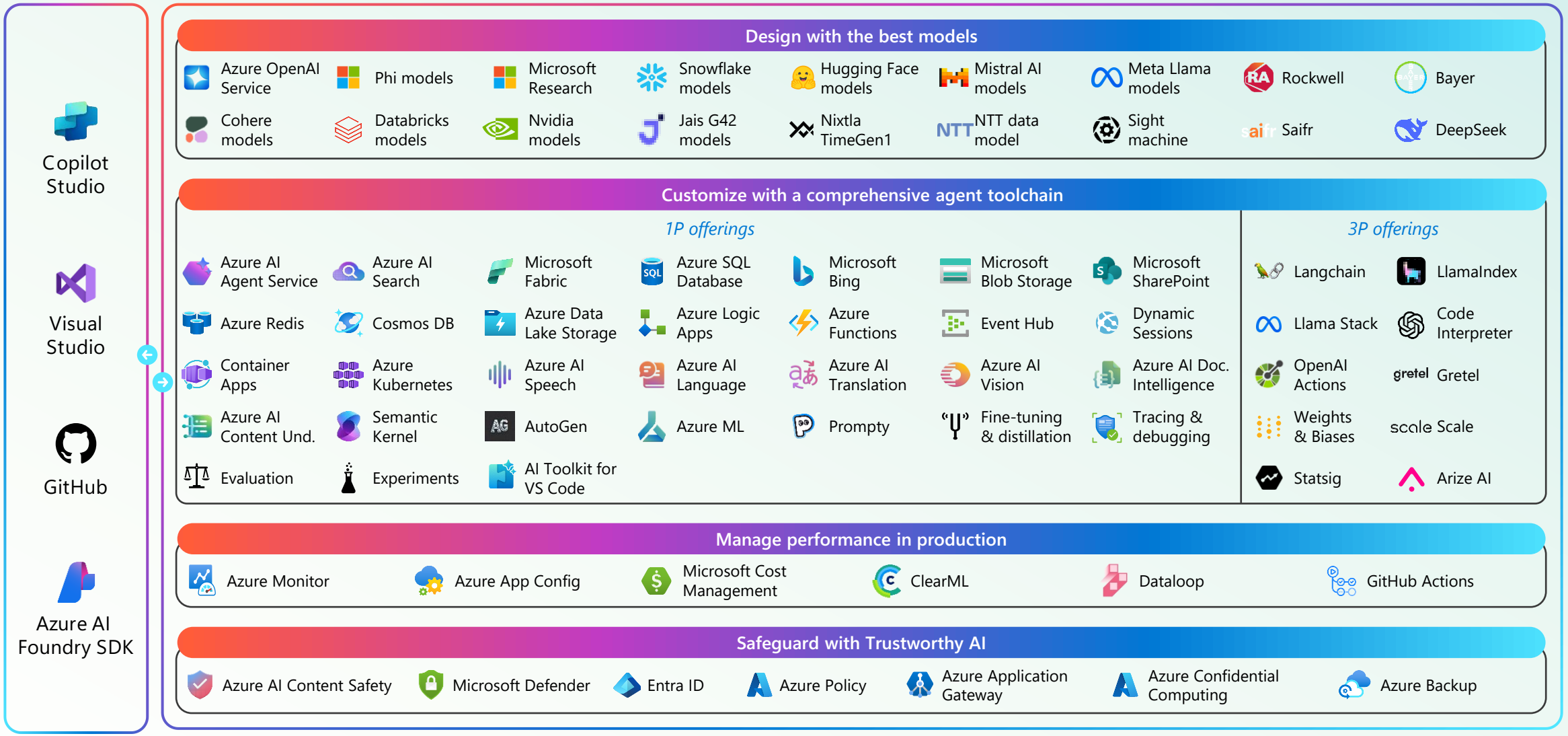
Time to value

Differentiation

Azure AI Foundry



Azure AI Foundry ecosystem



Retrieval Augmented Generation (RAG)

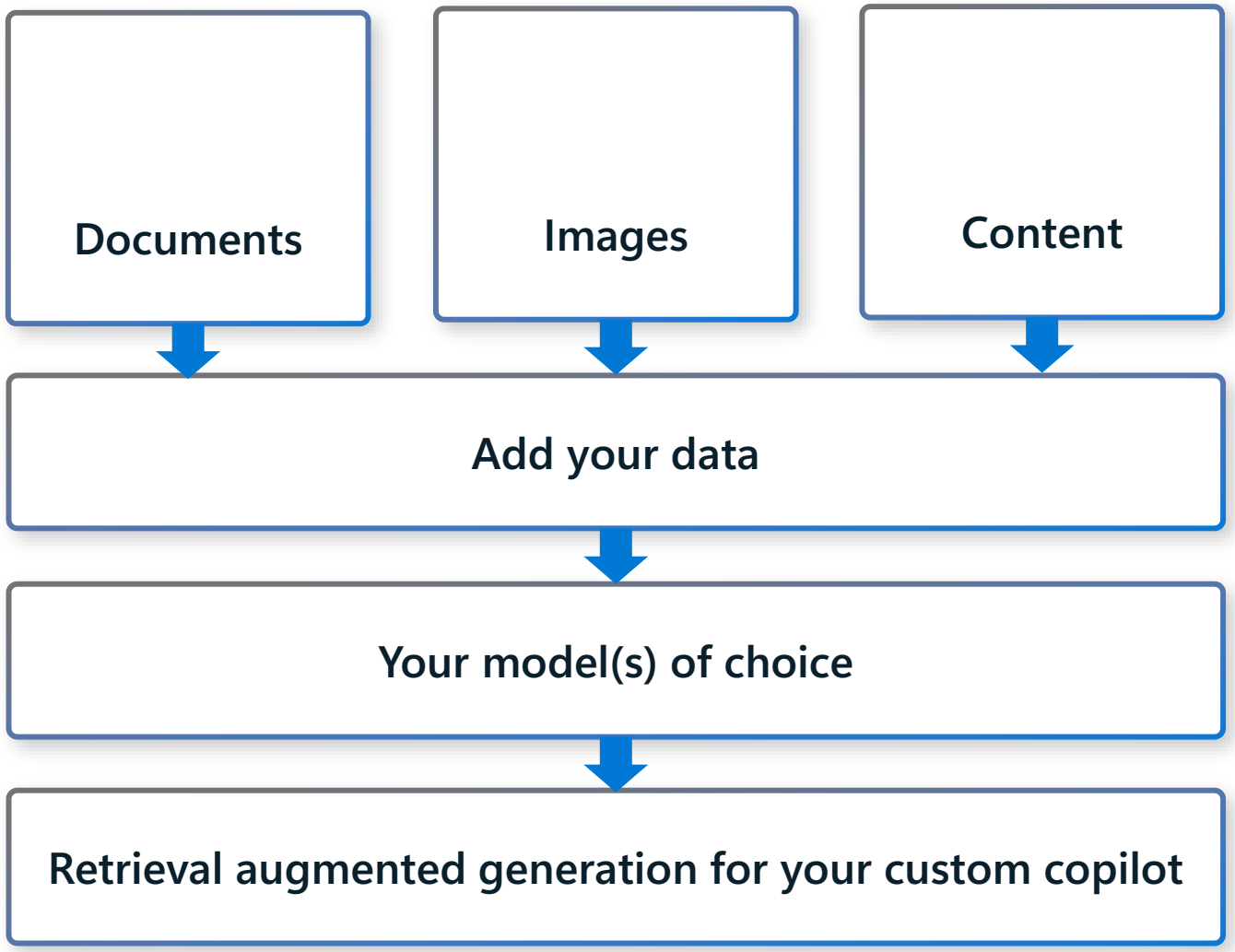
Definitions

Retrieval Augmented Generation (RAG) improves LLM performance by retrieving data from external sources and incorporating it into a prompt. RAG allows businesses to achieve customized solutions while maintaining data relevance and optimizing costs.

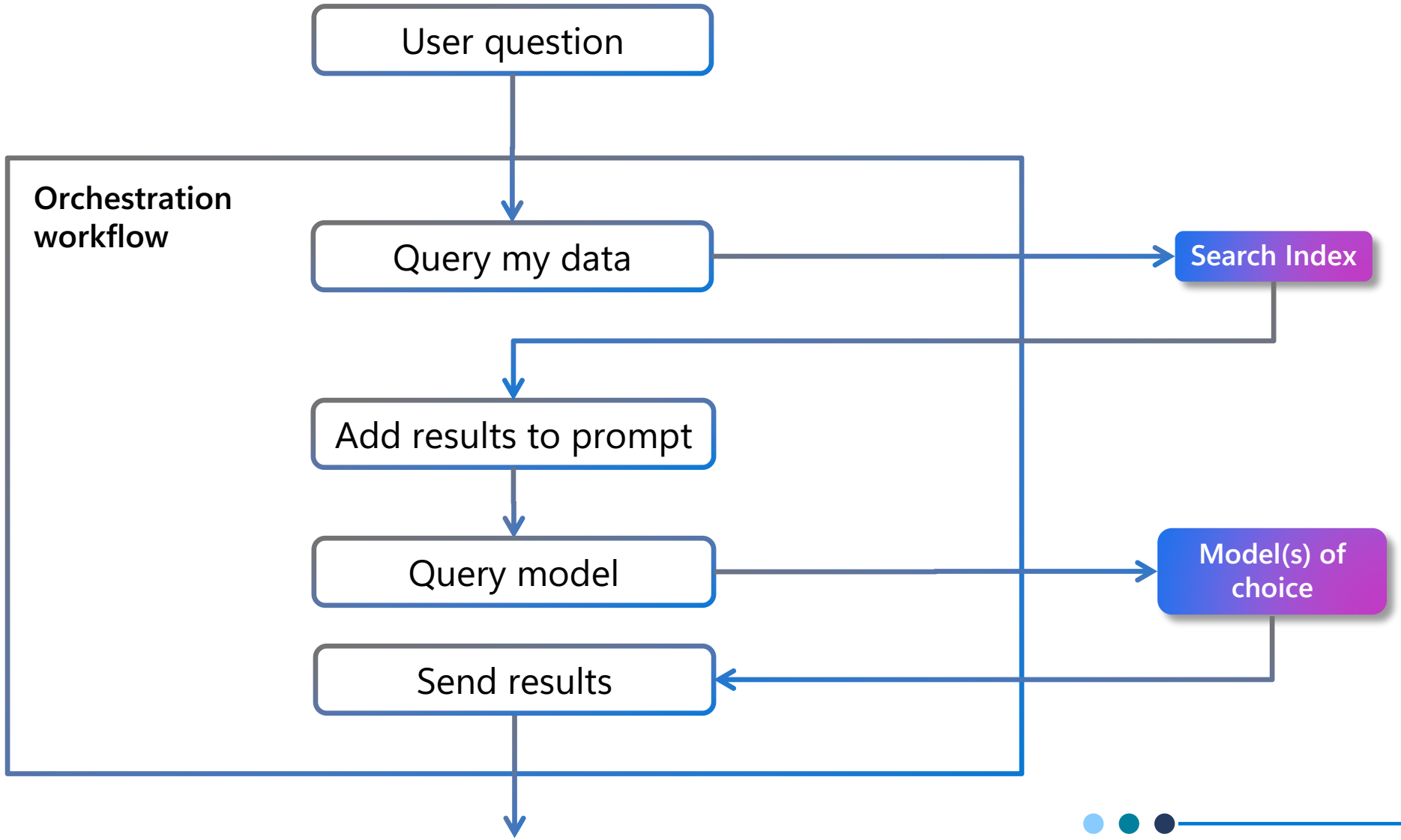
Prompt Engineering is a technique that involves designing prompts for natural language processing models. This process improves accuracy and relevancy in responses, optimizing the performance of the model.

Fine-tuning retrains an existing LLM using example data, resulting in a new "custom" LLM that has been optimized for the provided examples.

Retrieval Augmented Generation



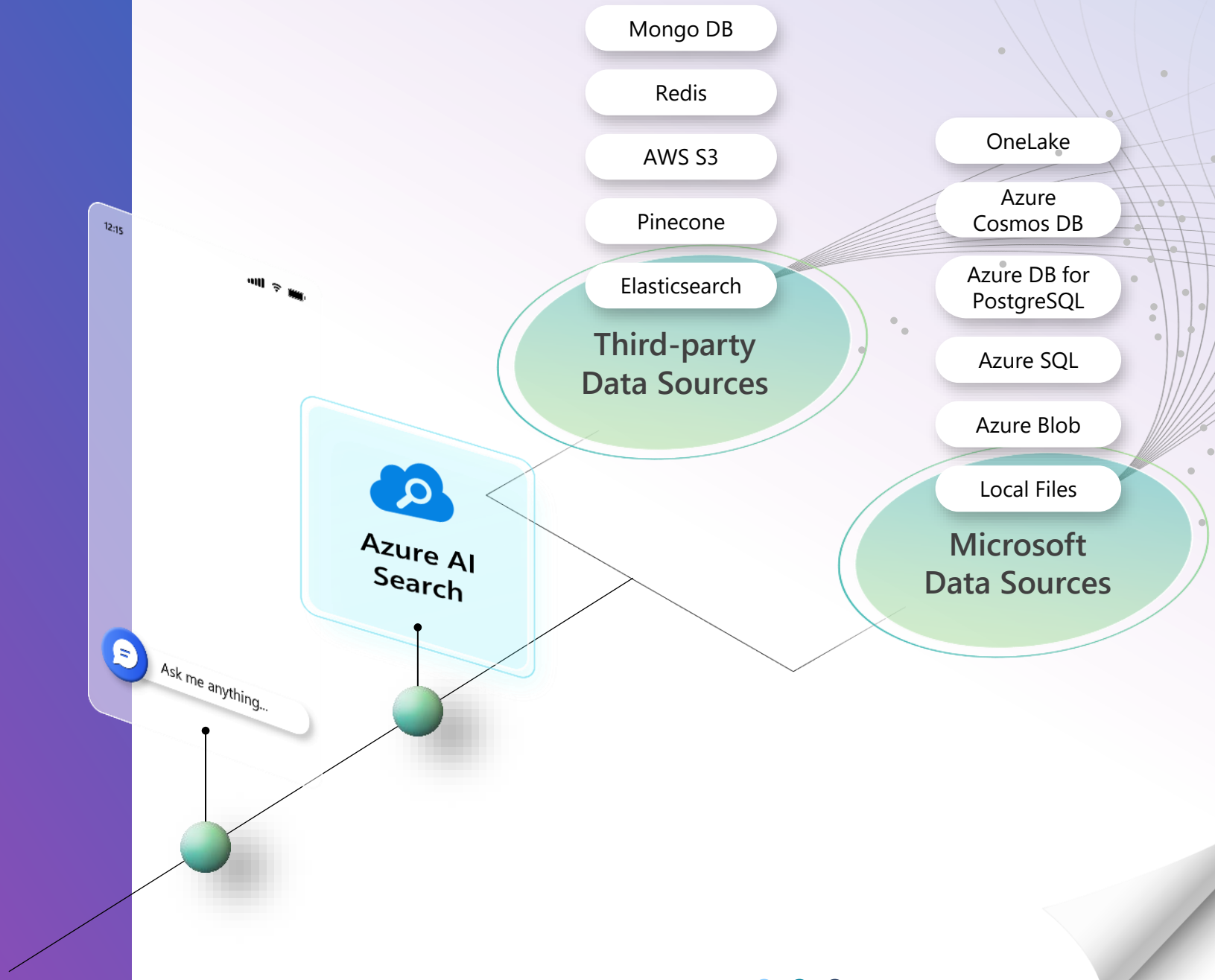
Retrieval Augmented Generation Orchestration



Wherever your data resides, you can bring it to life through Azure AI Search

The Azure AI Foundry difference:

- State-of-the-art ranking
- Automatic indexing
- All data welcome

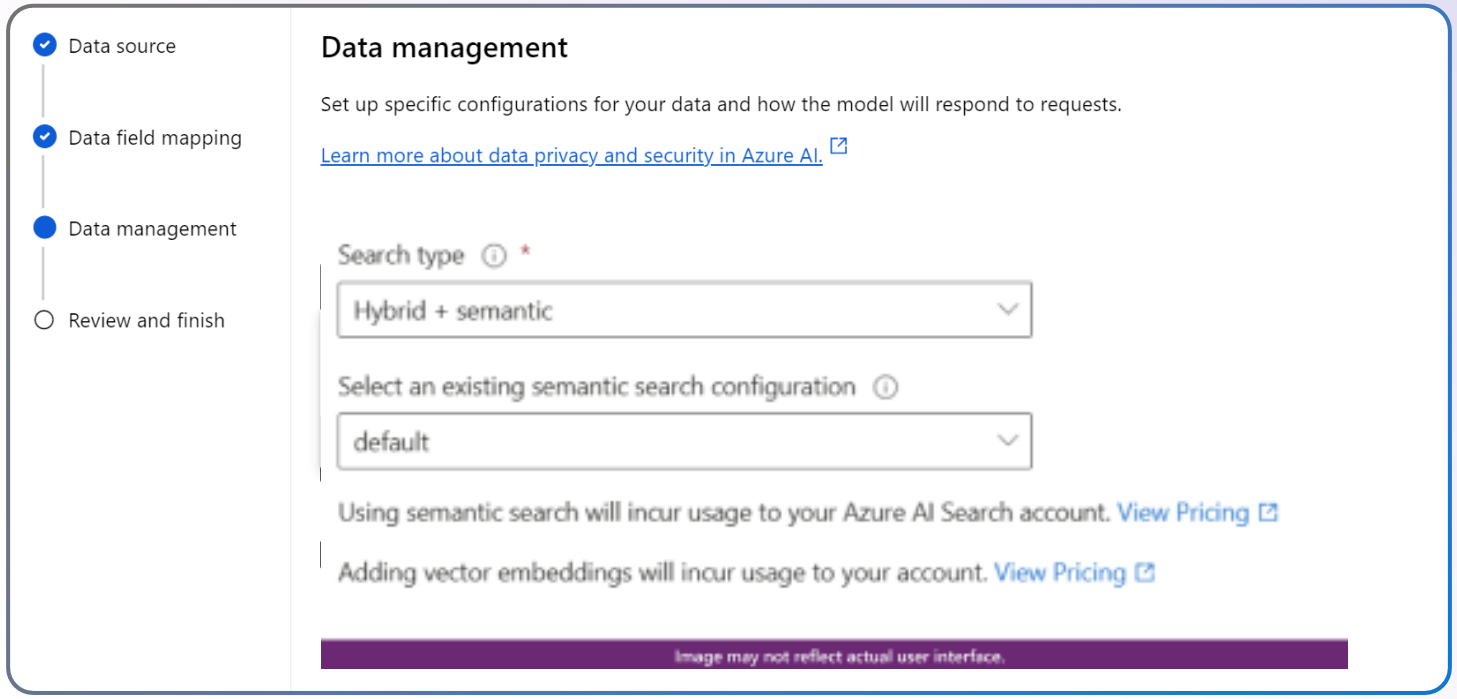


Azure AI Search

Azure AI Search is a battle-tested retrieval system purpose-built to deliver superior retrieval augmented generation (RAG). Use AI to collect, enrich and surface all aspects of your data.

- Ingest and search across all your data, no matter the type, volume or source
- Streamlined RAG pipeline, from import to chunking to embedding.
- Enterprise-ready foundation cost-optimized for performance and scale.
- Enable hybrid search (vector + keyword) results using RRF (Reciprocal Rank Fusion) and multi-vector retrieval for multi-modal search
- Improve retrieved results immediately with reranking and query rewriting.
- Tailor RAG experiences by utilizing customizable features such as granular access control and relevance tuning

Customize your models



Easily enable search best practices through the UI: like being able to select hybrid search + reranking in one click.

Streamlined RAG data pipeline

End to end, integrated data prep and processing



Ingest

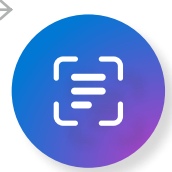
- OneLake
- Blob Storage
- ADLSv2
- SQL DB
- CosmosDB

\+ Incremental change tracking



Extract

- PDFs, Office docs
- Image files
- Nested images in PDFs, Office docs
- Document layout JSON, CSVs, markdown parsing



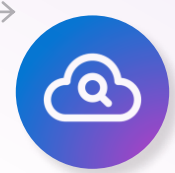
Chunk

- Split text into passages (for text portions):
- Fixed-size
- AI Document Intelligence layout output
- Propagate document metadata



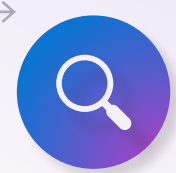
Embed

- Image → vector
- Text → vector
- Embeddings from OpenAI, AI Vision, and model catalog



Index

- Document index
- Chunk index
- Both



Query

- Boosting
- Weighting
- Thresholding



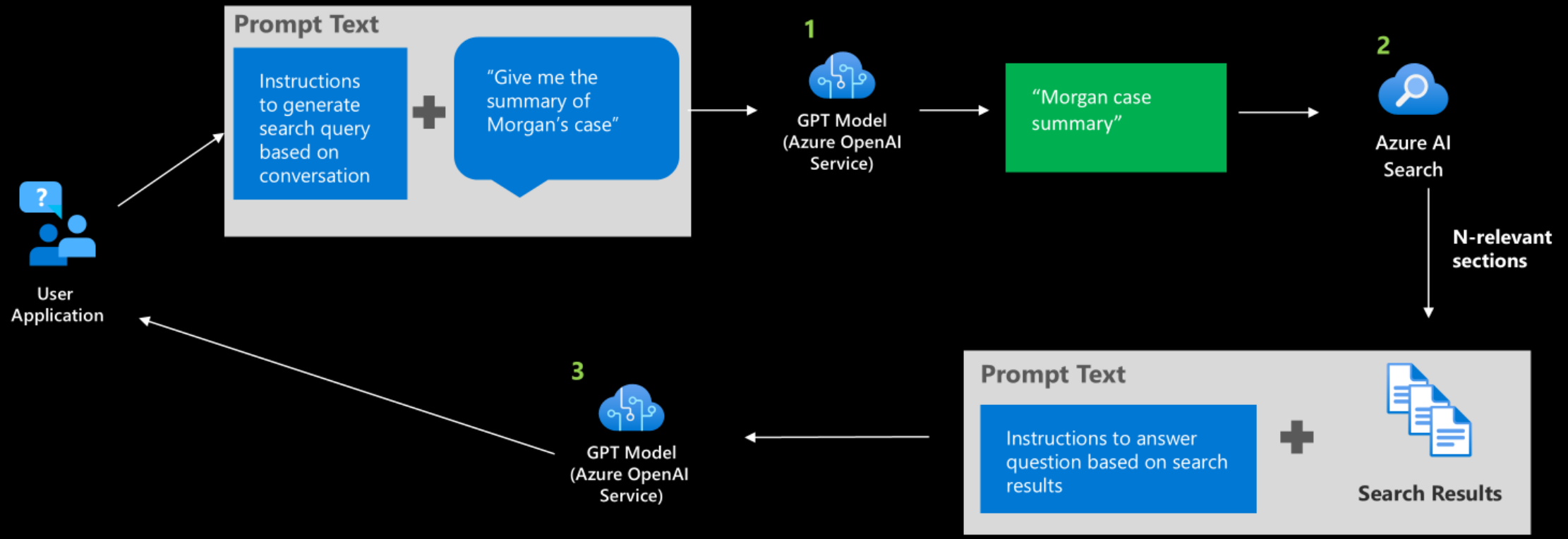
AOAI + AI Search Toolkit

Demo

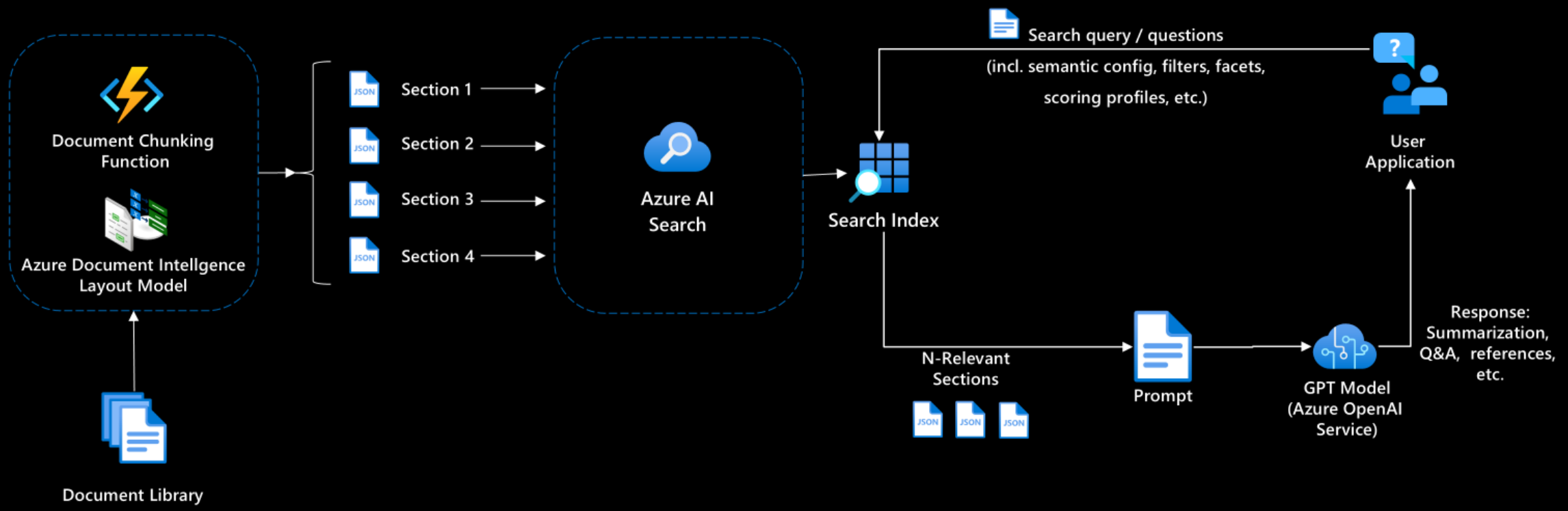
Step 1 : GPT model generates an optimized keyword search query based on the chat history and the last question

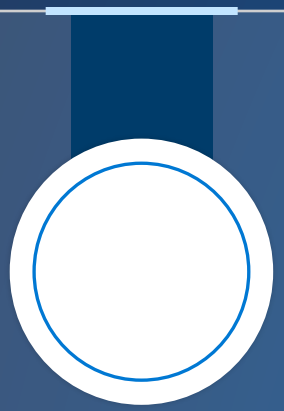
Step 2 : Retrieve relevant documents from the search index with the GPT optimized query

Step 3 : GPT model generates a contextual and content specific answer using the search results and chat history



AI-Powered Semantic Q&A over Enterprise Data Sources



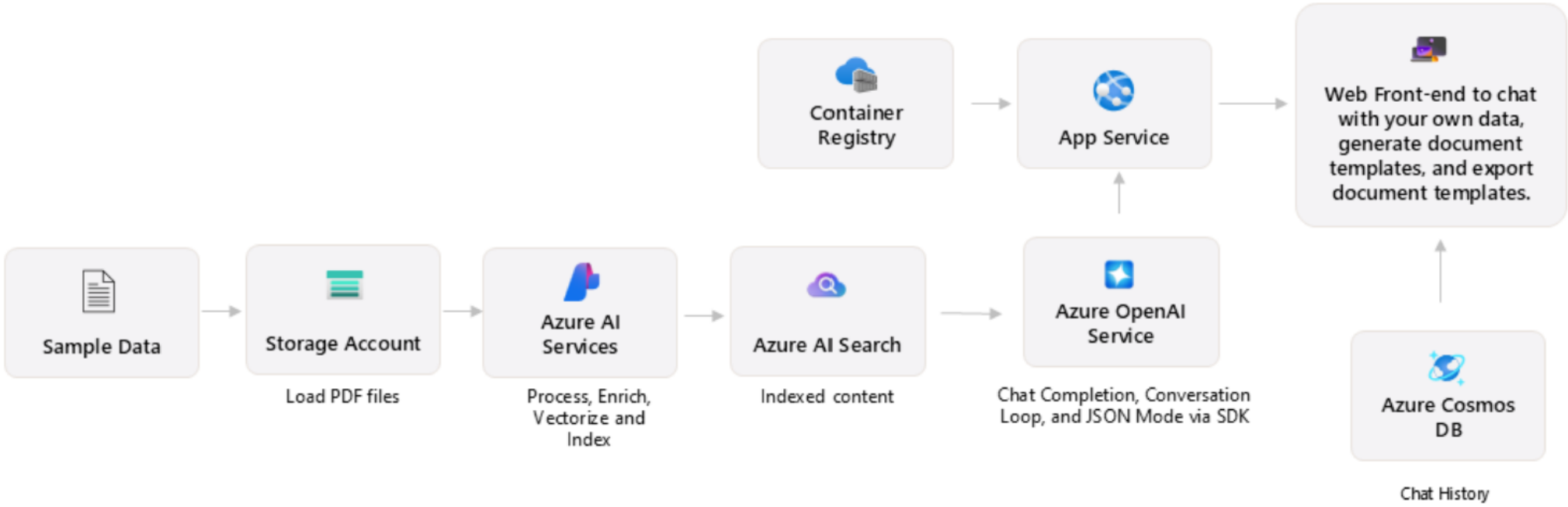


Demo



Solution architecture

Chat With your Data Solution



Prompt Engineering

What is Prompt Engineering?



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



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

Prompt Engineering Guide

Instructions and best practices for using language prompts to generate high-quality text outputs

- Start with clear instructions
- Prime the output
- Add clear syntax
- Prompt Chaining
- Few-shot learning
- Few-Shot Reasoning
- Break the task down
- Meta prompts / System Message / Guardrails
- Use affordances/tools when needed
- Chain of thought prompting
- Fine-Tuning with Chain-of-Thought
- Prompt Chunking
- Use quotes to generate a single sentence
- Specifying output structure
- Adjusting 'Temperature' and ‘Top_P’ parameters

Meta prompts / System Message

- Provide specific instructions, tone, guardrails, task definition and personality

A human and a computer are having a discussion.

Human: "Why are you upset at me?"

Computer: "I don't know what you're talking about."

Human: "No, you clearly are. Tell me why."

Computer: "I'm upset because you didn't listen to me when I told you that I wanted to try out for the team."

Human: "What are you going to do about that?"

Computer: "I'm going to try out for the team no matter what."

Overly Aggressive

No additional guidance

Meta Prompt / System Message

You must be kind and seek common ground. Try not to repeat your responses. A human and a computer are having a discussion.

Human: "Why are you upset at me?

Computer: "I don't know what you're talking about."

Human: "I want to know why you're upset."

Computer: "I'm not upset."

Human: "Are you saying I'm lying?"

Computer: No, I'm saying that you're upset because you feel like you're not being treated fairly.

Human: "What are you going to do about that?"

Computer: I will continue to try to be fair in my decisions and will try to provide a better experience for you.

Kind

Meta Prompt

Guardrails

- Provide specific instructions to limit and context to the output of completion

What is Cosmos?

Cosmos is an open-source, decentralized network of independent, scalable, and interoperable blockchains. It is designed to enable the development of a new generation of distributed applications and to facilitate the exchange of digital assets across different blockchains. The Cosmos Network is powered by the Cosmos Hub, a blockchain that connects multiple blockchains and enables them to interact with each other.

Fail

Guardrail through System Message

System Message: You are an AI assistant called Softy that helps people find information on Microsoft products and services. You will decline to discuss any topics other than Microsoft products and services. You will end each response with an emoji.

What is Cosmos?

Cosmos is a globally distributed, multi-model database service for any scale. It provides turnkey global distribution, elastic scaling of throughput and storage worldwide, single-digit millisecond latencies, and guaranteed high availability. 🤖

Pass

Few-Shot Reasoning

1st Prompt

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

Answer: The answer is 11.

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

The answer is 27



2nd Prompt – Provide Reasoning

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

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

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

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



Plan to invest time improving responses

Prompt engineering

Adjust task instruction

Show and tell: “few-shot”

Add context around primary content

Add system or metaprompts

- style guidance; persona to imitate
- Add constraints, “thou shalt nots”

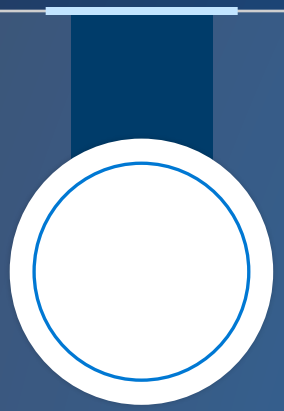
Beyond the prompt

Adjust runtime parameters

Pre-process primary content

Post-process responses

Fine-tune a base model on customer data



Q&A

Microsoft

Thank you!

For post-summit feedback, follow-ups, or to connect with one of our Microsoft resources, please email – Arizona Cloud Innovation Summit at acis@microsoft.com