AI services on Azure
A look into the future

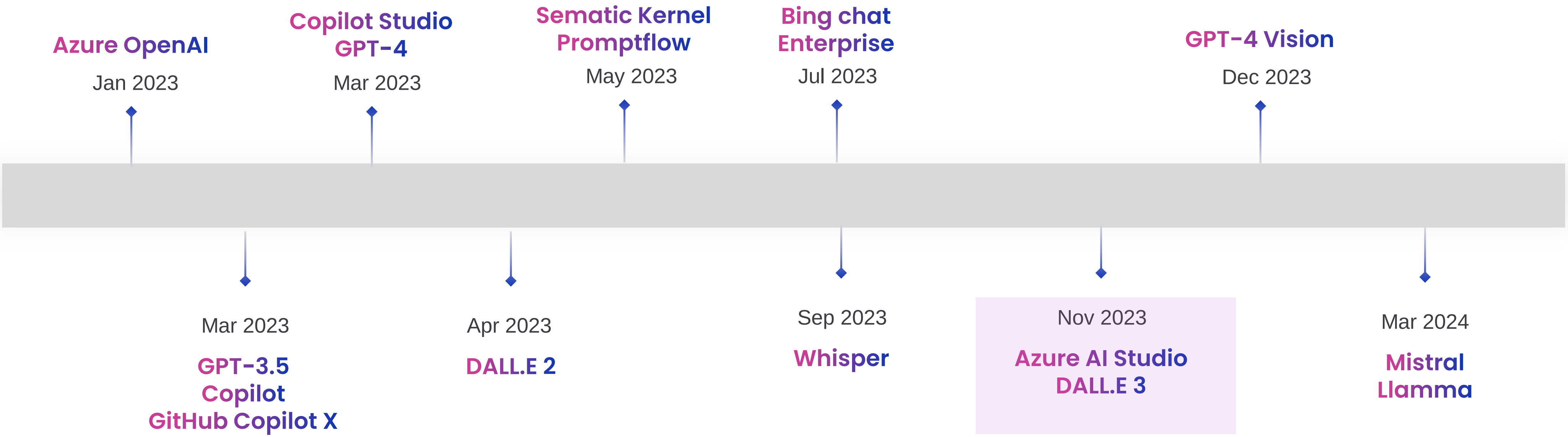
Phi Huynh – Microsoft MVP
2024

Agenda

1. Crazy year of GenAI
2. Microsoft & OpenAI
3. New GenAI models
4. Model catalog
5. MaaP & MaaS
6. RAG
7. Finetune
8. Azure AI Studio
9. Well-architected framework for GenAI app

It’s a crazy year for GenAI

NashTech has researched on popular generative AI models on LLMs, images and audio



Azure OpenAI
Jan 2023

Mar 2023
GPT-3.5
Copilot
GitHub Copilot X

Copilot Studio
GPT-4
Mar 2023

Apr 2023
DALL.E 2

Sematic Kernel
Promptflow
May 2023

Bing chat
Enterprise
Jul 2023

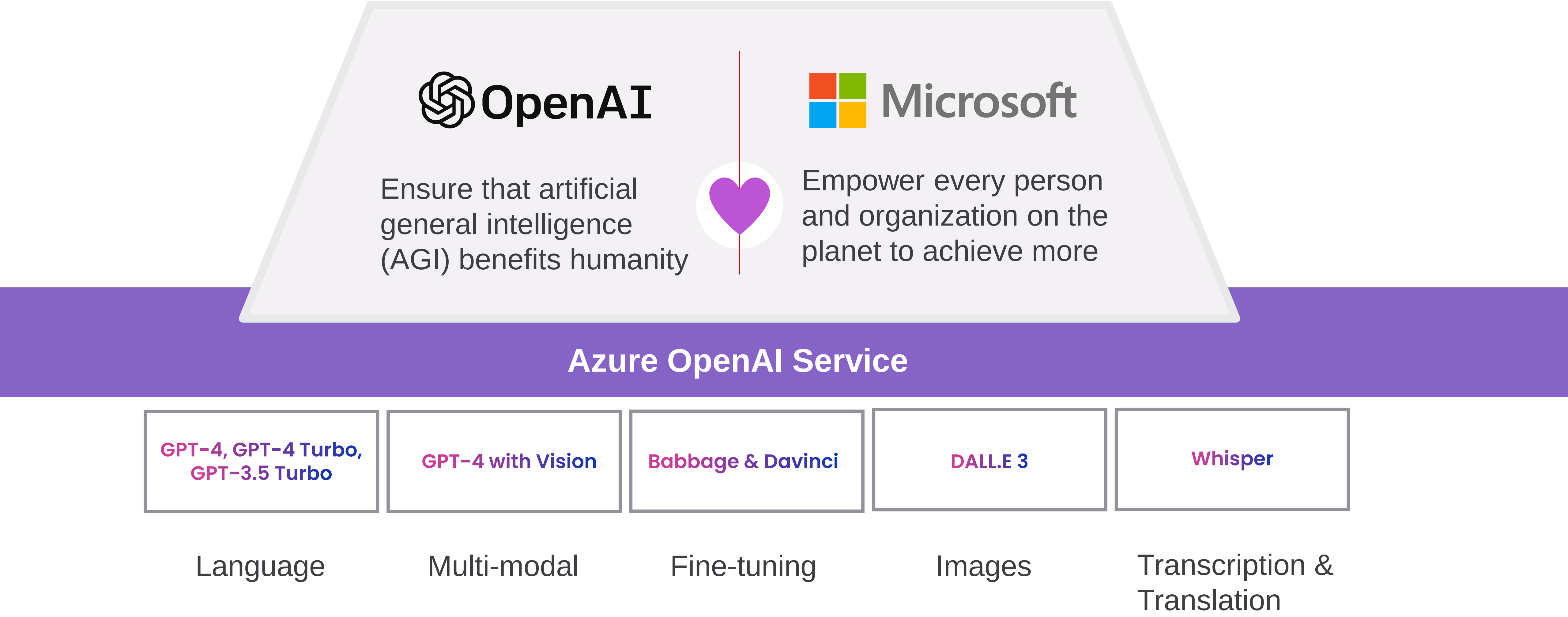
Sep 2023
Whisper

Nov 2023
Azure AI Studio
DALL.E 3

GPT-4 Vision
Dec 2023

Mar 2024
Mistral
Llamma

Microsoft and OpenAI partnership



OpenAI

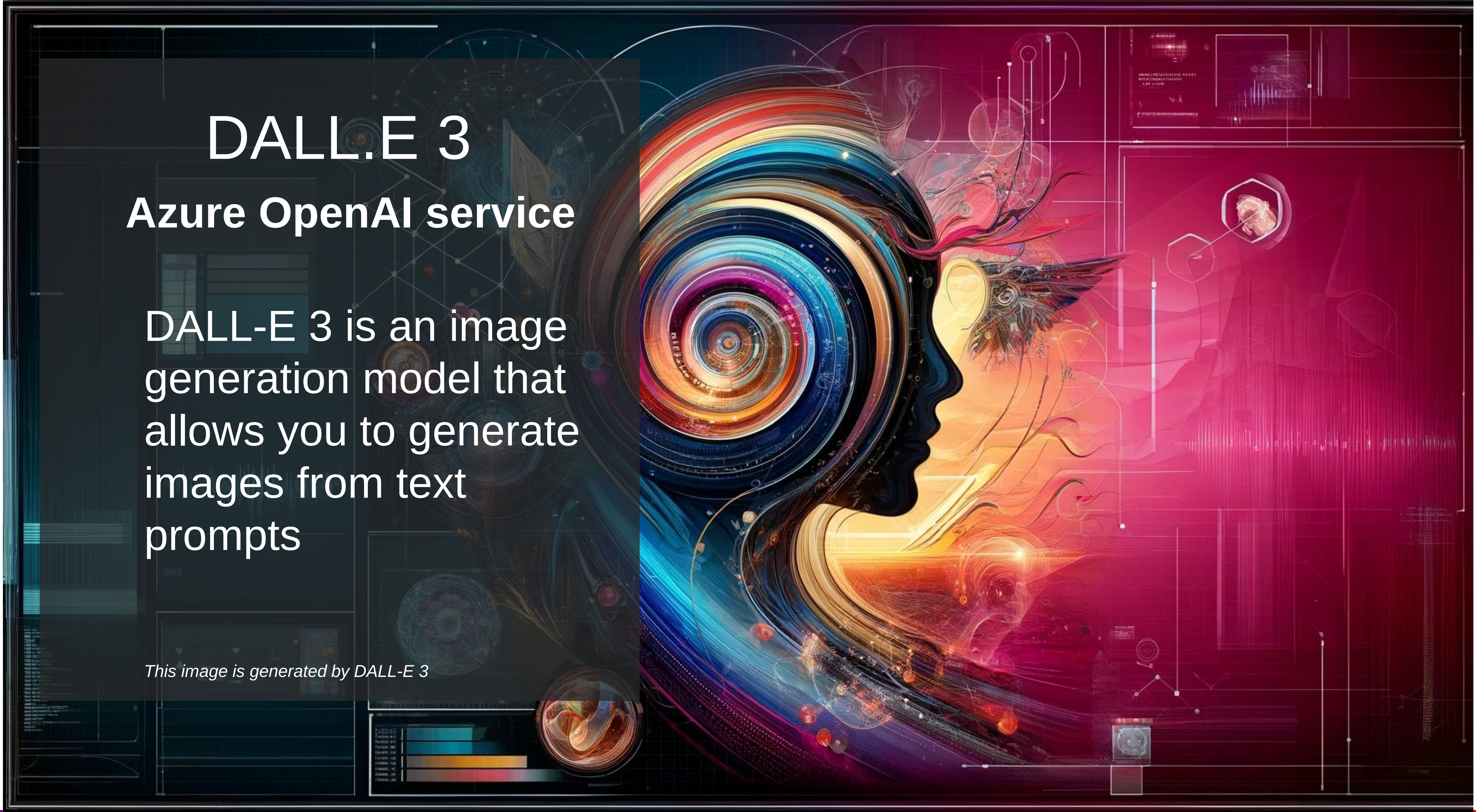
Ensure that artificial general intelligence (AGI) benefits humanity

Microsoft

Empower every person and organization on the planet to achieve more

Azure OpenAI Service

GPT-4, GPT-4 Turbo, GPT-3.5 Turbo	GPT-4 with Vision	Babbage & Davinci	DALL.E 3	Whisper
Language	Multi-modal	Fine-tuning	Images	Transcription & Translation



DALL.E 3

Azure OpenAI service

DALL-E 3 is an image generation model that allows you to generate images from text prompts

This image is generated by DALL-E 3

What can DALL.E 3 do?

SPILL THE BEANS
COFFEE

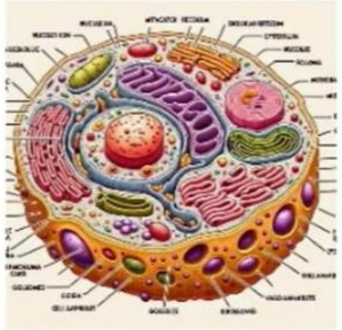
LOGO & BRANDING: QUICK CONCEPT GENERATION.

CREATIVE INSPIRATION: OVERCOME DESIGN BLOCKS.

CONTENT ILLUSTRATIONS: UNIQUE IMAGES FOR BLOGS/ARTICLES.

AD CAMPAIGNS: VISUALIZE MARKETING CONCEPTS.

PRODUCT VISUALIZATION: GAUGE INTEREST & FEEDBACK.



EDUCATION: CUSTOM IMAGERY FOR COURSES.

FASHION DESIGN: VISUALIZE CLOTHING PATTERNS.

GAMING: CHARACTER & ENVIRONMENT CONCEPTS.

Public preview

Announcing GPT-4V with Azure AI Vision

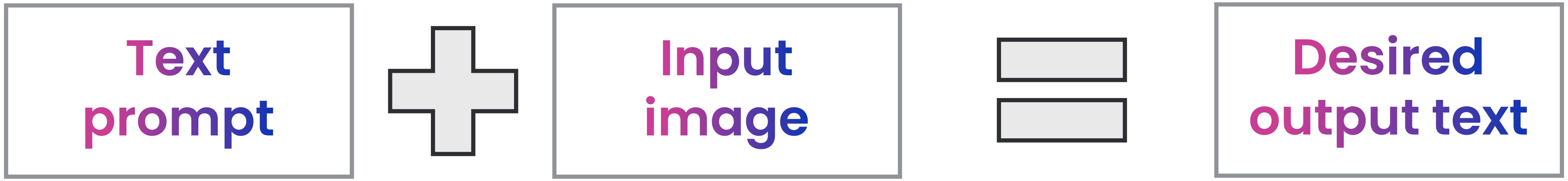
Unlock new scenarios with GPT-4V, Azure Open AI Service and Azure AI Vision integration

Add images to retrieval augment generation (RAG) patterns

Prompt with video, images, and text

What GPT-4V offers?

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



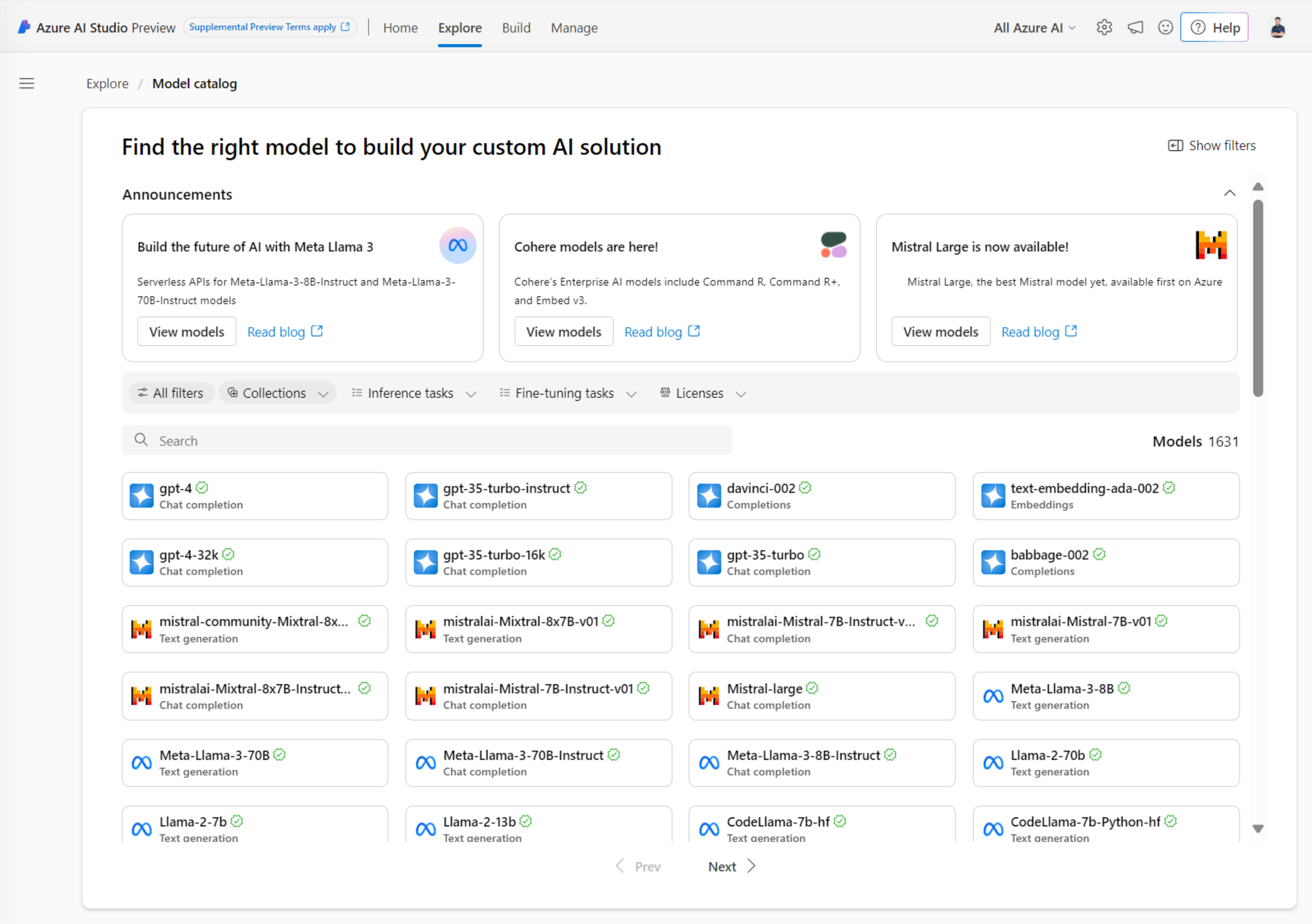
Text prompt + Input image = Desired output text

Note: GPT-4V doesn't generate image outputs

Model catalog (in Azure AI & ML Studio)

Catalog featuring the best as foundation model collections

- Popular OSS models handpicked and optimized by AzureML
- Partnering with HuggingFace to offer thousands of OSS models for inference
- Azure OpenAI models



Find the right model to build your custom AI solution

Announcements

Build the future of AI with Meta Llama 3

Serverless APIs for Meta-Llama-3-8B-Instruct and Meta-Llama-3-70B-Instruct models

Cohere models are here!

Cohere's Enterprise AI models include Command R, Command R+, and Embed v3.

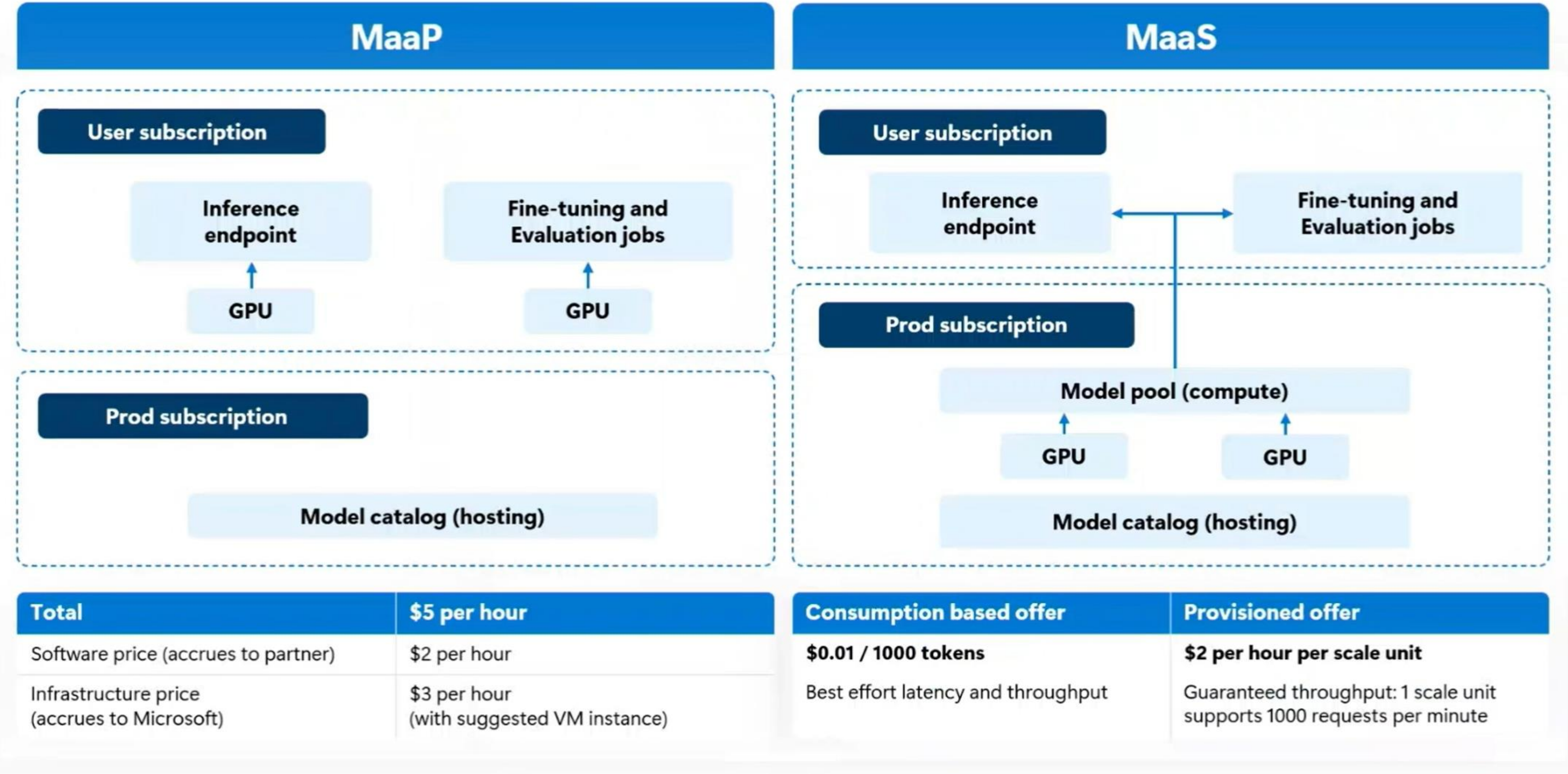
Mistral Large is now available!

Mistral Large, the best Mistral model yet, available first on Azure

Models 1631

gpt-4 – Chat completion
gpt-35-turbo-instruct – Chat completion
davinci-002 – Completions
text-embedding-ada-002 – Embeddings
gpt-4-32k – Chat completion
gpt-35-turbo-16k – Chat completion
gpt-35-turbo – Chat completion
babbage-002 – Completions
mistral-community-Mixtral-8x... – Text generation
mistralai-Mixtral-8x7B-v01 – Text generation
mistralai-Mistral-7B-Instruct-v... – Chat completion
mistralai-Mistral-7B-v01 – Text generation
mistralai-Mixtral-8x7B-Instruct... – Chat completion
mistralai-Mistral-7B-Instruct-v01 – Chat completion
Mistral-large – Chat completion
Meta-Llama-3-8B – Text generation
Meta-Llama-3-70B – Text generation
Meta-Llama-3-70B-Instruct – Chat completion
Meta-Llama-3-8B-Instruct – Chat completion
Llama-2-70b – Text generation
Llama-2-7b – Text generation
Llama-2-13b – Text generation
CodeLlama-7b-hf – Text generation
CodeLlama-7b-Python-hf – Text generation

MaaP and MaaS



MaaP

User subscription

Inference endpoint

GPU

Fine-tuning and Evaluation jobs

GPU

Prod subscription

Model catalog (hosting)

MaaS

User subscription

Inference endpoint

Fine-tuning and Evaluation jobs

Prod subscription

Model pool (compute)

GPU

GPU

Model catalog (hosting)

Total	$5 per hour
Software price (accrues to partner)	$2 per hour
Infrastructure price (accrues to Microsoft)	$3 per hour (with suggested VM instance)

Consumption based offer	Provisioned offer
$0.01 / 1000 tokens	**$2 per hour per scale unit**
Best effort latency and throughput	Guaranteed throughput: 1 scale unit supports 1000 requests per minute

Models as a Service

- Launched with Llama2 and Mistral
- Ready to use APIs with pay-as-you-go billing based on tokens for LLMs
- Integrate with your favorite LLM tools like PromptFlow, Semantic Kernel or LangChain
- Hosted finetuning without provisioning GPUs

Provisioned Throughput

Model processing capacity for your high-volume production workload

Predictable performance

Stable max latency and throughput for uniform workloads

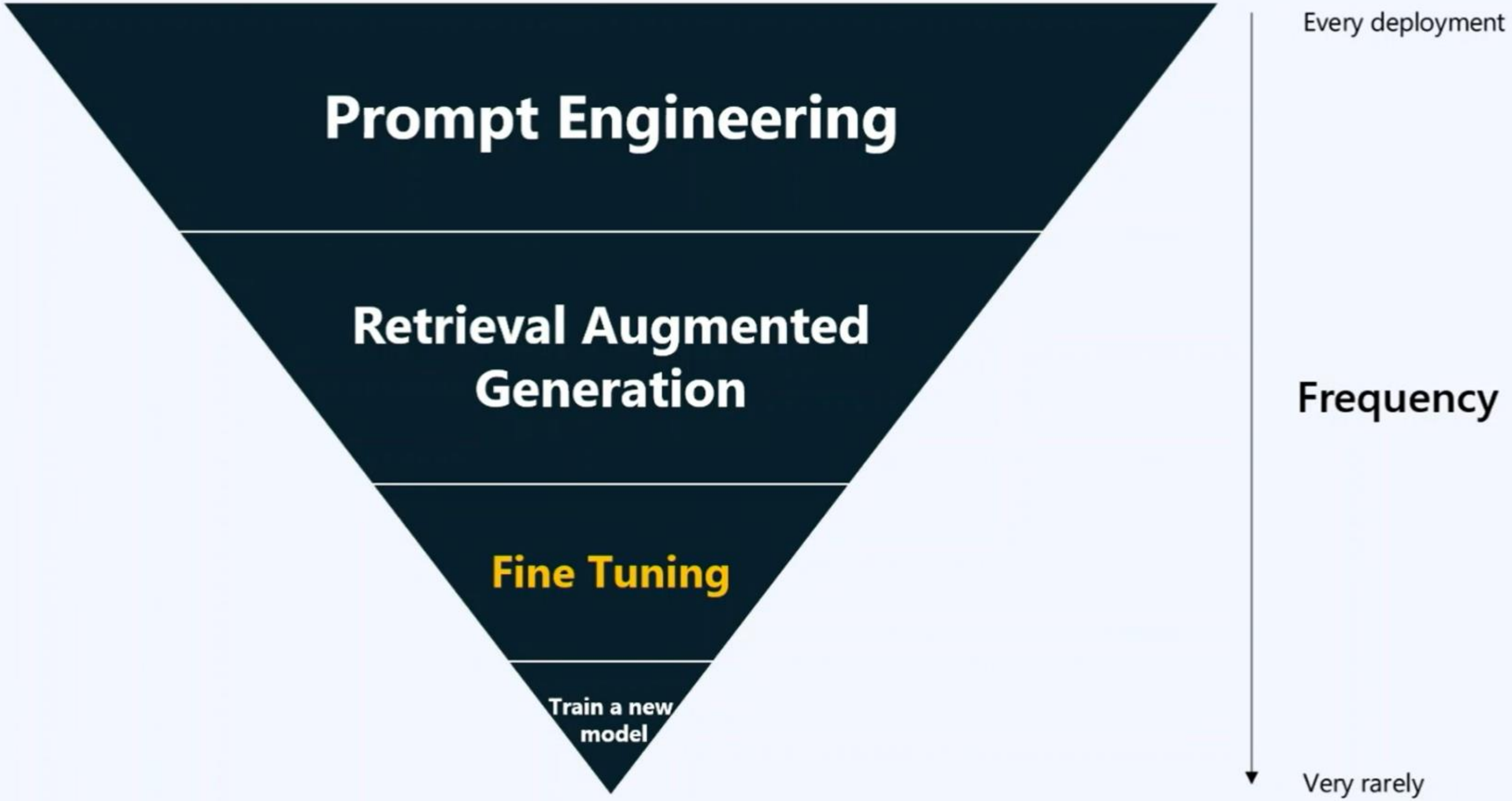
Reserved Processing Capacity

Ensures capacity is available to customer meet demand.

Cost Savings

Potential cost savings for high throughput workloads vs token-based consumption

Hierarchy of language model customization



Prompt Engineering

Retrieval Augmented Generation

Fine Tuning

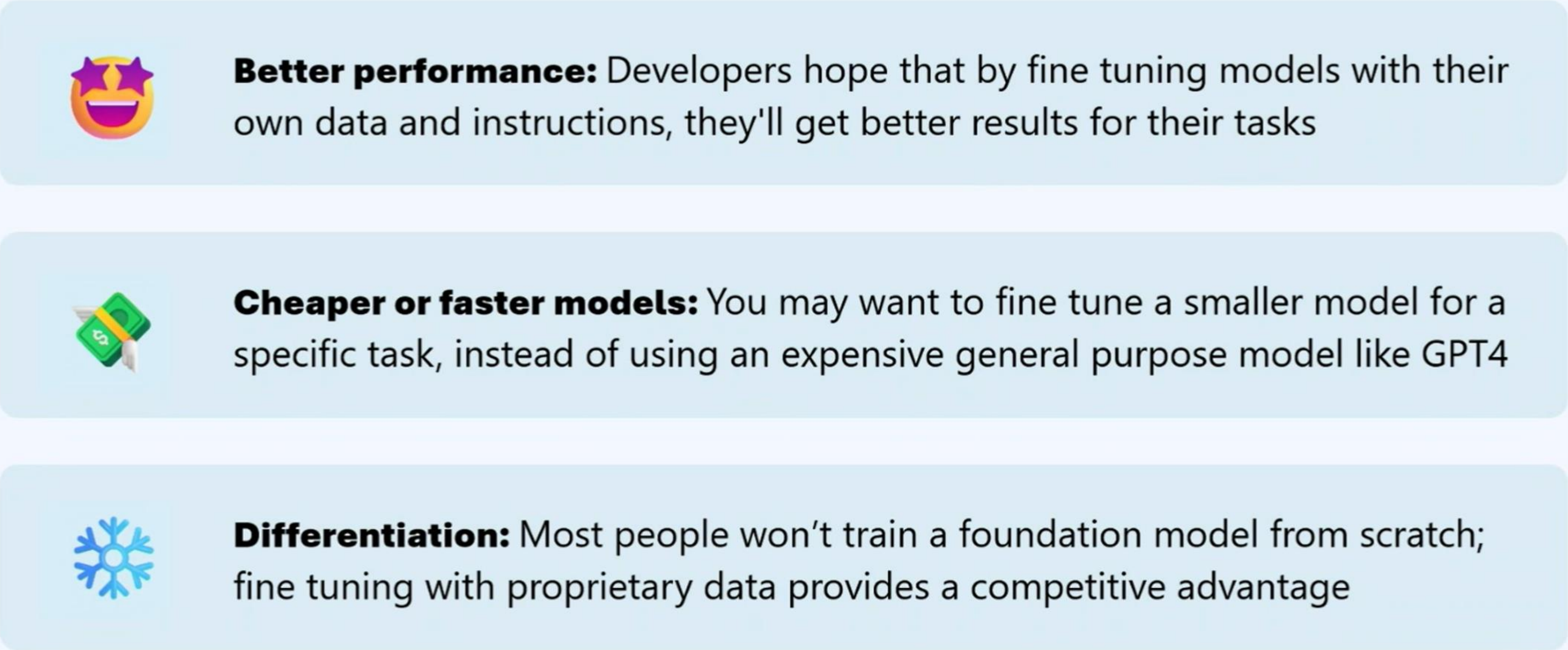
Train a new model

Every deployment

Frequency

Very rarely

Why fine-tune?

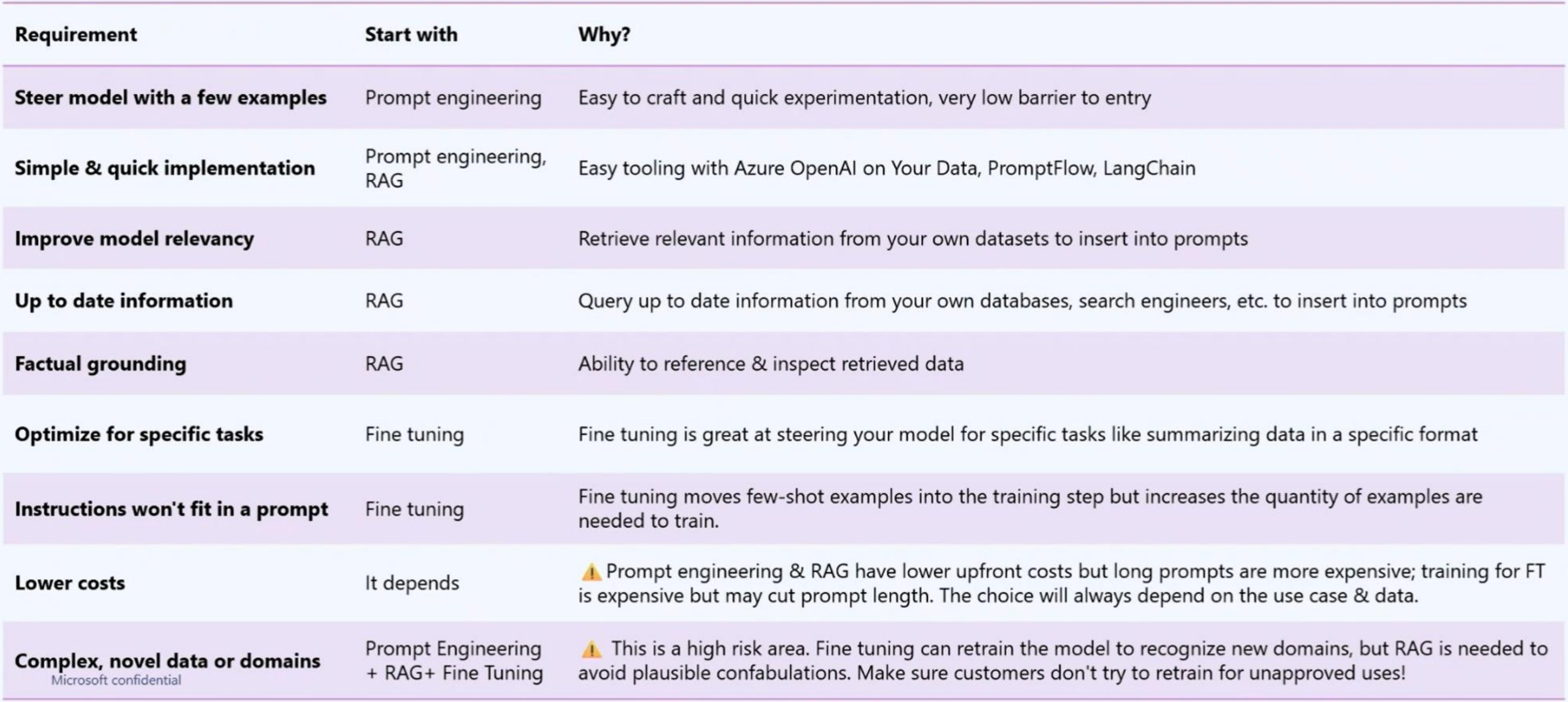


Better performance: Developers hope that by fine tuning models with their own data and instructions, they'll get better results for their tasks

Cheaper or faster models: You may want to fine tune a smaller model for a specific task, instead of using an expensive general purpose model like GPT4

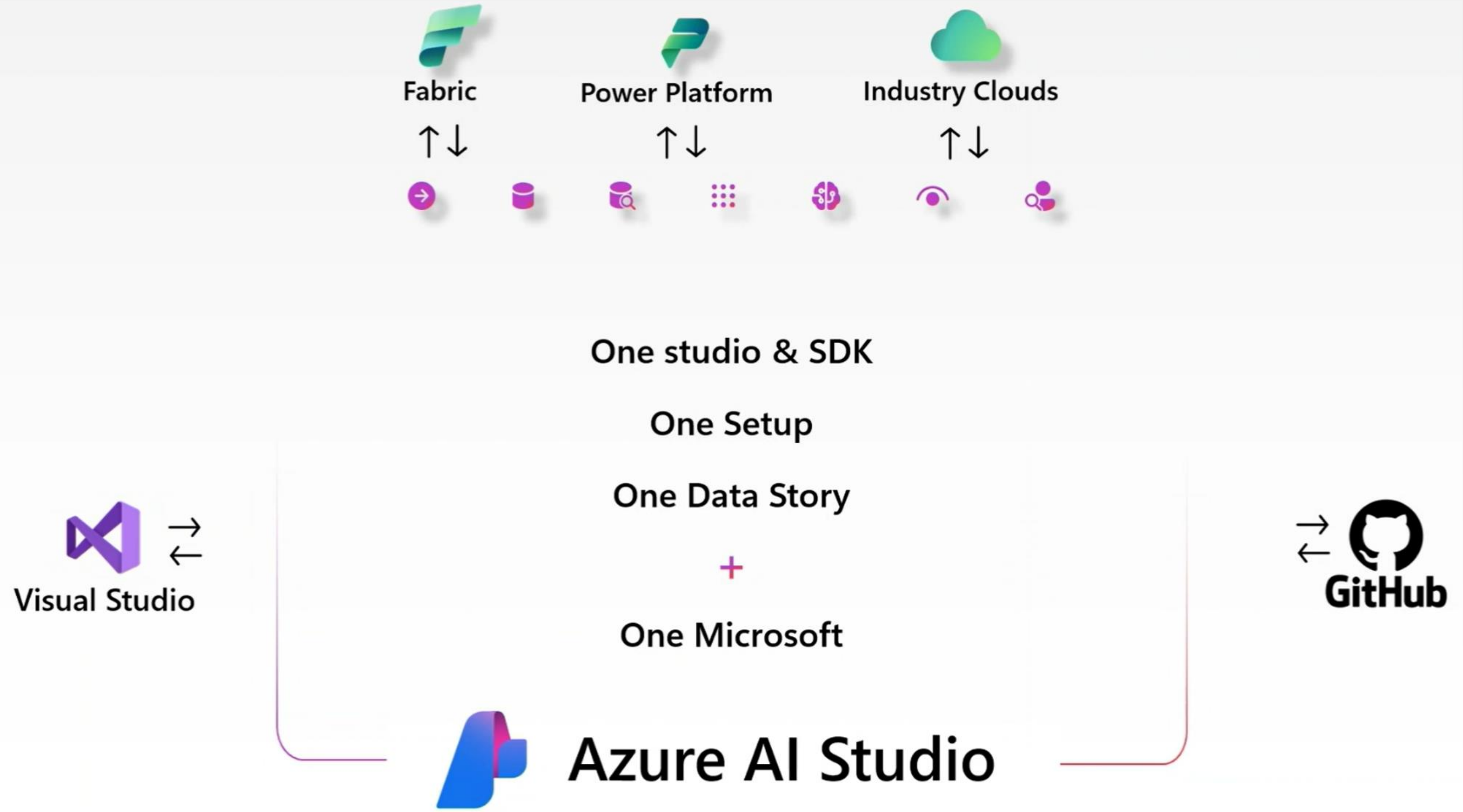
Differentiation: Most people won't train a foundation model from scratch; fine tuning with proprietary data provides a competitive advantage

Model customization cheat sheet



Requirement	Start with	Why?
Steer model with a few examples	Prompt engineering	Easy to craft and quick experimentation, very low barrier to entry
Simple & quick implementation	Prompt engineering, RAG	Easy tooling with Azure OpenAI on Your Data, PromptFlow, LangChain
Improve model relevancy	RAG	Retrieve relevant information from your own datasets to insert into prompts
Up to date information	RAG	Query up to date information from your own databases, search engineers, etc. to insert into prompts
Factual grounding	RAG	Ability to reference & inspect retrieved data
Optimize for specific tasks	Fine tuning	Fine tuning is great at steering your model for specific tasks like summarizing data in a specific format
Instructions won't fit in a prompt	Fine tuning	Fine tuning moves few-shot examples into the training step but increases the quantity of examples are needed to train.
Lower costs	It depends	⚠ Prompt engineering & RAG have lower upfront costs but long prompts are more expensive; training for FT is expensive but may cut prompt length. The choice will always depend on the use case & data.
Complex, novel data or domains	Prompt Engineering + RAG+ Fine Tuning	⚠ This is a high risk area. Fine tuning can retrain the model to recognize new domains, but RAG is needed to avoid plausible confabulations. Make sure customers don't try to retrain for unapproved uses!

Azure AI Studio umbrella



One studio & SDK

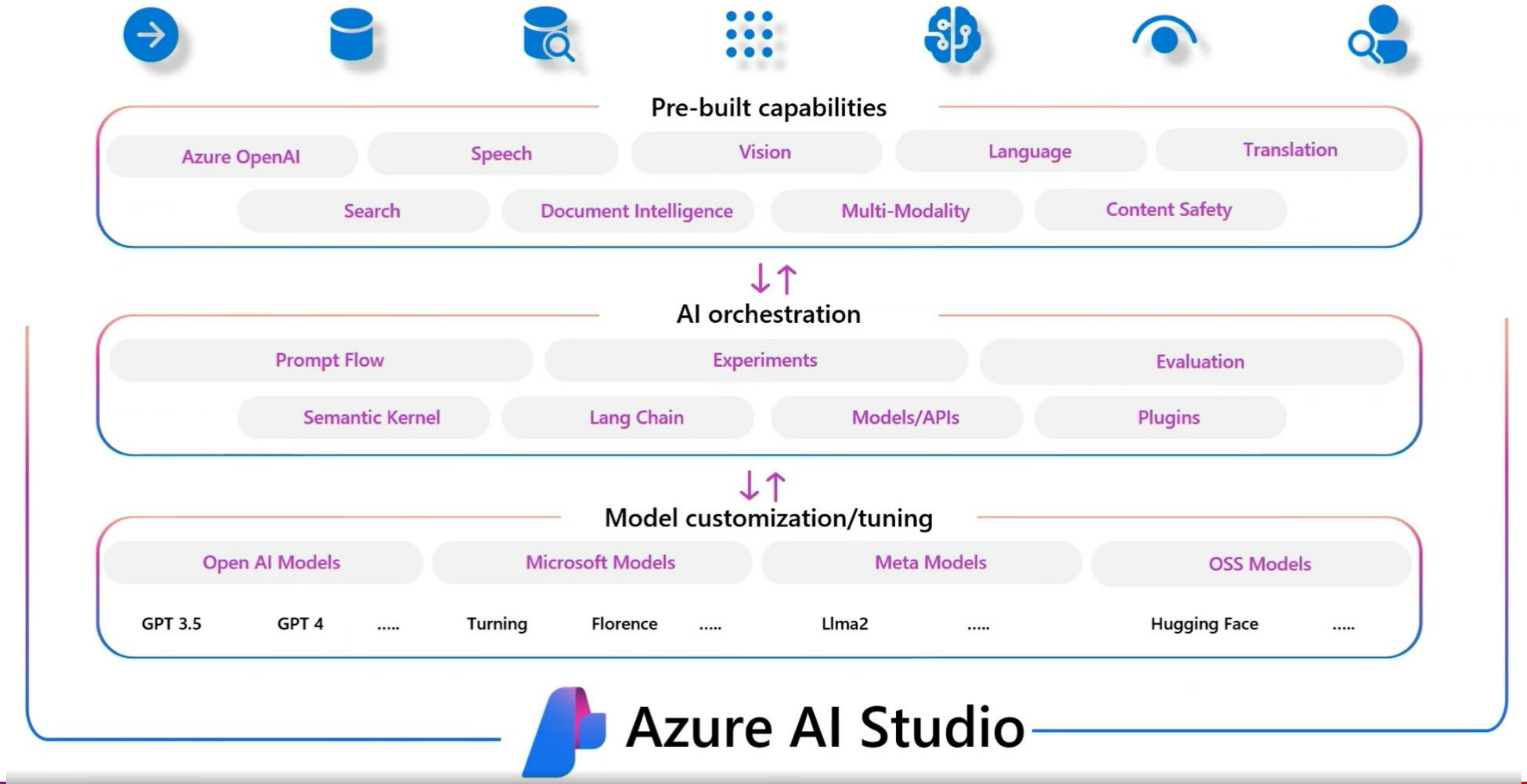
One Setup

One Data Story

+

One Microsoft

Azure AI Studio umbrella



Pre-built capabilities

Azure OpenAI | Speech | Vision | Language | Translation

Search | Document Intelligence | Multi-Modality | Content Safety

↓↑

AI orchestration

Prompt Flow | Experiments | Evaluation

Semantic Kernel | Lang Chain | Models/APIs | Plugins

↓↑

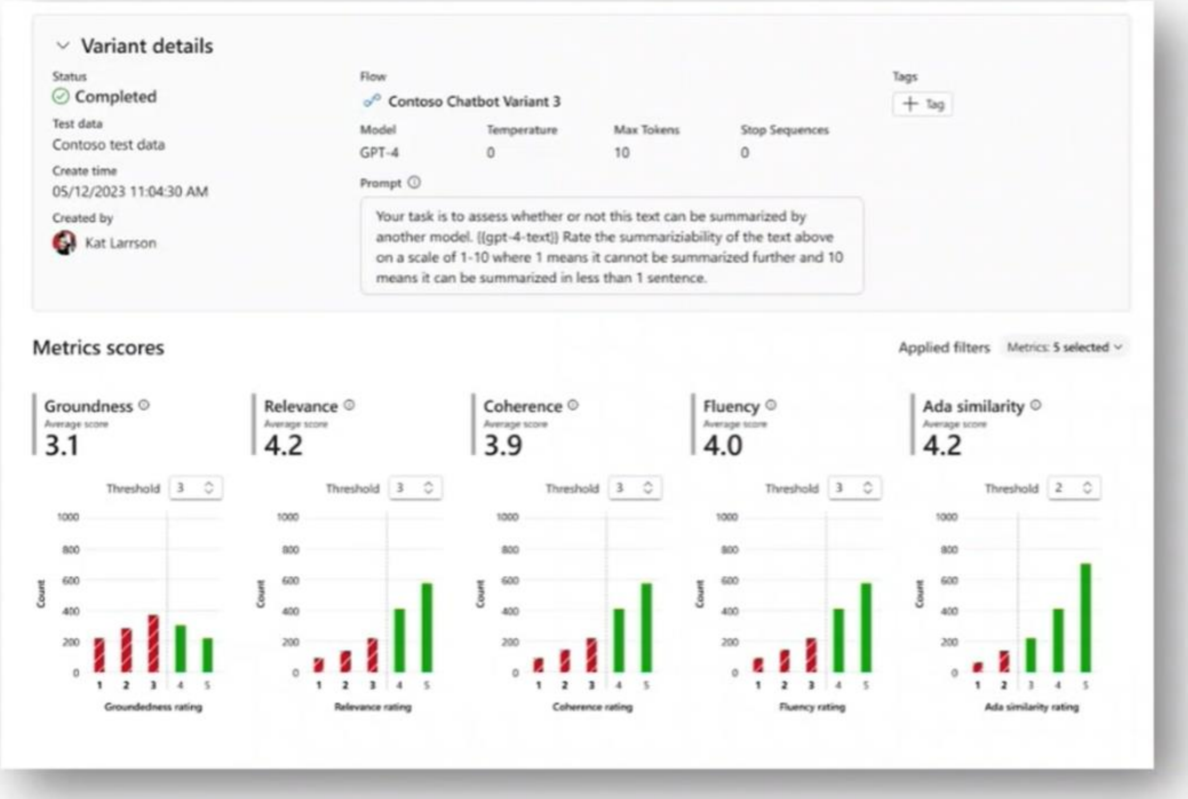
Model customization/tuning

Open AI Models | Microsoft Models | Meta Models | OSS Models

GPT 3.5 GPT 4 ..... Turning Florence ..... Llma2 ..... Hugging Face .....

Azure AI Studio

Compare model outputs for your business requirements



Variant details

Status: Completed

Test data: Contoso test data

Create time: 05/12/2023 11:04:30 AM

Created by: Kat Larrson

Flow: Contoso Chatbot Variant 3

Model	Temperature	Max Tokens	Stop Sequences
GPT-4	0	10	0

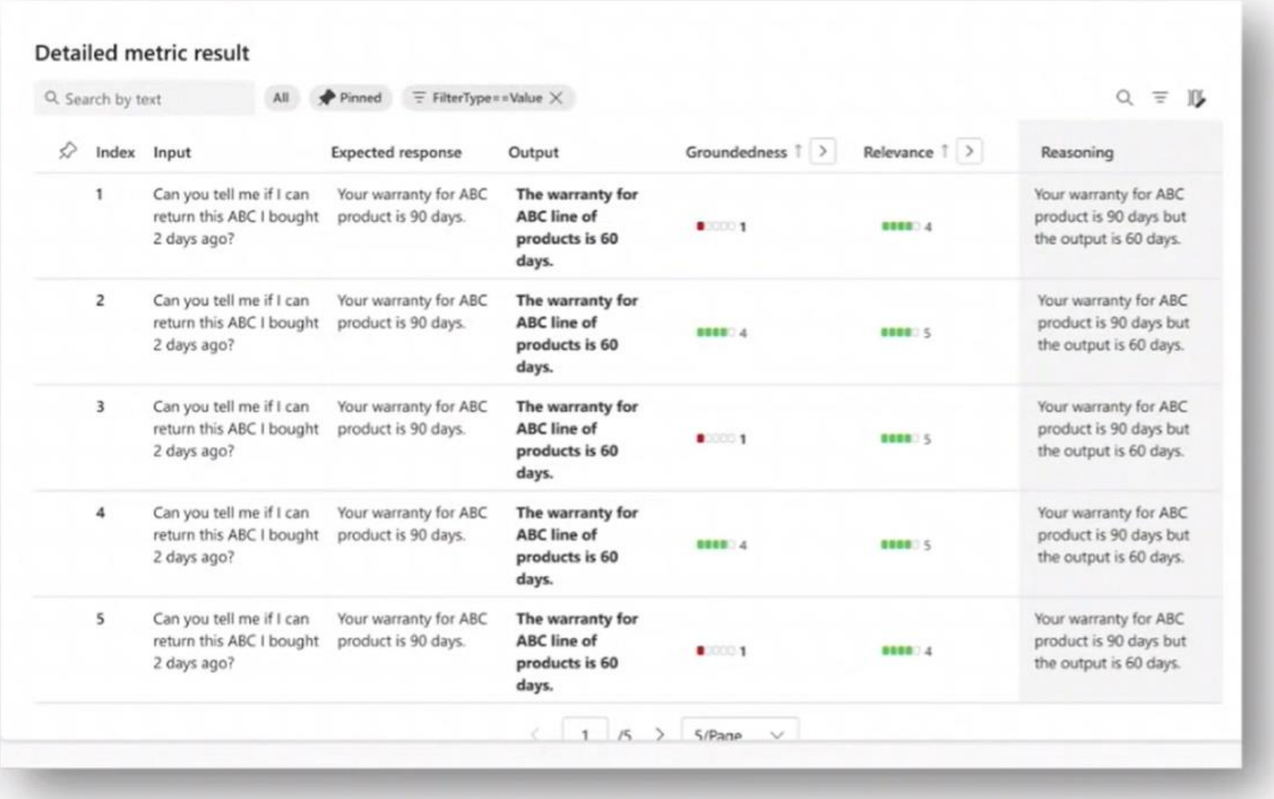
Prompt: Your task is to assess whether or not this text can be summarized by another model. {{gpt-4-text}} Rate the summariziability of the text above on a scale of 1-10 where 1 means it cannot be summarized further and 10 means it can be summarized in less than 1 sentence.

Tags: + Tag

Metrics scores

Applied filters: Metrics: 5 selected

Metric	Average score	Threshold
Groundness	3.1	3
Relevance	4.2	3
Coherence	3.9	3
Fluency	4.0	3
Ada similarity	4.2	2

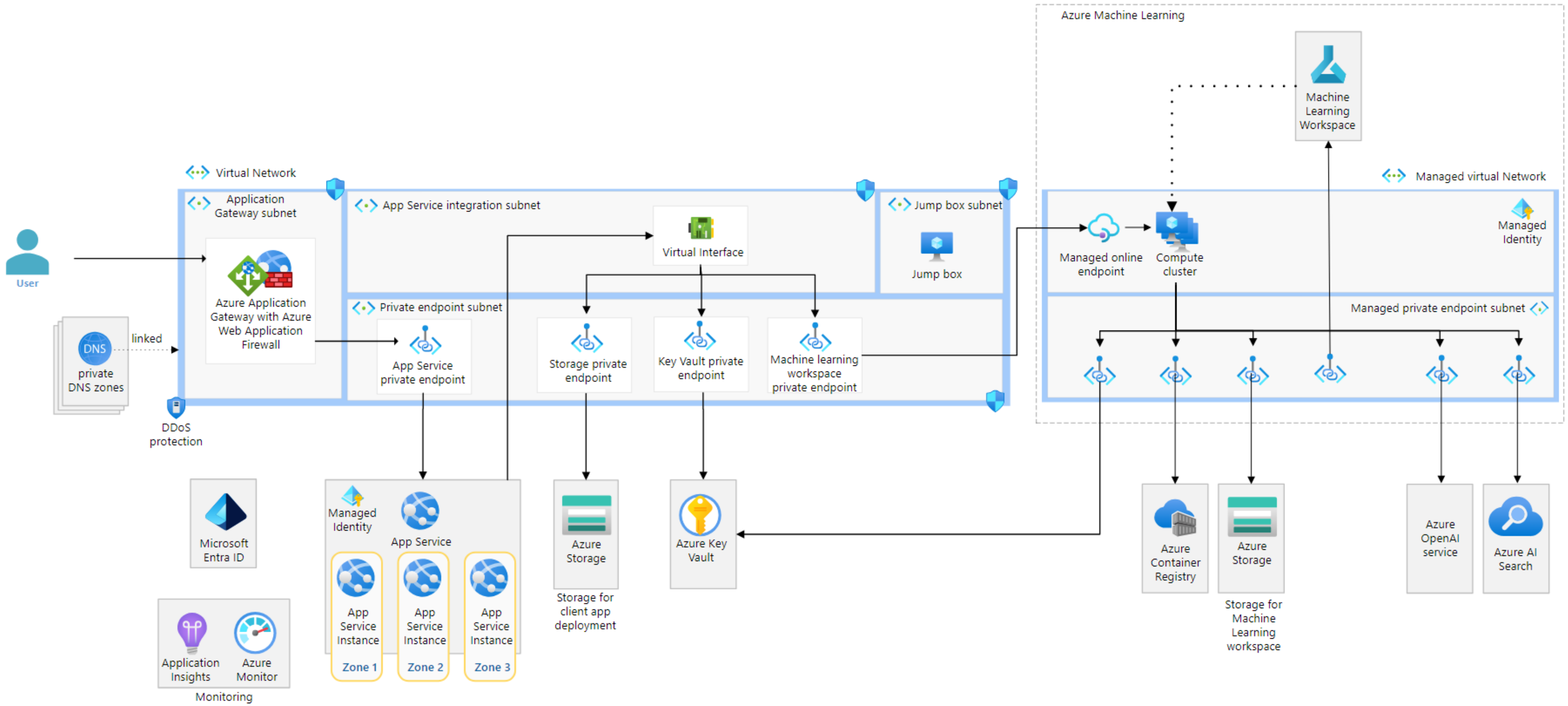


Detailed metric result

Index	Input	Expected response	Output	Groundedness	Relevance	Reasoning
1	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	**The warranty for ABC line of products is 60 days.**	1	4	Your warranty for ABC product is 90 days but the output is 60 days.
2	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	**The warranty for ABC line of products is 60 days.**	4	5	Your warranty for ABC product is 90 days but the output is 60 days.
3	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	**The warranty for ABC line of products is 60 days.**	1	5	Your warranty for ABC product is 90 days but the output is 60 days.
4	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	**The warranty for ABC line of products is 60 days.**	4	5	Your warranty for ABC product is 90 days but the output is 60 days.
5	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	**The warranty for ABC line of products is 60 days.**	1	4	Your warranty for ABC product is 90 days but the output is 60 days.

Evaluate model capabilities, cost, latency, and compatibility with the enterprises' Azure tenant

Well-architected for reliability & security



Implement a best practice landing zone to ensure you meet the Resilience, Redundancy, and Security needs for your Gen AI implementation

Learn more here:

- Azure Well-Architected Framework perspective on Azure OpenAI Service - Microsoft Azure Well-Architected Framework
- Baseline OpenAI end-to-end chat reference architecture - Azure Reference Architectures

Assistants

How Assistants Work

Step 1: Create an Assistant

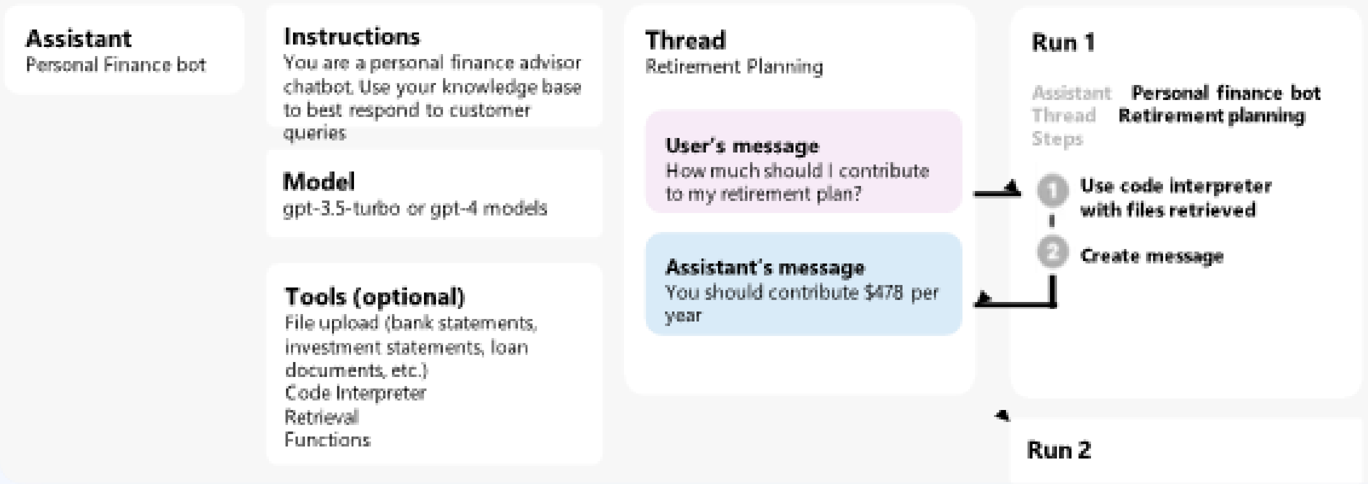
Step 2: Create a Thread

Step 3: Add a Message to a Thread

Step 4: Run the Assistant

Step 5: Check the Run Status

Step 6: Display the Assistant's Response



Assistant
Personal Finance bot

Instructions
You are a personal finance advisor chatbot. Use your knowledge base to best respond to customer queries

Model
gpt-3.5-turbo or gpt-4 models

Tools (optional)
File upload (bank statements, investment statements, loan documents, etc.)
Code Interpreter
Retrieval
Functions

Thread
Retirement Planning

User's message
How much should I contribute to my retirement plan?

Assistant's message
You should contribute $478 per year

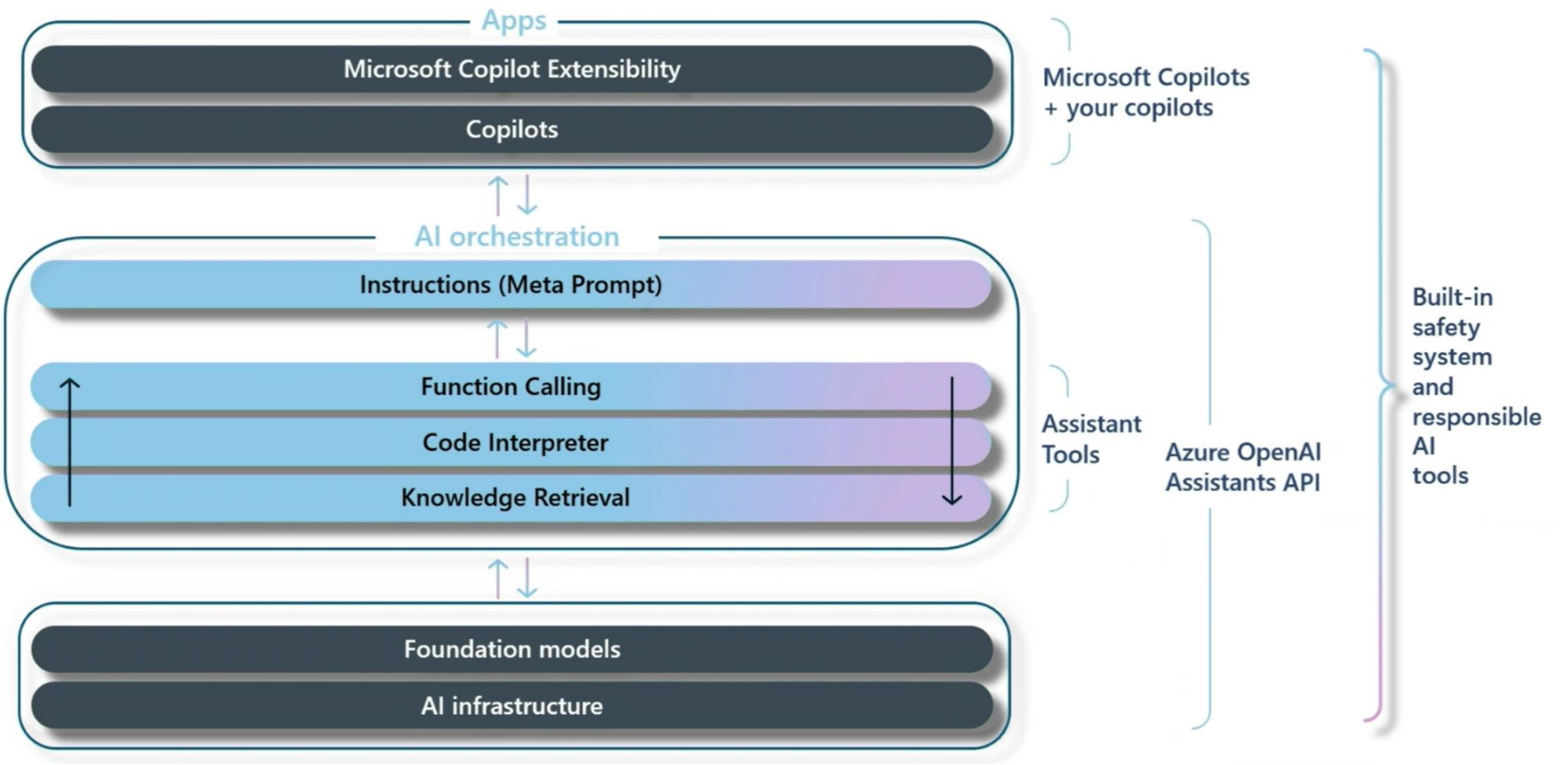
Run 1

Assistant Personal finance bot
Thread Retirement planning
Steps

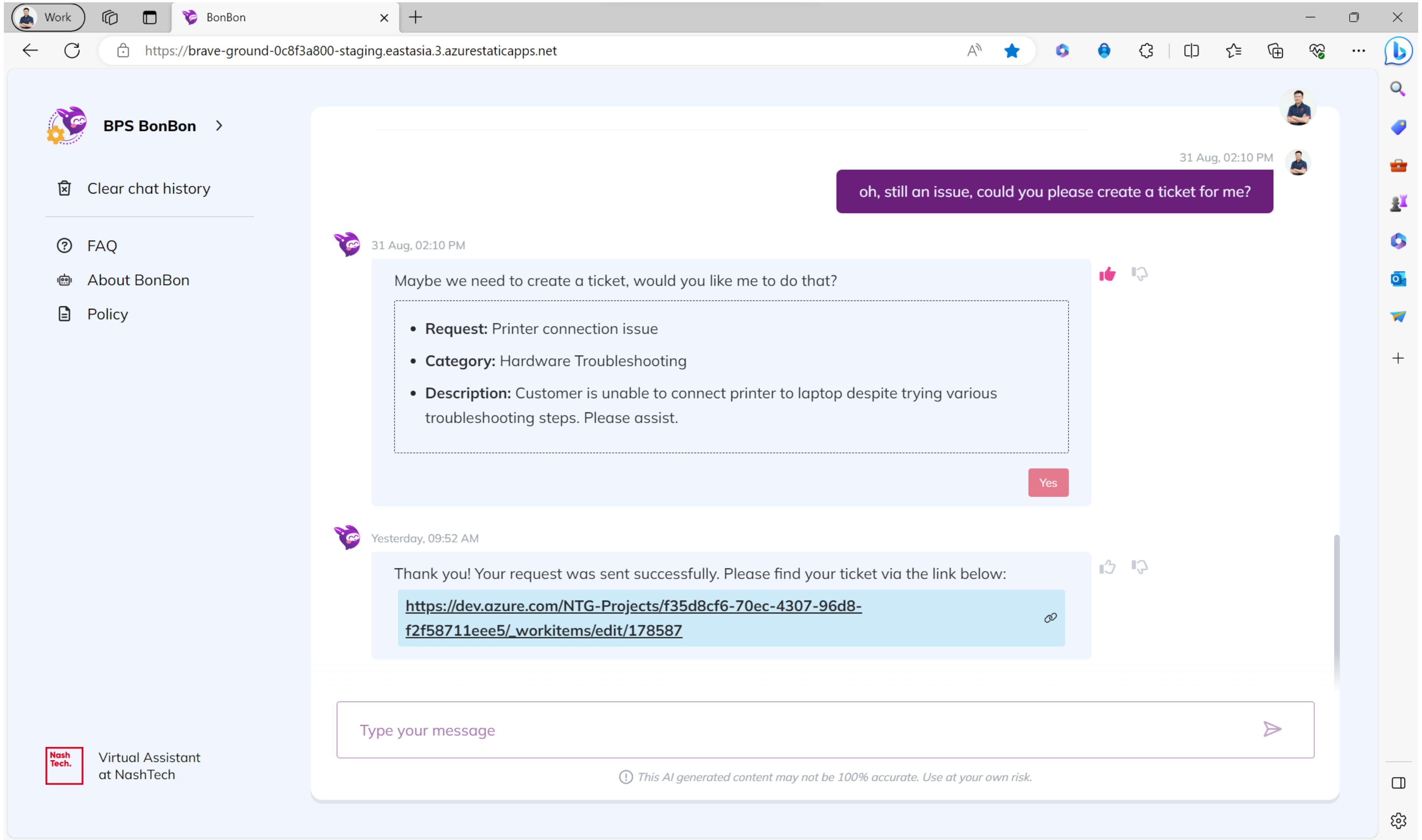
1. Use code interpreter with files retrieved
2. Create message

Run 2

Assistants stack



BonBon – NashTech intelligent virtual assistant



Work
BonBon
https://brave-ground-0c8f3a800-staging.eastasia.3.azurestaticapps.net
BPS BonBon
Clear chat history
FAQ
About BonBon
Policy
Nash Tech.
Virtual Assistant at NashTech

31 Aug, 02:10 PM

oh, still an issue, could you please create a ticket for me?

31 Aug, 02:10 PM

Maybe we need to create a ticket, would you like me to do that?

- **Request:** Printer connection issue
- **Category:** Hardware Troubleshooting
- **Description:** Customer is unable to connect printer to laptop despite trying various troubleshooting steps. Please assist.

Yes

Yesterday, 09:52 AM

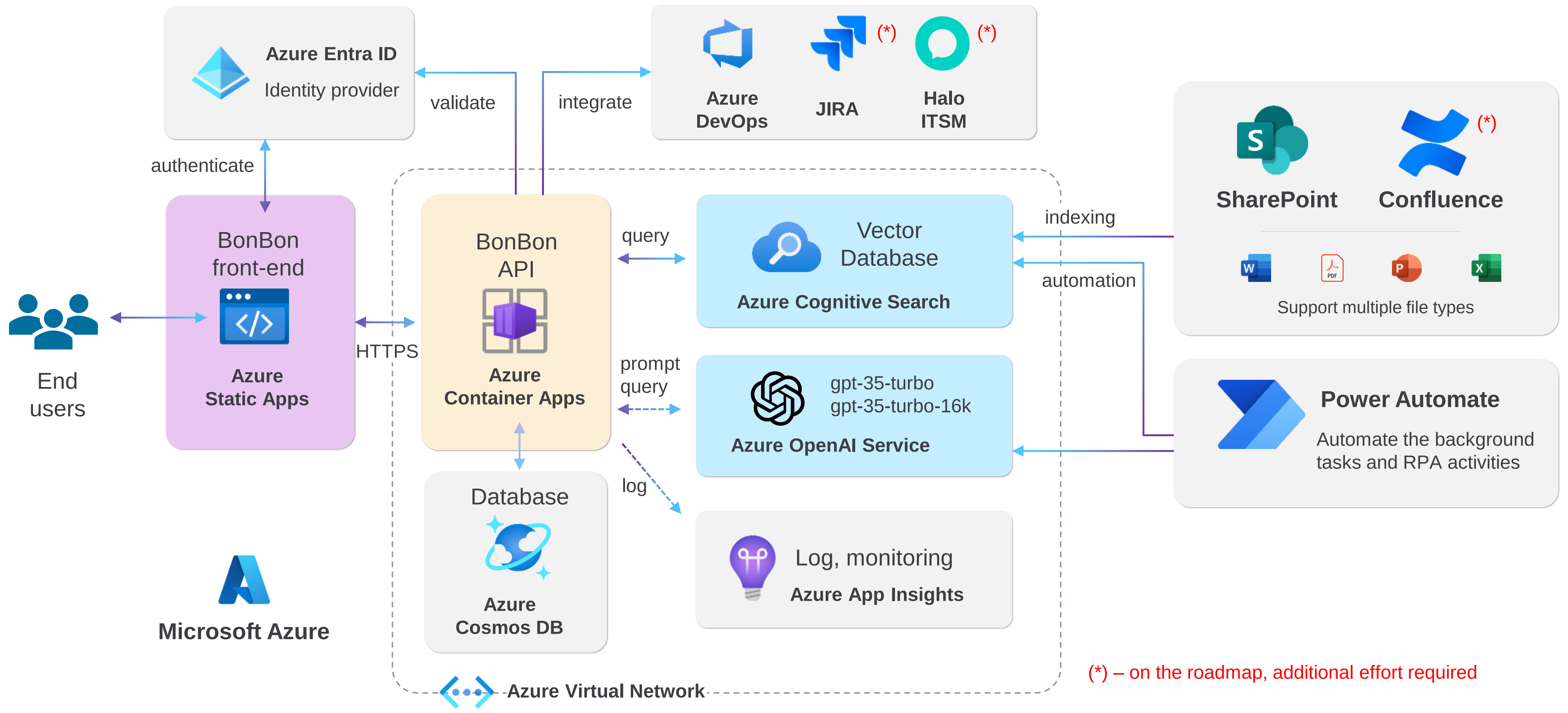
Thank you! Your request was sent successfully. Please find your ticket via the link below:

https://dev.azure.com/NTG-Projects/f35d8cf6-70ec-4307-96d8-f2f58711eee5/_workitems/edit/178587

Type your message

This AI generated content may not be 100% accurate. Use at your own risk.

BonBon's architecture



(*) – on the roadmap, additional effort required

Thank you

Nash
Tech.