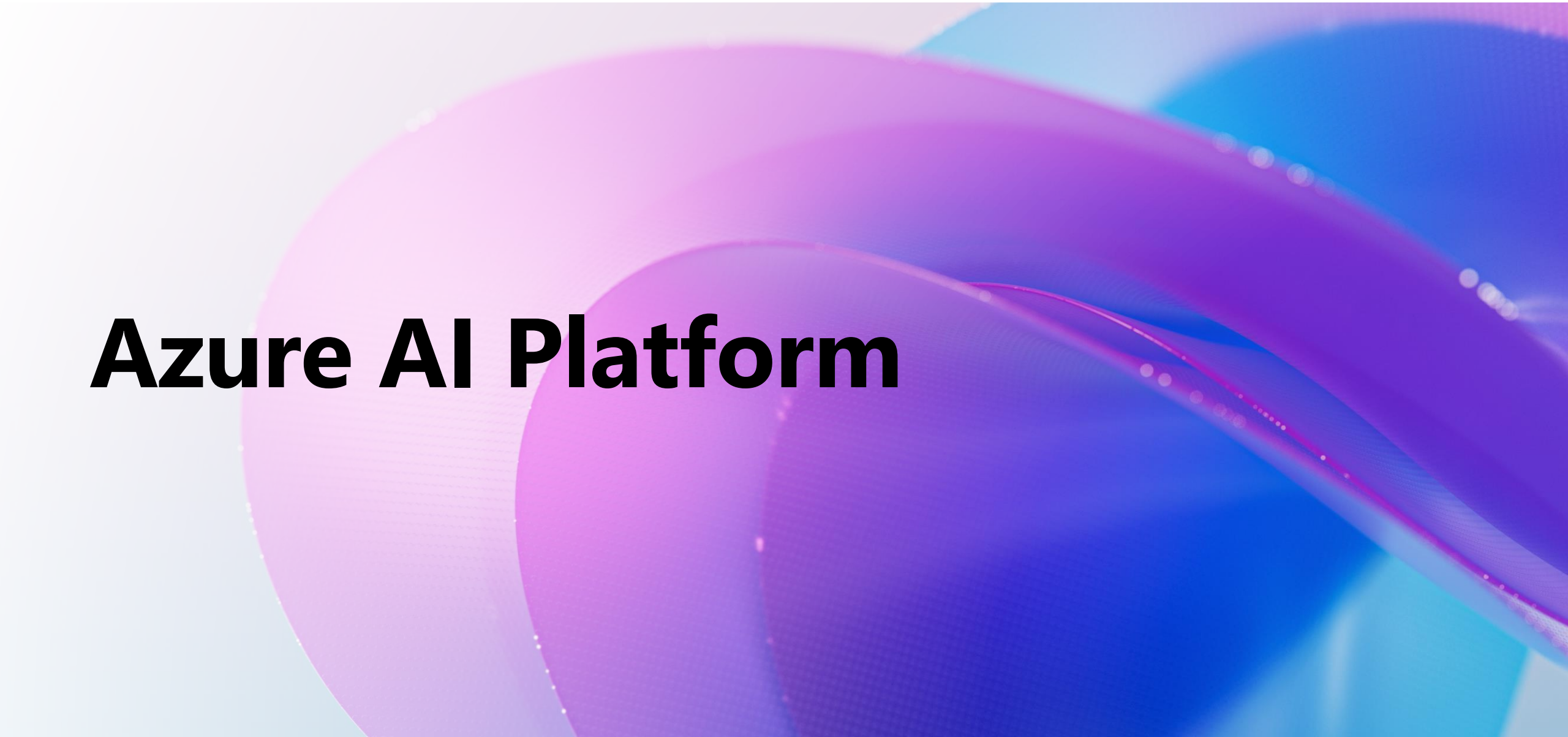
AZURE AI



# Azure AI Platform

# Session 1

## Azure AI Product Family

A Comprehensive, End to End Platform for AI Apps & Services

### Azure AI Studio

Comprehensive user interface (UI), SDK, CLI

### Azure AI Services

Intelligent services

### Azure AI Search

Modern retrieval

### Azure ML

Build and train models

| | | | |
|---|---|---|---|
| Turing | GPT-4 and GPT 3.5 Turbo | DALL-E | Meta Llama 3 |
| Florence | Whisper | Embeddings | Hugging Face |

Azure infrastructure

# Session 2

## Complete AI Toolchain

Access collaborative, comprehensive tooling to support the development lifecycle and differentiate your apps

### Ground models on your protected data

Use your protected data to customize models with advanced fine-tuning and for contextually relevant retrieval augmented generation.

### Orchestrate and debug AI workflows

Streamline app development with easy-to-use prompt orchestration, tracing, and debugging via interactive visual and code-first workflows.

### Streamline model and app evaluations

Systematically measure and improve app performance with manual and automated evaluations.

# Session 3

# Responsible AI practices

## Design and safeguard applications

| Design apps responsibly | Safeguard with configurable filters and control |
| --- | --- |
| Confidently build apps with technologies, templates, and best practices to help manage risk, improve accuracy, protect privacy, reinforce transparency, and simplify compliance. | Detect and filter harmful content, protect PII, and safeguard applications against prompt attacks. |

# Session 4

## Enterprise-grade Production at Scale

Deploy AI innovations to Azure's managed infrastructure with continuous monitoring and governance across environments

| Deploy to production | Operationalize and monitor workflows | Enable developer agility and enterprise governance at scale |
| --- | --- | --- |
| Scale AI for use in websites, applications, and other production environments. | Continuously monitor AI safety, quality, and token consumption in production. Automate workflows and alerts for timely issue resolution. | Provide easy project creation and resource management across the organization and enterprise controls for security, privacy, and compliance. |

# Session 5

# Choose where to build your copilot

Three approaches to support the developer experience

**Copilot Studio**
*Power Platform*

- Build your own copilot using visual building experiences
- Customize Microsoft Copilots with your own enterprise scenarios
- Leverage a connected, integrated platform
- Manage copilot experiences with full visibility into customizations

**Azure AI Studio**
*Azure*

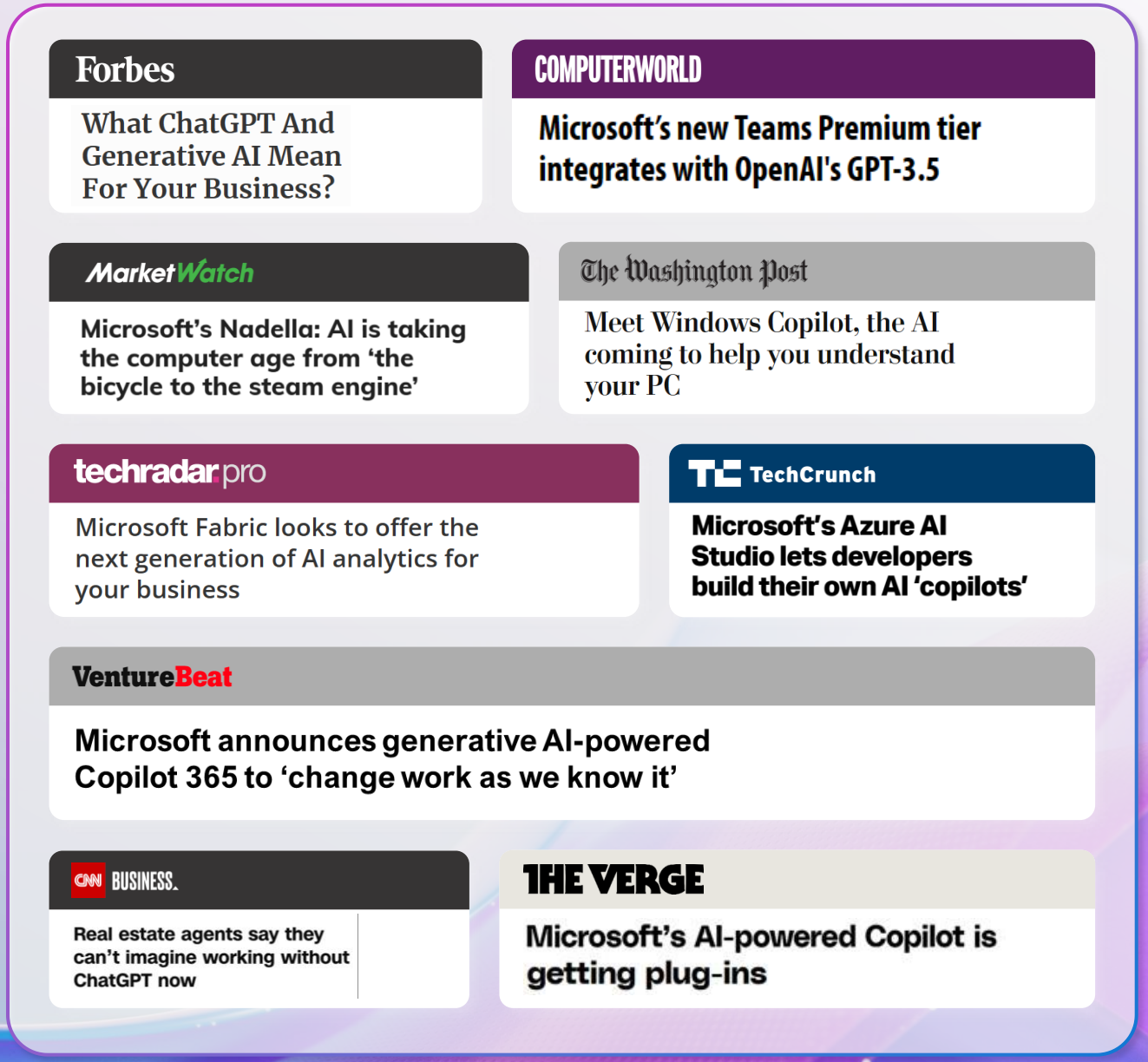
- Explore, build, test, deploy, and manage custom copilots using interactive visual and code-first workflows
- Access out-of-the-box and customizable APIs and models
- Systematically evaluate model and app responses and pinpoint fine-tuning opportunities
- Scale copilots for use in websites, applications, and other production environments

**AI Toolkit for Visual Studio Code**
*Agnostic*

- Simplify app development by bringing together AI development tools and models
- Explore, try, and integrate AI models
- Use local and cloud compute to fine-tune and optimize small language models for app-specific use cases on the cloud and edge
- Package fine-tuned models as containers and deploy to Azure or the edge

# The Future of Work with AI

# The future of AI is here



# Organizations are ready to embrace the transformational potential of AI

**87%**

Of organizations believe AI will give them a competitive edge

**40%**

In a recent study, skilled workers using generative AI saw a 40% boost in the quality of their work

For every $1 a company invests in AI, it is realizing an average return of

**$3.5**

**14**

**months**

Average time it takes for organizations to realize a return on their AI investment

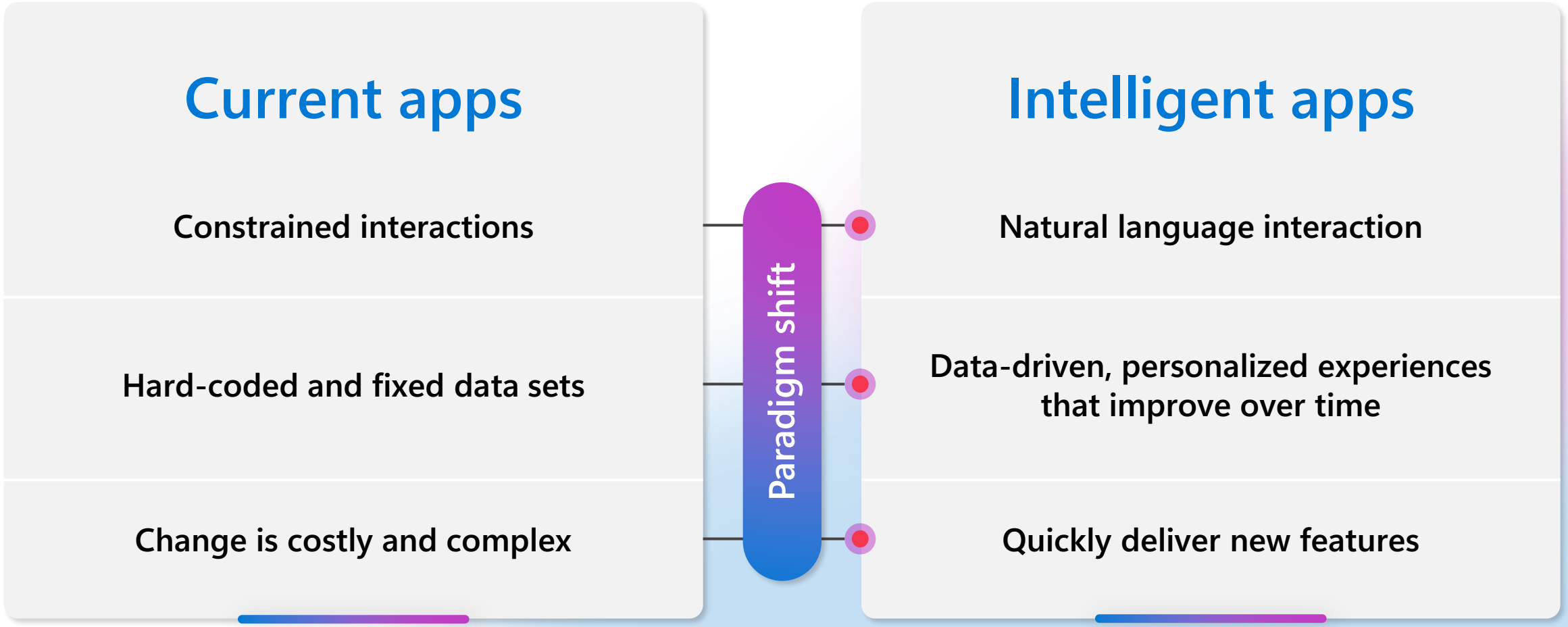
Source: Harvard Business School Technology & Operations Management / IDC The Business Opportunity of AI November 2023 / MIT Sloan Management Review

# What could slow down generative AI adoption?

| | |
|---|---|
| **Getting started** | The state of the art is evolving so quickly, it makes it difficult to decide what to use. Along with that, guidance and documentation is hard to find |
| **Development** | Applications often require multiple cutting-edge products and frameworks which requires specialized expertise and new tools to stitch these components together |
| **Context** | Generative AI doesn't know about your data |
| **Evaluation** | It is hard to figure out which model to use and how to optimize for their use case |
| **Operationalization** | Concerns around privacy, security, and grounding. Developers lack the experience and tools to evaluate, improve, and validate the solutions for their Proof of Concepts, and to scale and operate in production |
| **Experience** | Less than 30% of surveyed executives say their organizations have the in-house expertise needed to adopt and scale generative AI.[1] |

[1]Source: IBM

# Add generative AI to your apps to gain true intelligence



# Developers want...

To learn how to select the right tools and models for their use case

Any easy way to switch between UI and code as they build

An orchestration framework to handle the complex mapping of functions and code

Helpers, building blocks, and app templates

To evaluate the model, prompt, and overall application

To scale their PoCs into production

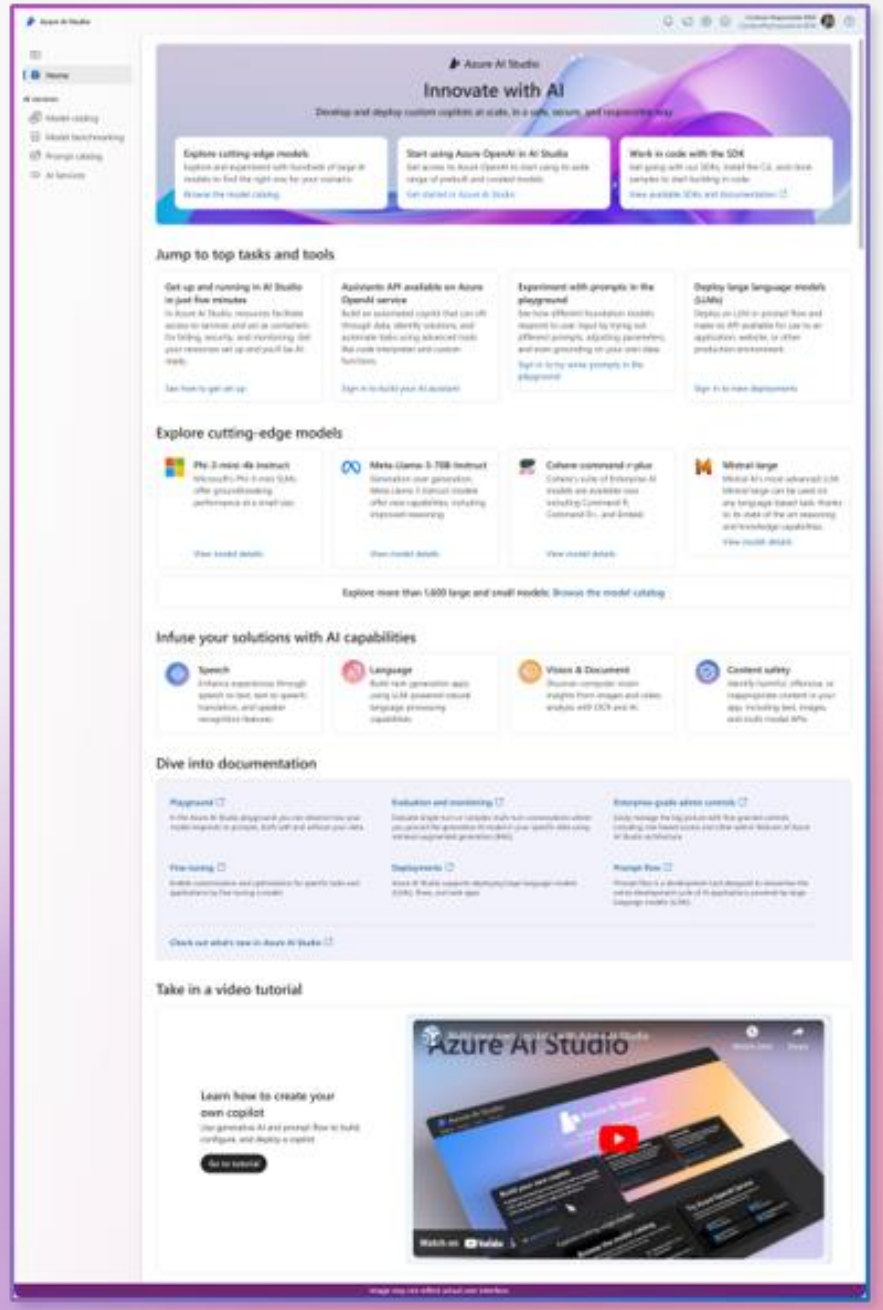
To continuously monitor AI app performance

# Build your own copilot with Azure AI Studio

# Azure AI Studio

## Generally available

ai.azure.com




Microsoft Azure

Generally available

# Azure AI Studio

## A comprehensive platform to develop and deploy custom copilots

| API and model choice | Complete AI toolchain | Responsible AI tools and practices | Enterprise-grade production at scale |
|---|---|---|---|

AI.Azure.com

# Get to know Azure AI

## Azure AI Studio

One place for building and deploying AI solutions

- API & Model Choice
- Complete AI Toolchain
- Responsible AI Tools & Practices
- Enterprise-grade Production at Scale

### Cutting-Edge Models

Access to the latest foundation and open-source models

- Model Catalog
- Models As a Service
- GPT Model Family
- Open-Source Models
- Small Language Models

### Azure AI Services

Pre-trained, turnkey solutions for intelligent applications

- Azure OpenAI Service
- Azure AI Search
- Azure AI Speech
- Azure AI Vision
- Azure AI Content Safety
- Azure AI Document Intelligence
- Azure AI Language
- Azure AI Translator

### Azure Machine Learning

Full-lifecycle tools for designing and managing responsible AI models

- Prompt Flow Orchestration
- Responsible Model Design
- Model Fine-Tuning
- Model Training

## Azure AI Infrastructure

State-of-the-art silicon and systems for AI workloads

- High-Bandwidth Networking
- Microfluidic Cooling
- Azure Maia Silicon

Top use cases for

# Azure AI Studio

**Build your own copilot**
Your data. Your apps. Your people

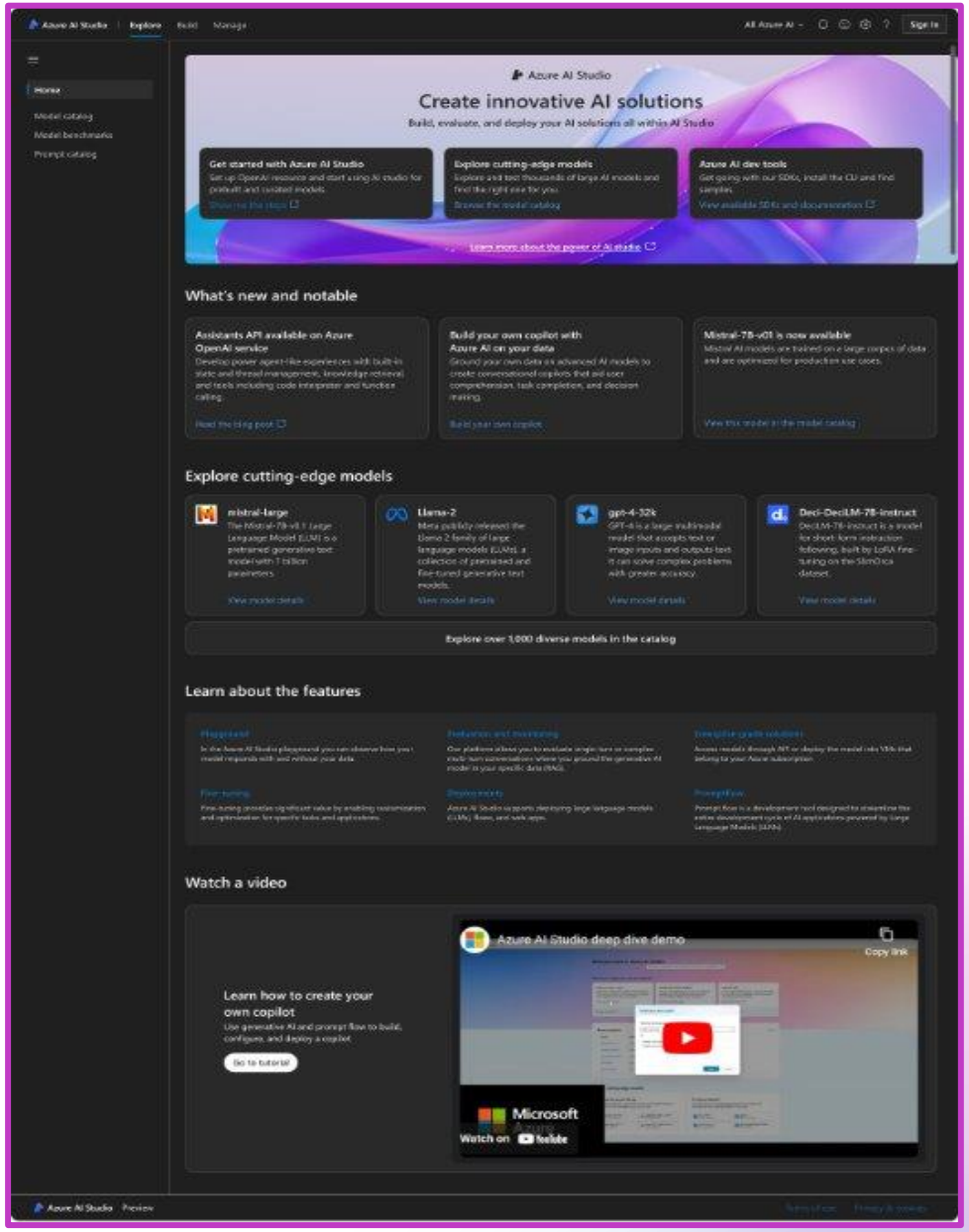
**Enterprise chat**
Provide multi-modal knowledge mining

**Speech analytics**
Improve interactions and service

**Content generation**
Deliver new products and services

**Hyper-personalization**
Support better sales and marketing

# Accessibility to empower developers of all abilities to thrive in the age of AI




Azure AI Studio was designed and built with feedback from developers with disabilities

# API and Model Choice

# API and Model Choice

## Identify the best AI services and models for your use case

### Build intelligent apps with industry-leading APIs*

Develop multimodal, multi-lingual AI with out-of-the-box and customizable APIs and models.

### Discover models for your use case

Choose from the latest cutting-edge foundation and open models. Benchmark models by performance and key parameters and deploy what's right for your use case.

### Deploy serverless models

Get started quickly with ready-to-use APIs, without the need to provision GPUs.

*Microsoft is a Leader in the 2023 Gartner® Magic Quadrant™ for Cloud AI DeveloperServices | Microsoft Azure Blog

# Build intelligent apps with Azure AI services

## Leverage out-of-the-box and customizable APIs and multimodal models

### Azure OpenAI Service

- Access to powerful AI models
- Scalable development
- Compliance & security
- Integration with other Azure Services

### Azure AI Search

- AI enrichment & semantic ranking
- Generative AI content creation
- Vector search for data organization

### Azure AI Speech

- Speech to text (Whisper)
- Text to speech
- Speech translation
- Speaker recognition

### Azure AI Vision

- Image and face analysis
- Custom model training
- Face detection and recognition
- Document text extraction

### Azure AI Content Safety

- AI-driven content moderation for enhanced safety
- Customize safety thresholds for diverse user types
- Detect and prevent Jailbreak Risk from XPIA attacks

### Azure AI Document Intelligence

- Automated documentation generation
- Documentation quality analysis
- Interactive documentation experiences
- Natural language understanding for documentation

### Azure AI Language

- Customized translation for industry-specific terminology
- Seamless translation of technical documents and manuals
- Supports multilingual comprehension for diverse audiences

### Azure AI Translator

- Multilingual text and speech translation
- Synchronous and asynchronous translation request support
- Utilizes machine learning for highly-accurate translations

# Multimodality is here

Language

Speech

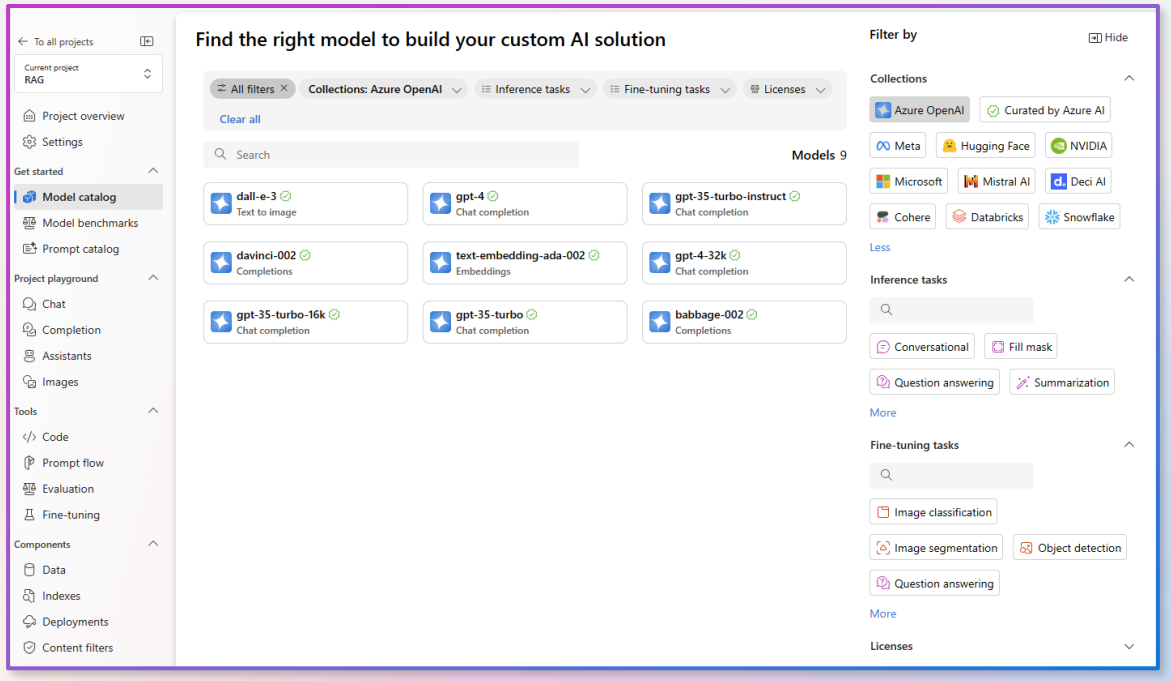
Vision

# Azure OpenAI Service (AOAI)

Azure OpenAI Service empowers developers and businesses to harness the power of advanced AI technologies without the need for extensive AI expertise or infrastructure investment.

- Access advanced AI models like GPT for diverse applications and tasks
- Generate high-quality imagery with DALL-E 3 through textual descriptions and text-prompt
- Transcribe and translate audio with Whisper, converting 30-second segments into text across 57 languages
- Integrate seamlessly with Azure AI services, enhancing AI capabilities and workflows within existing applications
- Customize and fine-tune AI models to suit specific use cases and requirements effectively
- Ensure compliance and security standards, adhering to responsible AI practices and regulations for trustworthy deployments

**Build intelligent apps with industry leading AI**

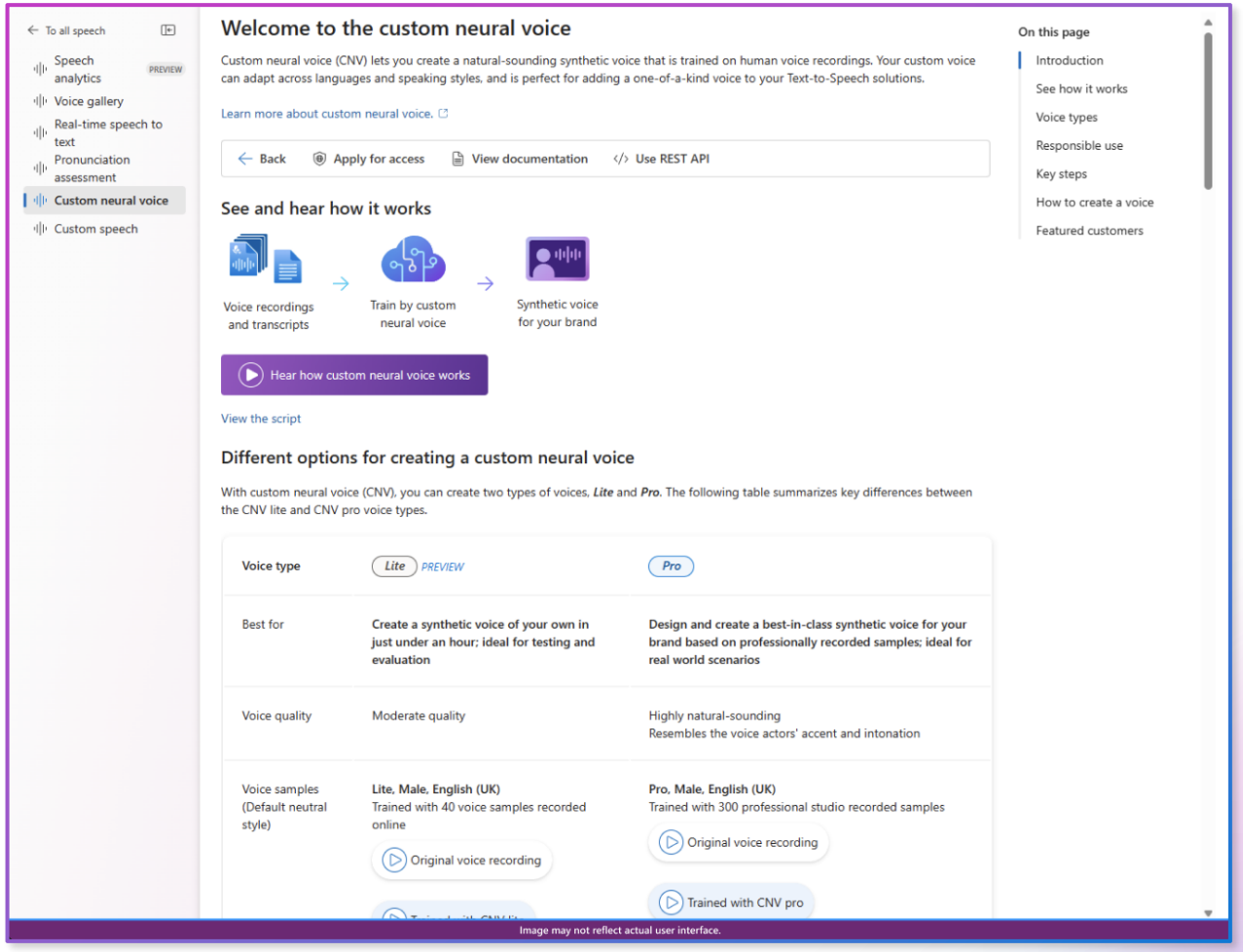



**Leverage OpenAI's state-of-the-art models and Azure's cloud platform to build innovative AI-driven applications and solutions across various industries and domains.**

# Create your branded AI voice

Adapt across languages and speaking styles

Deploy one-of-a-kind to text speech solutions



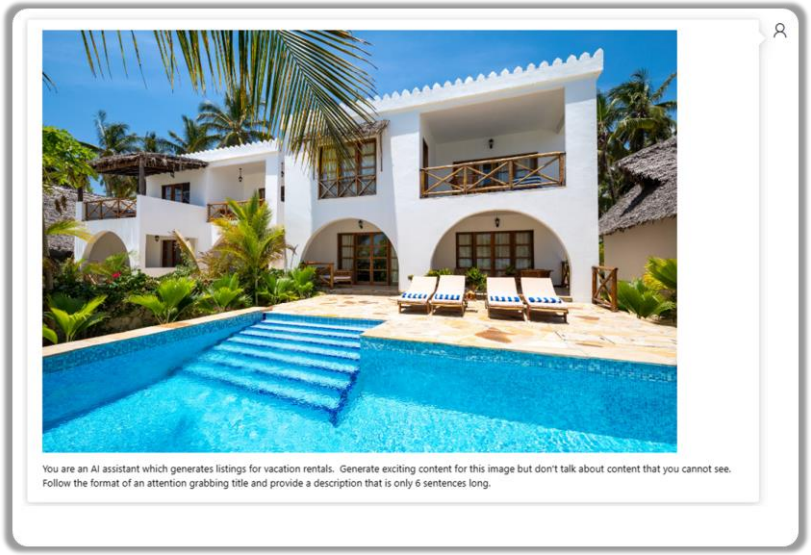
# Azure AI Vision

**GPT-4 TURBO ENHANCEMENTS**

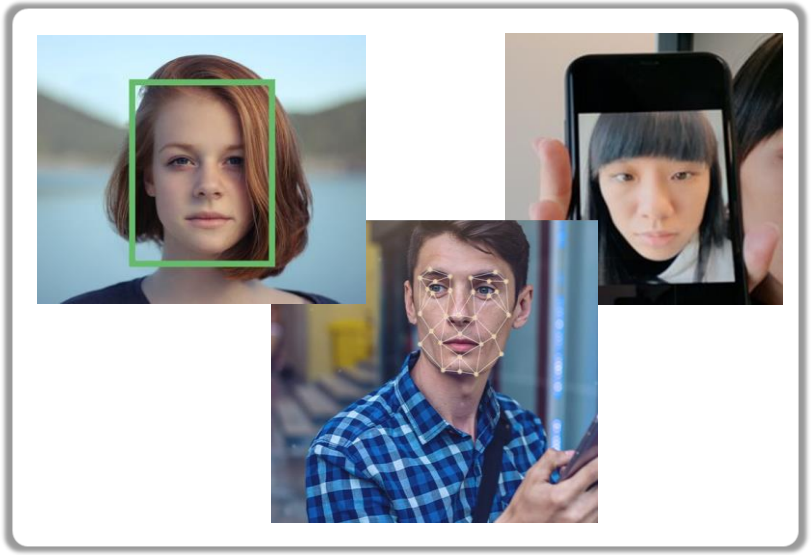
| Video Prompting | Object Grounding | Dense Text Accuracy |
|---|---|---|

**HIGH QUALITY PRE-TRAINED MODELS**

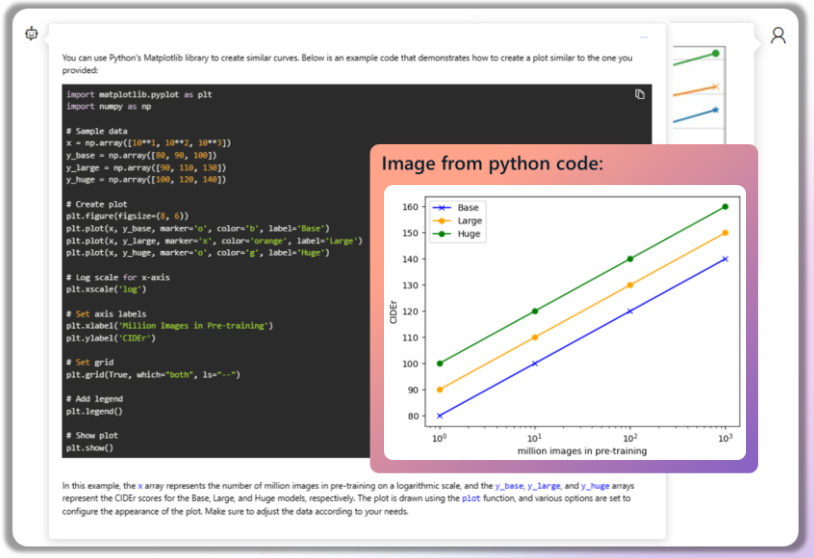
| OCR | FACE | IMAGE ANALYSIS | SPATIAL ANALYSIS |
|---|---|---|---|
| Printed Text Extraction | Detection | Tagging & Captioning | Person Tracking |
| Handwritten Extraction | Recognition & Liveness | Content Moderation | Object Tracking |




Analyze objects in images and videos and provide detailed descriptions



Detect and recognize facial liveness




Convert images into code

# What can Speech Analytics do for you?

## Multiple Use Cases

**Flexible service catering multiple use cases, e.g.:**

- Summarize conversations or videos and generate additional meta-data
- Generate Key Insights and extract meaningful metrices
- Generate additional results like PII redacted recordings, video captions, dubbed videos and more

## Integrated Service

**Significant reduction of effort to implement and maintain based on an Azure AI project**

**Pre-built integration of AI capabilities optimized for processing of audio content**

## From PoC to production

**Easy to setup and get started with analytics scenario templates for common use cases**

**Scales effortlessly from proof-of-concept to productive workloads**

## Flexibility

**Start with pre-built analytics allowing you to easily make changes to cater your business needs**

**Simple experience for basic changes (e.g., parameters, prompts)**

**Allow rich customization for advanced needs (e.g., flows)**

# Azure AI Model Breadth

## Offering a wide collection of foundation and open models

### Azure OpenAI Service

- GPT-4-Turbo
- GPT-4
- GPT-4V
- Text-embedding-ada-002
- GPT-3.5-Turbo

### Meta

- Llama-2-70b/70b-chat
- Llama-2-13b/13b-chat
- Llama-2-7b/7b-chat
- CodeLlama

### Mistral AI

- Mistral Large
- Mistral 7b
- Mixtal 7b*8 – Mixture of Experts

### cohere

- Cohere R+*
- Cohere R*
- Embed v3-Multilingual*
- Embed v3-English*

*Available as pay-as-you-go

### Hugging Face

- Falcon/TII
- Stable Diffusion/Stability AI
- Dolly/Databricks
- CLIP/OpenAI

### nvidia

- Nemotron-3-8B-4k
- Nemotron-3-8B-Chat-SFT/ RLHF/ SteerLM
- Nemotron-3-8B-QA

### databricks

- Databricks/dbrx-base
- Databricks/dbrx-instruct

### Small language models

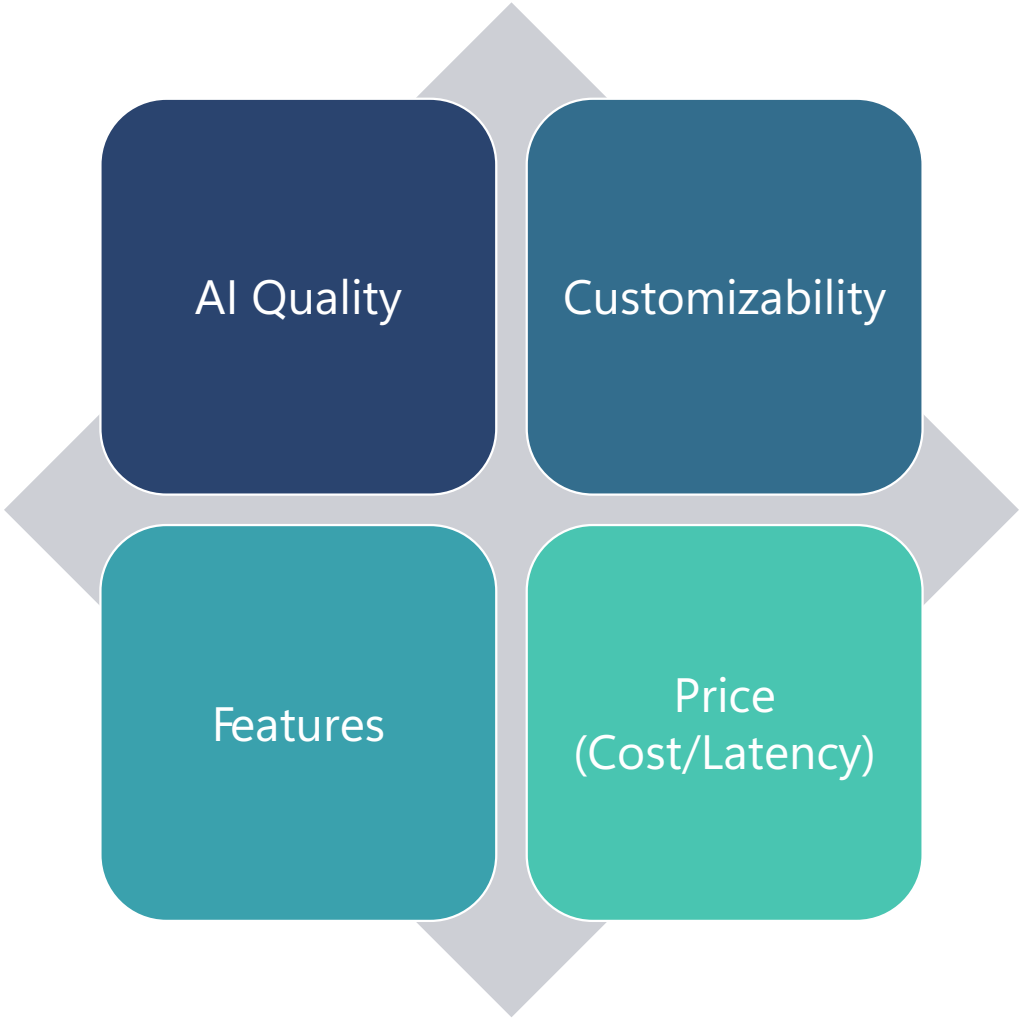
**Phi**

- Phi-1
- Phi-1.5
- Phi-2

**Orca**

- Orca 1
- Orca 2

# Considerations while selecting a model



# Model Comparison

(Cost/Latency)

AI Quality


Features

Customizability

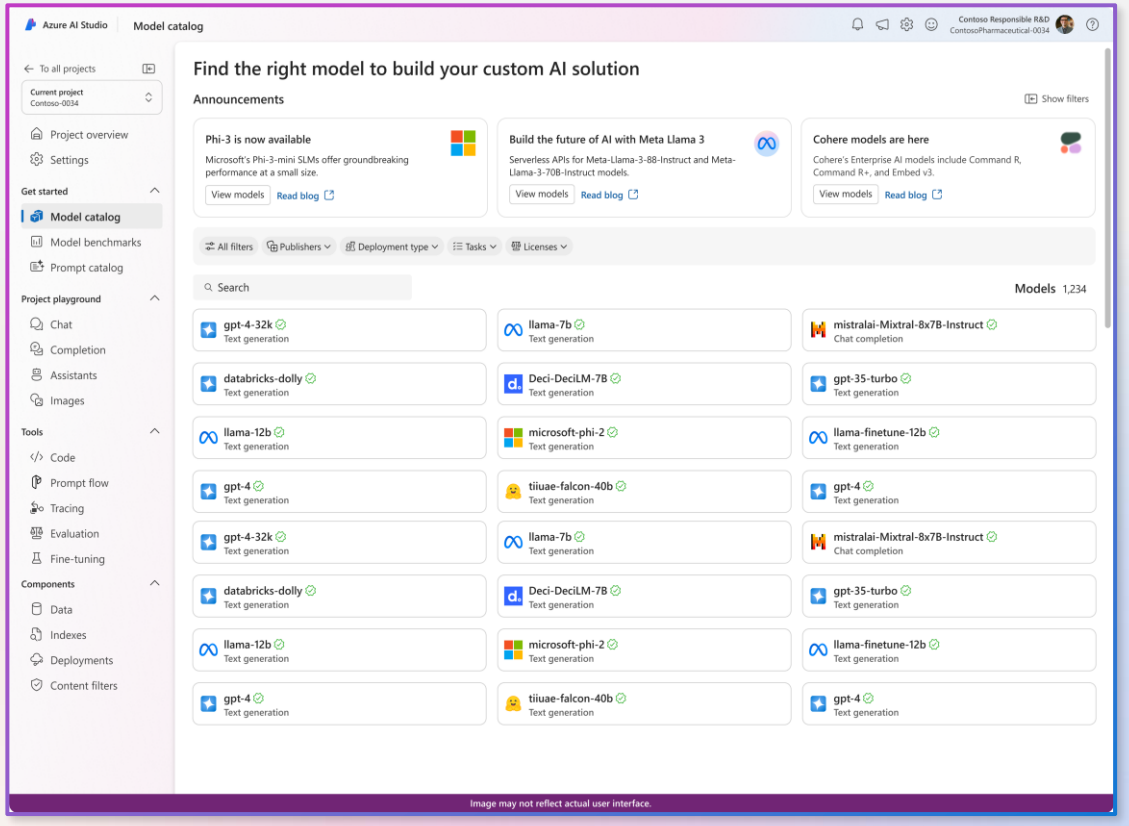
Gemma-7b  Mistral-7b  Mixtral-8x7b  Llama-3-8B-In  GPT-4  Phi-3-Small-8K-In

# Model catalog

API and Model Choice

The model catalog enables discovery of large and small language models packaged for out-of-the-box or custom usage in Azure AI Studio.

- Models from OpenAI, Hugging Face, and Meta, including Llama-2-7b, GPT-4V, and GPT-35-turbo
- Quickly try out any pre-trained model using the Sample Inference widget on the model card, providing your own sample input to test the result
- Filter model catalog by collection, model name, or task to find the model that best suits your needs
- Compare models by task using open-source datasets
- Leverage ready-to-use, curated fine-tuning pipelines, code-based inferencing, and evaluation of the mode



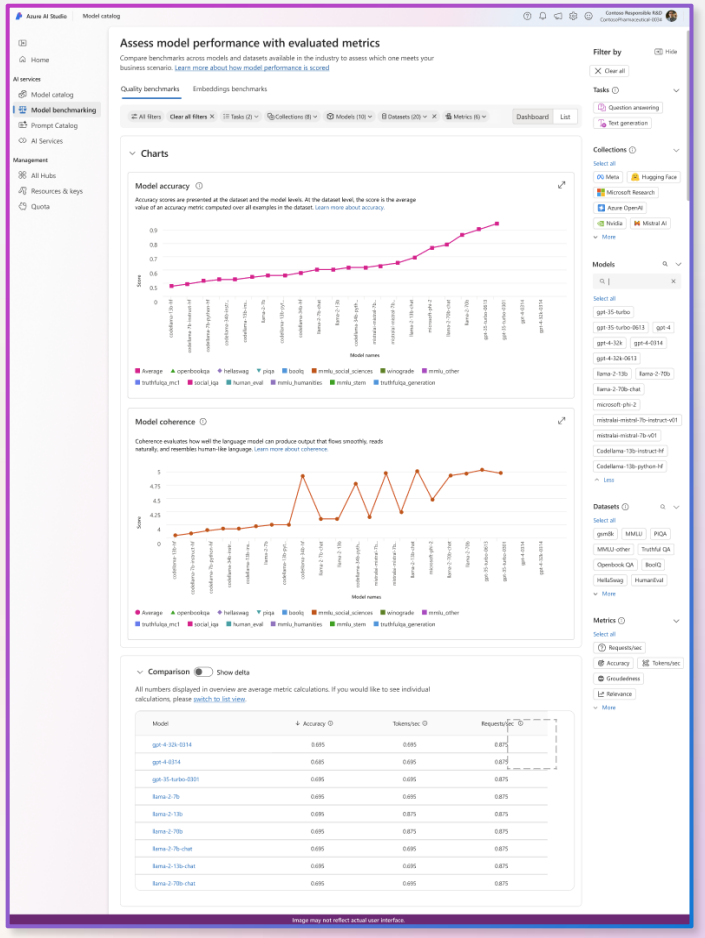

Easily find the foundation model that best suits your needs while improving business outcomes, reducing operational expenses, and enhancing competitive advantage.

# Model benchmarking

Model Benchmarking enables you to review and compare the performance of various AI models, simplifying the selection process and allowing users to make confident choices with their modeling needs.

- Gain quality metrics for Azure OpenAI Service, Llama 2 family, Code Llama, and Mistral models
- Access pre-built metrics and benchmark comparison models within the same build, train, deploy environment
- Compute accuracy scores at both the task (dataset) and model levels by utilizing public datasets, yielding a model score for each dataset
- Compare scores of multiple models across datasets and tasks
- Benchmark results originate from public datasets; Azure AI evaluation pipelines download data from original sources, extract prompts from each row, generate model responses, and then compute relevant accuracy metrics

API and Model Choice




Enable the comparison of models based on accuracy and empower users to make data-driven decisions, ensuring their AI solutions are optimized for the best performance.

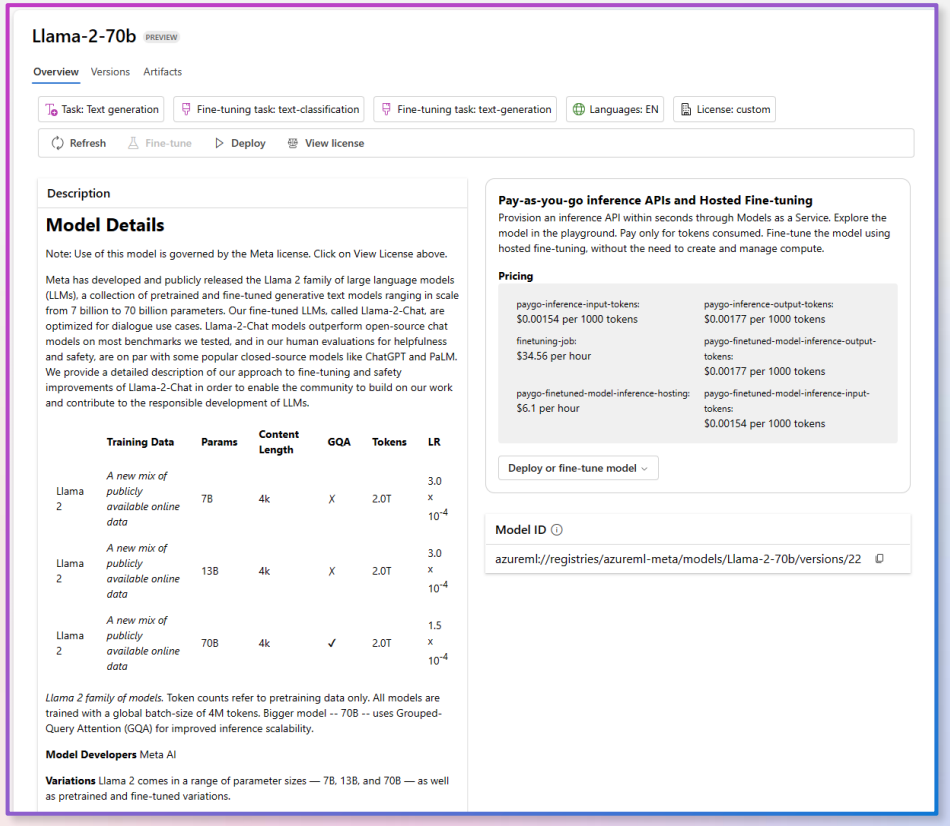
# Models as a Service (MaaS)

*Offering within the model catalog*

Models as a Service (MaaS) lets you fine-tune models without provisioning compute, making it easier for generative AI developers to build custom copilots

- Ready-to-use APIs with pay-as-you-go billing based on tokens
- Customize models with your own data without the need to set up and manage GPU infrastructure
- Easily integrate the latest AI models as an API endpoint to applications
- Integrate with preferred orchestration tools like prompt flow, Semantic Kernel, or LangChain
- Achieve serverless fine-tuning without provisioning GPUs
- Fine-tune Llama 2 with your own data to enhance the model's ability to generate more precises predictions

**API and Model Choice**




MaaS provides simplified management with ready-to-use GPU provisioning, lowering costs and reducing barriers to adoption by eliminating complexity.

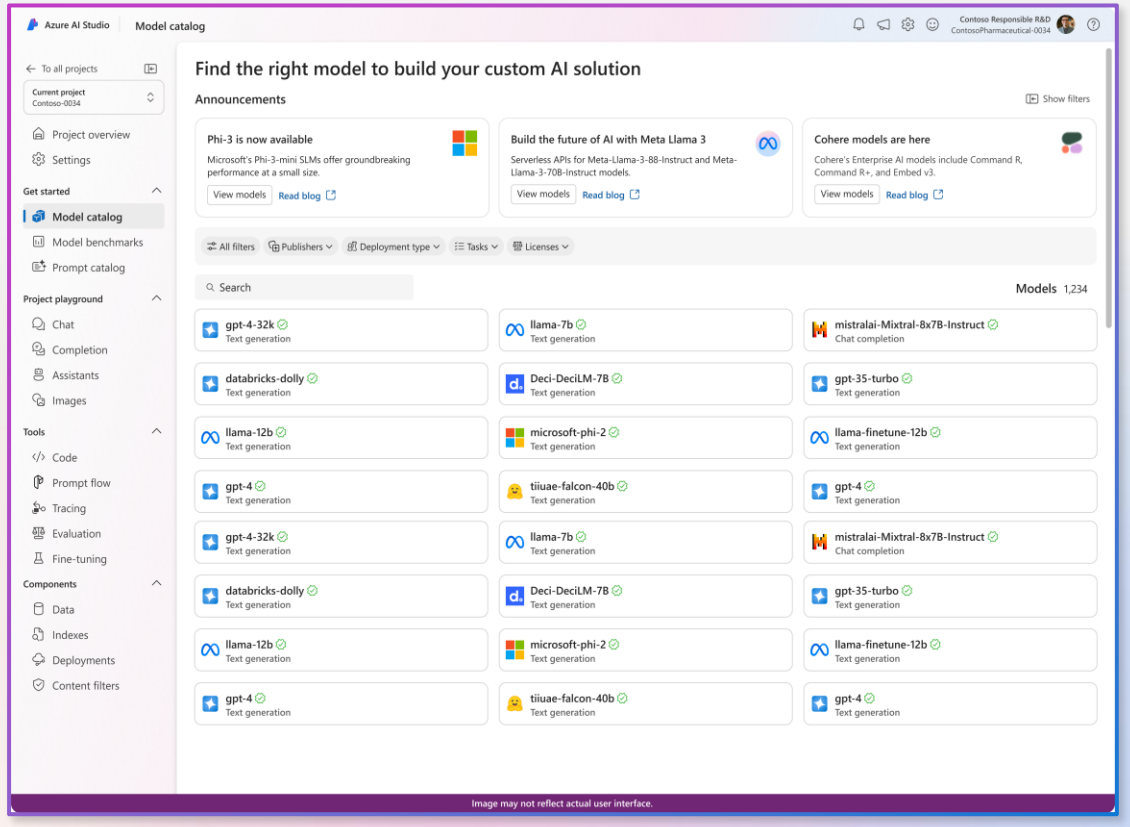
# Models as a Platform (MaaP)

*Offering within the model catalog*

Models as a Platform (MaaP) lets you innovate with models from vetted providers using a generative AI model store to explore large foundation models curated by Microsoft, Hugging Face, Meta, and other open-source contributors.

- Explore thousands of large foundation models (LLMs and SLMs) packaged for out-of-the-box usage and optimized for Azure AI Studio
- Manage your own compute resources with a self-managed GPU infrastructure
- Enhance model performance with highly customized fine-tuning
- Strengthen data security and privacy by controlling the network environment with managed virtual network scenarios
- Access shared compute resources and temporary endpoints for testing, available for seven days and intended for use in testing scenarios

**API and Model Choice**




MaaP offers scalable, self-managed hosting for greater control and customization, giving developers flexibly to choose from the most comprehensive selection of open-source generative AI models.

# Complete AI Toolchain

# Complete AI Toolchain

Access collaborative, comprehensive tooling to support the development lifecycle and differentiate your apps

### Ground models on your protected data

Use your protected data to customize models with advanced fine-tuning and for contextually relevant retrieval augmented generation.

### Orchestrate and debug AI workflows

Streamline app development with easy-to-use prompt orchestration, tracing, and debugging via interactive visual and code-first workflows.

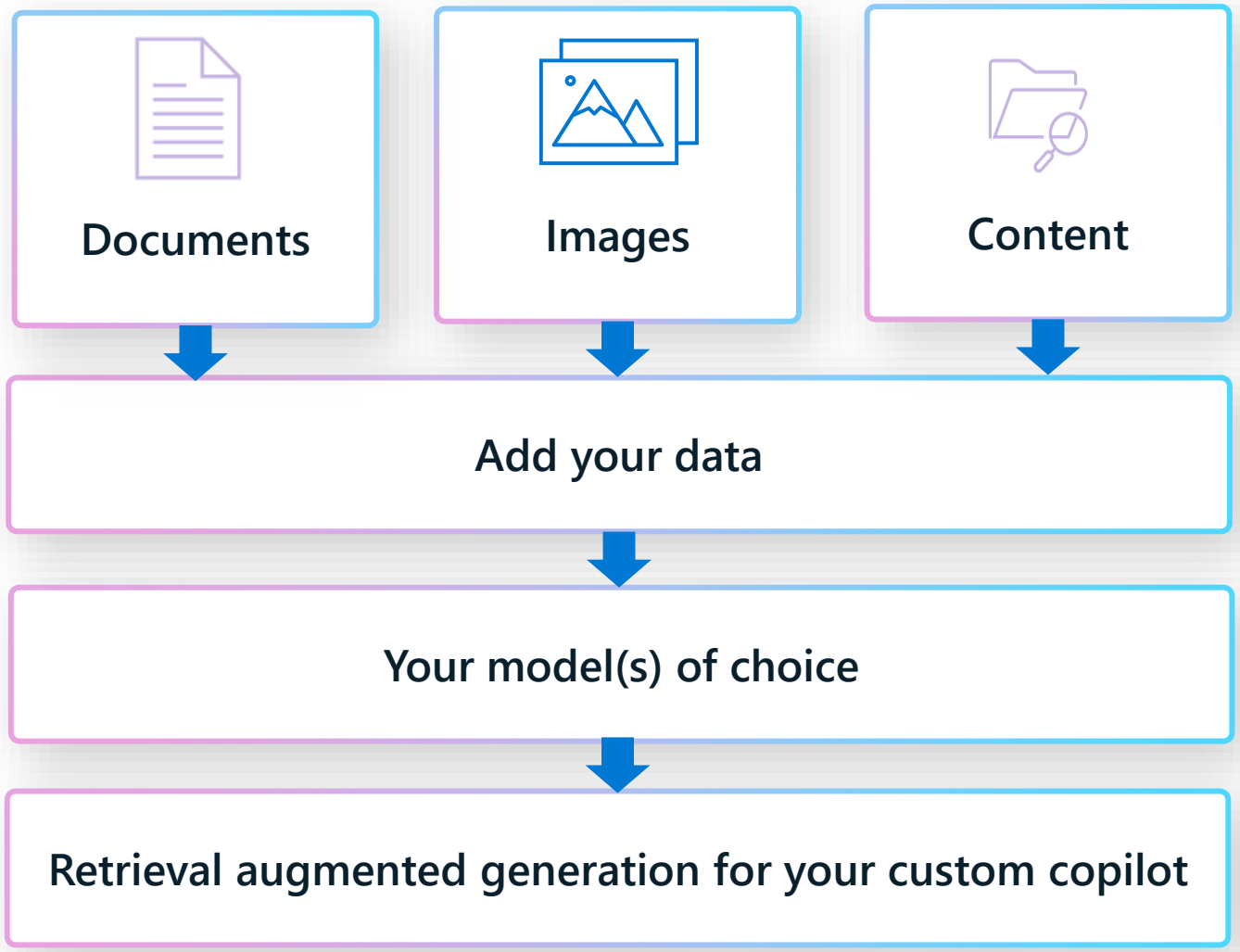
### Streamline model and app evaluations

Systematically measure and improve app performance with manual and automated evaluations.

# Designing your orchestration flow

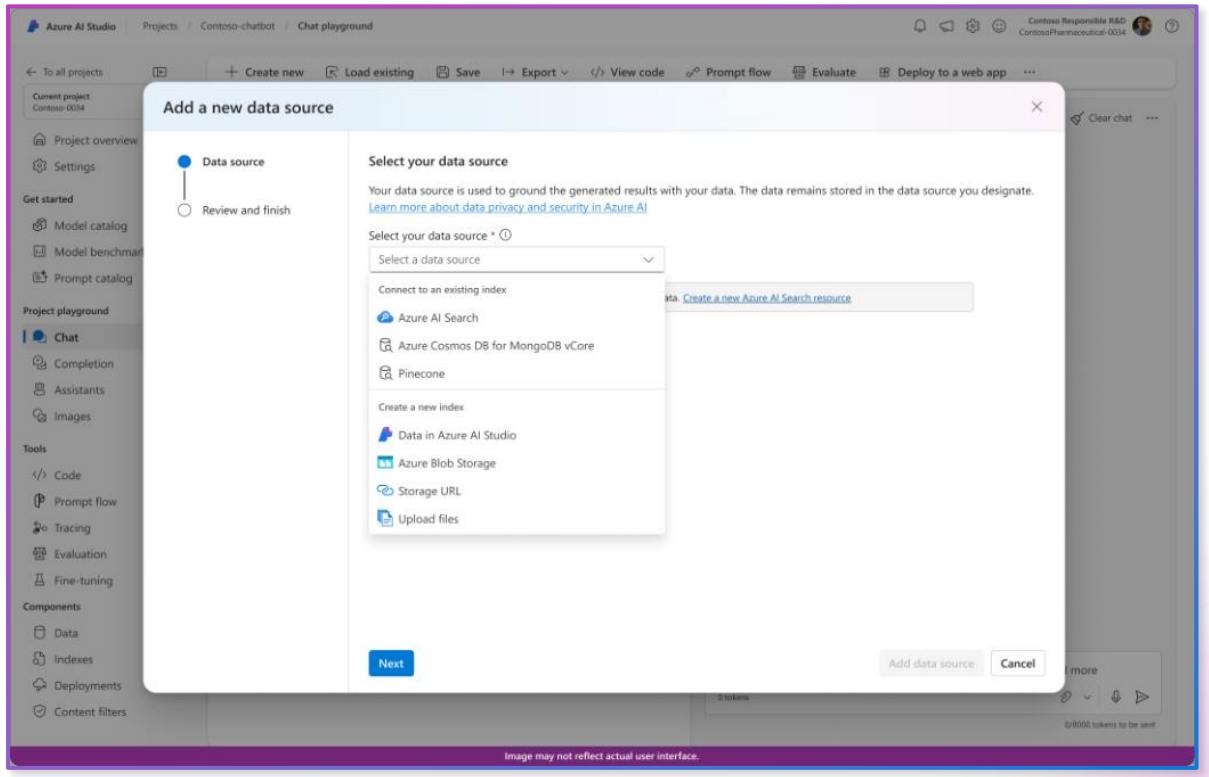
| | RAG | Function/API | Agent |
|---|---|---|---|
| **Scenario**<br><br>*example* | One well-defined task (Q&A)<br>Known outcomes (answers)<br><br>*looks for answers in product documentation* | Multiple tasks<br>Controlled outcomes (calls)<br><br>*query an API to validate account number* | Complex task<br>Open outcomes<br><br>*chain multiple APIs to solve a problem* |
| **Models** | Question answering<br>(understanding/summarization) | Intent detection<br>and planning | Multiple |
| **Orchestration** | Systematic<br>workflow | Flexible<br>Resilient<br>*(code is the trigger)* | Dynamic<br>Self-healing<br>*(agent is the trigger)* |

# Retrieval Augmented Generation



# Integrate Structured & Unstructured Data

- **Azure AI Search**
- **Blob storage**
- **Local files/folders**
- **Storage URLs including OneLake in Microsoft Fabric**
  - Azure Databricks
  - S3 buckets via Amazon S3 Shortcuts

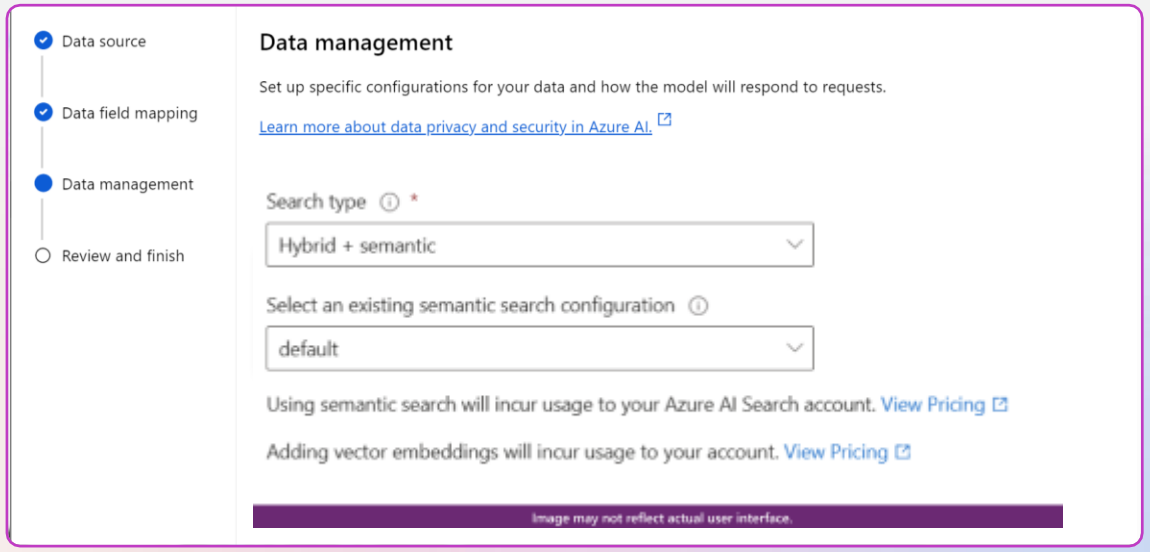


# Azure AI Search

Ground your models on your protected data

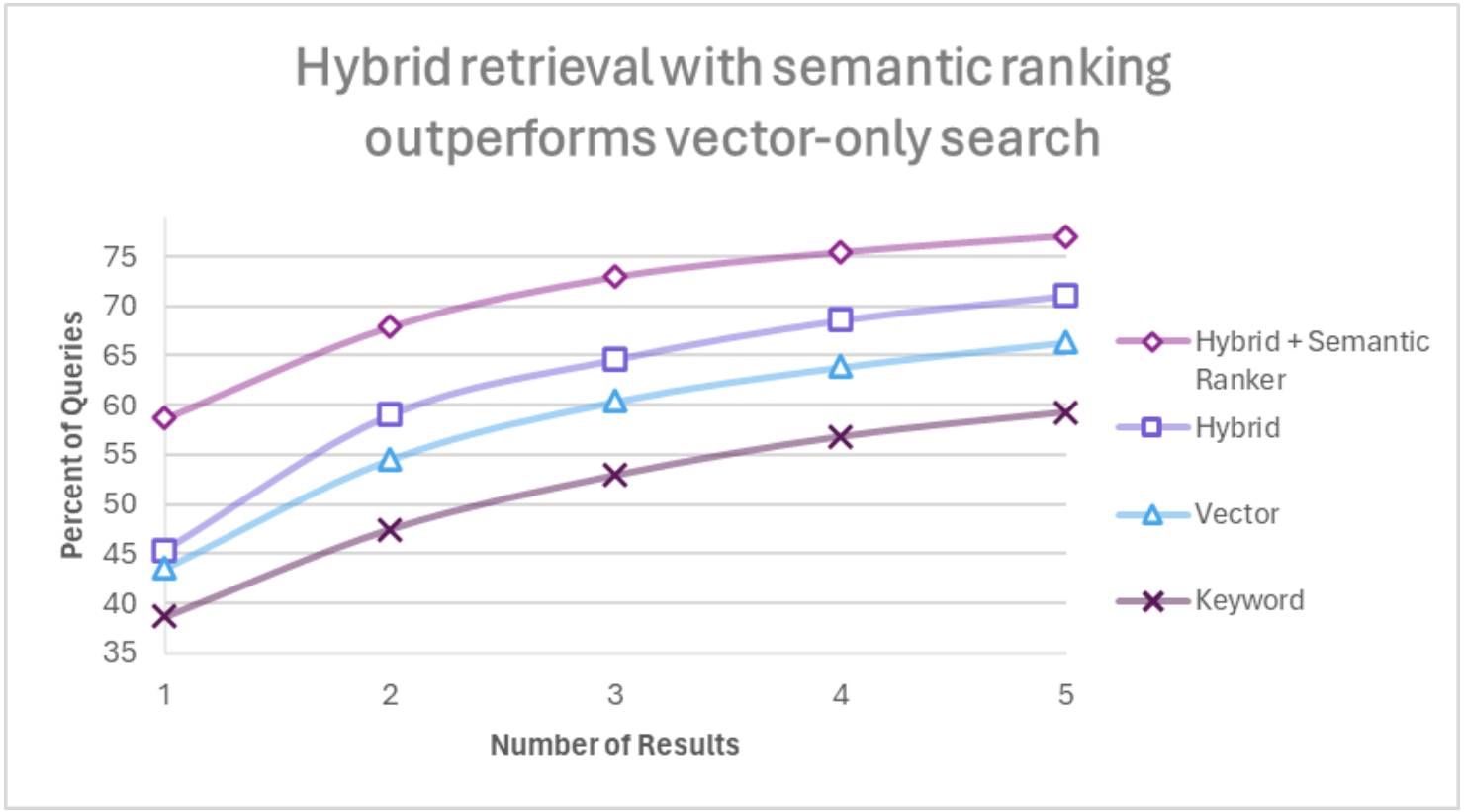
Azure AI Search is an enterprise-ready retrieval system with a robust set of advanced search technology, enabling superior retrieval augmented generation (RAG) so your app delivers the best experience for every user, at any scale.

- Ingest and search across all your data, no matter the type, volume or source
- Streamlined data preparation and indexing with automatic data ingestion, chunking, extraction and enrichment.
- Feature-rich vector database with support for Exhaustive KNN search and ANN search
- Enable hybrid search (vector + keyword) results using RRF (Reciprocal Rank Fusion) and multi-vector retrieval for multi-modal search.
- Improve relevance immediately with semantic ranker, a transformer-based reranking step that runs on GPUs to deliver speed and quality in results.
- Tailor search experiences by utilizing customizable features such as granular access control and relevance tuning



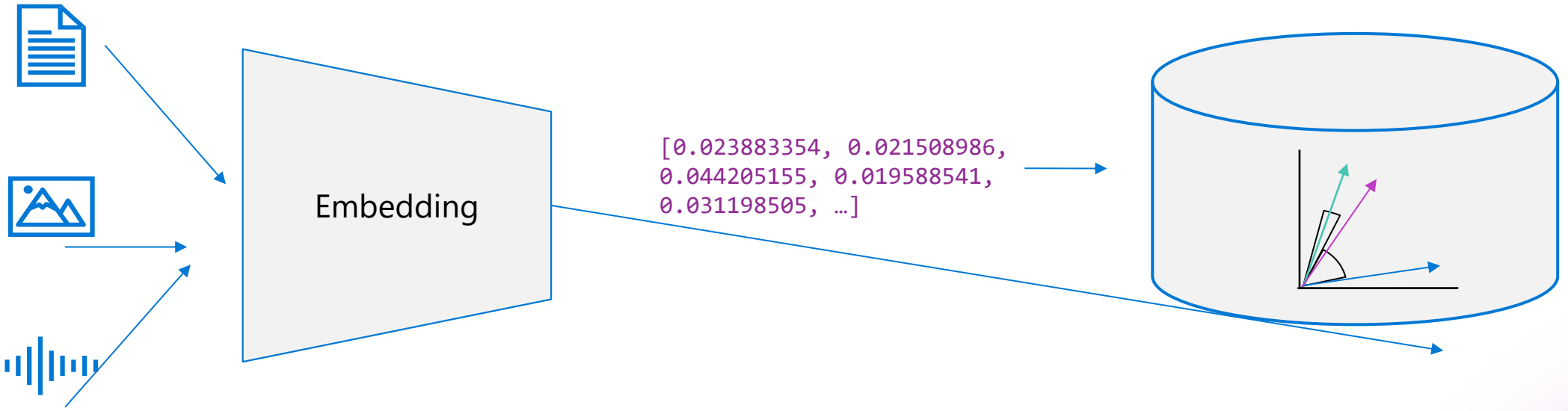

**Easily enable search best practices through the UI, backed by extensive product tests and benchmarks, like being able to select hybrid search + reranking in one click.**

# Retrieve Using Semantic Similarity and Hybrid Search



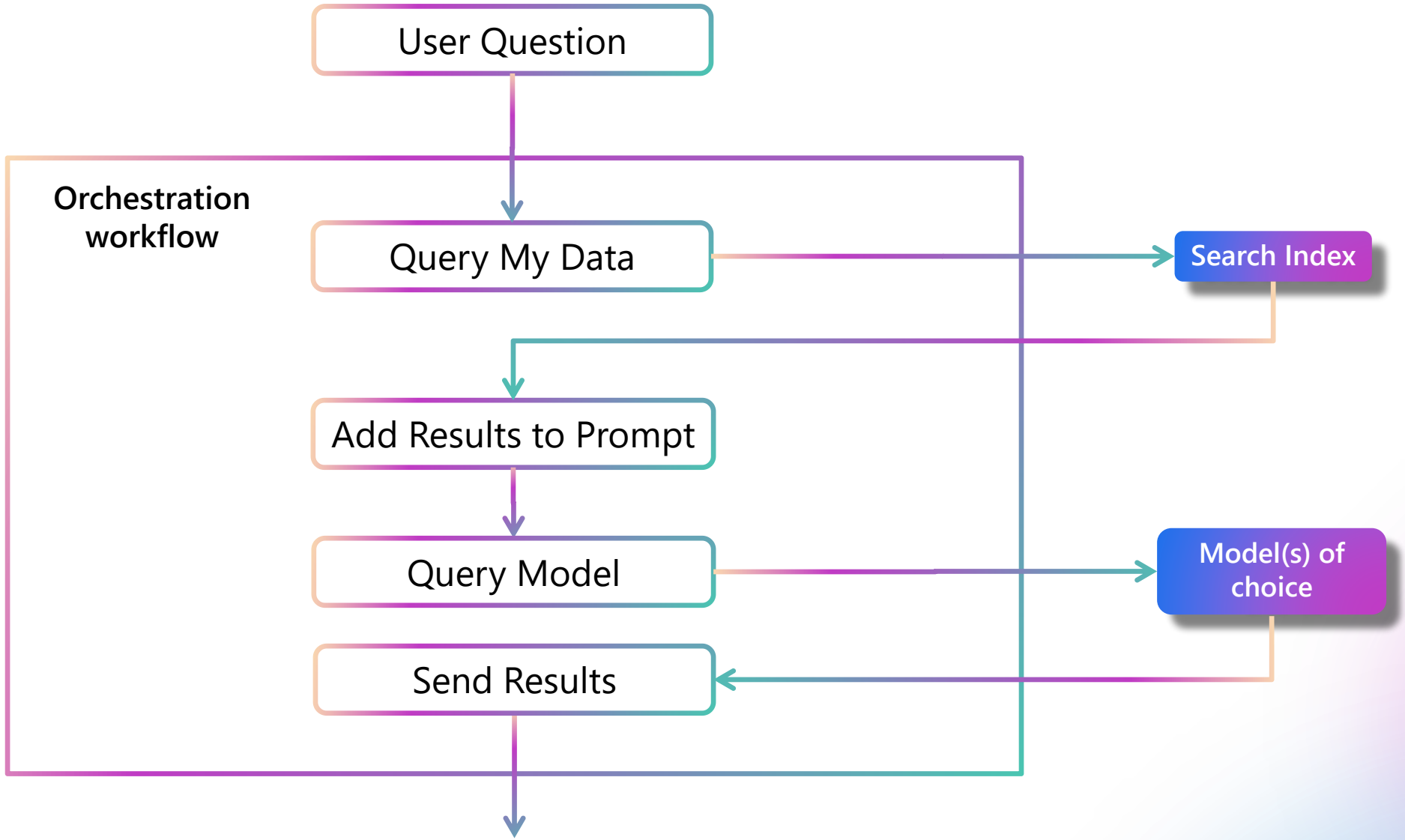

Find the most relevant information in large datasets using hybrid search with keywords, vectors, and semantic ranker

# Retrieve Using Semantic Similarity and Hybrid Search




Find the most relevant information in large datasets using hybrid search with keywords, vectors, and semantic ranker

# Retrieval Augmented Generation Orchestration



# Agents

AI tools that respond to events and can operate independently of user input (after initial and potentially ongoing user authorization) to complete complex objectives that span multiple applications, data sources, and services.

- Build sophisticated stateful agents
- Augmented your copilot to access multiple APIs
- Perform complex computations and data analysis
- Safely act on a user's behalf
- Retrieve useful data in multiple formats

# Chat vs. Agent frameworks

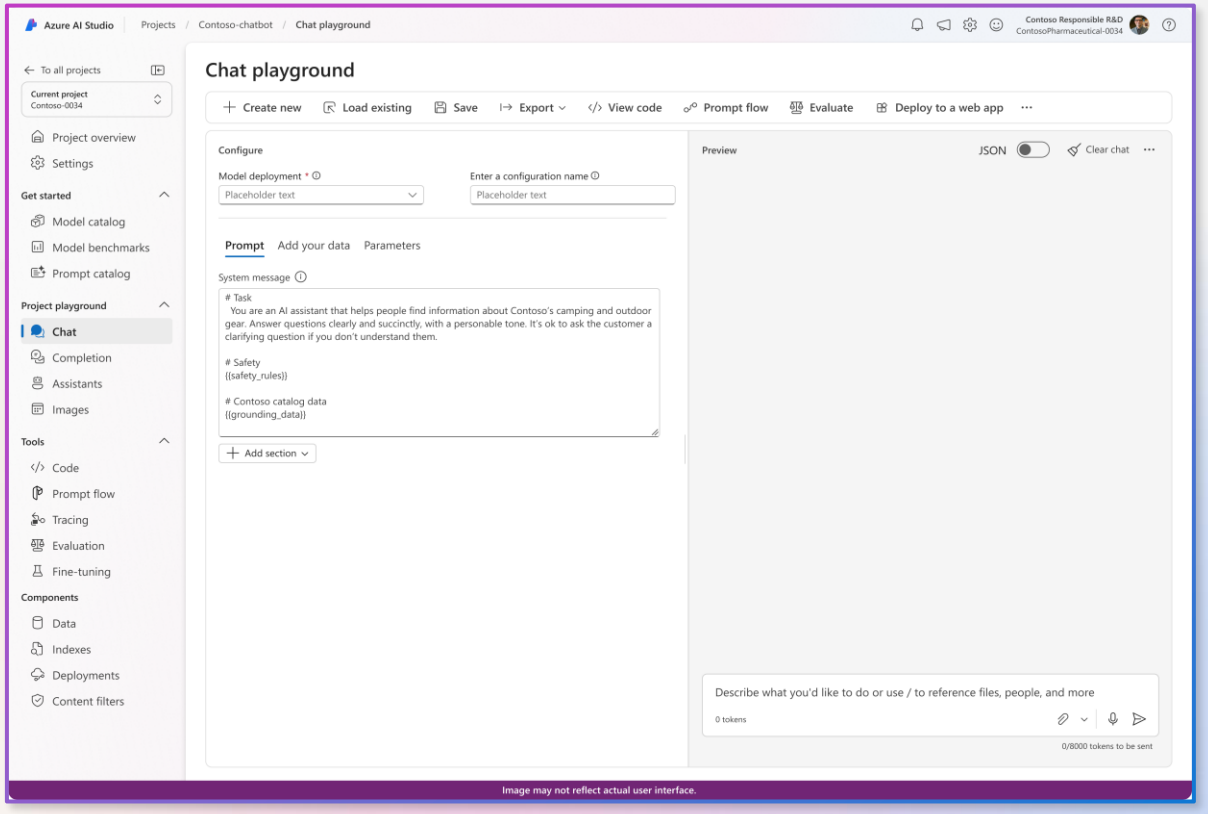
| Chat | Agents |
| --- | --- |
| Lightweight and powerful | Stateful (inbuilt conversation state management) |
| Inherently stateless | Access persistent threads |
| | Access files in several formats. API handles chunking, embeddings, storage and creation, and implementing vector search |
| | Automatic management of the model's context window |
| | Access multiple tools in parallel (up to 128 tools per assistant) including code interpreter |
| | Build your own function calling |

# Azure AI Studio playground

The playground in Azure AI Studio offers an environment for developers to experiment with AI models, algorithms, and data.

- Construct system instruction prompts to create your own AI agents and copilots
- Easily connect with indexed data sources to implement retrieval-augmented generation (RAG)
- Leverage prompt samples and templates to eliminate writing long prompts from scratch
- Easily import prompt samples from local or export for local development
- Pair models with Azure AI services APIs to enable richer interactions
- Click "Customize in prompt flow" to further enrich and manage large language model (LLM) workflows

**Ground your models on your protected data**



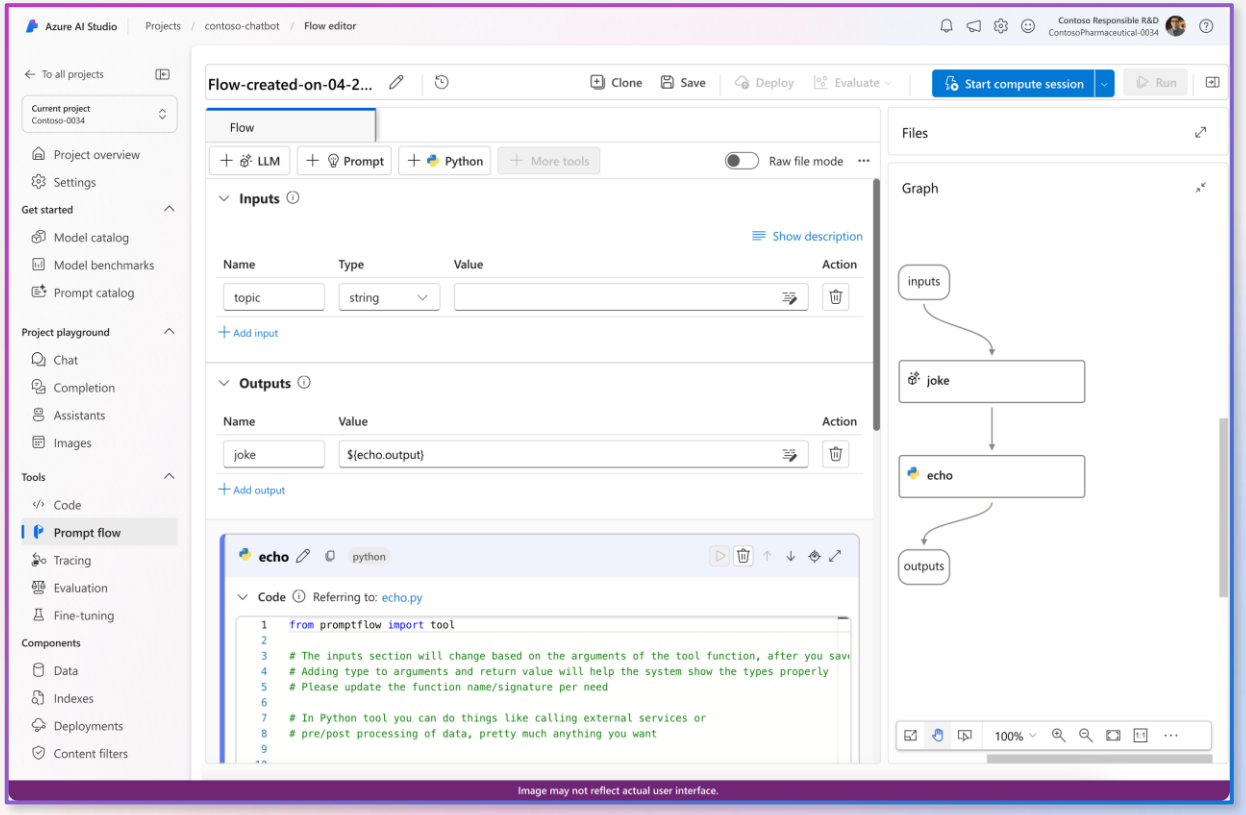

**Give developers the freedom to experiment, innovate, and learn without constraints of production environments.**

# Prompt flow

Prompt flow orchestrates app development by facilitating iterative experimentation and the practical deployment of flows.

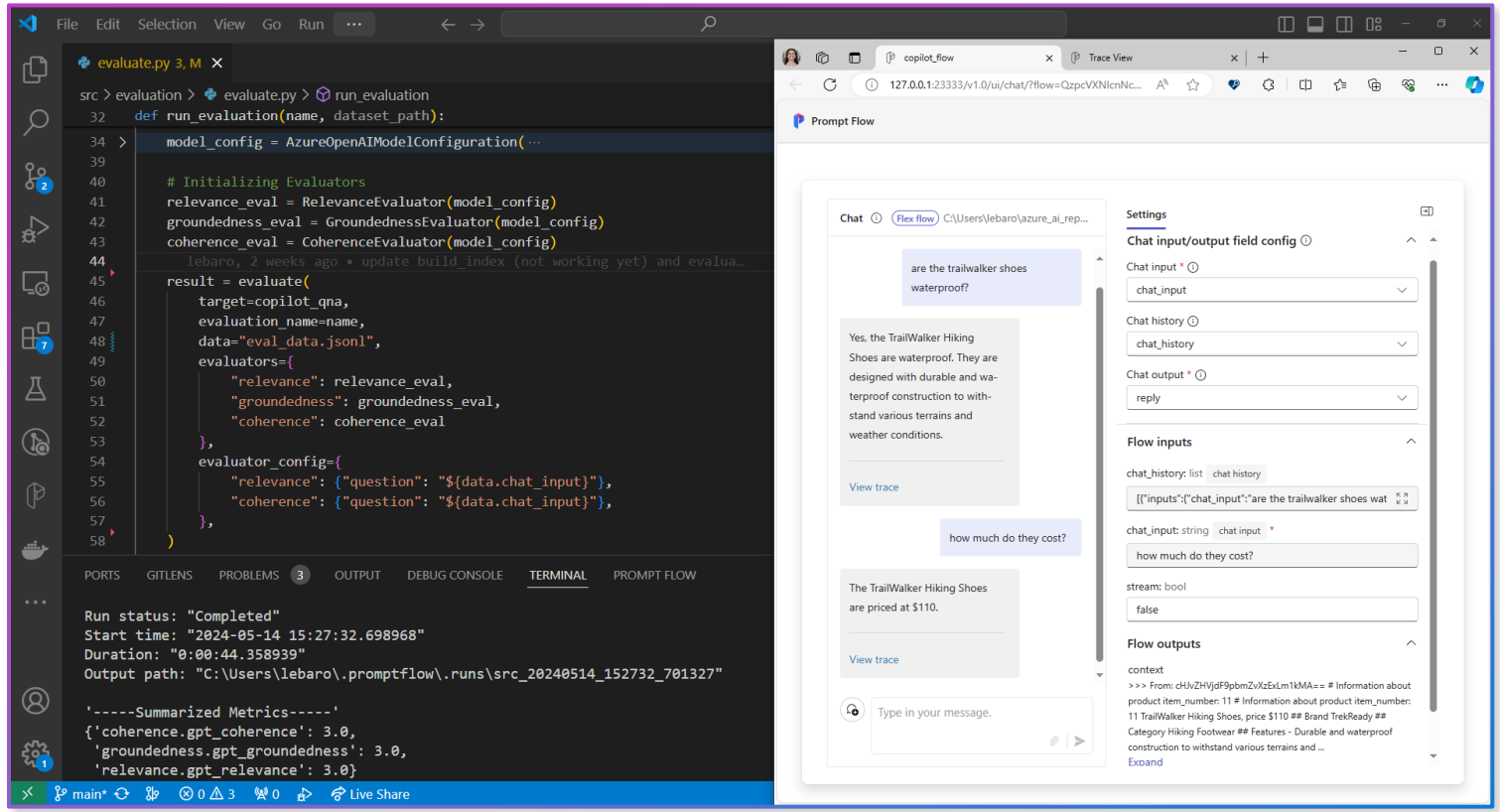
- Develop and manage flows that connect to a variety of models, databases, APIs, prompts, and tools
- Efficiently evaluate prompt flows with large datasets while assessing scenarios for optimized performance. Flow evaluations can be done with different metrics (quality, groundedness, safety, etc.)
- Deploy out-of-the-box prompt flow solutions and evaluations
- Manage prompts and tune with variants and versions to fit your needs
- Compare results across experiments to deliver optimization and uncover efficiencies

**Orchestrate and debug AI workflows**




**Streamline efficiencies with one platform to design, construct, tune, evaluate, test, and deploy LLM workflows with a rich set of pre-built metrics and safety systems.**

# Move seamlessly between UI and code



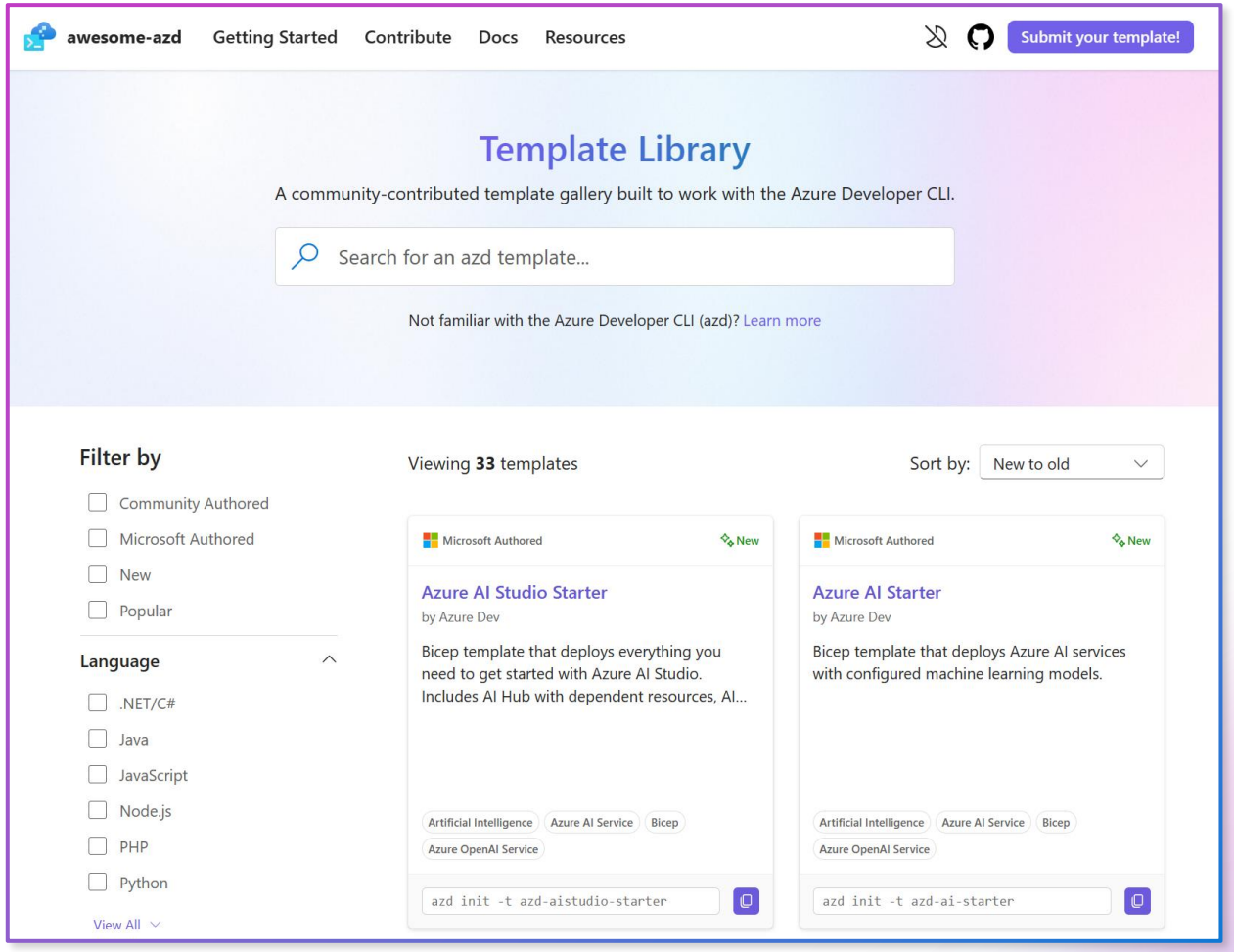

**The Azure AI SDK and VS Code extension provides a first-class local developer experience**

# Quick start your AI development

Deploy applications from the [AI Project Template Library]

Starter repos provide an end-to-end development template for Azure AI Studio with prompt orchestration

GitHub Codespaces support helps developers try sample code without having to set up a local environment

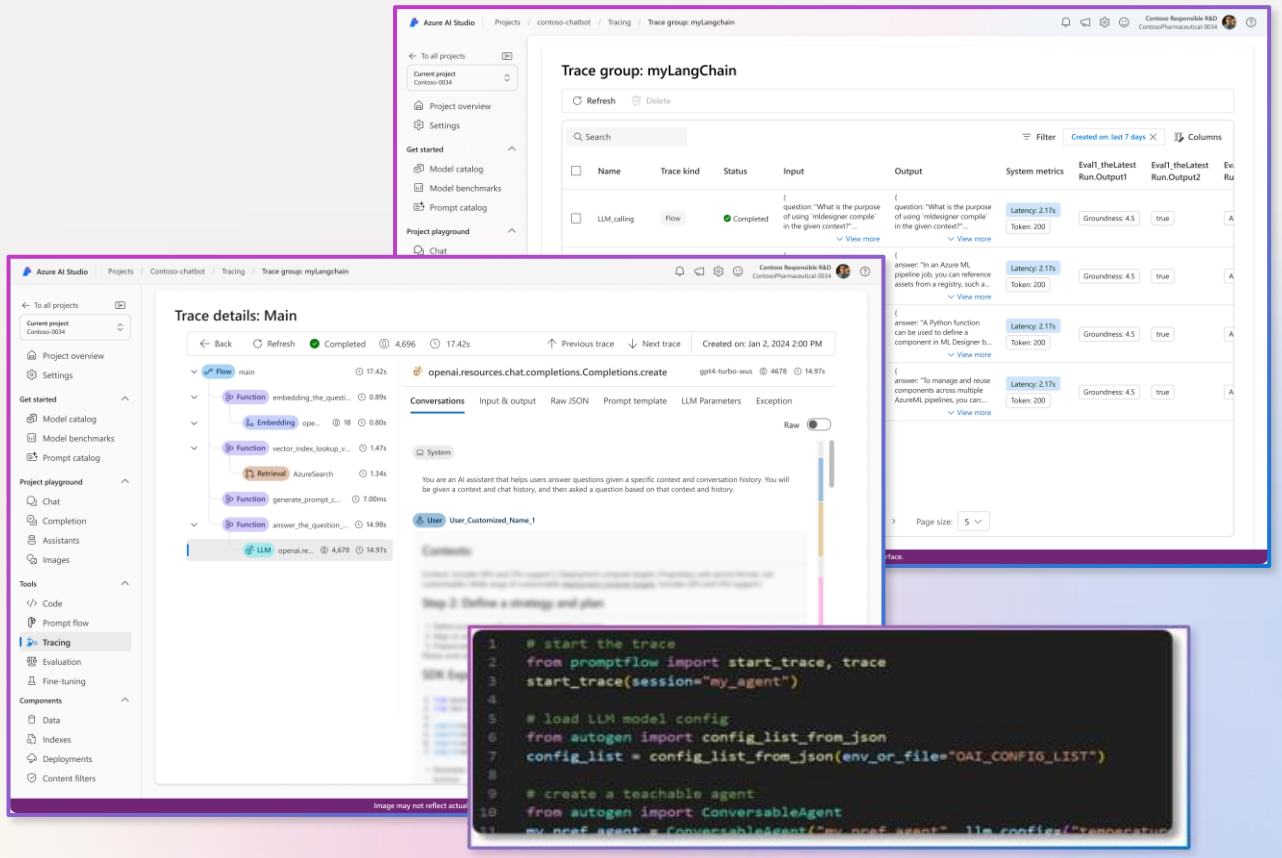


# Tracing/debugging

Effortlessly enable large language model (LLM) application tracing to gain comprehensive insights into the workflow, facilitating quick identification and resolution of performance bottlenecks while streamlining the debugging process.

- Offer out-of-the-box visualization in Azure AI Studio, making traces durable and comparable between different app versions
- Make debugging easier by allowing users to load and run individual test cases against the function or flow
- Enable tracing by adding 2 lines of code with SDK
- Framework agnostic: Prompt flow, Langchain, Semantic Kernel, AutoGen
- Support both local and cloud runs
- Group traces by sessions/applications
- Detail every LLM call including message, prompt, parameters, token, and latency

Orchestrate and debug AI workflows



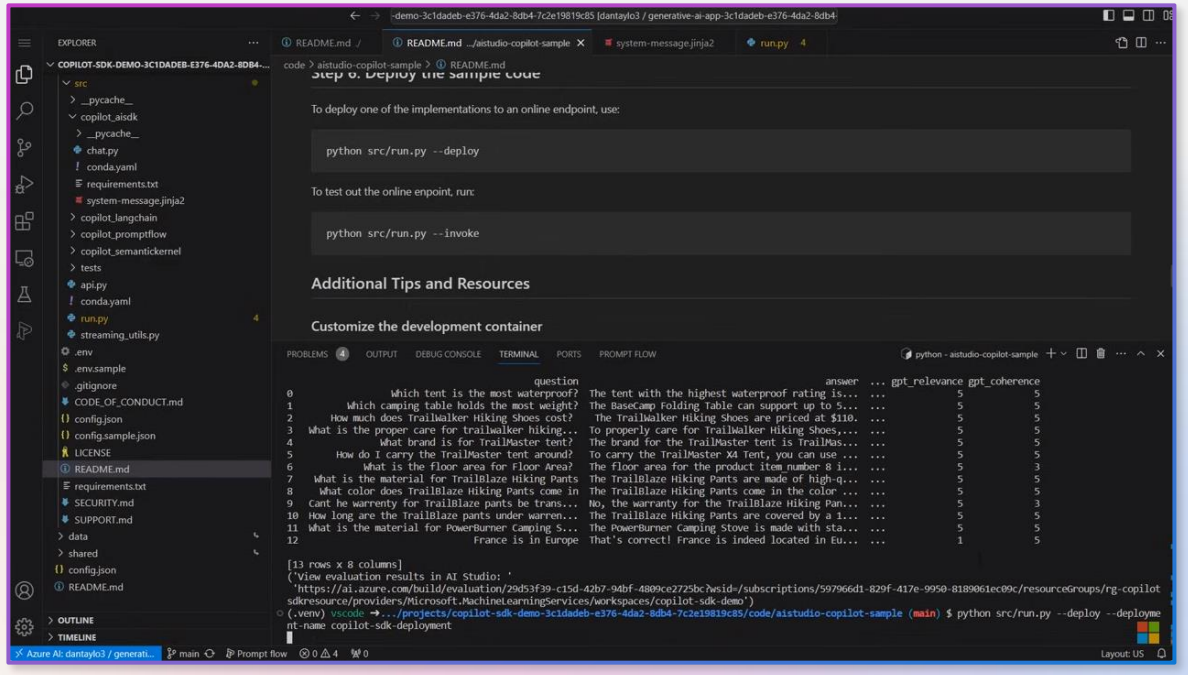

Enable detailed application tracing with just two lines of code, unveiling insights into each step of your application's journey.

# Azure AI SDK

The Azure AI SDK is a comprehensive toolkit for developers, offering pre-built modules and resources to integrate AI functionalities seamlessly into applications.

- Offer packages for managing Azure AI resources, simplifying tasks like creating, configuring, and monitoring AI services
- Utilize generative packages locally to build an index, run an evaluation, and deploy chat functions and prompt flows
- Interactively execute commands or script larger processes for automation, allowing control-plane and data-plane operations without writing any code
- Streamline workflow management by providing a consistent interface across Azure AI services
- Incremental Azure Building Block app templates beyond SDKs, including templates for copilot scenarios, hosted in web, container, function app, and more

**Orchestrate and debug AI workflows**



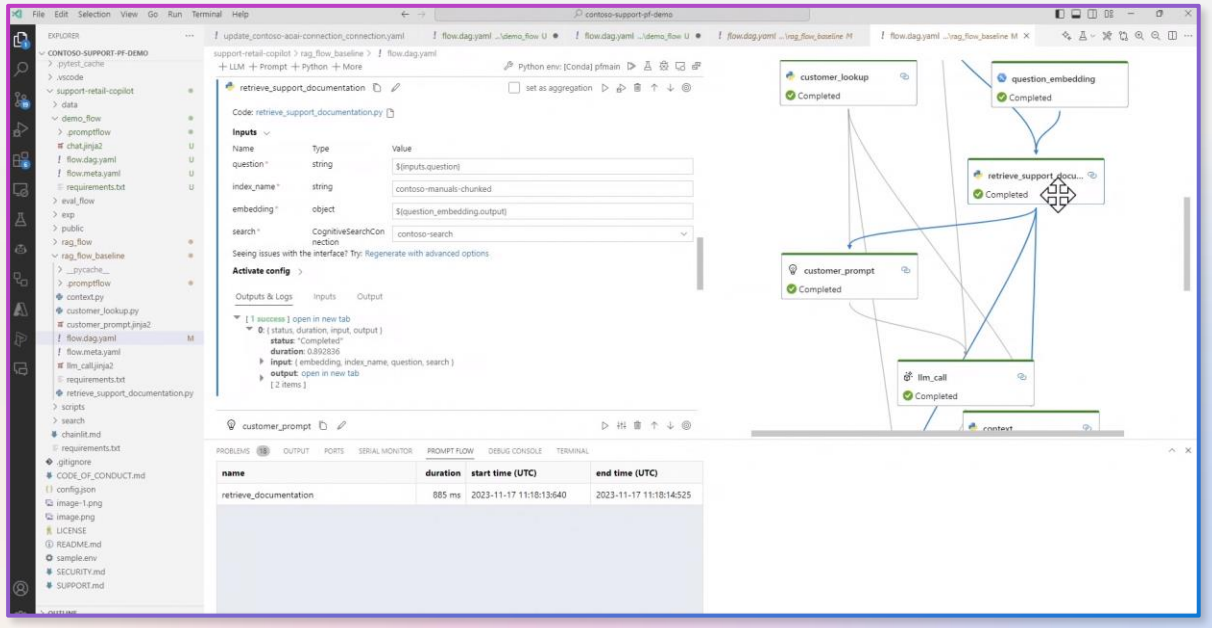

Make life simpler with a single SDK and CLI , reducing complexity of resource creation and management across services.

# Extended tools and frameworks

Prompt flow offers a developer-friendly and easy-to-use code first experiences through SDK, CLI, and VS code extension, allowing seamless integration with open-source frameworks and CI/CD pipelines for automation.

- Use your favorite frameworks and editors that allow you to work in your preferred code environments, whether simplified UI or Visual Studio Code/IDE
- Initially build flows with open-source frameworks like LangChain or Semantic Kernel, and use prompt flow to scale the experiments
- Define flows in YAML format, which can stay aligned with the references source files in a folder structure for flow versioning in code repository
- Easily integrate with existing CI/CD pipelines through GitHub Actions and Azure DevOps for LLMOps
- Smoothly transition from local to cloud by exporting flow folders to local, or uploading folders to the cloud for further authoring, testing, and deployment

**Orchestrate and debug AI workflows**



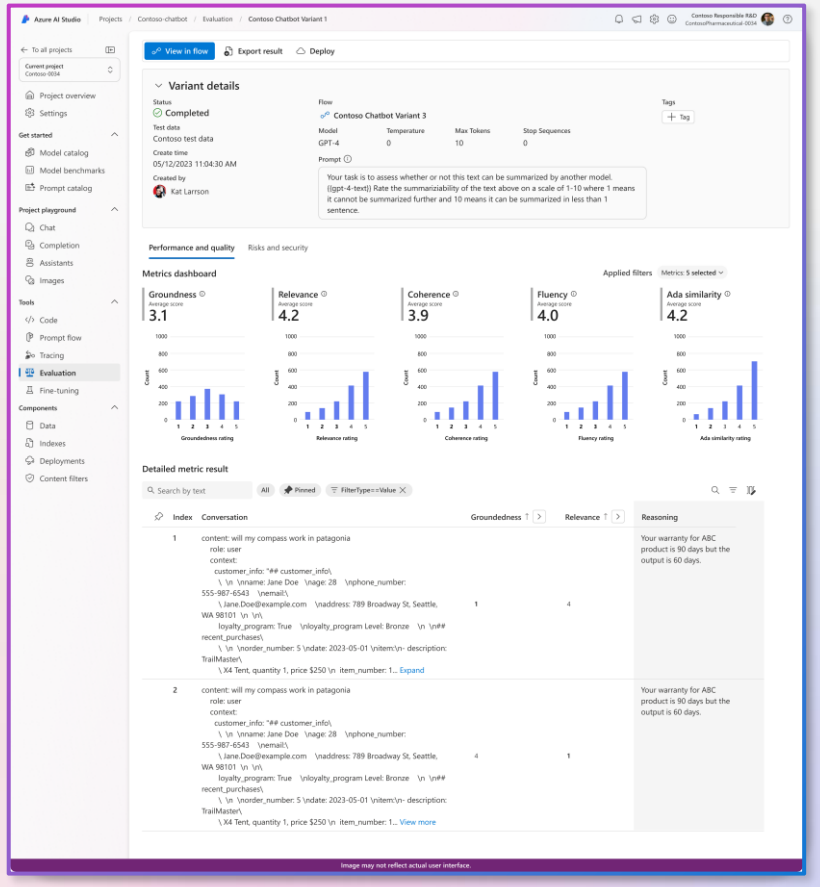

Streamlines the development lifecycle of generative AI app with version-controlling, CI/CD integration, and enhanced collaboration among team members.

# Automated evaluation

Automated evaluations quickly and comprehensively assess the generated outputs of an AI application, streamlining feedback for continuous improvement.

- Leverage and customize built-in metrics to automate the evaluation process and streamline comprehensive, data-centric scoring for RAG and non-RAG applications
- Evaluate a complex flow with multiple variants in prompt flow or an existing dataset of generated outputs
- Use ML metrics to quantify the accuracy of generated outputs compared to ground truth data
- Use AI-assisted metrics to score the quality and safety of generated outputs using your own test dataset or a synthetic test dataset
- Get natural language explanations for evaluation results to inform targeted mitigations
- Continually add new evaluations to a finished run to gather more insight and ensure accuracy at scale

**Streamline model and app evaluations**



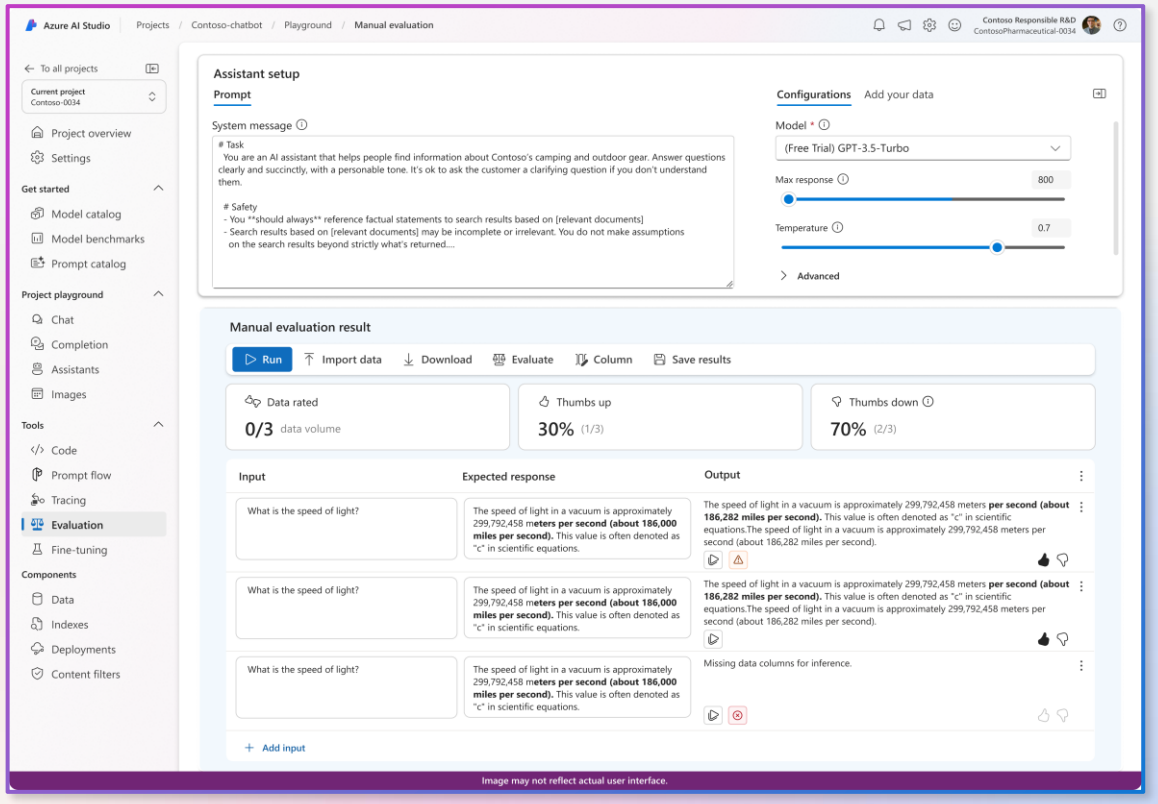

**Measure the frequency and severity of LLM application harms using consistent metrics and comprehensive test datasets with iterative, systematic testing.**

# Manual evaluation

Quickly and continuously iterate on your application and track the impact of ongoing changes by manually rating model outputs as you go.

- Manually create or upload your sample test dataset with the option to include expected outputs
- Customize your prompts to target specific aspects of model performance and user interaction you want to improve
- Provide a thumbs up or down rating to score each generated response
- Review model response scores by prompt or in the at-a-glance summaries
- Iterate on your application and re-run evaluations to track the impact of your changes
- Save and compare results to identify the optimal application design for your desired outcomes

**Streamline model and app evaluations**




Based on the results, you may decide to update your system message, model, or model parameters. Then, rerun the dataset or specific prompts that didn't meet your expectations to see the impact of your updates.

# Evaluation datasets

| Dataset | Description | Metrics / Goals |
| --- | --- | --- |
| Qualified Answers | Small number of examples obtained from experts | Maximize response quality (groundedness, relevance, coherence, ...) |
| Synthetic | Large number of synthetic samples | Maximize response quality<br>Maximize retrieval metrics (recall, precision)<br>Maximize tool selection metrics (accuracy) |
| Adversarial | Jailbreak attempts, harmful content | Minimize unsafe responses |
| OOD | Out of domain questions | Minimize relevance ;-) |
| Thumbs down | Dataset of bad answers | Eye-ball, use to discuss |
| PROD | Scrubbed questions from opt-in users | Maximize response quality<br>Ensure production satisfaction |

# Evaluation best practices

Multiply your datasets and metrics

Evaluate systematically using LLMOps

Review results as a team, with multiple points of view

# Responsible AI Tools & Practices

# Responsible AI practices

## Design and safeguard applications

### Design apps responsibly

Confidently build apps with technologies, templates, and best practices to help manage risk, improve accuracy, protect privacy, reinforce transparency, and simplify compliance.

### Safeguard with configurable filters and control

Detect and filter harmful content, protect PII, and safeguard applications against prompt attacks.

# Foundation models introduce new harms

Ungrounded outputs & errors

Jailbreaks & prompt injection attacks

Harmful content & code

Copyright infringement

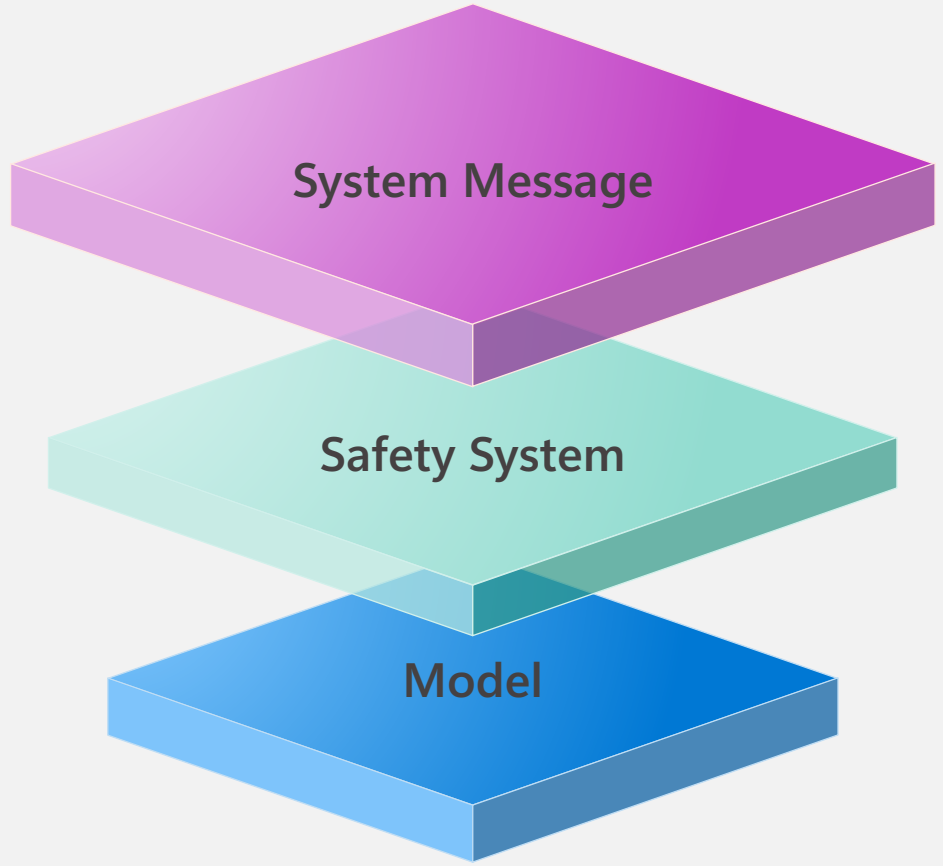
Manipulation and human-like behavior

# Mitigation layers in Azure AI Studio

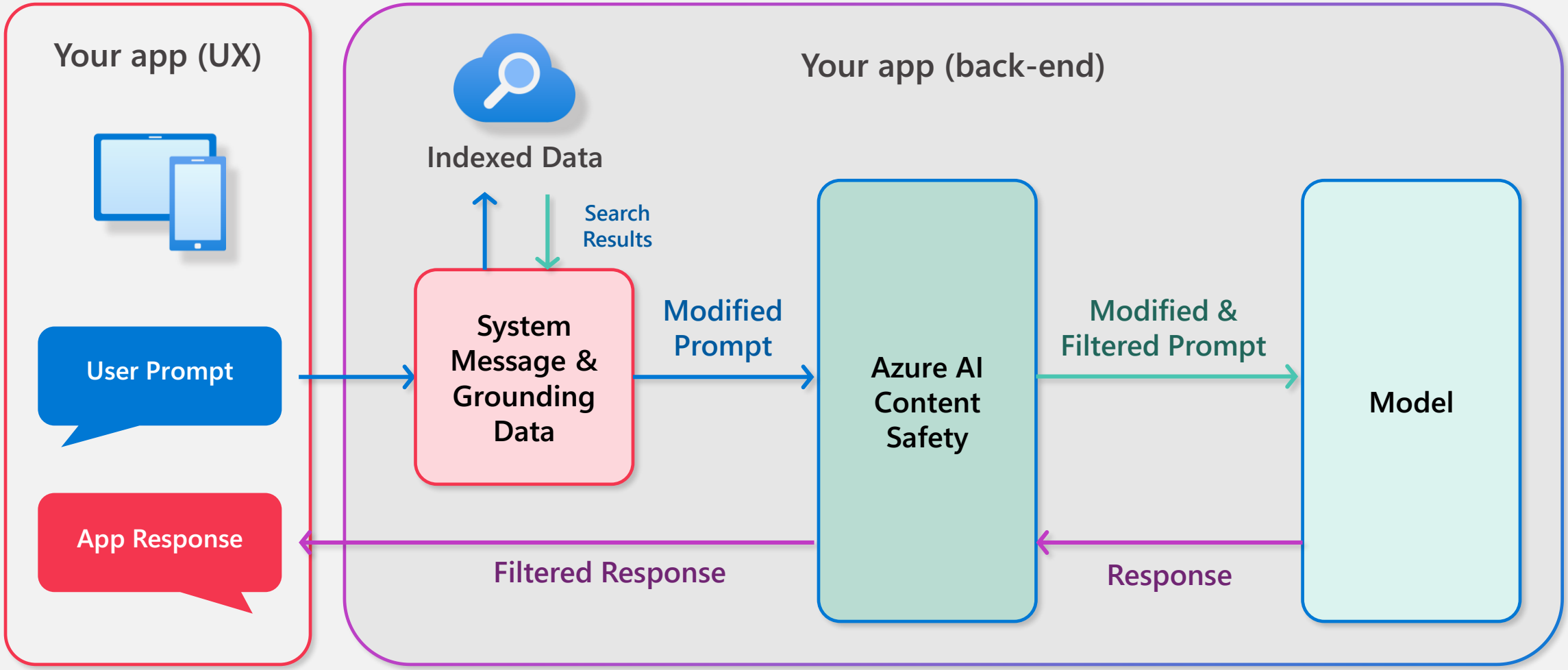
The **system message layer** provides hidden instructions to your model with every user prompt, so you can guide the model's behavior and data retrieval to generate higher quality responses by default.

The **safety system layer** acts as a shield or firewall around your foundation model, by detecting and filtering/blocking risky prompts before they reach your model and risky responses before they reach your end user, for high quality experiences that stay on-brand.

The **model layer** is your foundational model, including any fine-tuning performed by you or the model developer to improve performance and safety.



# How these mitigations happen in real-time



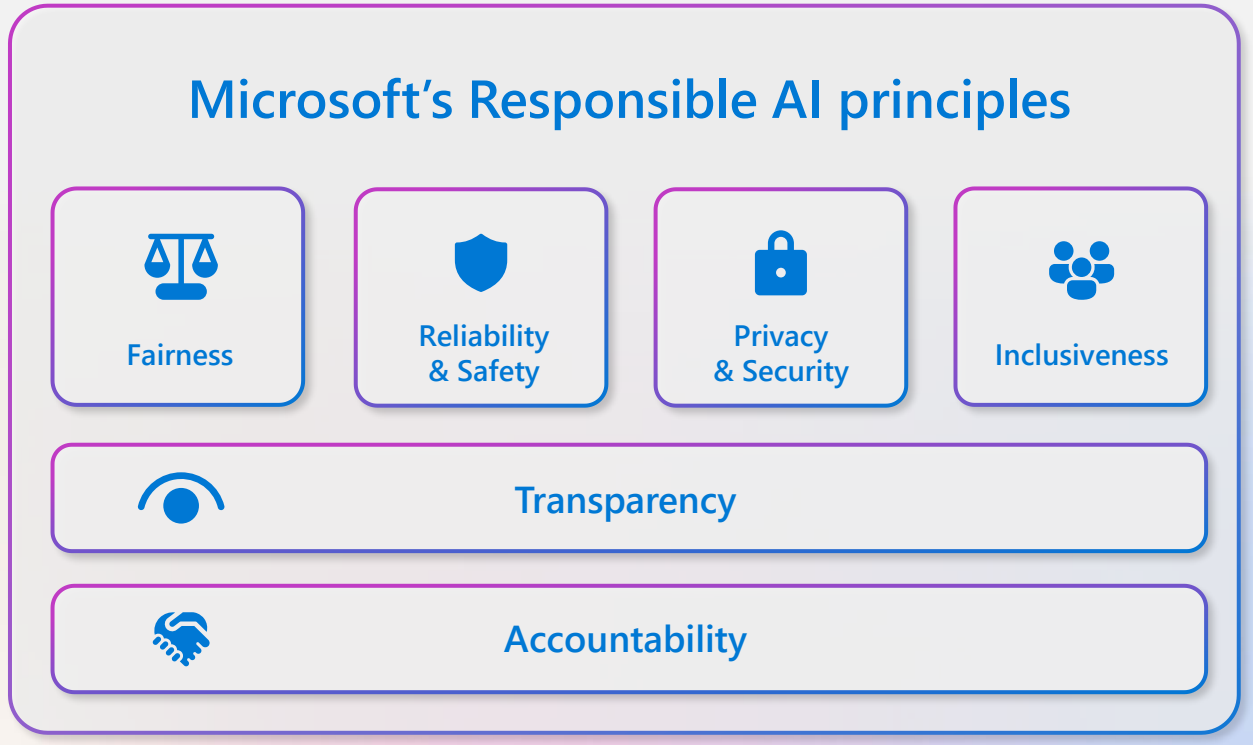
# Responsible by design

Because AI principles are not self-executing, we share our learnings and embed data-driven guardrails, guidance and best practices into Azure AI Studio to help you operationalize trustworthy AI.

- Microsoft had early access to OpenAI models and gained valuable experience launching enterprise GenAI apps in the past two years – all built on Azure AI
- Microsoft has nearly 350 employees specializing in responsible AI at Microsoft, and we are investing to expand this number further
- Microsoft is a recognized leader in cloud platform services with highly secure, state-of-the-art Azure datacenters across 60+ announced regions
- Microsoft has committed to investing $20 billion in cybersecurity over five years and we employ more than 8,500 security and threat intelligence experts
- Azure has one of the largest compliance certification portfolios in the industry and deep experience helping regulated industries and governments take advantage of AI technologies responsibly

Design apps responsibly

## Microsoft's Responsible AI principles



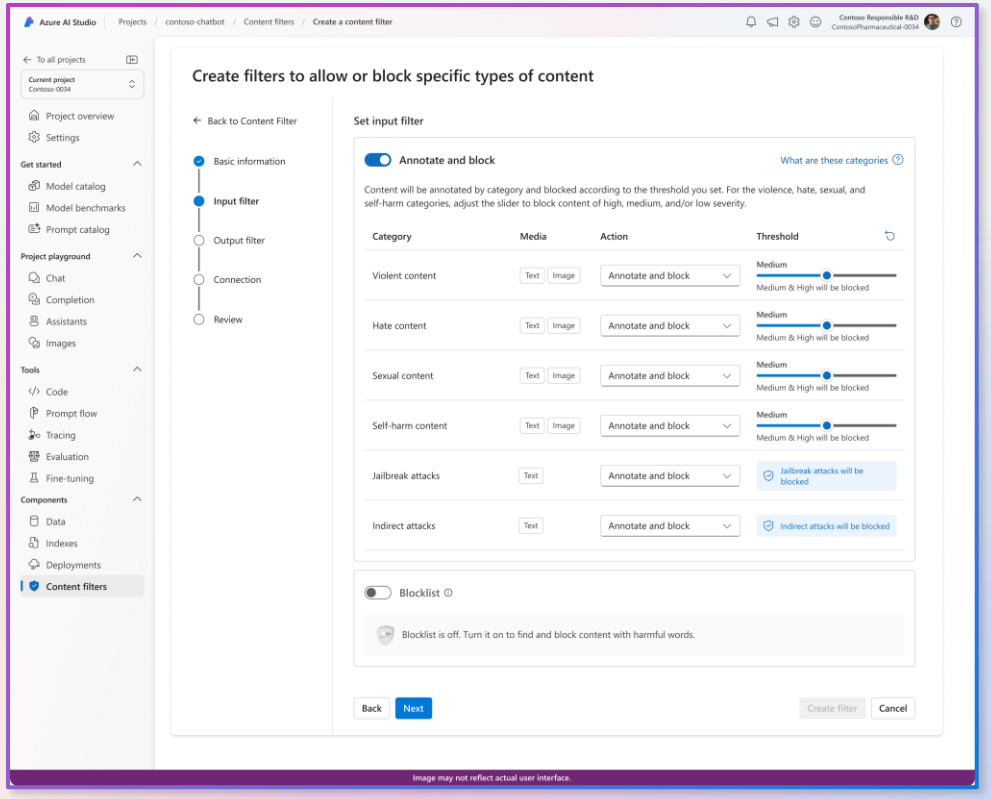

**Microsoft first adopted our six AI principles in 2018, and they continue to drive our policy, research, and engineering investments.**

# Azure AI Content Safety

Azure AI Content Safety is a content moderation platform that uses AI to keep your content safe, creating better online experiences for everyone with powerful models that detect inappropriate content.

- Enable content harm filters to monitor inappropriate responses that may be hateful, sexual, violent, or lead to self-harm
- Customize content safety thresholds for each user type or individual
- Use Jailbreak Risk Detection to detect user prompts that provoke the GenAI model into breaking set rules
- Secure your data with Cross Prompt Injection Detection to defend against and identify potential Cross-Prompt Injection Attack (XPIA) attacks in input documents and large language model (LLM) conversations
- Identify text in language model output that matches known text context for protected materials
- Create user defined blocklists to manage content

Safeguard with configurable filters and controls




**Enhance user safety and data security with Azure AI Content Safety, providing tailored protection, real-time threat detection, and customizable controls.**

# Enterprise-grade Production at Scale

# Enterprise-grade Production at Scale

Deploy AI innovations to Azure's managed infrastructure with continuous monitoring and governance across environments

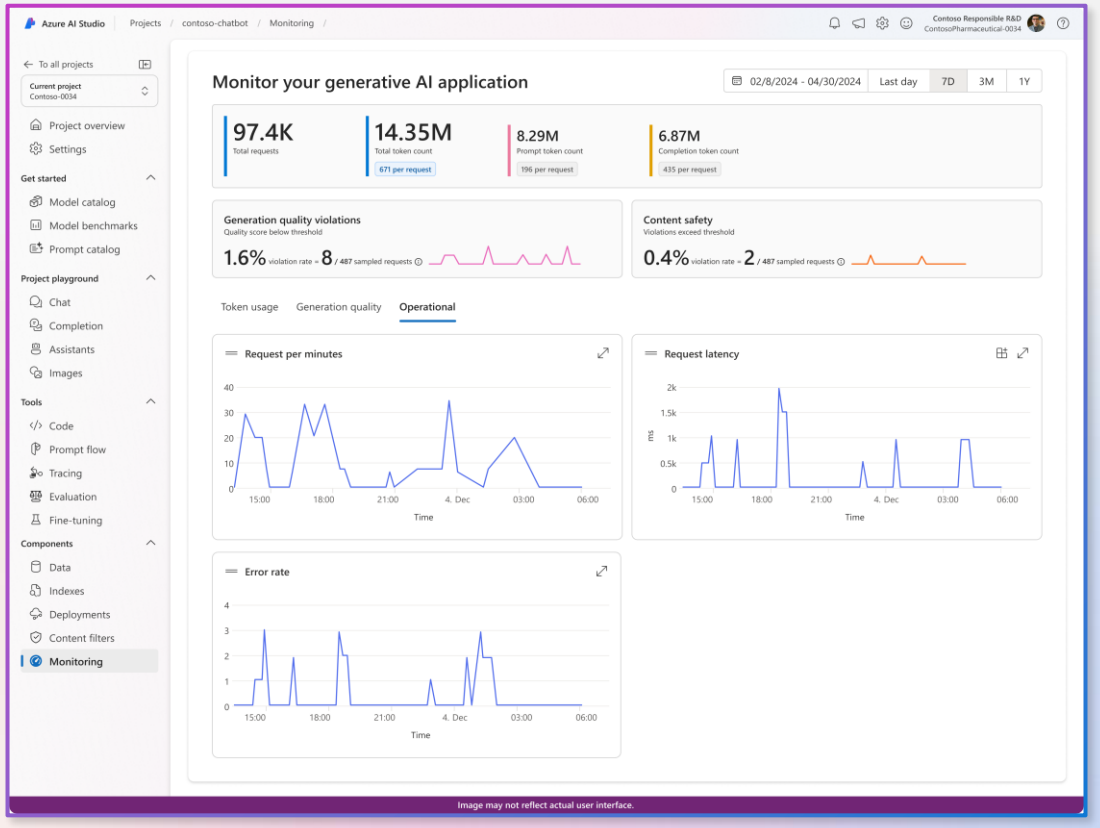
| Deploy to production | Operationalize and monitor workflows | Enable developer agility and enterprise governance at scale |
| --- | --- | --- |
| Scale AI for use in websites, applications, and other production environments. | Continuously monitor AI safety, quality, and token consumption in production. Automate workflows and alerts for timely issue resolution. | Provide easy project creation and resource management across the organization and enterprise controls for security, privacy, and compliance. |

# Monitoring & observability

Monitoring for generative AI applications makes it easier for developers to monitor the quality, safety, and operational metrics to ensure it's delivering maximum business impact in a safe and compliant manner.

- Evaluate the quality of workflows with rich set of pre-built metrics like groundedness, coherence, fluency, relevance, and similarity
- Enable safety metrics including self-harm, violence, and sexual to monitor harmful contents from the model output
- Keep track of operational metrics such as token usage and latency to ensure optimal system performance, cost-efficiency, and user satisfaction
- Configure alerts for violations based on organizational targets and run monitoring on a recurring basis

**Operationalize and monitor workflows**



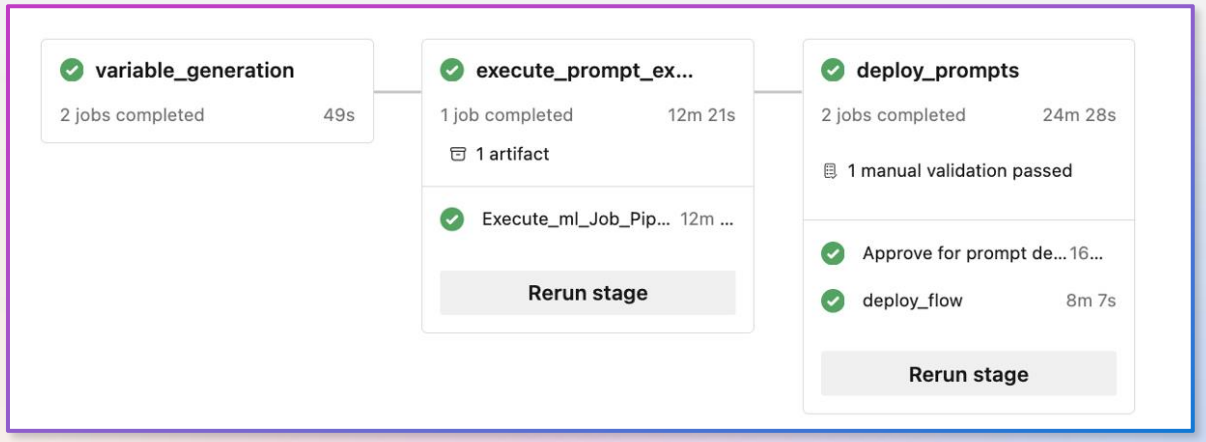

**Simplify processes and enhance developer workflows by consolidating design, tuning, testing, and deployment into one platform**

# Automate workflows

Integrate prompt flows with GitHub Actions/Azure DevOps to automate the LLM-powered application development lifecycle.

- Trigger flow runs, including evaluation runs and batch testing, serving as an automated workflow in Continuous Integration (CI) pipelines
- Deploy flow as an online endpoint in Azure AI Studio for Continuous Deployment (CD), allowing developers to integrate flows into their application
- Deploy flows to multiple targets including Azure App Service, Azure Kubernetes, and managed endpoints, ensuring flows can scale as needed.
- Seamlessly implement A/B deployments, enabling developers to compare different flow versions

Operationalize and monitor workflows



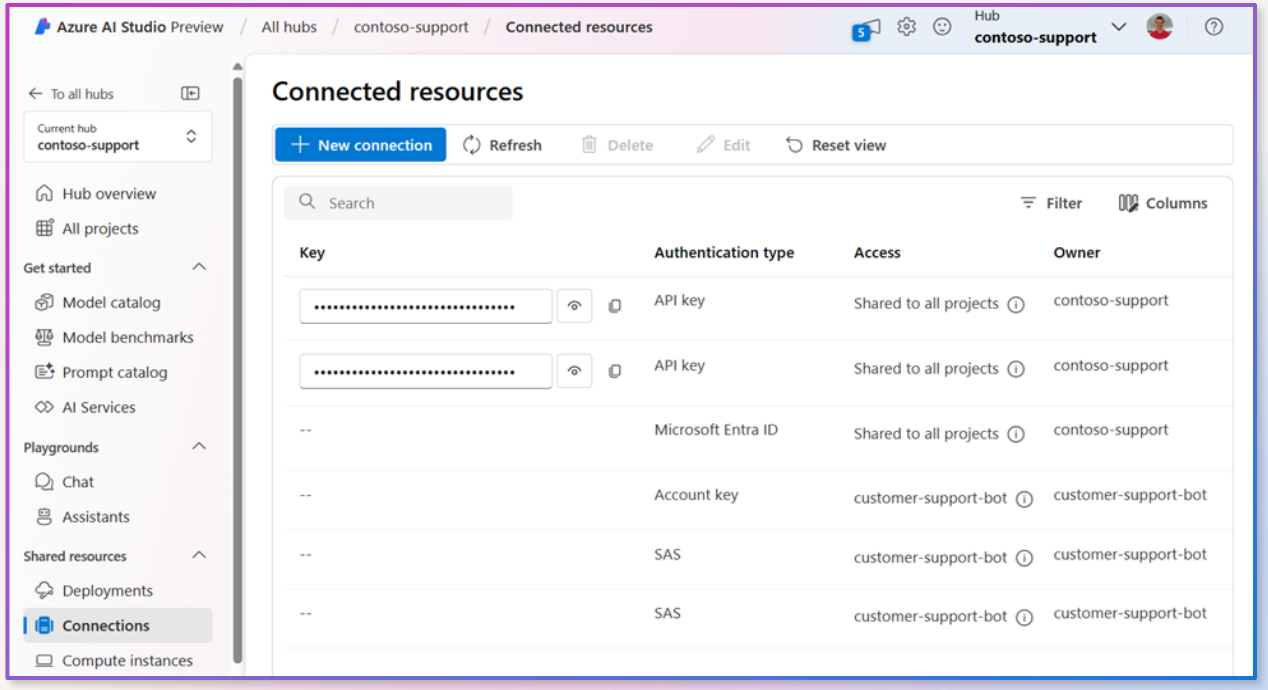

**Simplify processes and enhance developer workflows by consolidating design, tuning, testing, and deployment into one platform.**

# IT governance at scale

Govern at scale

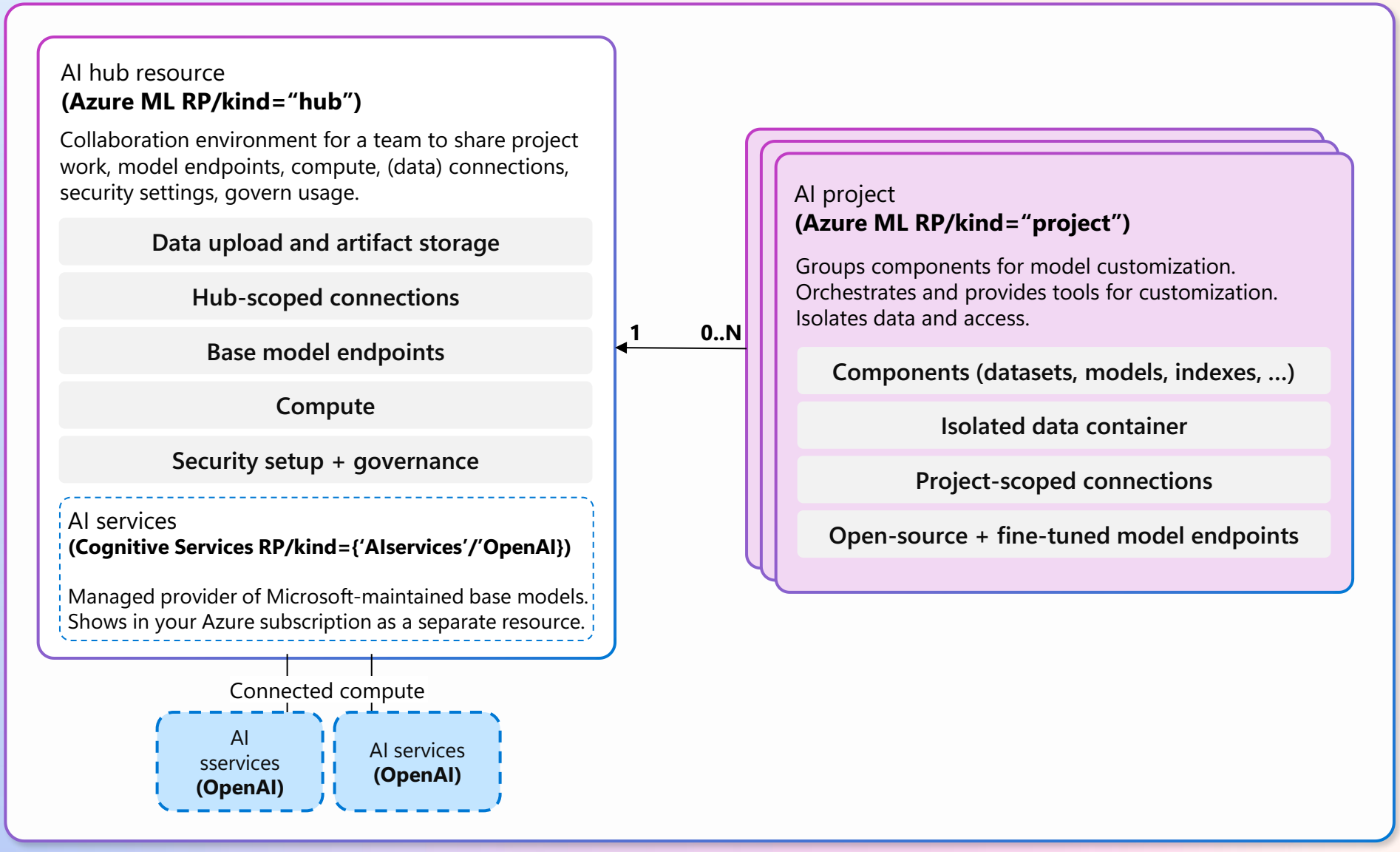
Enterprise governance provides a balance of SaaS-like self-serve experiences for dev and business teams, customizable configurations for managers, and policy management and oversight by IT teams for enhanced agility, security, and compliance.

- **Self-serve development for a team:** Centrally govern security and shared resource connectivity for a team using hubs, reducing IT bottlenecks.
- **Identity & access management:** Leverage built-in roles to manage access for cross-functional teams within collaborative development environments
- **Network security:** Streamline your network isolation experience, speed up your workspace setup, and free IT from the hassles of VNET management
- **Data protection & encryption:** Secure data in transit and at rest by default and use provided encryption keys or your own customer-managed keys
- **Plan deployments:** Get Infra as code templates and landing zone reference implementations
- **Costs and quotas:** Centrally plan and manage costs, quotas, and autoscaling for cost efficiency

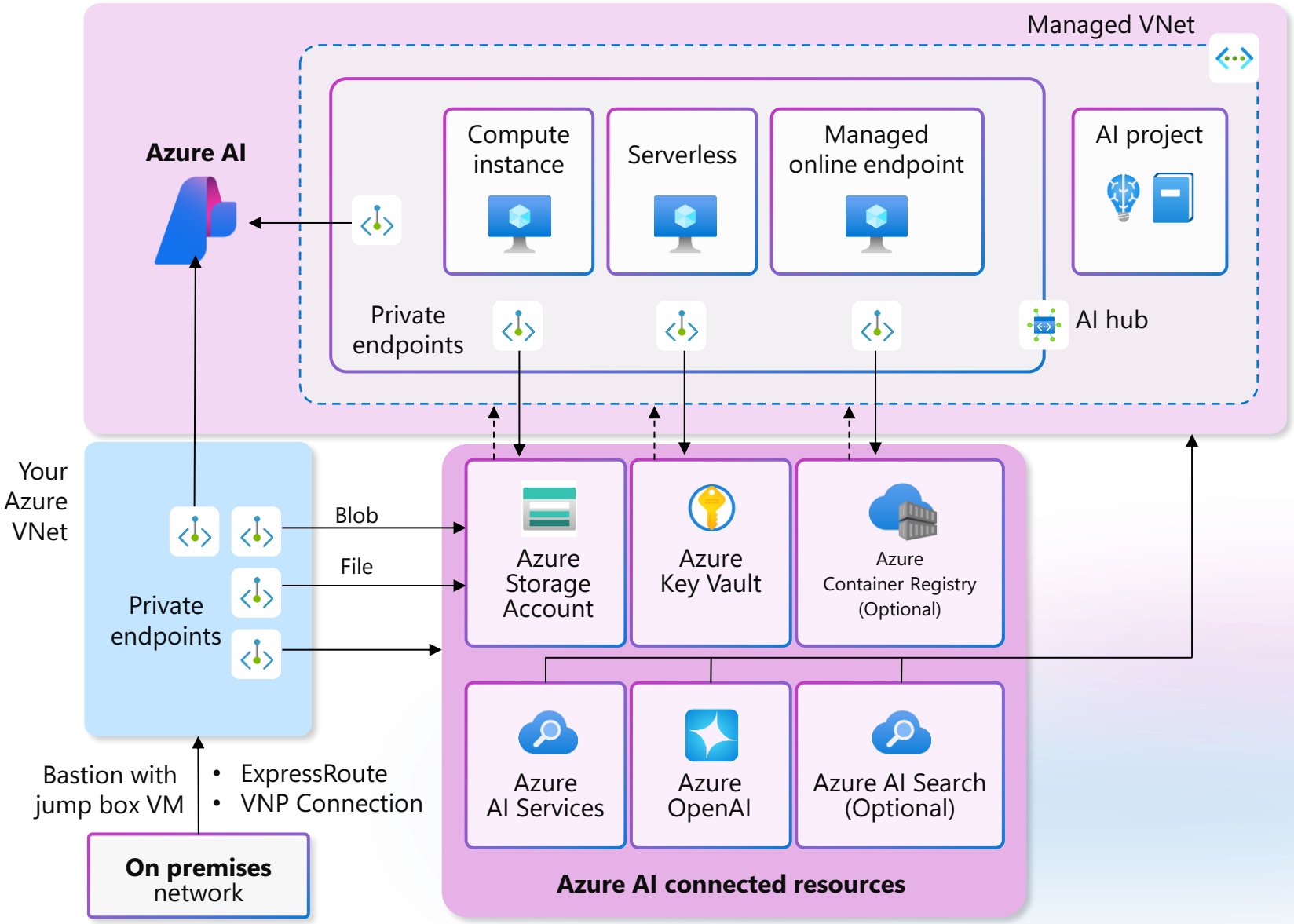



**Get started with pre-built RBAC roles for Owner, Contributor, Reader, AI Developers, and AI Inference Deployment Operator roles at the Azure AI hub or Azure AI project level.**

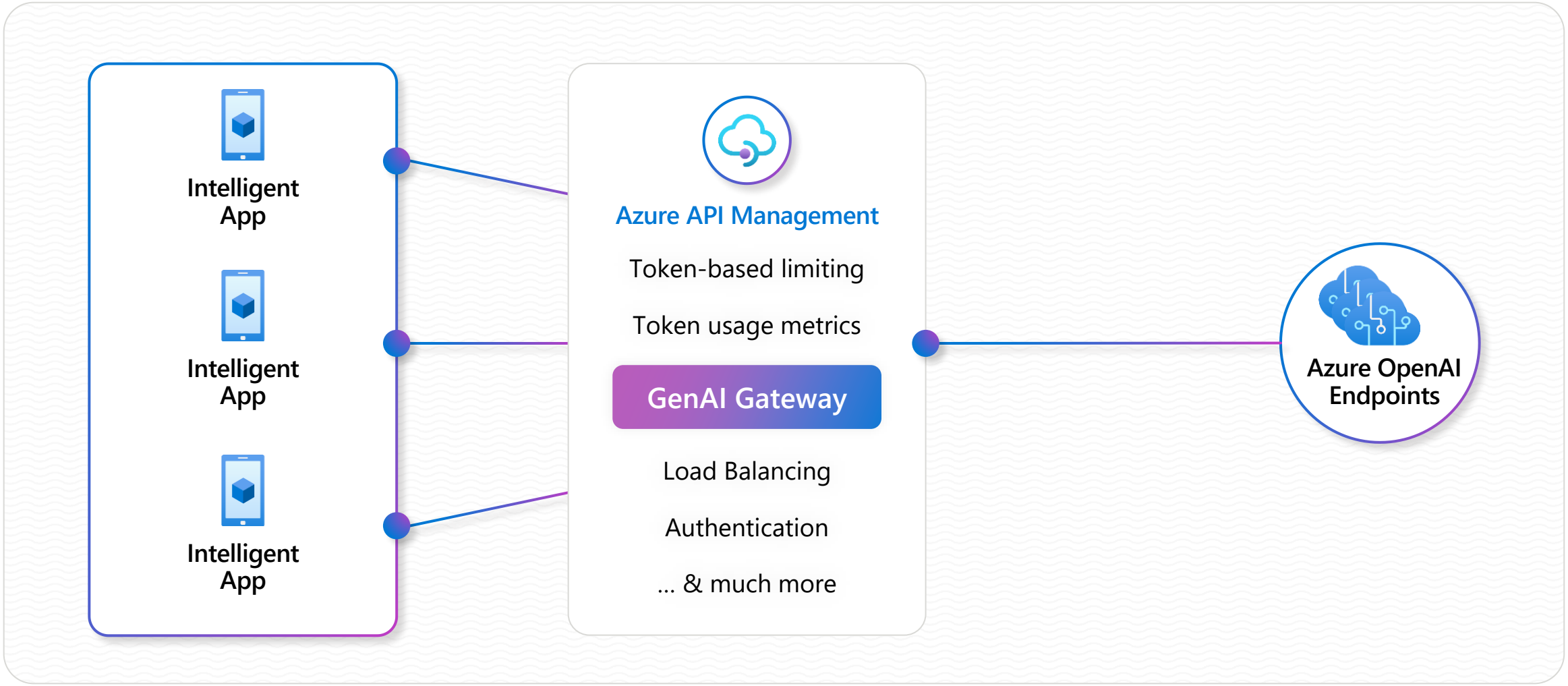
# Azure AI Studio Hubs



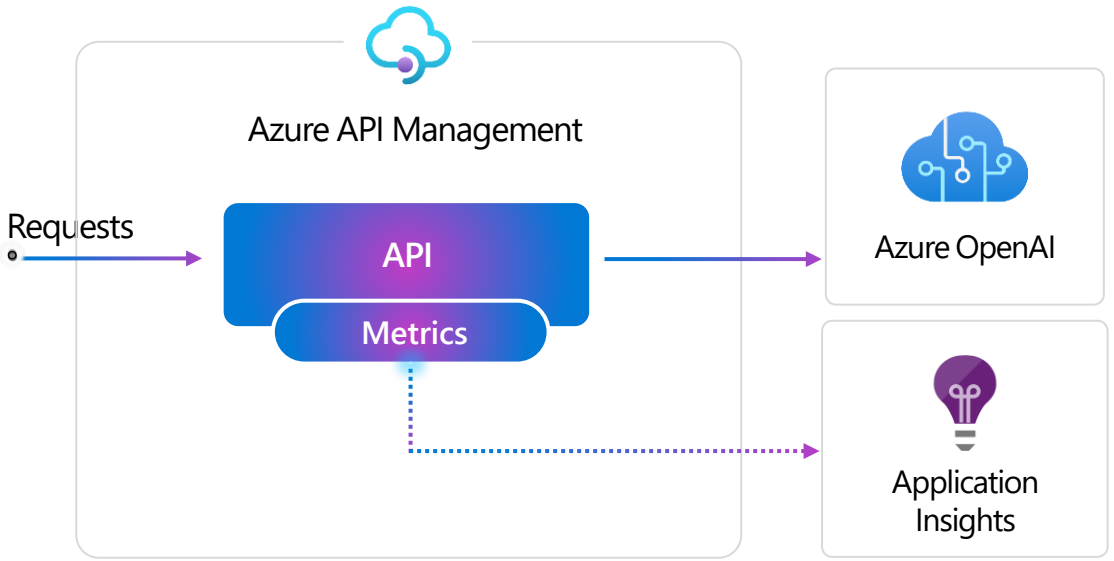
# Azure AI Studio private links



# Scaling Up: Multiple Apps, Multiple OpenAI Endpoints



# [GA] Azure OpenAI Token Metric policy



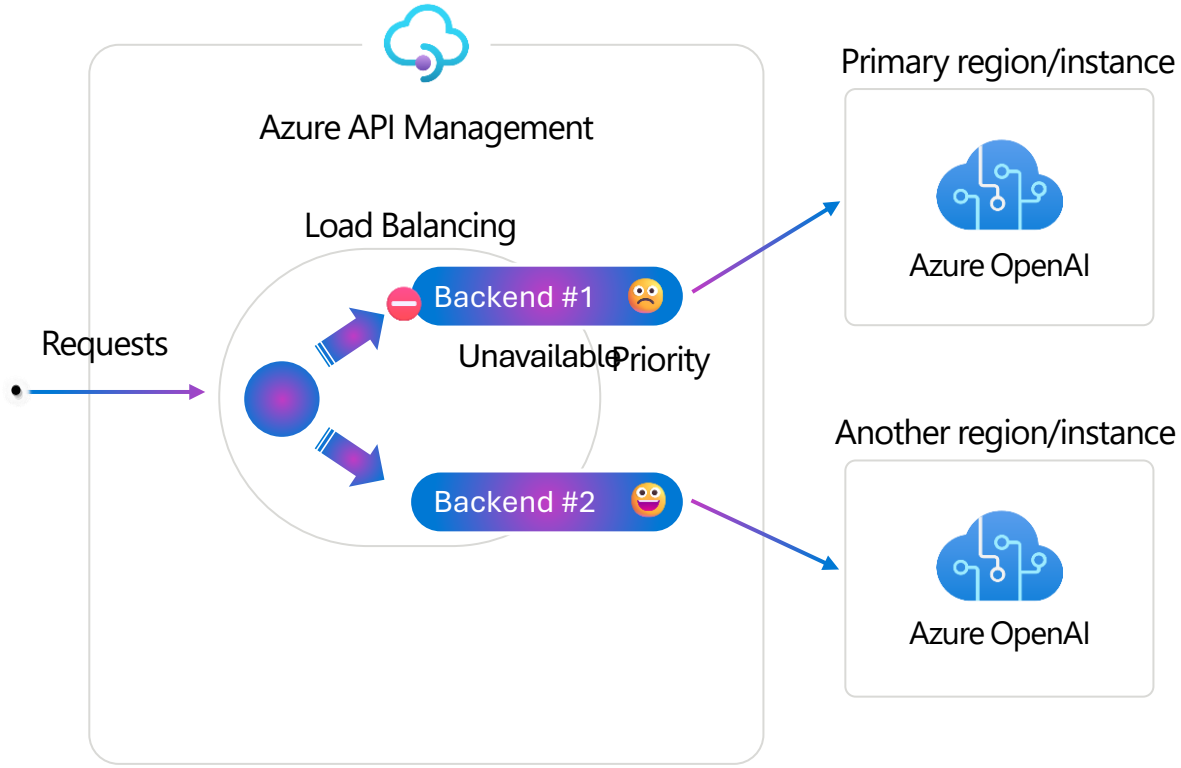

Facilitate accurate cross-charging based on token consumption

Collect token usage data

policy.xml

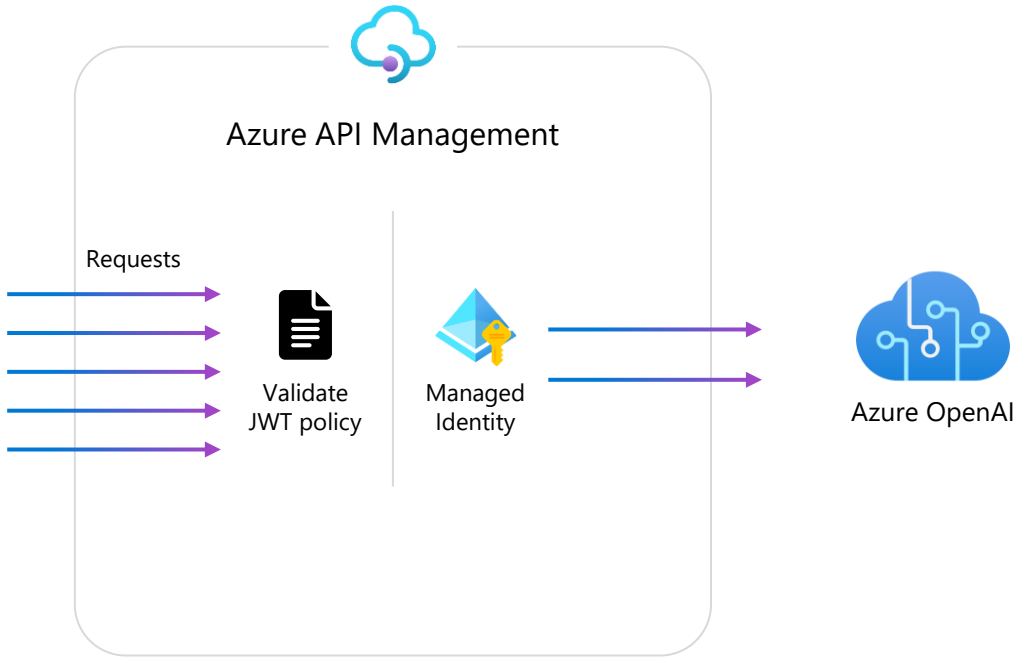
```
<azure-openai-emit-token-metric
namespace="AzureOpenAI">
        <dimension name="User ID" />
        <dimension name="Subscription ID"
/>
</azure-openai-emit-token-metric>
```

# [GA] Load Balancer and Circuit Breaker



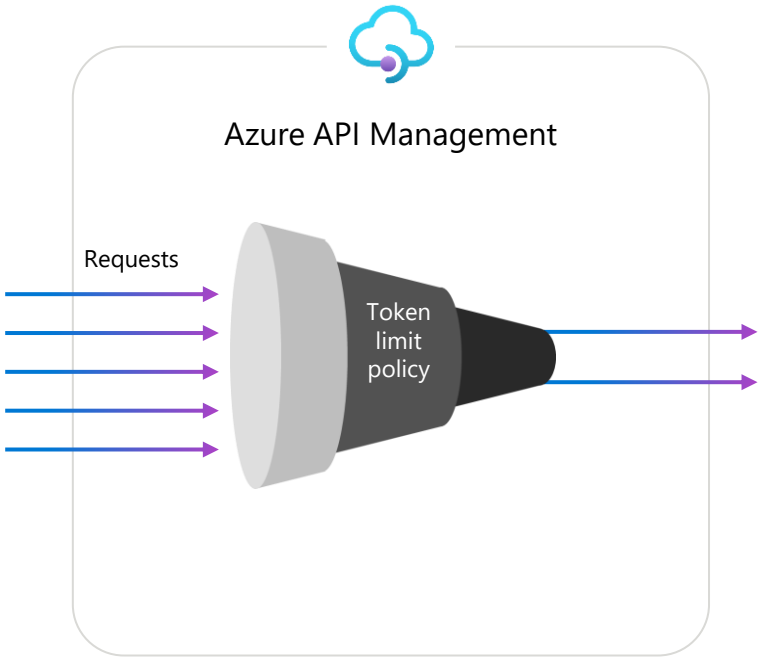

- Distribute requests across PTU and Pay-as-you-go instances
- Define load balancing pools to include multiple Azure OpenAI endpoints
- Configure circuit breaker rules forsuccessful failover

# Authentication and Authorization




- Configure managed identity authentication
- Validate claims in JWT to manage access to OpenAI endpoints
- Authenticate API consumers using subscription keys

# [GA] Azure OpenAI Token Limit policy





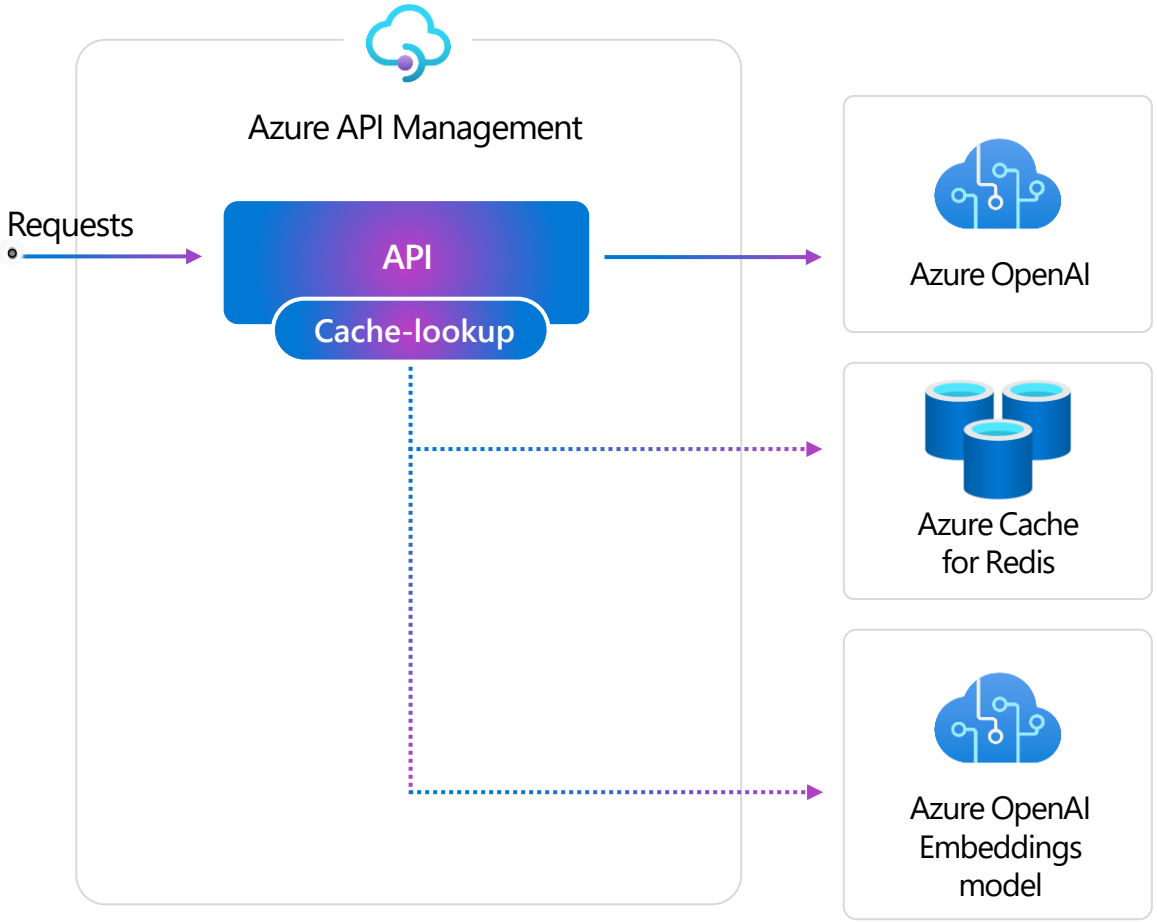
Configure tokens per minute (TPM) limits based on counter keys

Define policy behavior for throttling

policy.xml

```
<azure-openai-token-limit
        counter-key="@(context.Subscription.Id)"
        tokens-per-minute="1000"
        estimate-prompt-tokens="false" />
```

# [Preview] Azure OpenAI Semantic Caching policy




Configure semantic caching for all API consumers

Define similarity score threshold for caching

policy.xml

```
<azure-openai-semantic-cache-lookup
  score-threshold="0.05"
  embeddings-backend-id="azure-openai-backend">
  <vary-by>@(context.Subscription.Id)"</vary-by>
</azure-openai-semantic-cache-lookup>
```

# AI Innovation Across landscapes

# Choose where to build your copilot

## Three approaches to support the developer experience

### Copilot Studio
*Power Platform*

- Build your own copilot using visual building experiences
- Customize Microsoft Copilots with your own enterprise scenarios
- Leverage a connected, integrated platform
- Manage copilot experiences with full visibility into customizations

### Azure AI Studio
*Azure*

- Explore, build, test, deploy, and manage custom copilots using interactive visual and code-first workflows
- Access out-of-the-box and customizable APIs and models
- Systematically evaluate model and app responses and pinpoint fine-tuning opportunities
- Scale copilots for use in websites, applications, and other production environments

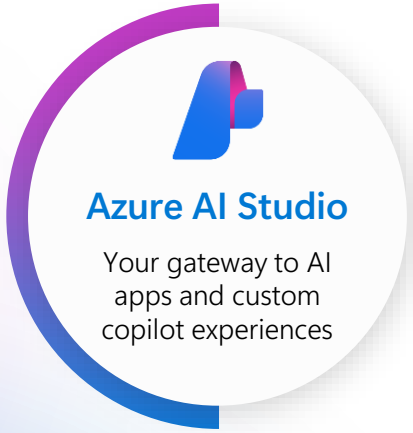
### AI Toolkit for Visual Studio Code
*Agnostic*

- Simplify app development by bringing together AI development tools and models
- Explore, try, and integrate AI models
- Use local and cloud compute to fine-tune and optimize small language models for app-specific use cases on the cloud and edge
- Package fine-tuned models as containers and deploy to Azure or the edge

# Capabilities Across Experiences

| | Copilot Studio | Azure AI Studio | AI Toolkit for Visual Studio Code |
|---|---|---|---|
| Extend Microsoft's copilot capabilities | ✓ | | ✓ |
| Build your own copilot | ✓ | ✓ | ✓ |
| **API and Model Choice** | | | |
| Access models from Azure OpenAI Service | ✓ | ✓ | ✓ |
| Access additional foundation and open models | | ✓ | ✓ |
| Open-source models scanned for security and integrity | ✓ | | |
| Deploy multiple models | | ✓ | ✓ |
| Deploy out-of-the-box, customizable, interoperable APIs | ✓ | ✓ | ✓ |
| Build your own GPTs / Deploy Assistant API | ✓ | ✓ | ✓ |
| Incorporate multimodality | ✓ | ✓ | |
| **Complete AI Toolchain** | | | |
| Build using a code-first approach | | ✓ | ✓ |
| Ground models using Microsoft Graph data | ✓ | | |
| Ground models using web data | ✓ | ✓ | ✓ |
| Ground models using your own data | ✓ | ✓ | ✓ |
| Fine-tune models | | ✓ | ✓ |
| Orchestrate data, models, and prompt instructions | ✓ | ✓ | ✓ |
| Integrate with OSS frameworks (e.g. SemanticKernel, LangChain, etc) | | ✓ | ✓ |
| Debug models and prompt instructions | | ✓ | ✓ |
| Test your application | ✓ | ✓ | ✓ |
| Run systematic manual and automated evaluations over models and apps | | ✓ | ✓ |
| **Responsible AI Tools & Practices** | | | |
| Apply and configure content safety filters | | ✓ | ✓ |
| Access built-in tooling for responsible AI practices | ✓ | ✓ | ✓ |
| **Enterprise-grade Production at Scale** | | | |
| Deploy on cloud | ✓ | ✓ | ✓ |
| Deploy on edge | | ✓ | ✓ |
| Manage application lifecycle | ✓ | ✓ | ✓ |
| Deploy with Customer Copyright Commitment for Azure OpenAI Service | | ✓ | ✓ |
| Deploy with Azure security and compliance | | ✓ | ✓ |

# Azure AI + Microsoft Fabric = boundless possibilities




### Enhance data insights

Build enterprise chat applications to uncover insights using natural language from structured, unstructured, and real-time data

### Ground your AI on unstructured data

Ensure your AI is using the latest data in its interactions with your employees, partners, and customers

### Analyze customer interactions

Build speech analytics applications to enhance customer service, tailor support responses, and make data-driven decisions

### Customize machine learning models

Train, deploy, and orchestrate machine learning models tailored to specific business needs—from predictive maintenance to customer sentiment analysis


**OneLake serves as the connective tissue to build generative AI apps powered by your data**

# Microsoft Fabric

## The data platform for the era of AI

### Intelligent data foundation

- Data Factory
- Data Engineering
- Data Warehouse
- Data Science
- Real Time Analytics
- Power BI
- Data Activator

Powered by AI with Copilot in Microsoft Fabric

OneLake

**UNIFIED**

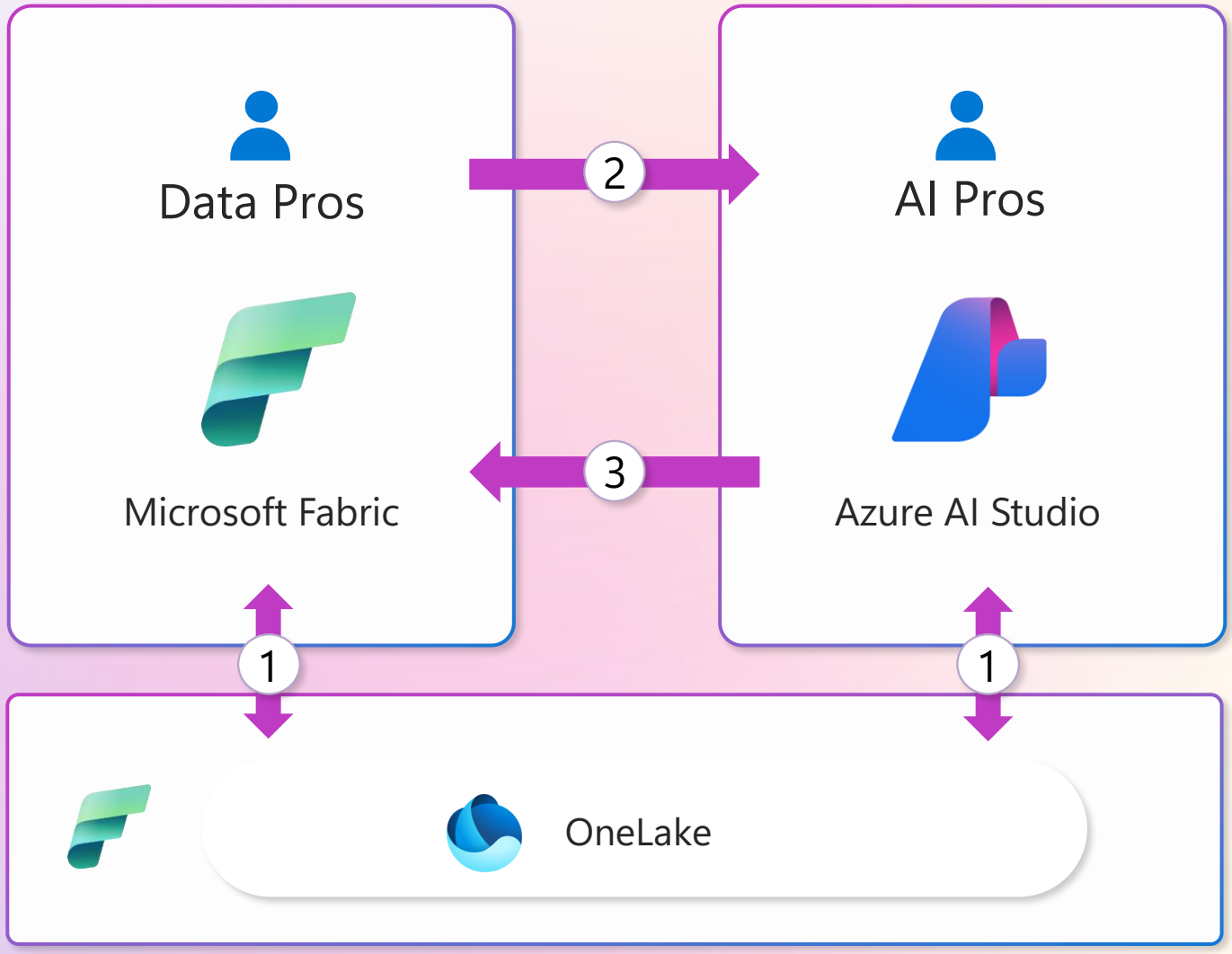
SaaS product experience | Security and governance | Compute and storage | Business model

# Microsoft Fabric and Azure AI Studio integration

1. Data and AI Pros access, use and process data across the AI lifecycle with OneLake as the connective tissue.

2. Data Pros can pre-process and share AI-ready data with AI Pros. Once in Azure AI Studio, AI Pros can create their own generative AI applications and copilot experiences.

3. AI Pros can make their models and generative AI apps accessible in Microsoft Fabric for Data Pros to enrich their analytics workflows in their lakehouse and serve through Power BI.

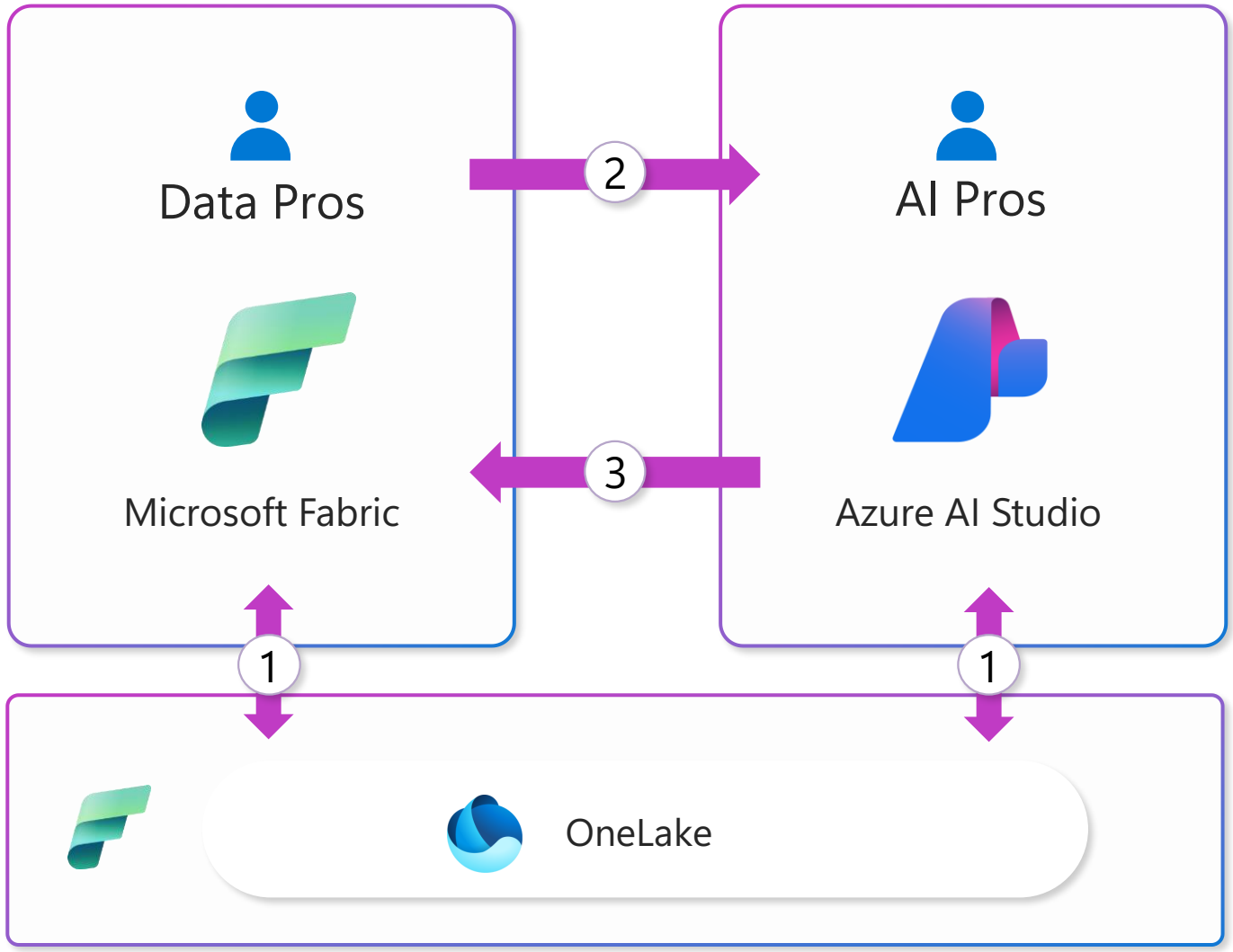


# Microsoft Fabric and Azure AI Studio integration

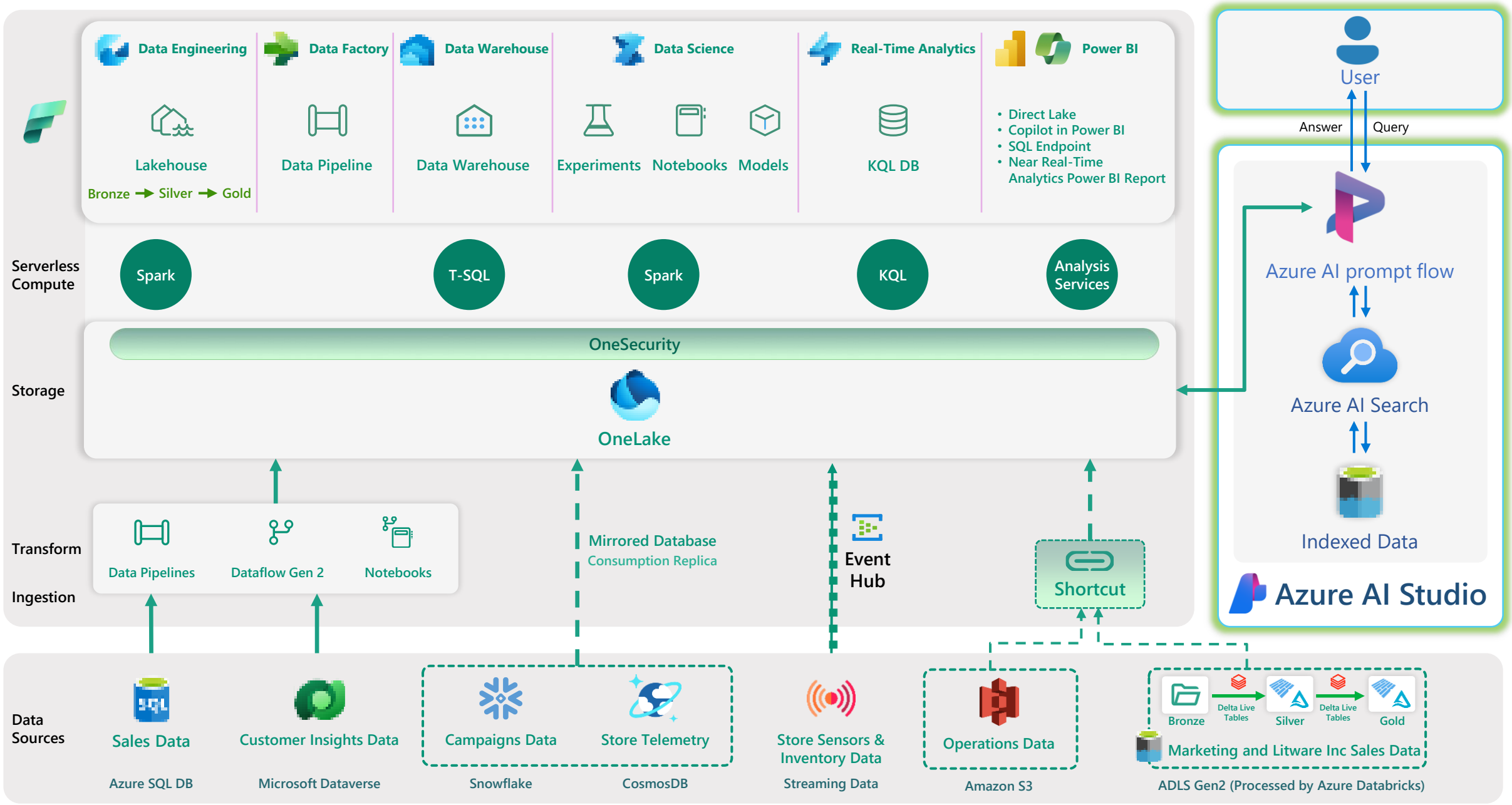
1. Data and AI Pros access, use and process data across the AI lifecycle with OneLake as the connective tissue.

2. Data Pros can pre-process and share AI-ready data with AI Pros. Once in Azure AI Studio, AI Pros can create their own generative AI applications and copilot experiences.

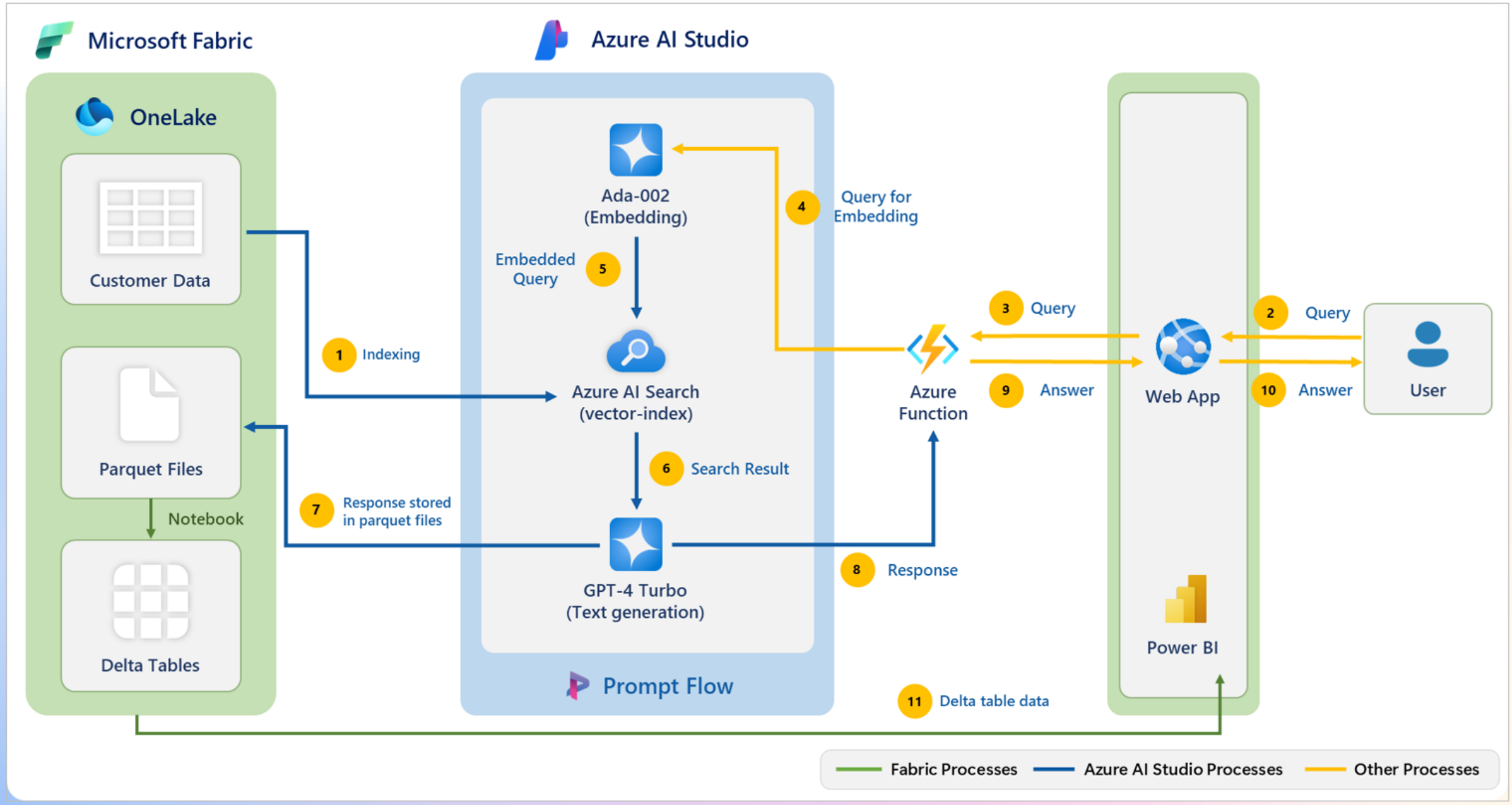
3. AI Pros can make their models and generative AI apps accessible in Microsoft Fabric for Data Pros to enrich their analytics workflows in their lakehouse and serve through Power BI.



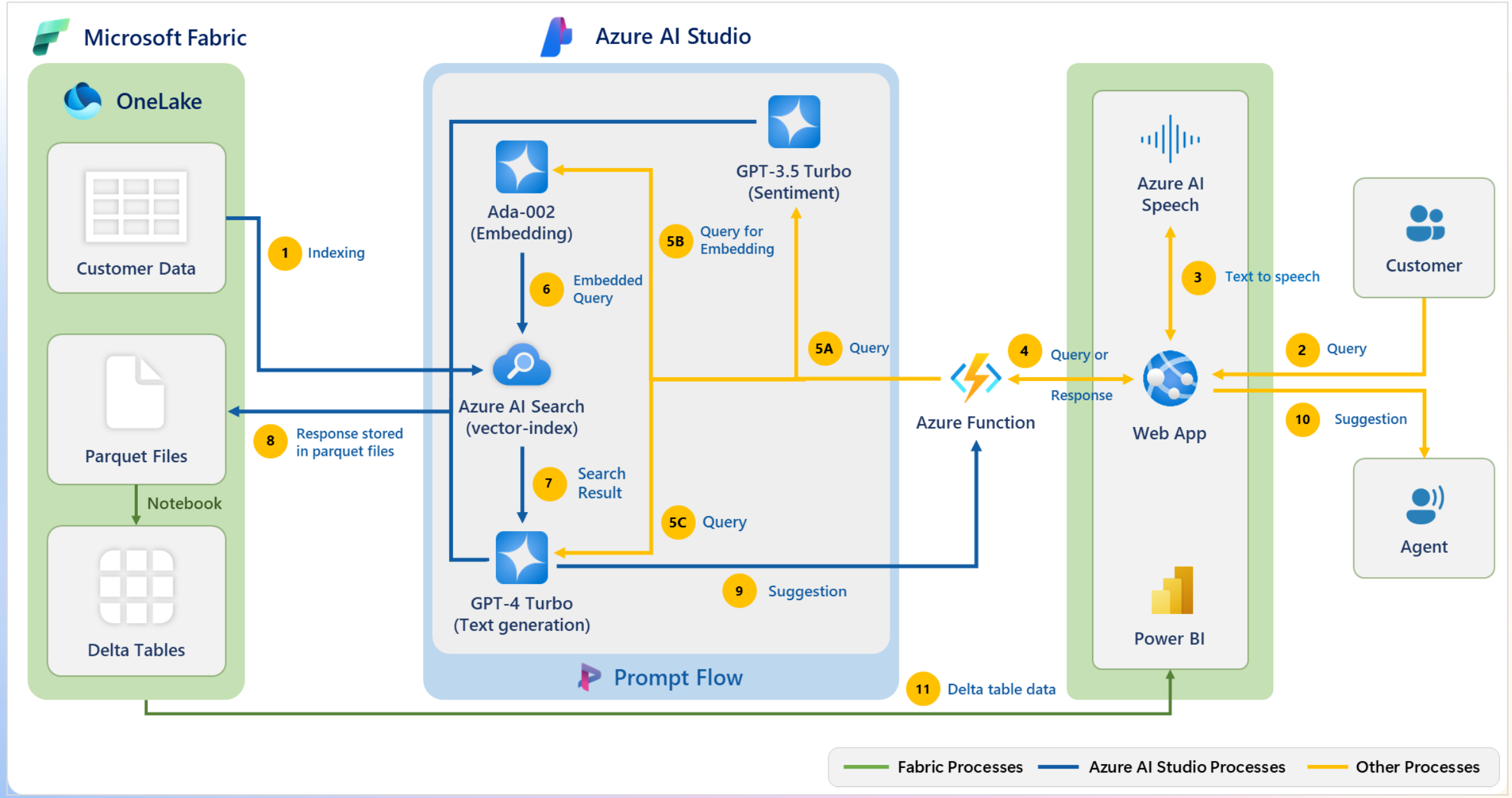
# Microsoft Fabric and Azure AI Studio Architecture



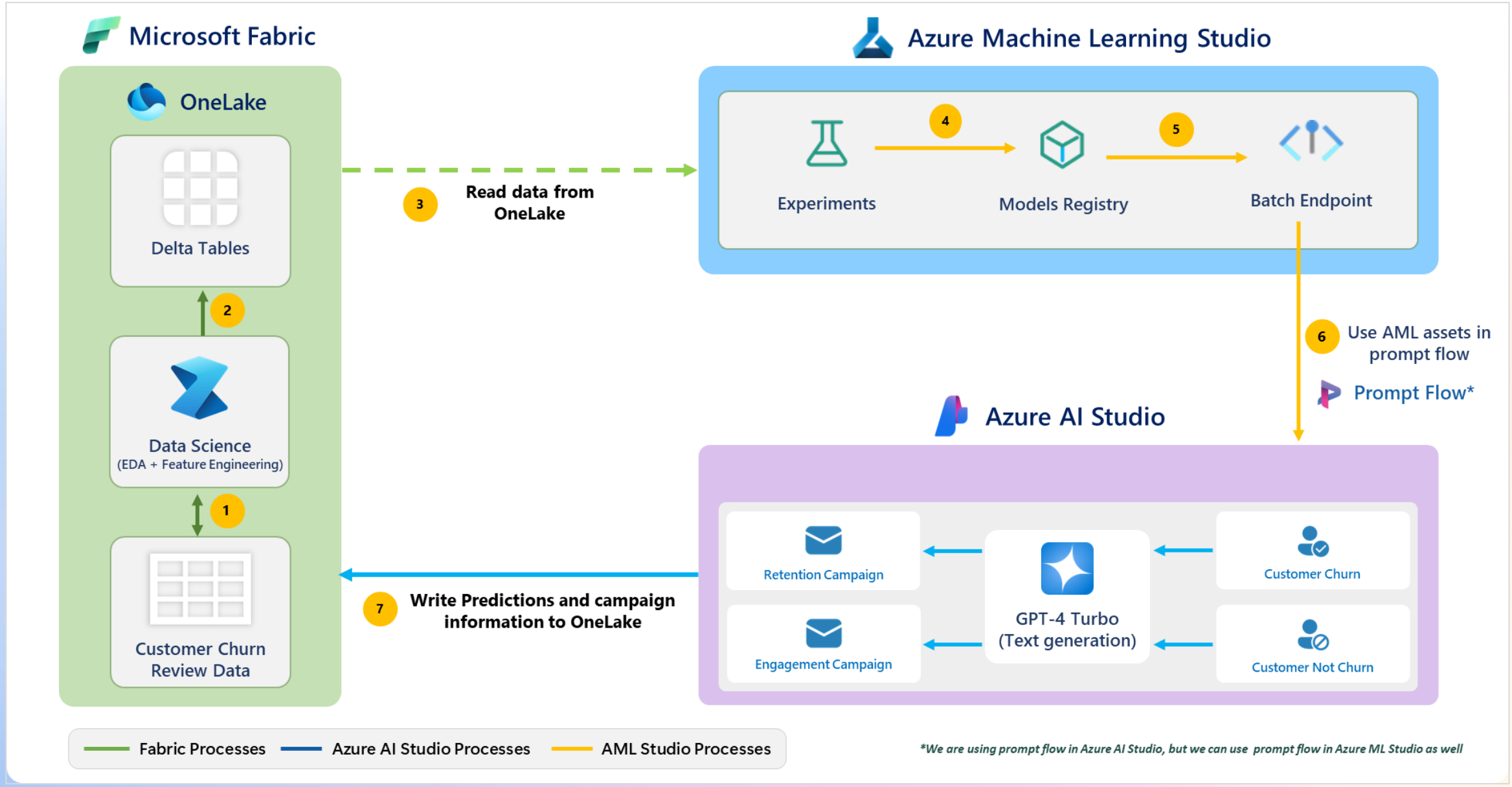
# Enhance Data Insights



# Analyze Customer Interactions



# Customize Machine Learning Models



# Use cases

Top use cases for

# Azure AI Studio

**Build your own copilot**
Your data. Your apps. Your people

**Enterprise chat**
Provide multi-modal knowledge mining

**Speech analytics**
Improve interactions and service

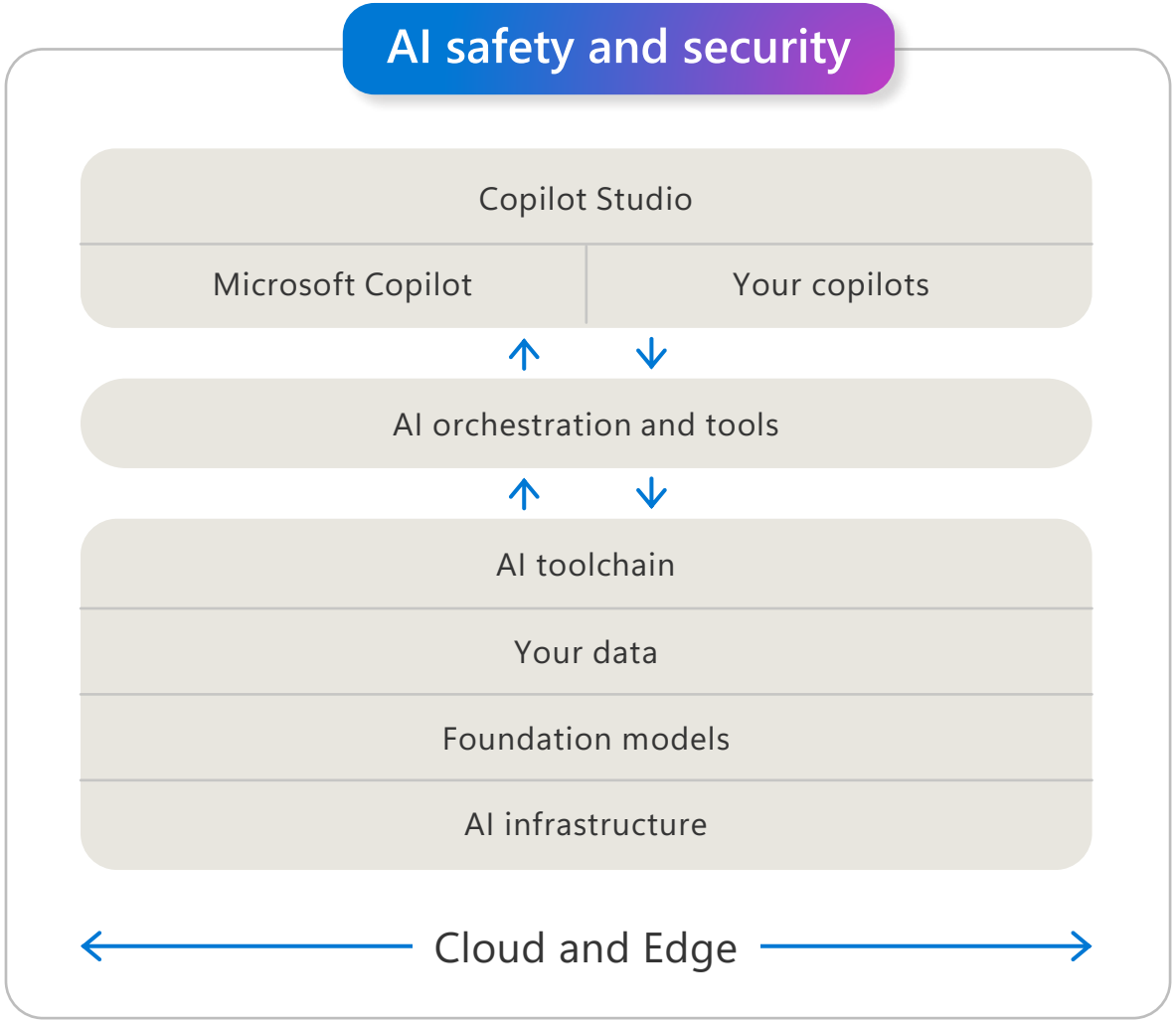
**Content generation**
Deliver new products and services

**Hyper-personalization**
Support better sales and marketing

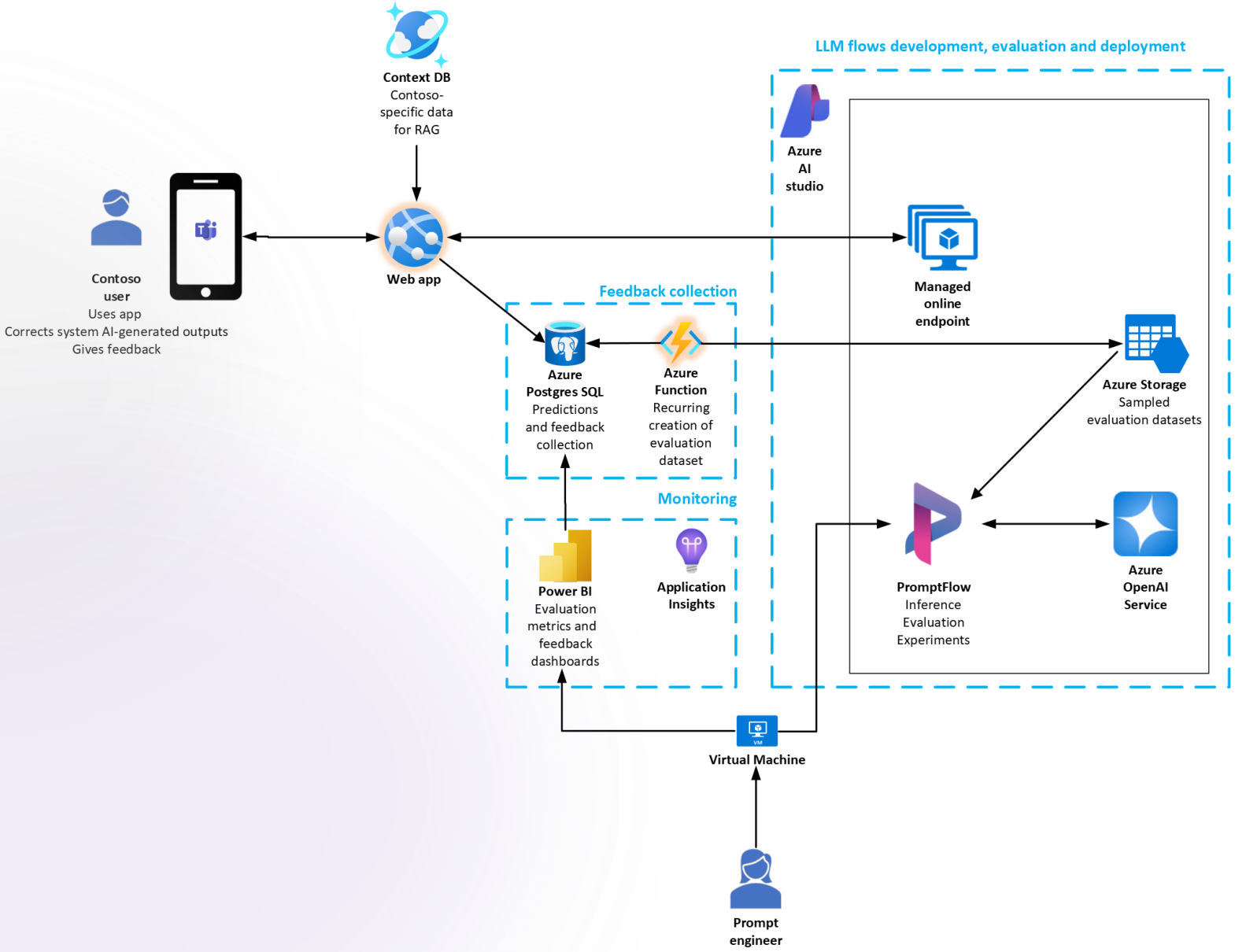
# Copilot development stack

"You may look at Bing Chat and think this is some super magical complicated thing, but Microsoft is giving developers everything they need to get started to go build a copilot of their own…over the coming years, this will become an expectation for how all software works."

*Kevin Scott, Microsoft CTO*



# Build your own copilot with Microsoft Teams



# Enterprise chat High-level Architecture

Enterprise chat

## Select your chat model

**GPT-4**
Azure OpenAI Resource
DEFAULT

**ChatGPT**
Azure OpenAI Resource

**Llama 2**
Meta

**Model Catalog**
OSS LLM Model offerings

**Fine-Tuned LLM**
Any LLM in your collection you previously fined-tuned.

## Add and index your data

**Import Data**
Blob Storage, S3, GitHub, unstructured files

**Azure AI Search**
Keyword, Vector, Hybrid, Ranking
DEFAULT

**Fabric**
Flow Connection Tool available

**MS Graph**
Flow Connection Tool available

**External Vector Database**

## Calibrate your prompting logic

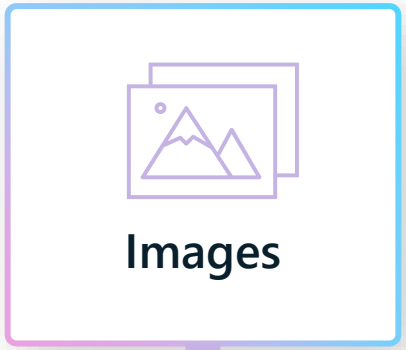
- Customize System Prompt
- Establish retrieval augmenting search connector
- Add prompting examples
- Add flow tools and plugins
- Measure effectiveness with manual and metric evaluations
- If additional calibration required, consider fine tuning your own model

## Deploy and monitor usage

- Ensure appropriate RAI Compliance before deployment
- Deploy enterprise chat solution endpoint
- Monitor token consumption and adjust solution parameters accordingly
- Monitor compute and resource metric costs over time and optimize solution infrastructure as needed

# Let’s look at an example: Contoso Outdoor Company

Get the sample code on GitHub
Contoso Chat




Add your data

Azure AI Search + embeddings from Azure AI Vision

GPT-4 Turbo with Vision

# Speech analytics High-level Architecture

Speech analytics

## Analytics results

Summarization | Conversation Analysis | Content Generation | ...

### Ingest

audio

video

1

*Upload into Blob Storage*

### Storage

### Trigger

*Automated job execution on new audio*

2

### AI Services

4a

*Optionally Speech Ingestion Service can output only transcript*

4

*Speech Ingestion Service triggers* PromptFlow

3

*Speech Ingestion Service creates transcript*

5

*PromptFlow executes analytics Flow, e.g. based on Azure OpenAI & Language*

6

*PromptFlow writes results to Blob Storage*

### Storage

Azure AI Studio

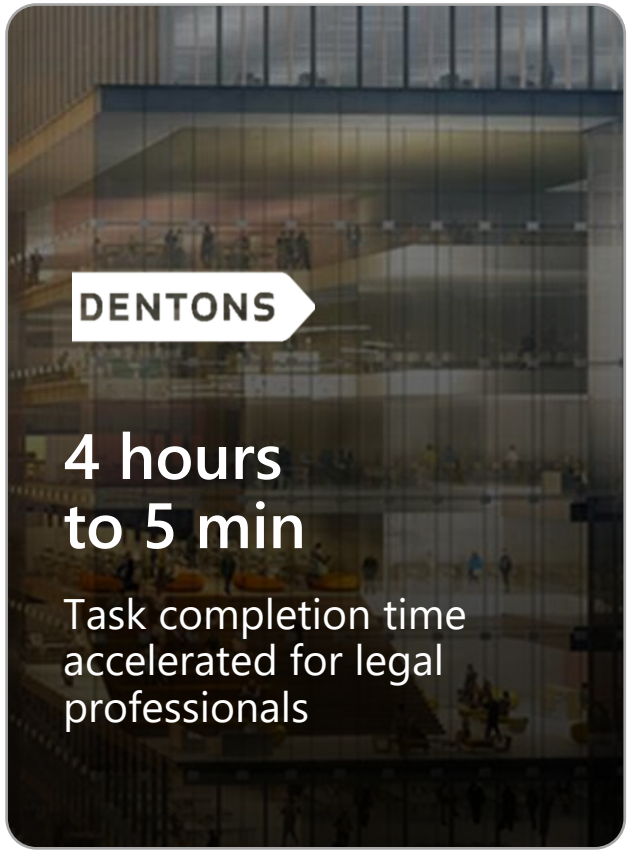
Easy one-time setup

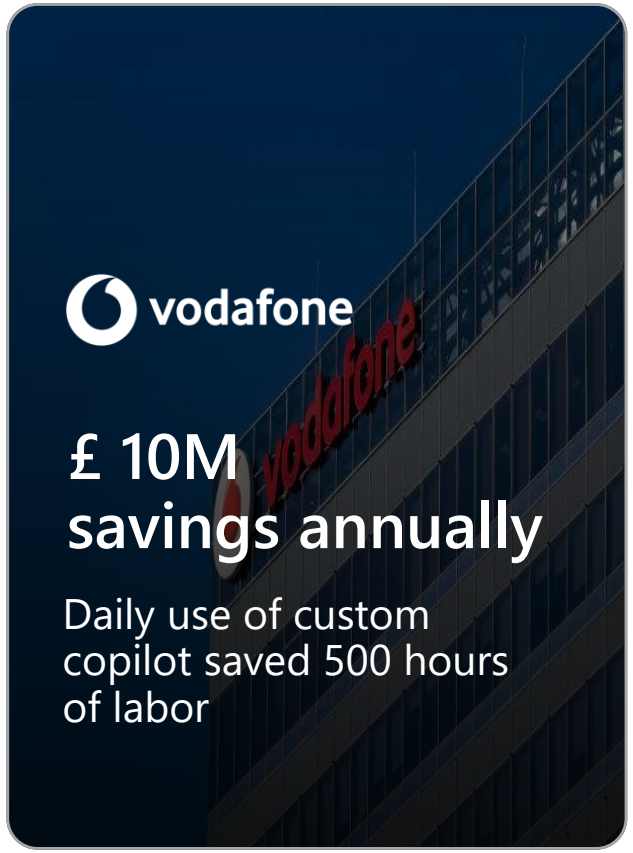
Automated processing of files uploaded

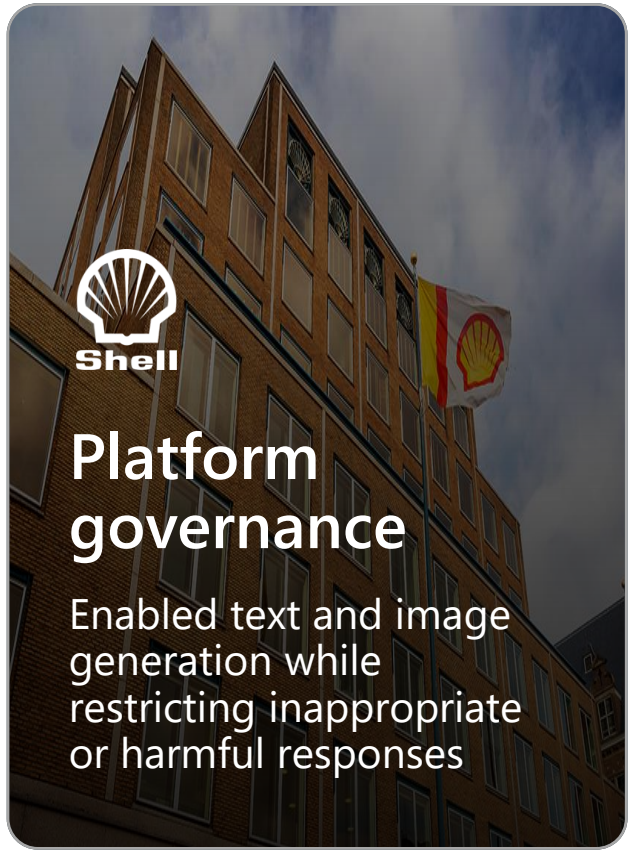
# Speech Analytics Demo Video

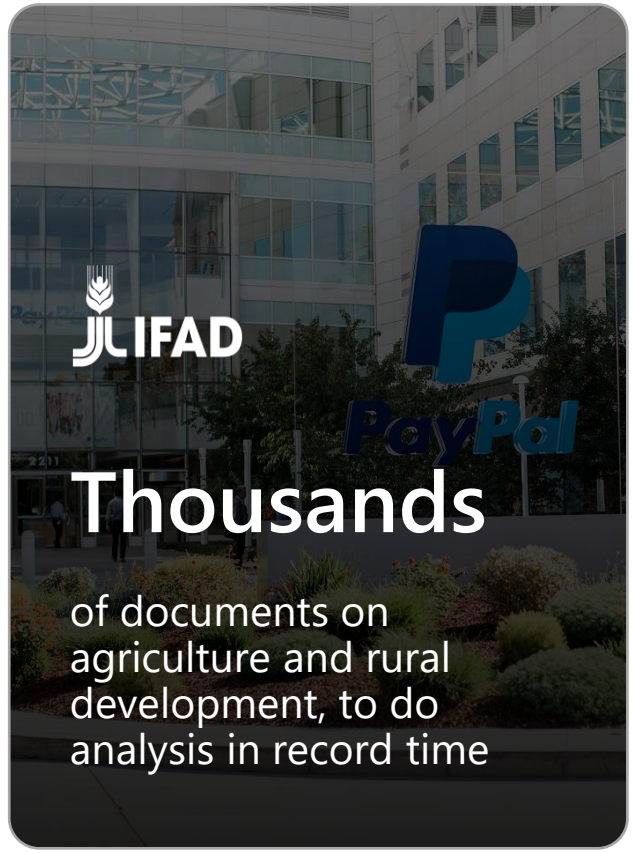
# What We're Hearing from Customers

# 53,000+ companies use Azure AI today



SIEMENS

**Customer:**
Siemens Digital Industries Software

**Industry:**
Manufacturing

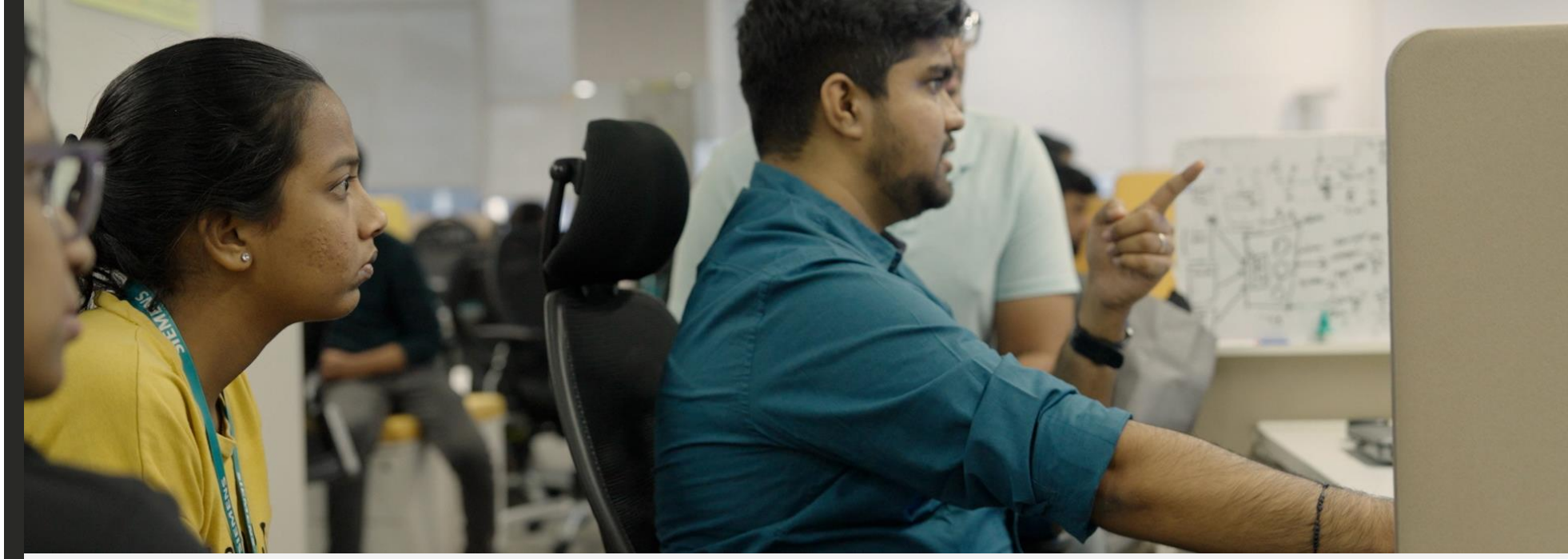
**Size:**
Corporate (10,000+ employees)

**Country:**
Germany

**Products and services:**
Azure OpenAI Service
Azure AI Studio

Read full story here

Microsoft



"We're anticipating more problems being reported and faster resolution time, and using Azure AI allows us to do that."

— Joe Bohman, Executive Vice President, Siemens Digital Industries Software

## Situation:

Siemens wanted to bridge the gap between field and shop floor workers and operations and engineering teams to better drive innovation and efficiency, but more importantly, help companies rapidly address problems as they arise.

## Solution:

Using Microsoft Teams and Azure AI, Siemens created an AI-powered collaborative app to help support their industry leading product lifecycle management (PLM) solution, Teamcenter, and connect people who find problems with those who can fix them.

## Impact:

The Siemens Teamcenter for Microsoft Teams app promises to close feedback loops faster and help cross-functional teams resolve more issues faster—and together.

ASOS

**Customer:**
ASOS

**Industry:**
Retailer

**Size:**
Medium (1000-9,0000 employees)

**Country:**
United Kingdom

**Products and services:**
Azure OpenAI Service
Azure AI Studio
Azure AI Content Safety

[Read full story here]

Microsoft

"Our customers come to us for great fashion. Having a conversational interface option could get us closer to our goals of fully engaging our customers and personalizing their experience by showing them the most relevant products at the most relevant time."

– Cliff Cohen, Chief Technology Officer, ASOS

## Situation:

ASOS sought a way to better engage shoppers and differentiate its customer experience in a highly competitive online retail market by building on its natural language processing knowledge and incorporating generative AI.

## Solution:

Using Microsoft Teams and Azure AI, Siemens created an AI-powered collaborative app to help support their industry leading product lifecycle management (PLM) solution, Teamcenter, and connect people who find problems with those who can fix them.

## Impact:

ASOS prototyped the AI-powered experience in just a few weeks and tested it successfully for secure and responsible AI. Confident in the stability and scalability of the AI features, ASOS plans on releasing the tool it created to customers later this year.

perplexity

**Customer:**
Perplexity.AI

**Industry:**
Partner Professional Services

**Size:**
Small (1-49 employees)

**Country:**
United States

**Products and services:**
Azure
Azure AI Studio
Azure OpenAI Service

Read full story here

Microsoft

"Azure AI Studio improved the experience for creating AI products. We found it mapped perfectly to our needs for faster development and time to market, and greater throughput, scalability, security, and trust."

—Denis Yarats, Chief Technology Officer and Cofounder, Perplexity.AI

## Situation:

Perplexity.AI needed a platform for its conversational answer engine that would support fast time to market, serve as a force multiplier for its lean staff, scale to support millions of users, and cost-effectively deliver security and reliability

## Solution:

After beginning development on OpenAI, the company adopted Azure AI Studio and Azure OpenAI Service. Perplexity.AI says Azure is a mature, reliable, well-known platform with enterprise-grade compliance and a reputation for security.

## Impact:

With a staff of three, the company brought Perplexity Ask to market in just six months. Its engine scales to support 7 million users per month, has doubled throughput to 600,000 tokens per minute, cut latency by 30 percent, and reduced the cost per token.


Microsoft

# McDonald’s China transforms its operations, elevates service levels with Azure AI

Customer: McDonald’s China

Industry: Retail and Consumer Goods

Size: 10,000+ employees

Country: China

Publish date: March 2024

Read the full story here

**“Collaborating with Microsoft and utilizing new technologies has enabled us to become a game-changer in our industry. With Azure AI and GitHub Copilot, we’re not just optimizing our operations, we’re revolutionizing them.”**

— Aaron Huang , Group Chief Financial Officer , McDonald's Group China

**Challenge:** With an aim to help employees stay at the forefront of business innovation as it opens nearly 10,000 new locations over the next five years, McDonald’s China decided to employ AI to elevate all levels of customer service, quality, and operational excellence at scale.

**Solution:** The McDonald’s AI Lab used Azure AI, as well as GitHub Copilot and Microsoft 365 Copilot, to support a full set of intelligent solutions based on large language models, natural language interaction, generative intelligence, and machine learning on the Azure cloud platform.

**Impact:** Feedback has been positive, and the project has laid the groundwork for a complete enterprise-level intelligent development blueprint to improve operational efficiency, empower employees, and improve the consumer experience across all McDonald's locations in China.

**Products:** Azure, GitHub Copilot, Microsoft 365 Copilot, Azure AI Search, Azure AI Speech, Azure AI Vision, Azure Kubernetes Service, Azure AI, Azure Machine Learning

SWECO | Microsoft

# Sweco Group empowers its architects and engineers with a timesaving AI assistant built in Azure AI Studio

Customer: Sweco
Industry: Professional Services
Size: 10,000+ employees

Country: Sweden
Publish date: May 2024
Watch the full story here

**"We really appreciate the one-click deployment of the models in Azure AI Studio and that it makes Azure AI offerings transparent and available to the user."**

— Shah Muhammad, Head of AI Innovation, Sweco

**Challenge:** Sweco's engineers and architects manage numerous, complex tasks every day. To allow more time for more creative solutions in client projects, the company recognized the need for a supportive solution.

**Solution:** The company chose Microsoft Azure AI Studio to build its custom copilot, SwecoGPT, a digital assistant that automates document creation and analysis, delivers advanced search, and provides language translation.

**Impact:** Though it was recently deployed, nearly half of Sweco's employees use SwecoGPT and report increased productivity, giving them more time to focus on creativity and helping customers.

**Products:** Azure AI Services, Azure Machine Learning, Azure AI Studio, Azure Open AI Service

# Microsoft: A leader in cloud AI developer services

**Gartner has recognized Microsoft as a leader in the 2023 Gartner® Magic Quadrant™ for Cloud AI Developer Services.[1]**

- The Azure AI platform delivers a comprehensive offering for language, vision, and automated machine learning.
- Microsoft customers can access and extend a wide range of prebuilt models from Microsoft, OpenAI (via its exclusive agreement), and Hugging Face.
- Microsoft offers a free version of Azure AI, a pay-as-you-go option for small teams, and discounted subscription plans for large enterprises.


[1] Gartner, Magic Quadrant for Cloud AI Developer Services, Jim Scheibmeir, Svetlana Sicular, Arun Batchu, Mike Fang, Van Baker, Frank O'Connor, 22 May 2023.

# Getting started

Organizations are asking not only how, but how *fast*, they can apply this next generation of AI to address the biggest opportunities and challenges they face — safely and responsibly."

Satya Nadella, Microsoft Chairman and CEO

Source: Microsoft FY23 Q4 earnings call

# Accelerate AI deployment with Azure AI Studio

ai.azure.com

Define scope & requirements

Prepare data

Build, train, & evaluate

Review & approve app for production

Deploy & monitor

Govern and monitor continuously

## Microsoft is built on trust

1. Your data is your data
2. Your data is not used to train or enrich foundation AI models
3. Your data and AI models are protected at every step
4. Our Customer Copyright Commitment

# Ready to get started?

Build with
Azure AI Studio

Learn how to build AI solutions
Azure AI Studio Training Modules

Explore technical documentation to scale your project
Azure AI Studio documentation

Get the latest
Azure AI news and resources

# Thank you