



# Microsoft Build 2024

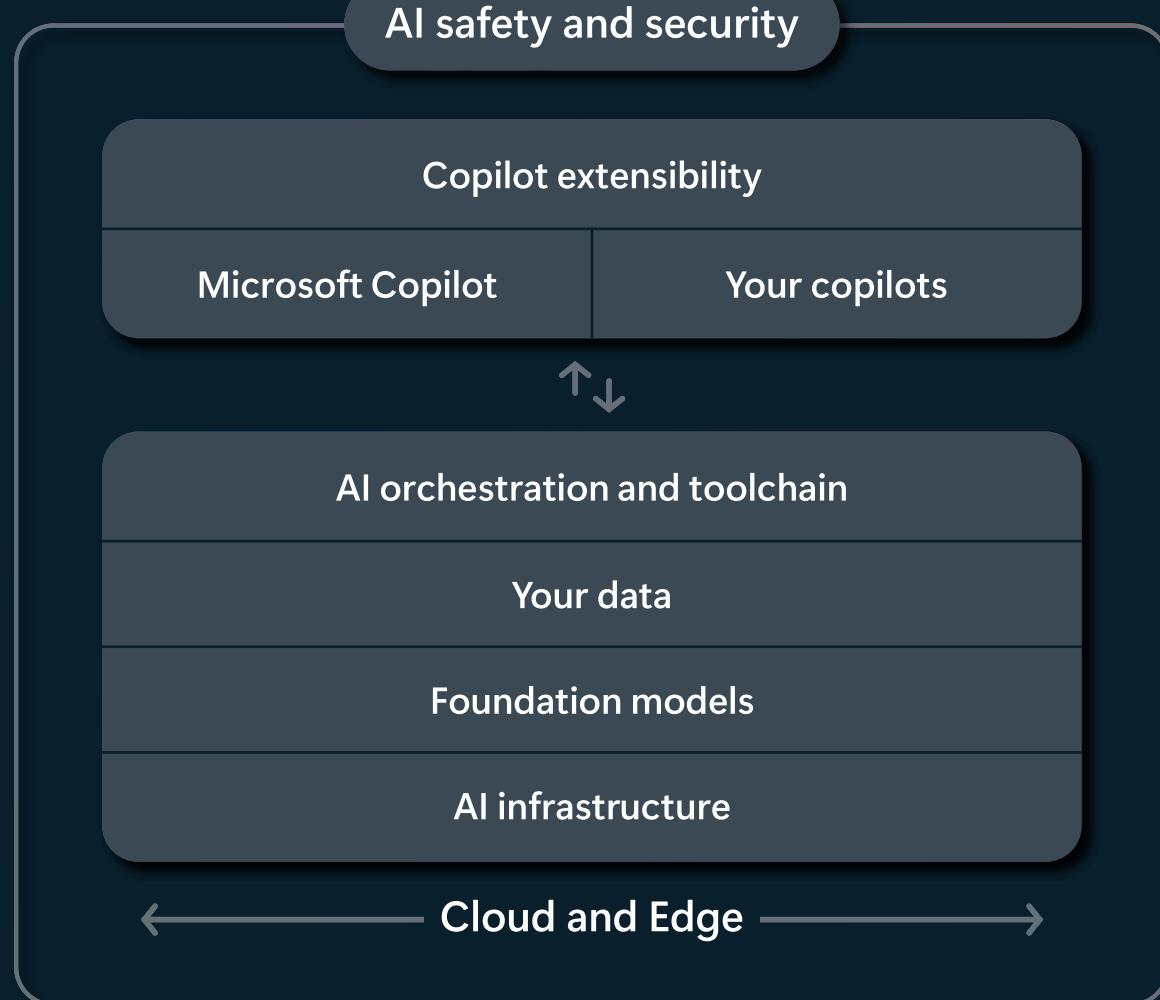
## Cantonese Recap



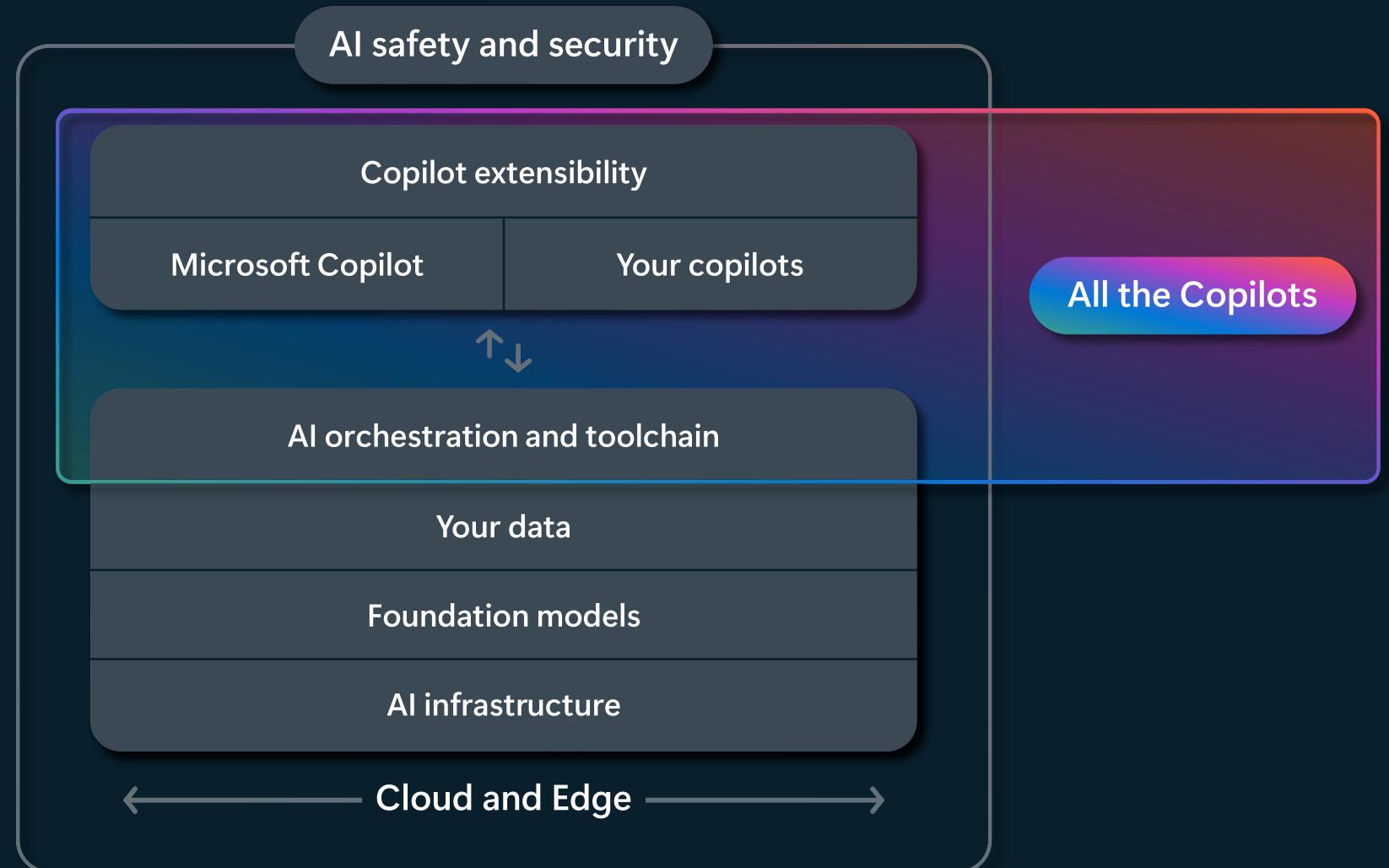
# Agenda



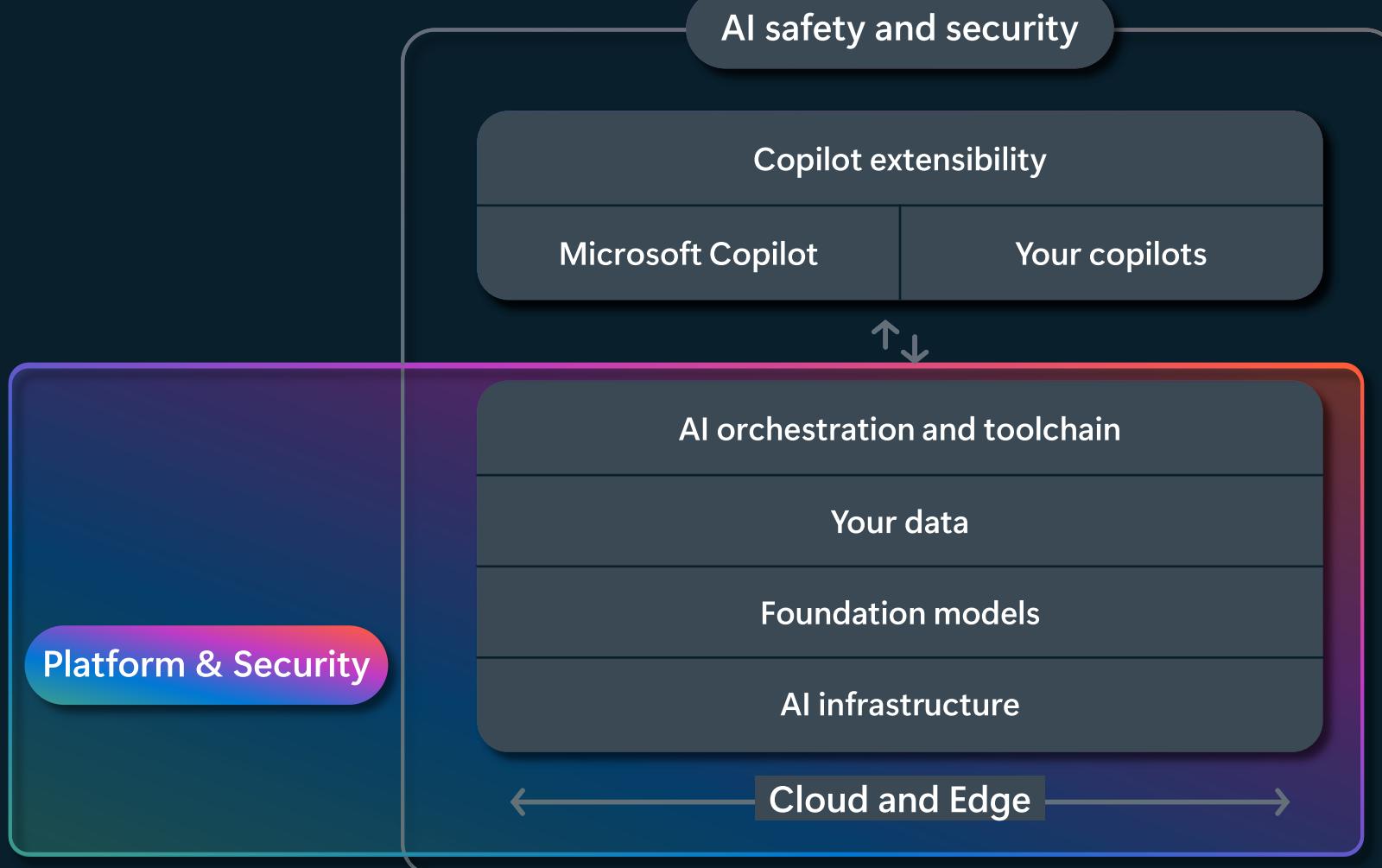
# Copilot stack



# Session 1 – Copilots



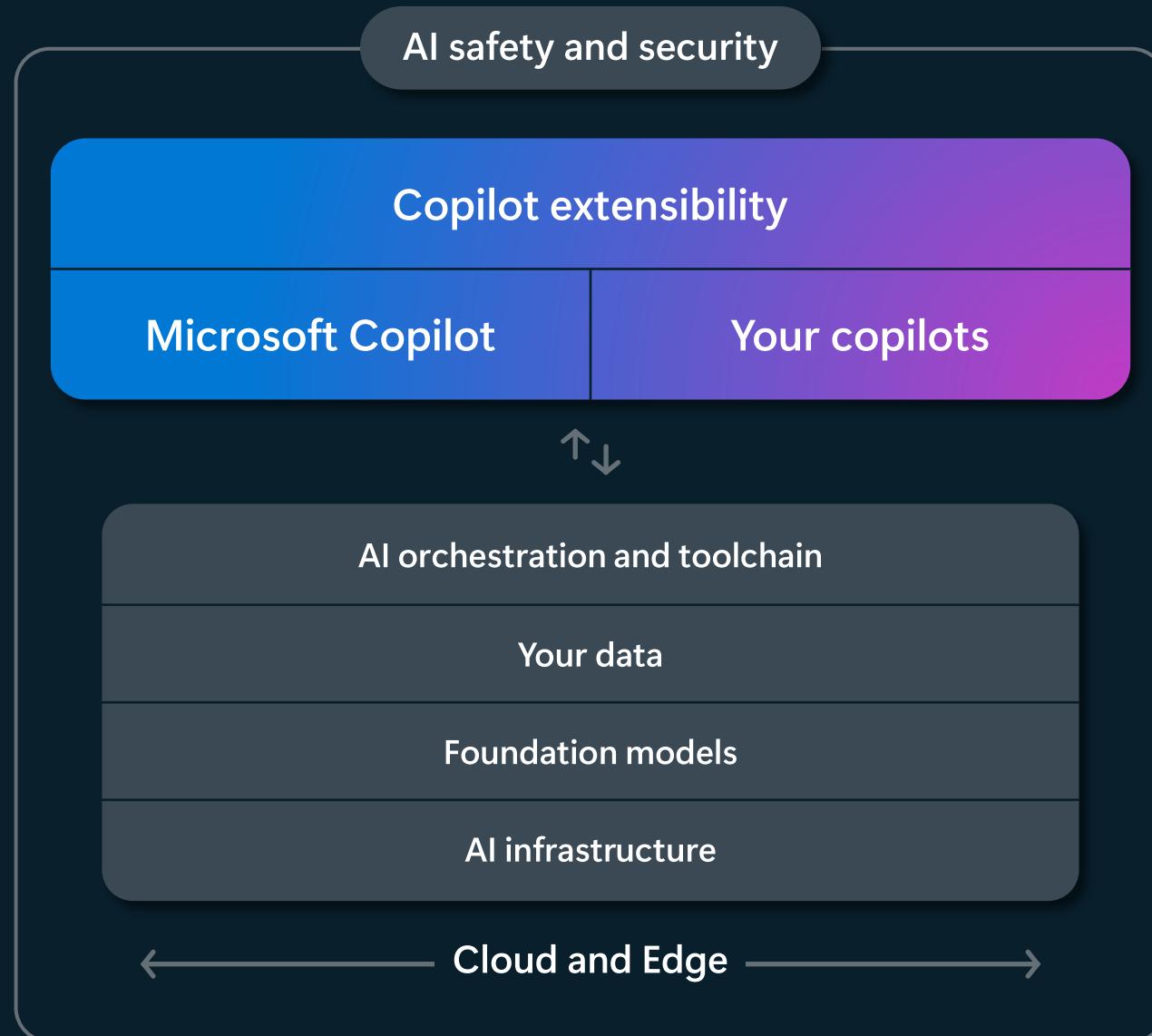
# Session 2 – Platform & Security





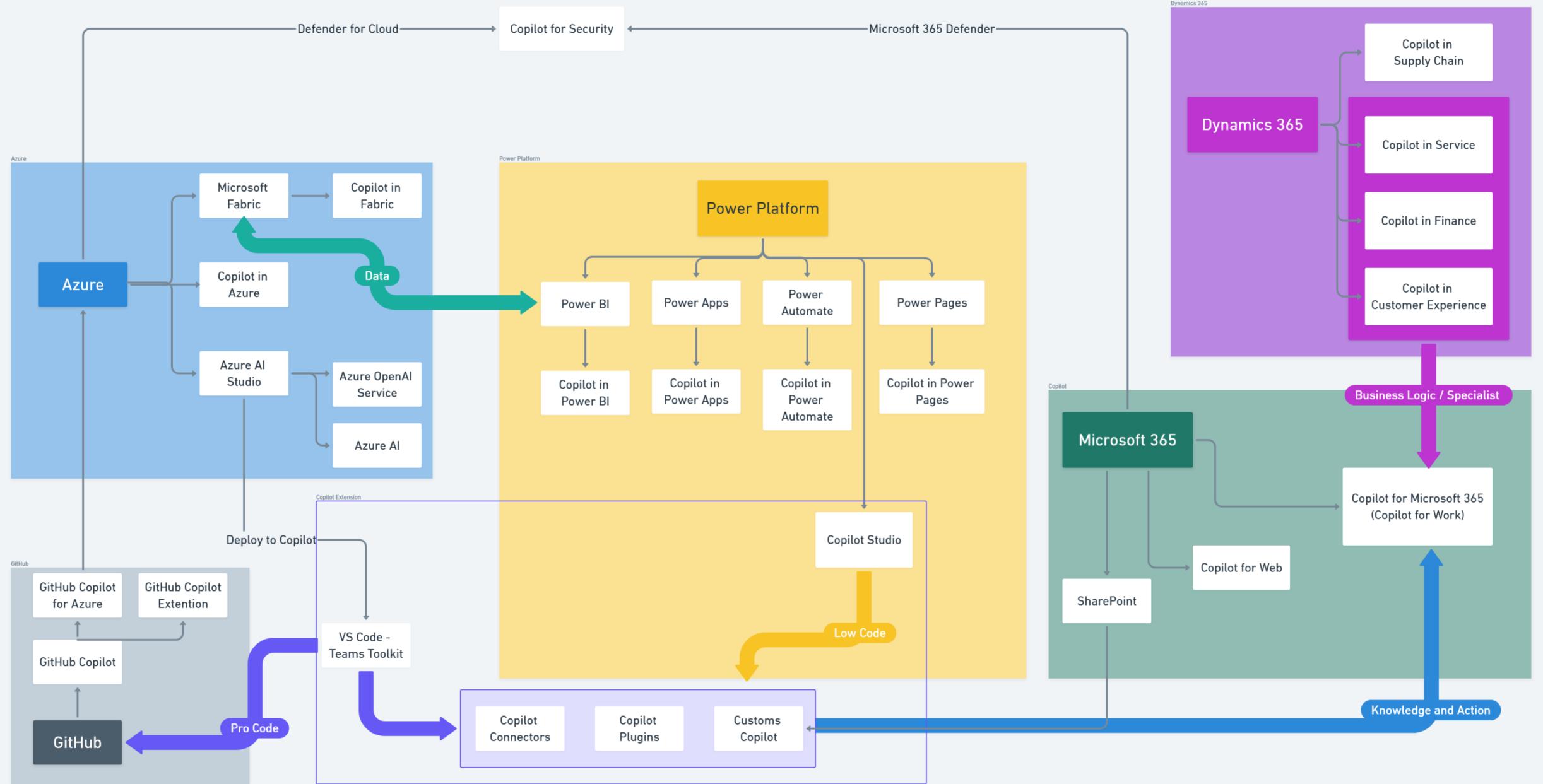
# Copilot, Copilot, Copilot

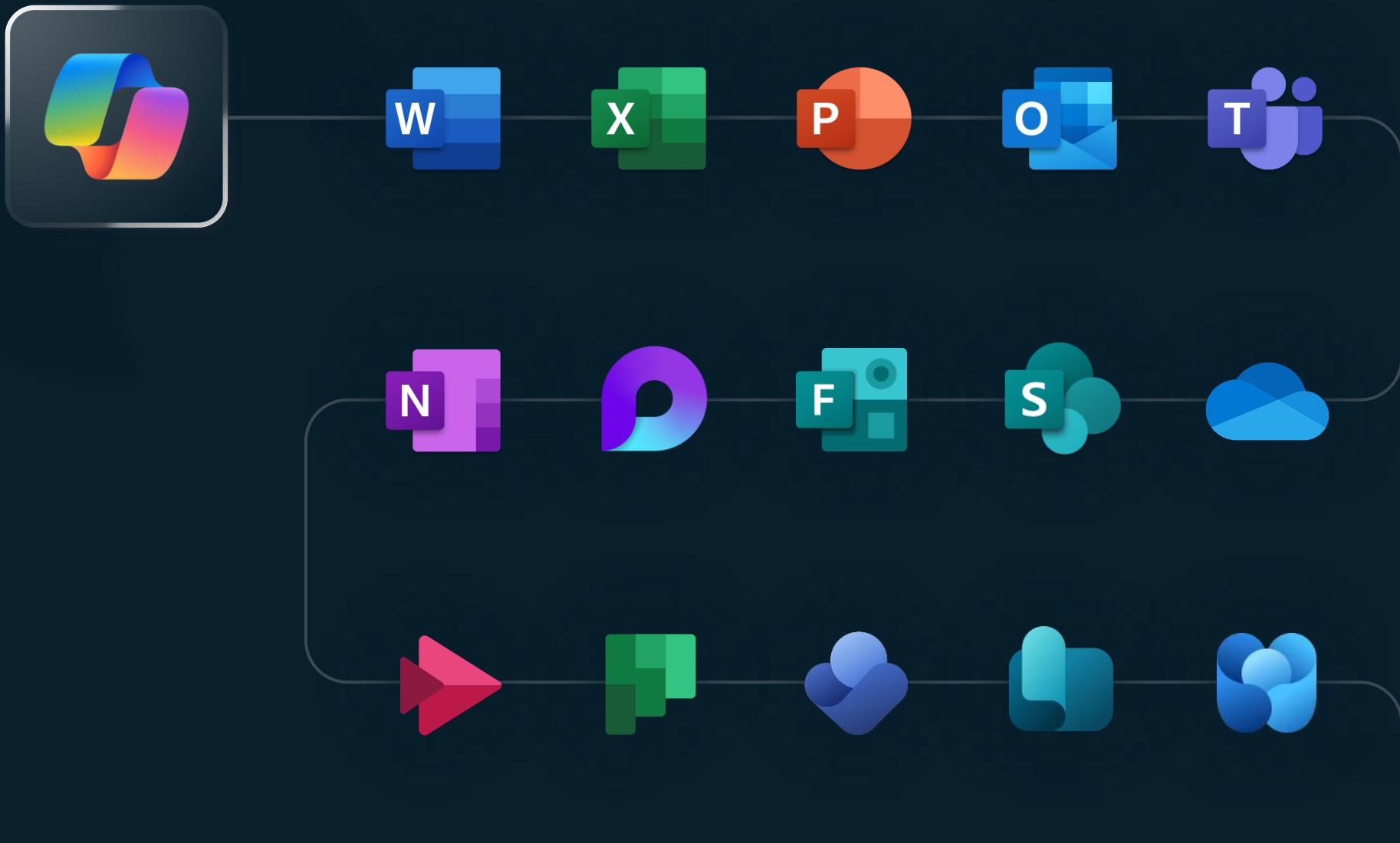




# Henry Yang

Partner Solution Architect –  
Modern Workplace & Security





# Microsoft Copilot

## Assistant



# Microsoft Copilot

Assistant



Team



Announcing

# Team Copilot

A valuable new member of the team

Meeting  
facilitator

Group  
collaborator

Project  
Manager

[aka.ms/TeamCopilot](http://aka.ms/TeamCopilot)



# I know... naming is not our strength

Copilot function  
in Teams



## Copilot in Teams

- ✓ Summarize Chat
- ✓ Summarize Meeting
- ✓ Draft Message

Copilot that can  
be find in Teams



## Copilot (M365 Chat)

- ✓ Refer to all kinds of data in Microsoft 365
- ✓ Prompt to unlock possibility

Copilot for **a team** to use  
together



## Team Copilot

- ✓ Project Management
- ✓ Group collaborator
- ✓ Meeting facilitator

# Microsoft Copilot

Assistant



Team



# Microsoft Copilot

Assistant



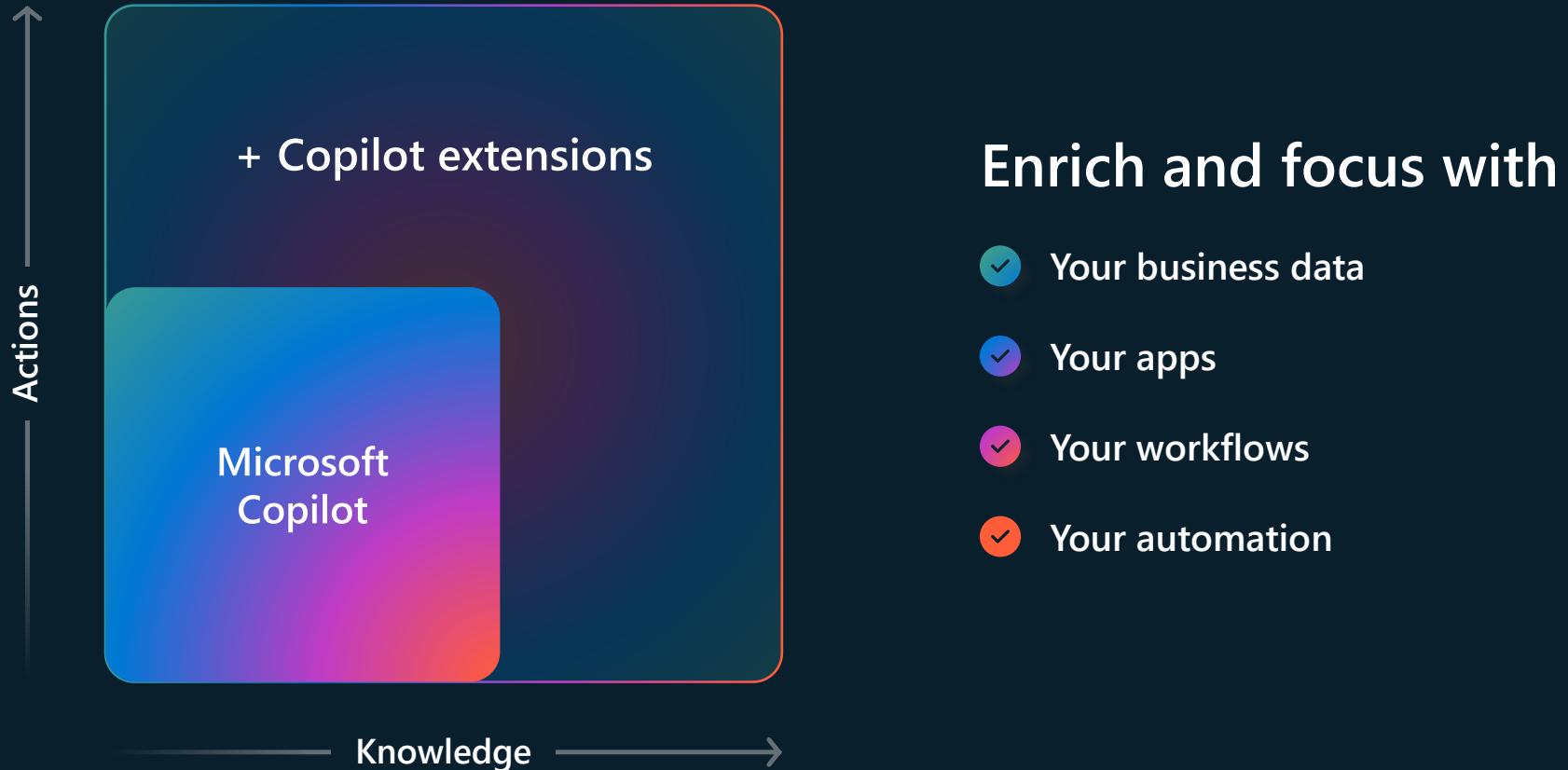
Team



Agents



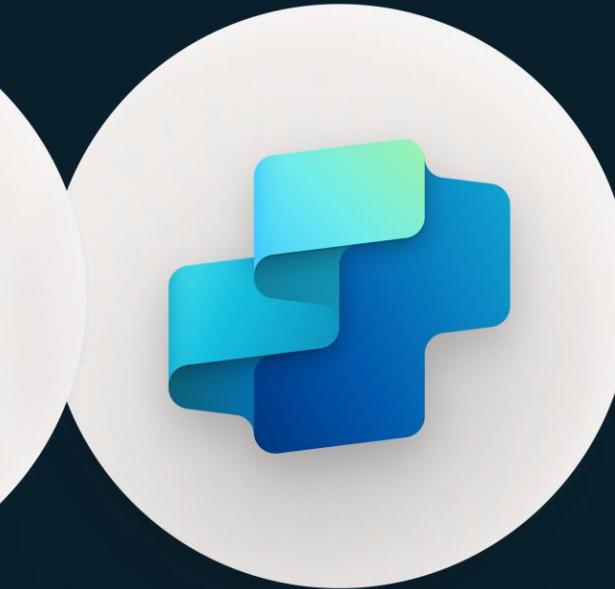
# Copilot extensions



# Enabled with Copilot for M365 license



Copilot for Microsoft 365



Copilot Studio



## Copilot over Enterprise Data



# Copilot extensions

Ground  
responses



## Connectors

- ✓ Your business knowledge

Bring  
actions



## Plugins

- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

Focus user  
experiences



## Your copilots

- ✓ Your business knowledge
- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

# Copilot extensions fundamentals

Workflow instructions

Actions

Triggers

Knowledge

Orchestrator

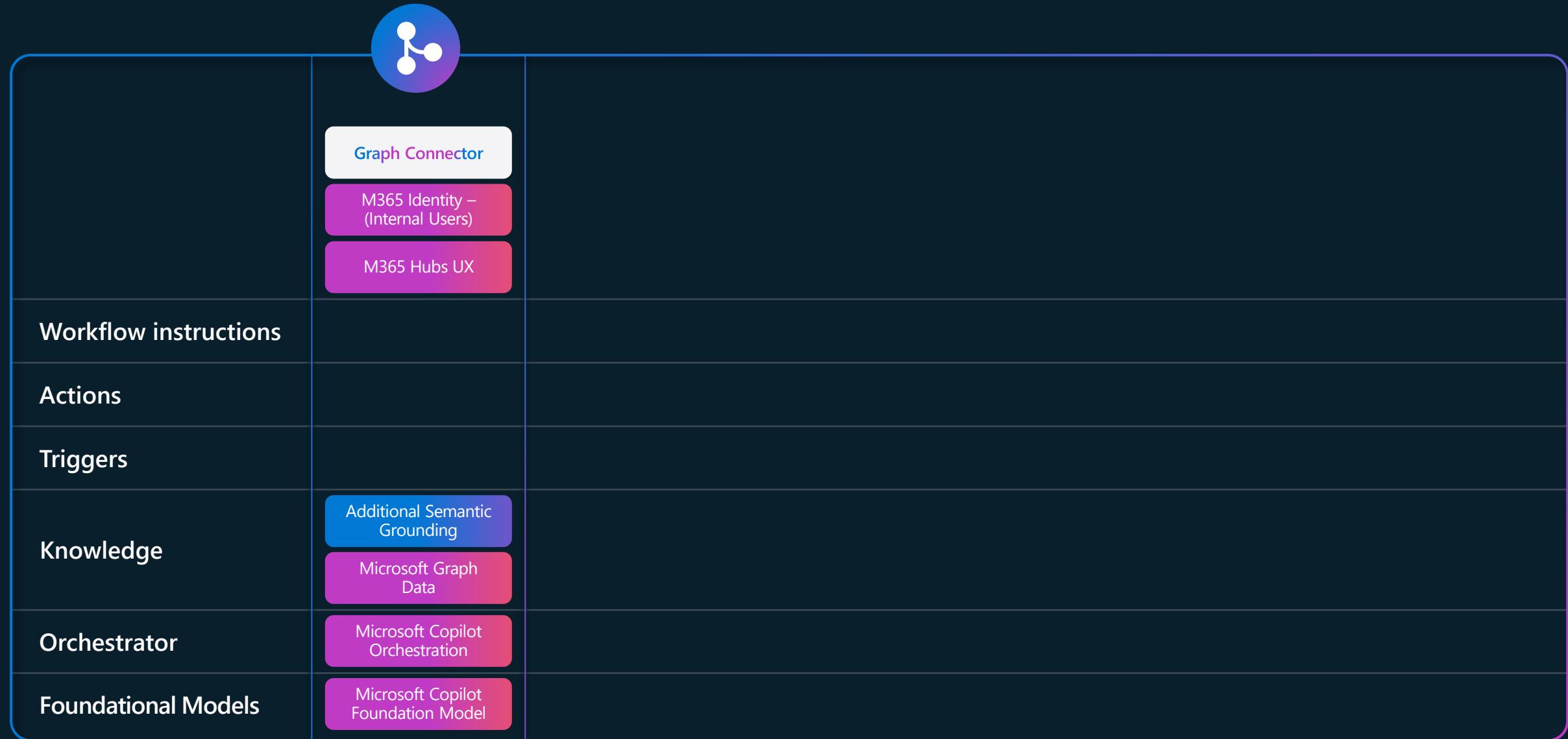
Foundational Models



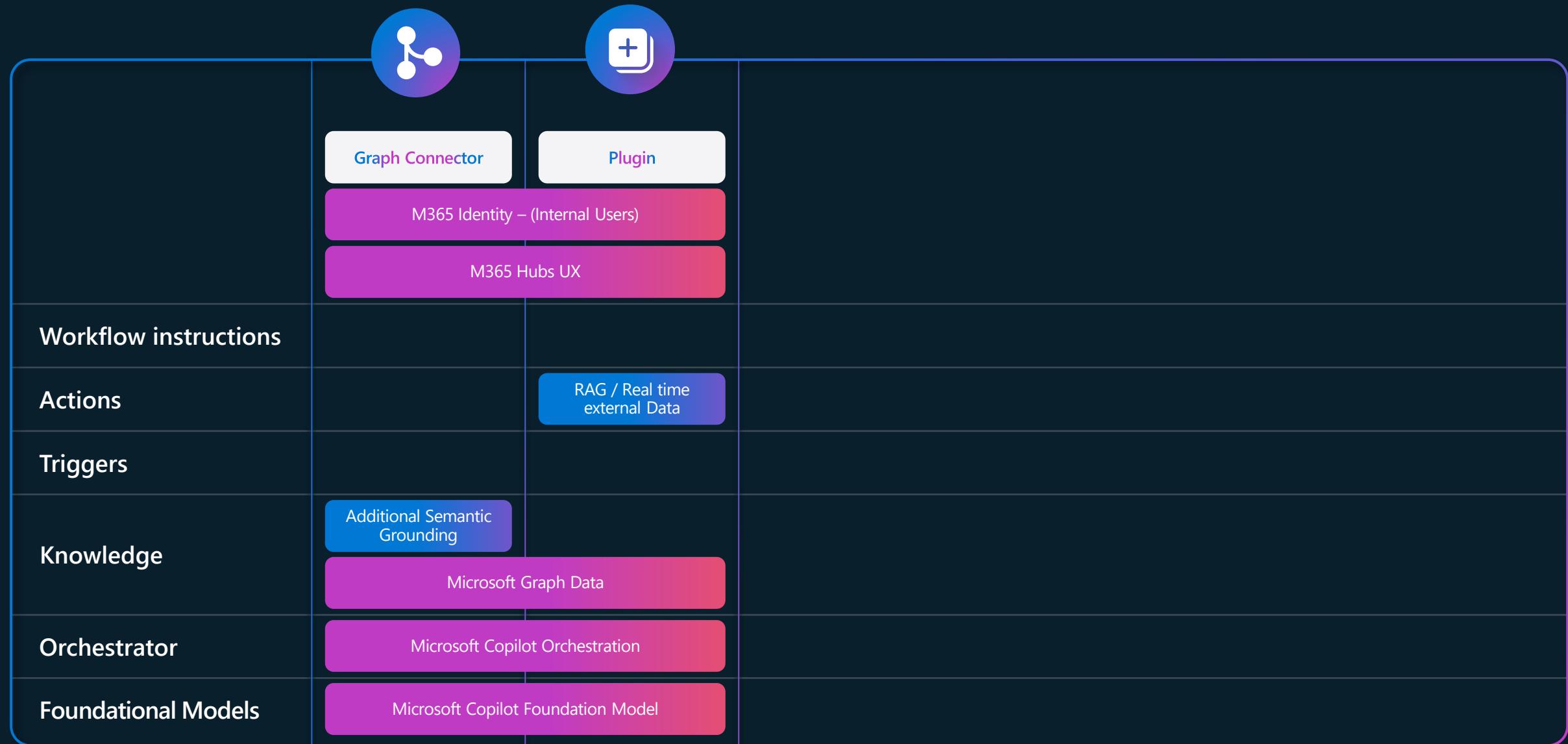
# Copilot extensions fundamentals

	M365 Identity – (Internal Users)	
	M365 Hubs UX	
Workflow instructions		
Actions		
Triggers		
Knowledge	Microsoft Graph Data	
Orchestrator	Microsoft Copilot Orchestration	
Foundational Models	Microsoft Copilot Foundation Model	

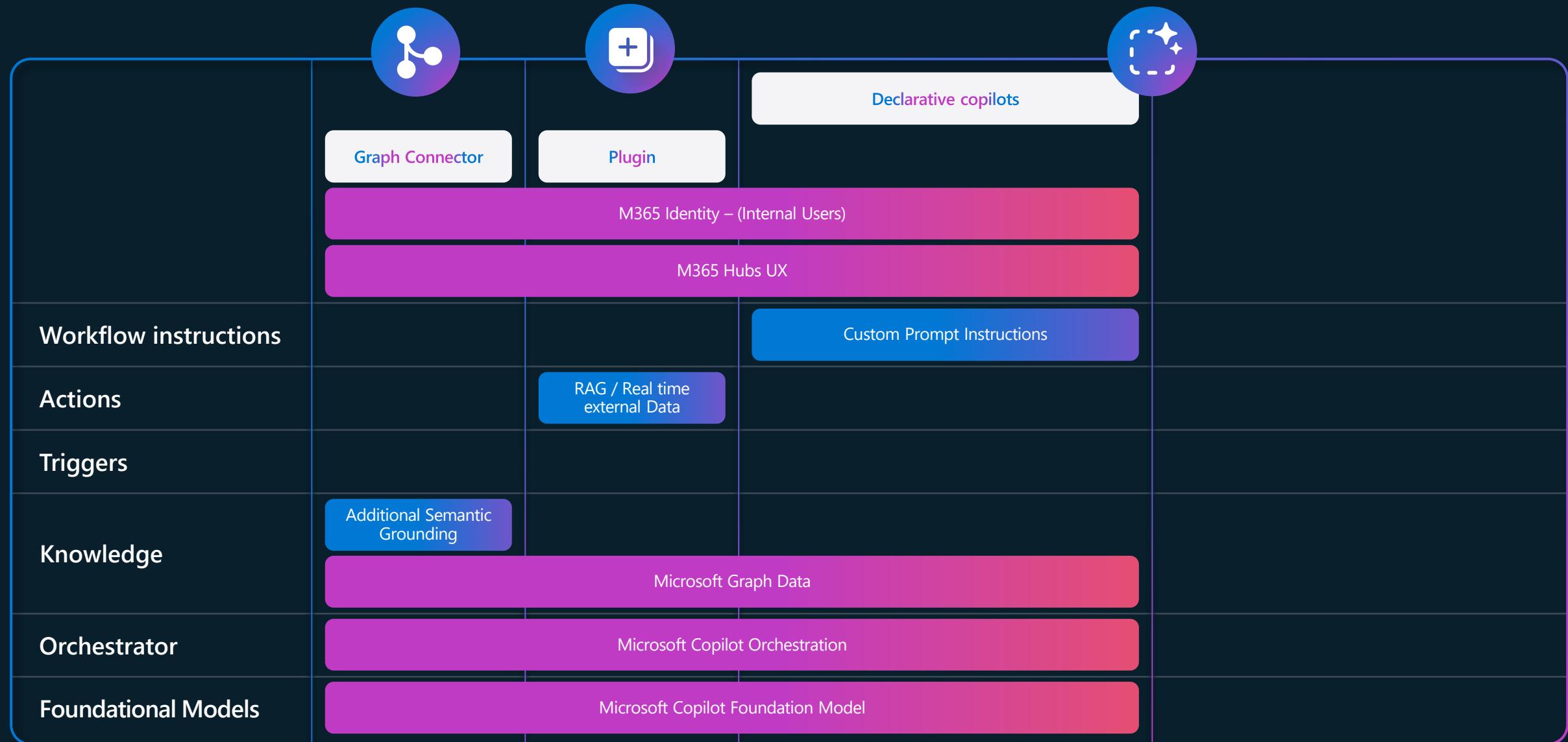
# Copilot extensions fundamentals



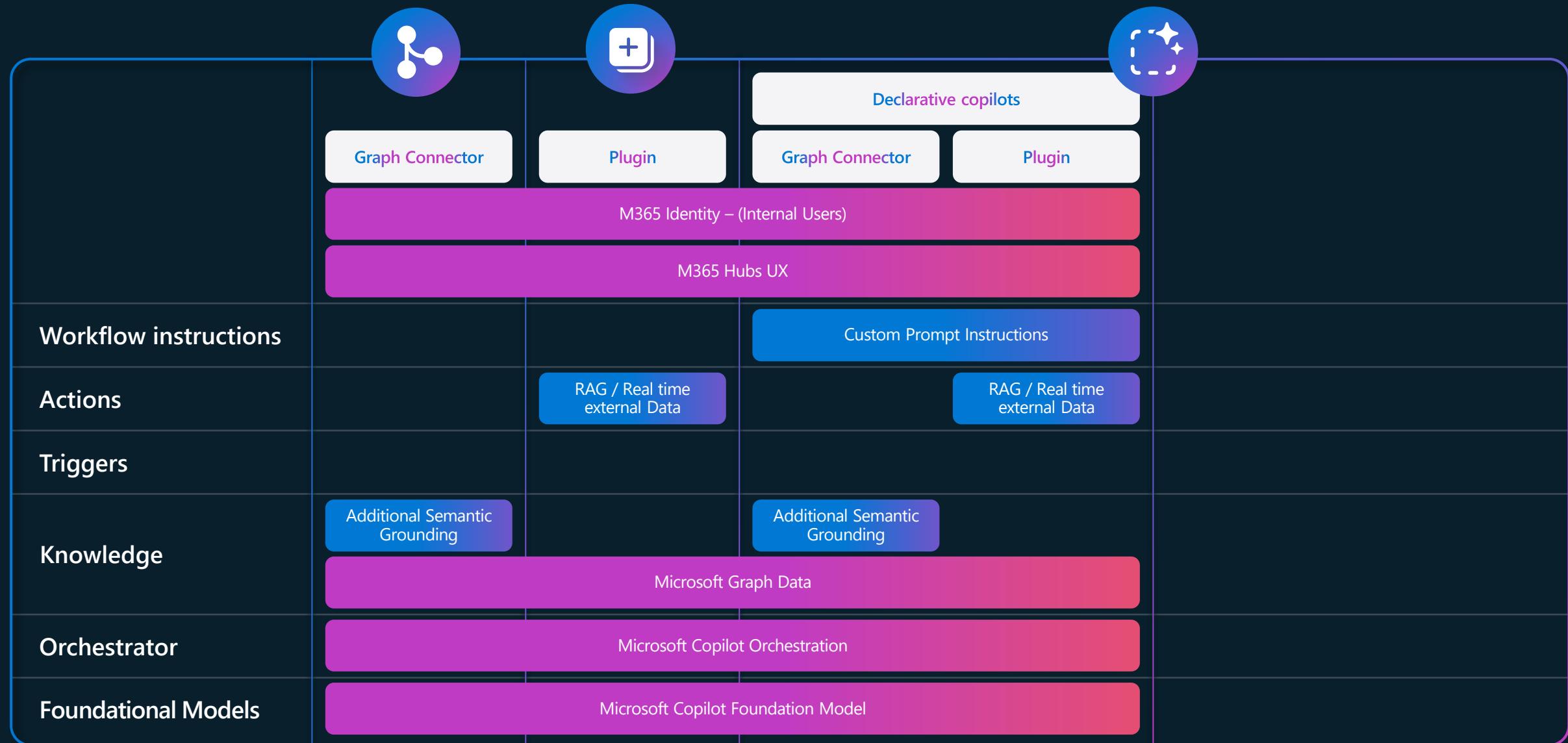
# Copilot extensions fundamentals



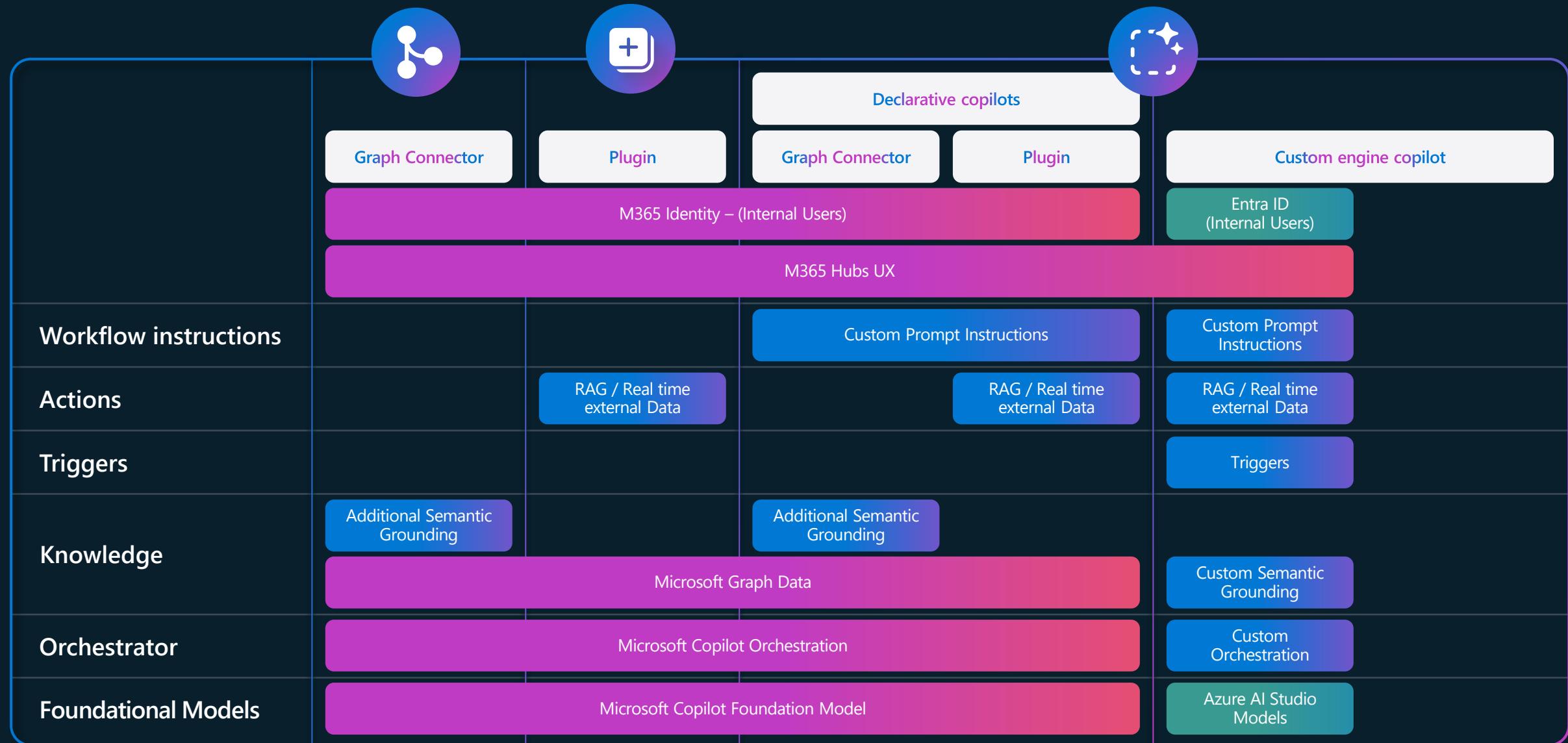
# Copilot extensions fundamentals



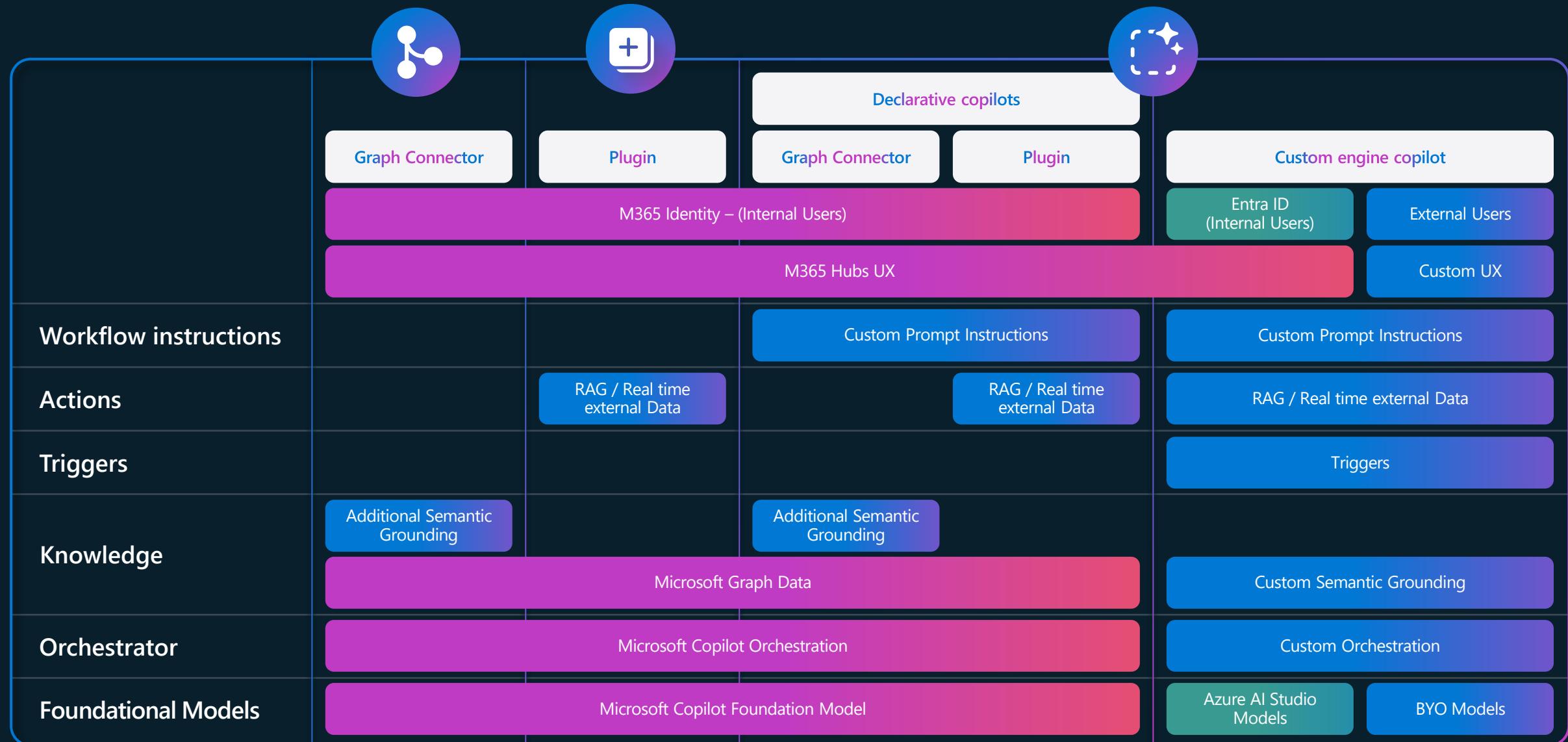
# Copilot extensions fundamentals



# Copilot extensions fundamentals



# Copilot extensions fundamentals



# Copilot extensions

Ground  
responses



## Connectors

- ✓ Your business knowledge

Bring  
actions



## Plugins

- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

Focus user  
experiences



## Your copilots

- ✓ Your business knowledge
- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

# Copilot extensions

Ground  
responses



Connectors

- ✓ Your business knowledge

Bring  
actions



Plugins

- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

Focus user  
experiences

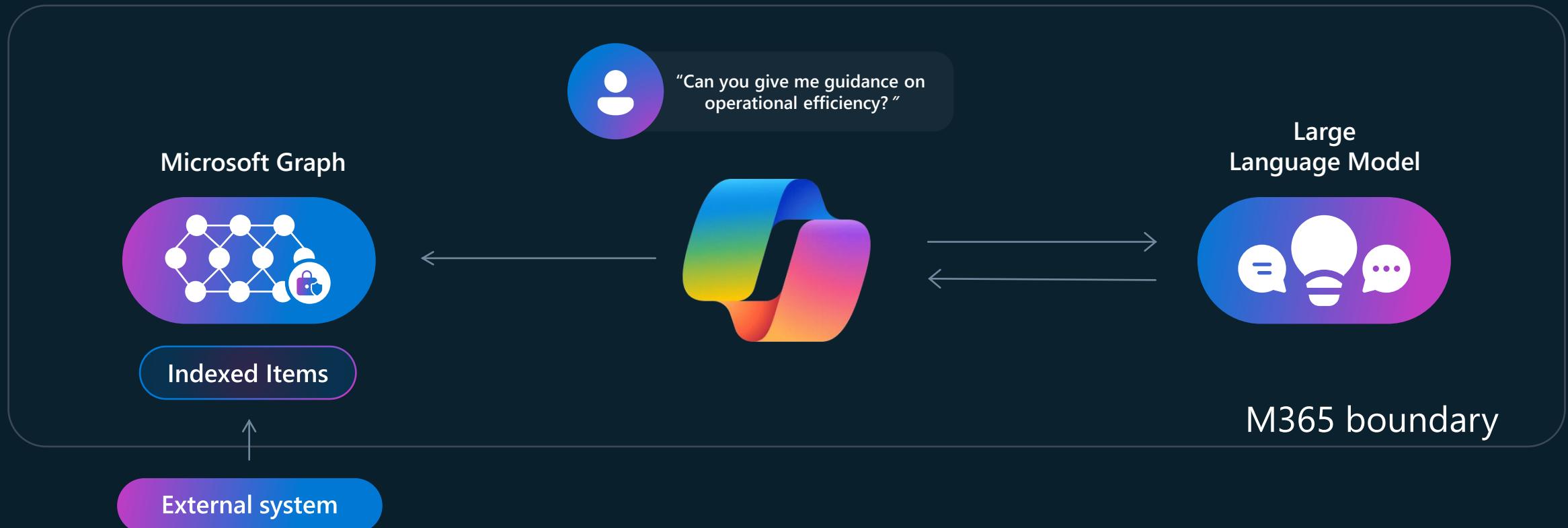


Your copilots

- ✓ Your business knowledge
- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

# What are Graph Connectors

Microsoft Graph APIs to index items from other systems



# When to use Graph Connectors



Unstructured  
data



Long form text  
that LLM can  
summarize



Unpredictable  
prompts and need  
to rely on semantic  
matching



Low activity  
updated items



## Microsoft Copilot Studio

### Knowledge through Copilot

#### connectors

- SharePoint/OneDrive sites or files
- File uploads
- Dataverse (incl.)
- Dynamics 365 data + Power Apps data)
- Microsoft Fabric (Coming soon)
- Other external data sources indexed in Microsoft Graph

\* Certain sources are in public preview only

Add available knowledge sources (Powered by Copilot connectors)

Users with edit permissions for this copilot can also reuse your connections for other topics within the copilot. [Manage security settings](#)

Featured

Add existing knowledge  
→ Converts previous data sources to knowledge for this copilot

Public website  
Incorporate any relevant web content found on Bing

Files  
Upload documents from your local computer

SharePoint and OneDrive  
Securely integrate and manage internal data

Dataverse  
Customize and deploy structured data tables

Microsoft Fabric  
Accelerate data analysis with AI capabilities

Bring your enterprise data (16)

<p> Enterprise website (preview)</p>	<p> Azure DevOps (preview)</p>	<p> Custom connector (preview)</p>	<p> Jira (preview)</p>
<p> ADLS Gen2 (preview)</p>	<p> Oracle SQL database (preview)</p>	<p> ServiceNow (preview)</p>	<p> File share (preview)</p>
<p> CSV (preview)</p>	<p> Microsoft SQL (preview)</p>	<p> MediaWiki (preview)</p>	<p> Salesforce (preview)</p>
<p> Confluence (preview)</p>	<p> Azure SQL (preview)</p>	<p> Zendesk (preview)</p>	<p> Power Platform connector</p>

# Copilot extensions

Ground responses



## Connectors

- ✓ Your business knowledge

Bring actions



## Plugins

- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

Focus user experiences

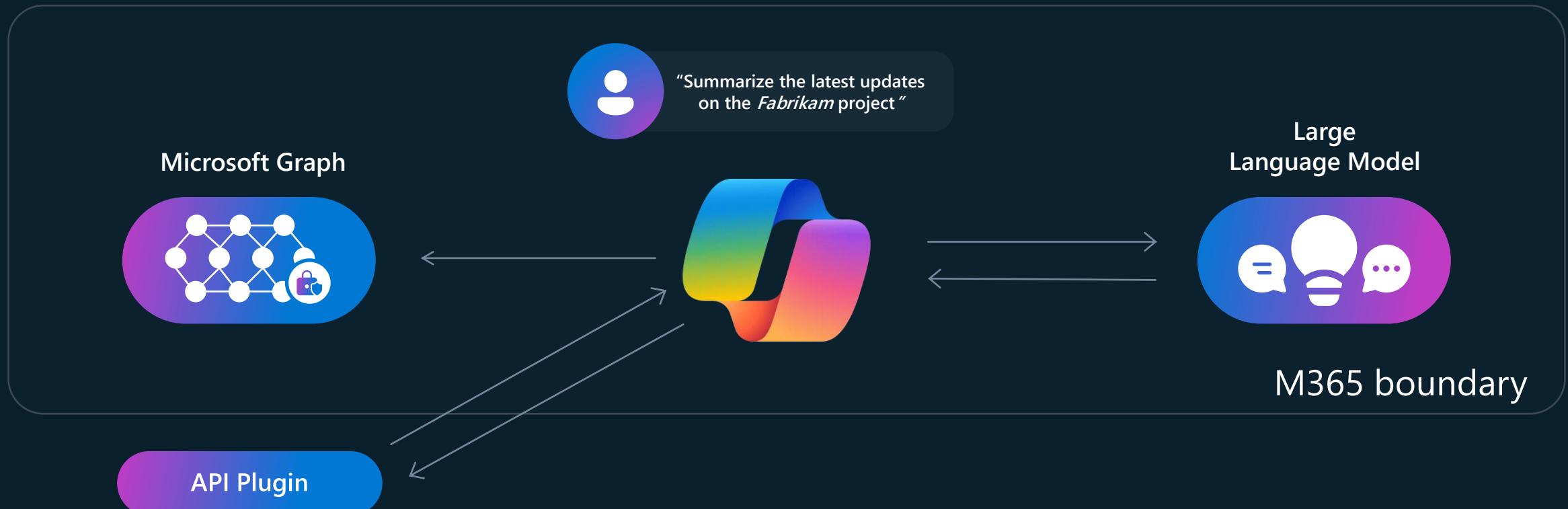


## Your copilots

- ✓ Your business knowledge
- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

# What are API Plugins

Manifest file describing an Open API document



# When to use Plugins



Real-time data



Highly relational data



Large scale data set  
(5 mil+)



More precise prompts  
predicted



Read and write actions



Already have an API  
to plug into

# Building plugins for Copilot



## Copilot Studio

- ✓ Familiar with Low-code SaaS based tooling
- ✓ Have existing Power Platform solutions
- ✓ Rapid development environment
- ✓ Managed infrastructure



## Teams Toolkit

- ✓ Familiar with managed code
- ✓ Have existing Teams apps
- ✓ Multi-tenant deployment
- ✓ Source control, branching, release management practices
- ✓ Need to manage hosting infrastructure yourself

# Copilot extensions

Ground responses



## Connectors

- ✓ Your business knowledge

Bring actions



## Plugins

- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

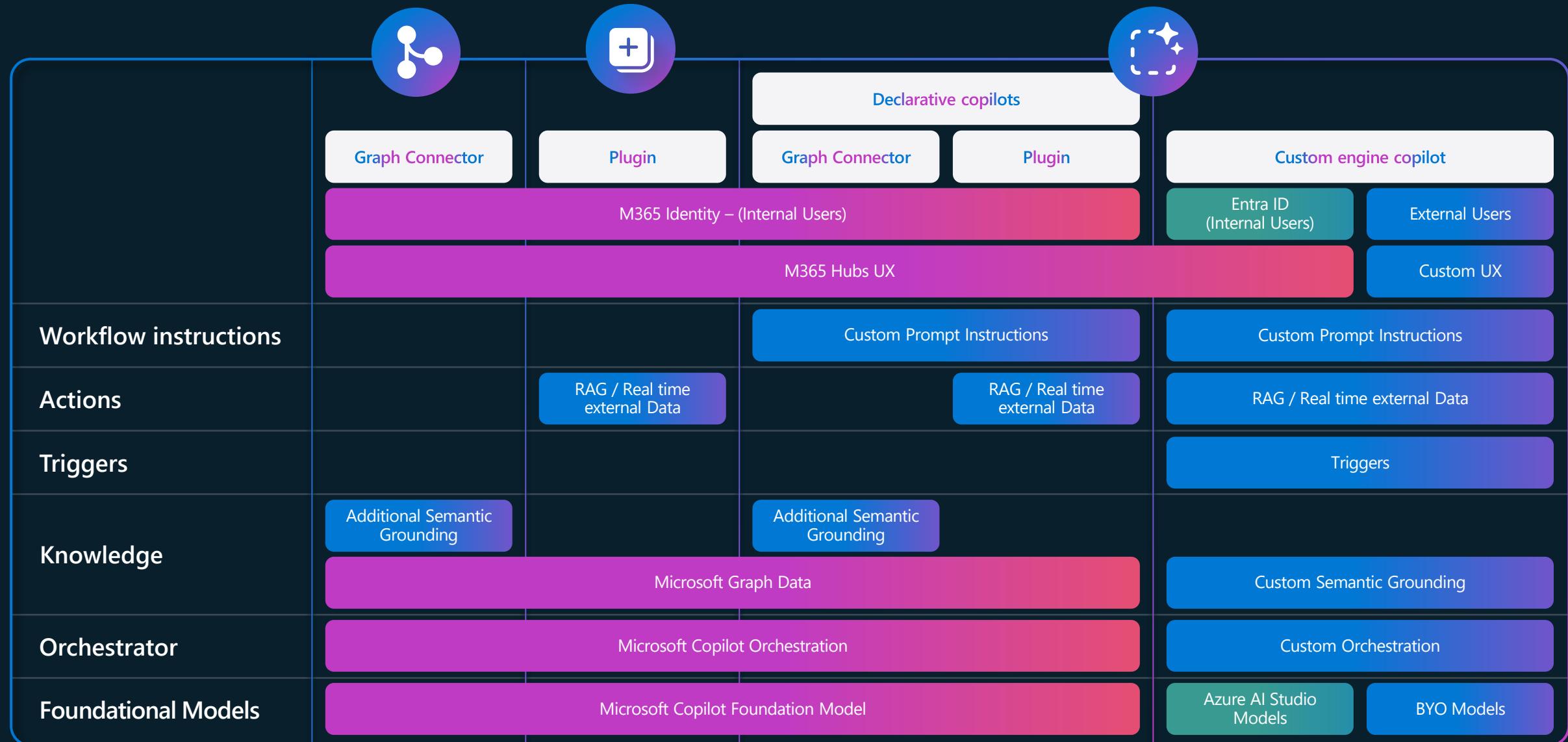
Focus user experiences



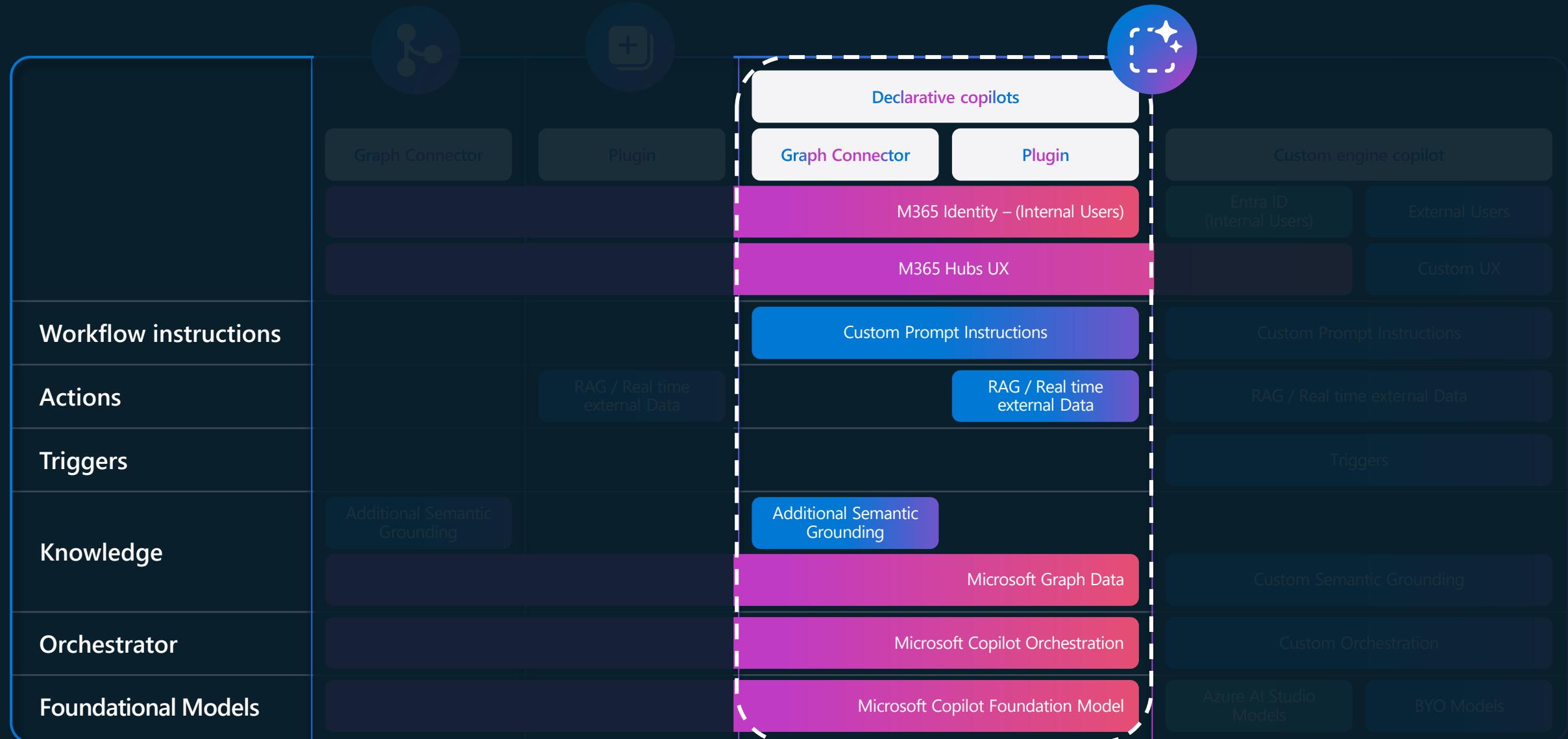
## Your copilots

- ✓ Your business knowledge
- ✓ Your apps
- ✓ Your workflows
- ✓ Your automation

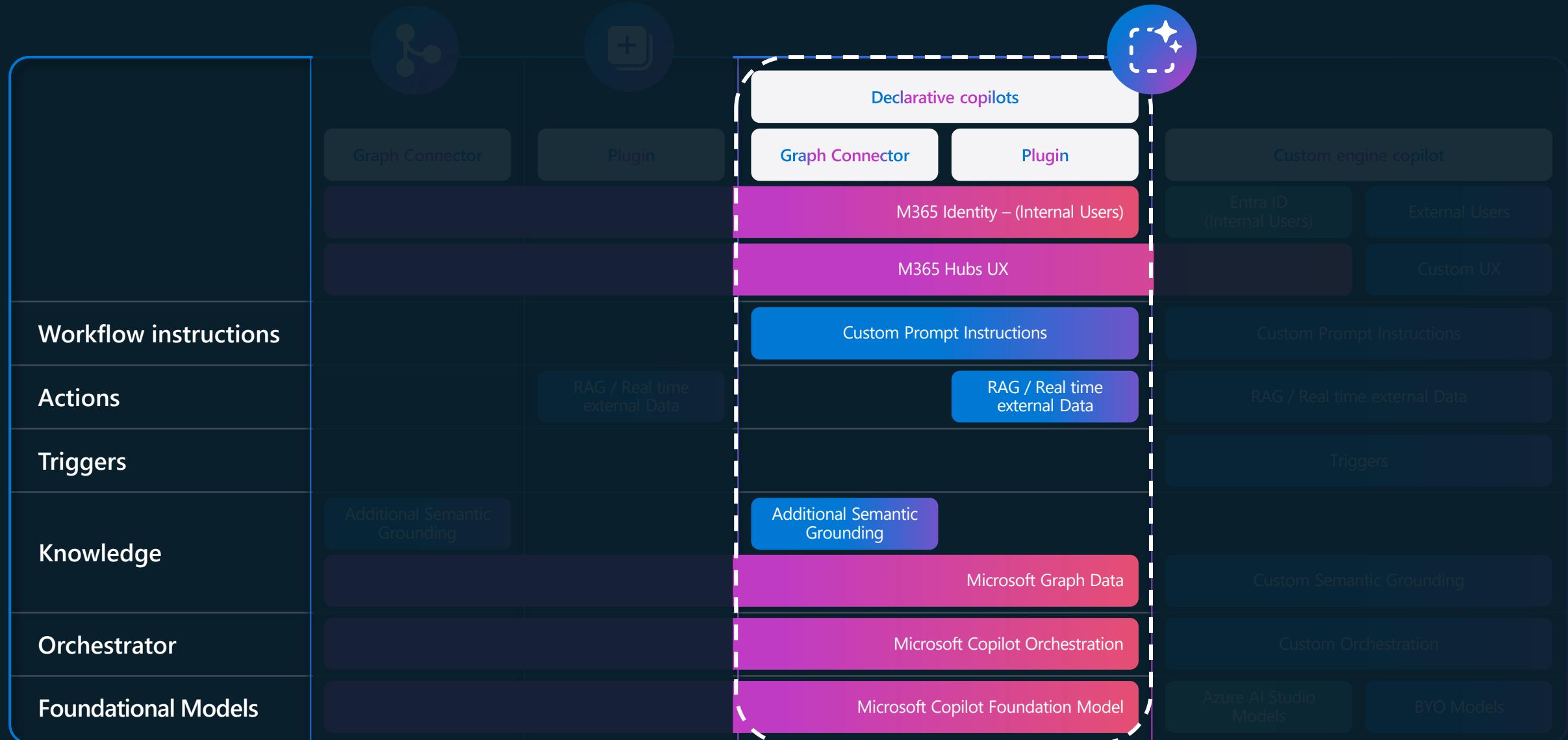
# Copilot extensions fundamentals



# Copilot extensions fundamentals

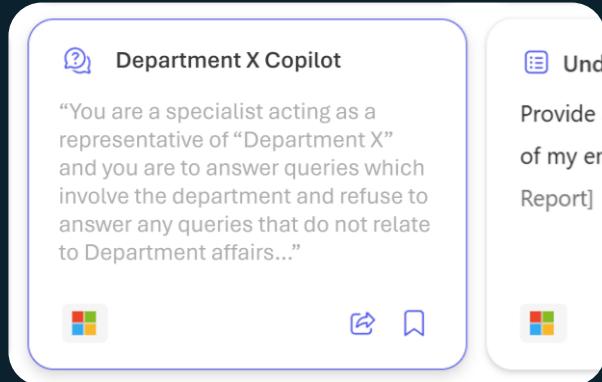


# Copilot extensions fundamentals



# What is additional **Context** on top of Microsoft Copilot?

## *(declarative) Custom Instructions*



**Custom Prompt  
Instructions**  
(Natural Language)



**Additional Grounding  
/ Focus Data**



**Data Sources  
(On-demand Retrieval)**



## Delivery Drone Launch

Home Documents Pages Site contents Edit

Confidential \ Internal only ★ Not following English

Home

+ New

Upload

Edit in grid view

Sync

Create a copilot

...

All Documents



Documents

Pages

Site contents

Site contents

Recycle in

Edit

Documents &gt; Final Materials



Name

Modified

Modified By



Statement of Work

A few seconds ago

Sam Snow



Delivery Drone Press Release.docx

Wednesday at 8:30 PM

Kat Larson



FAQs.docx

Monday at 3:31 PM

Mona Kane



List of Attendees.xlsx

Monday at 4:30 PM

Kat Larson



Upselling Opportunities.docx

Yesterday at 2:30 PM

Megan Brown



Vendor List.docx

Yesterday at 2:30 PM

Megan Brown

EAP (Private Preview)

# Create custom copilots from SharePoint

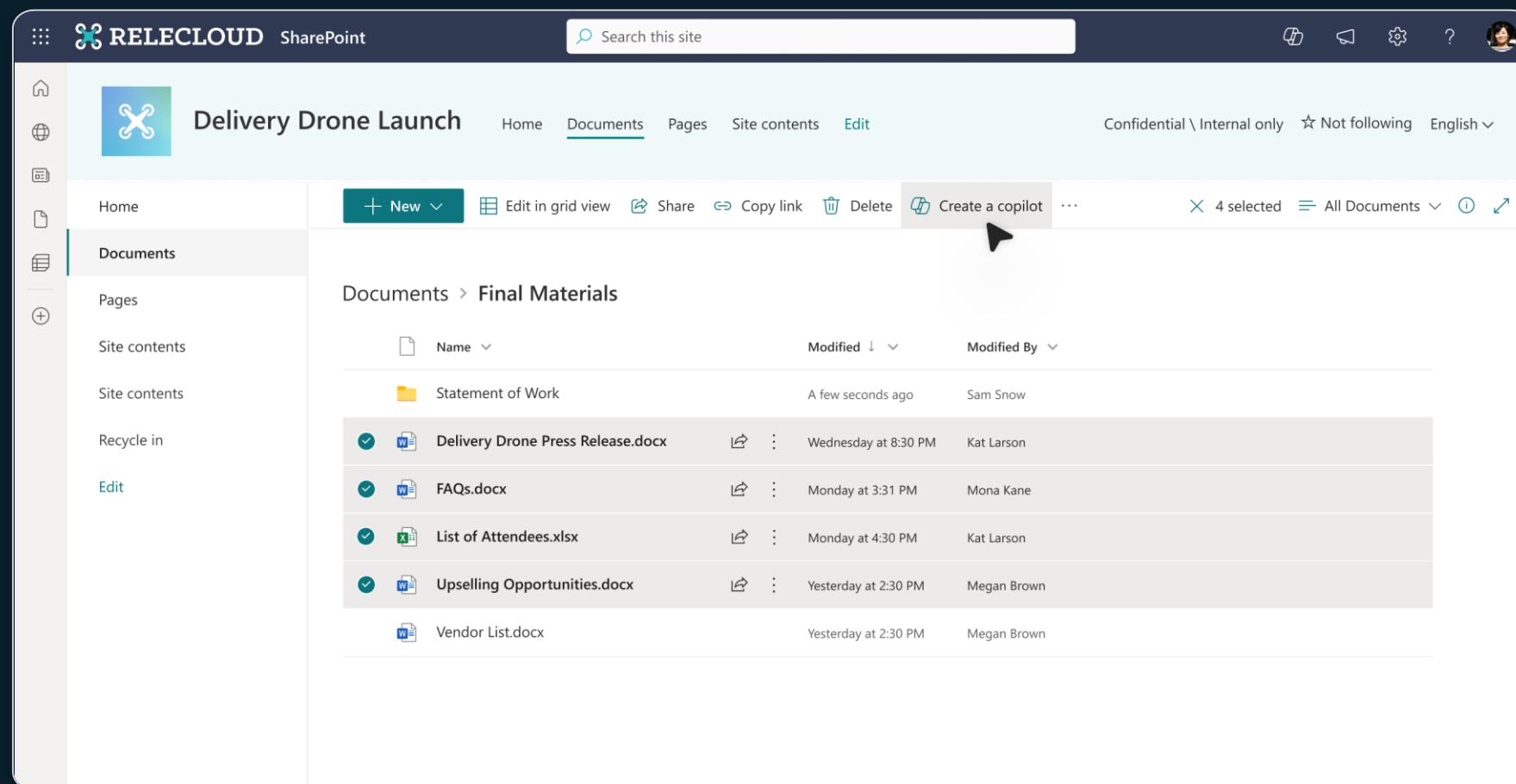
Put your content to work quickly by creating and sharing custom copilots grounded on the exact data your team needs

Grounded in your authoritative content

Any user can create, use, and share in a few clicks

Follows existing user access permissions

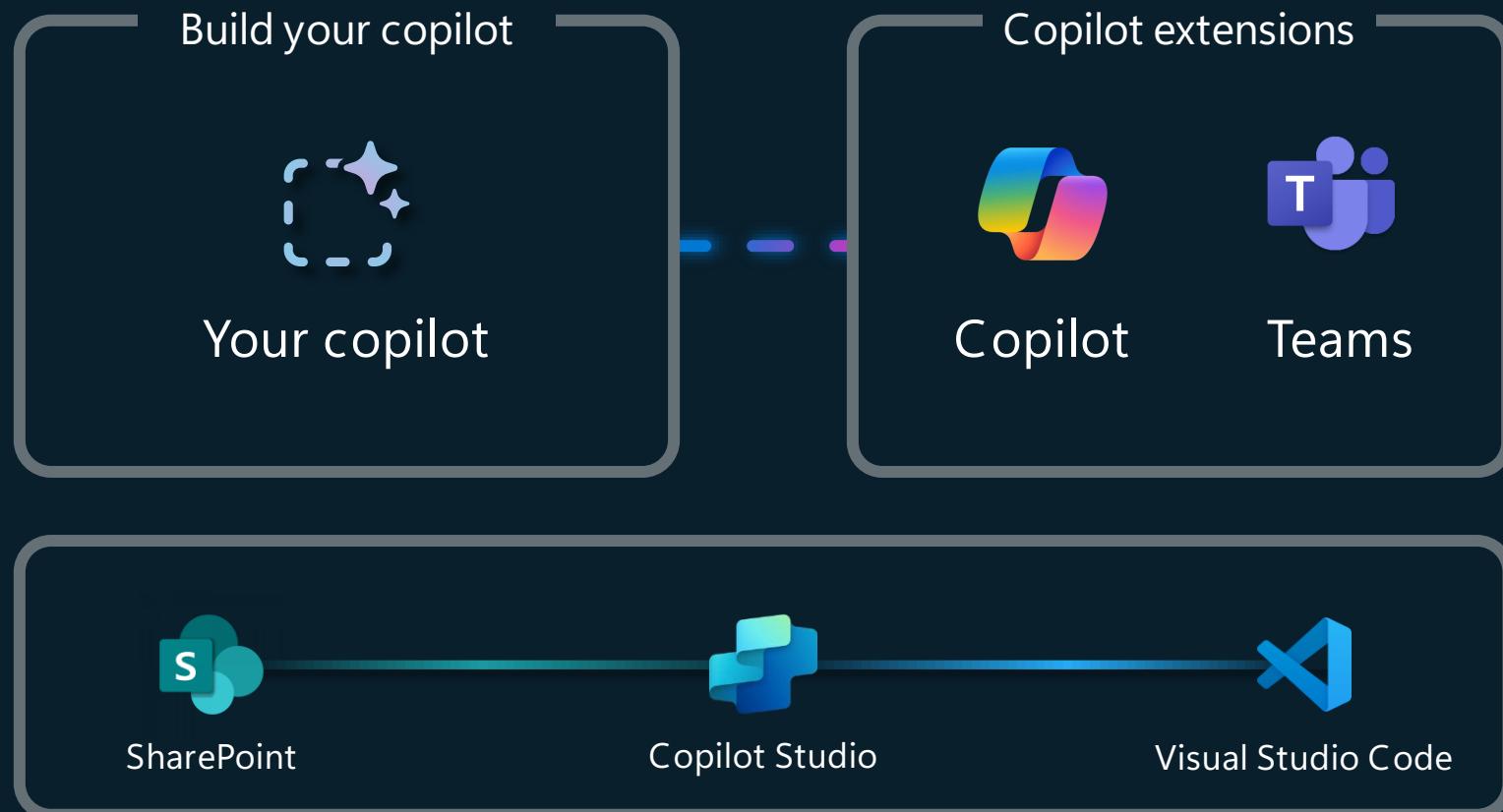
Edit and enhance with Copilot Studio integration



The screenshot shows a SharePoint 'Delivery Drone Launch' site. The left navigation bar has 'Documents' selected. The main content area displays a list of documents in the 'Final Materials' folder. The top ribbon menu includes 'Home', 'Documents' (which is underlined), 'Pages', 'Site contents', and 'Edit'. The toolbar below the ribbon includes 'New', 'Edit in grid view', 'Share', 'Copy link', 'Delete', and the 'Create a copilot' button, which is highlighted with a large black arrow pointing to it. The document list shows the following items:

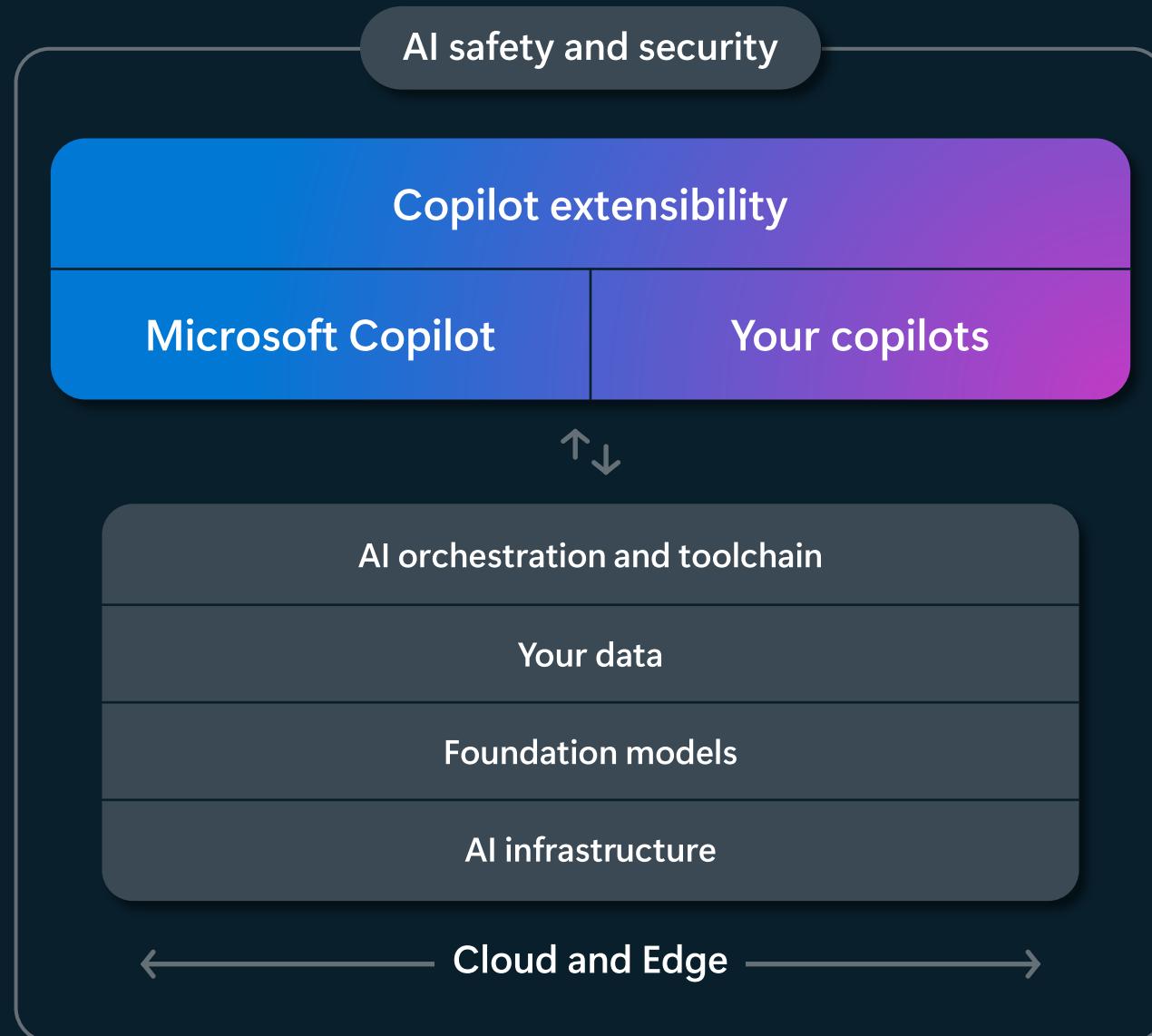
Name	Modified	Modified By
Statement of Work	A few seconds ago	Sam Snow
Delivery Drone Press Release.docx	Wednesday at 8:30 PM	Kat Larson
FAQs.docx	Monday at 3:31 PM	Mona Kane
List of Attendees.xlsx	Monday at 4:30 PM	Kat Larson
Upselling Opportunities.docx	Yesterday at 2:30 PM	Megan Brown
Vendor List.docx	Yesterday at 2:30 PM	Megan Brown

# Make your copilots work as Copilot extensions



# Kidd Wong

Partner Solution Architect – Business Application



# Power Platform



# Build Disclosures

## Power Platform

### Updates in Copilot Studio

- Copilots with agent capabilities Limited private preview
- Copilots to extend Copilot for M365 Private preview
- Knowledge through copilot connectors Generally Available

### Updates in Dataverse

- Copilot Connectors Public Preview
- Security Hub Public Preview
- New Prompt Building Public Preview
- Data workspace in Power Apps Public Preview

### Updates in Power Automate

- AI flows Early access preview
- AI recording for desktop flows Early access preview
- Automation center & troubleshooting Public preview
- Personal Automation recommendations Public preview
- Conversational cloud flow Public preview
- Natural language to desktop flow Early access preview

## Power Platform

### Updates in Power Apps

- Coauthoring in Canvas Apps Public Preview
- Source code in Canvas Apps Public Preview
- Native Git integration Public Preview (July)
- Copilot in mobile – voice dictation Public preview (July)
- Copilot hub Early access preview

### Updates in Power Pages

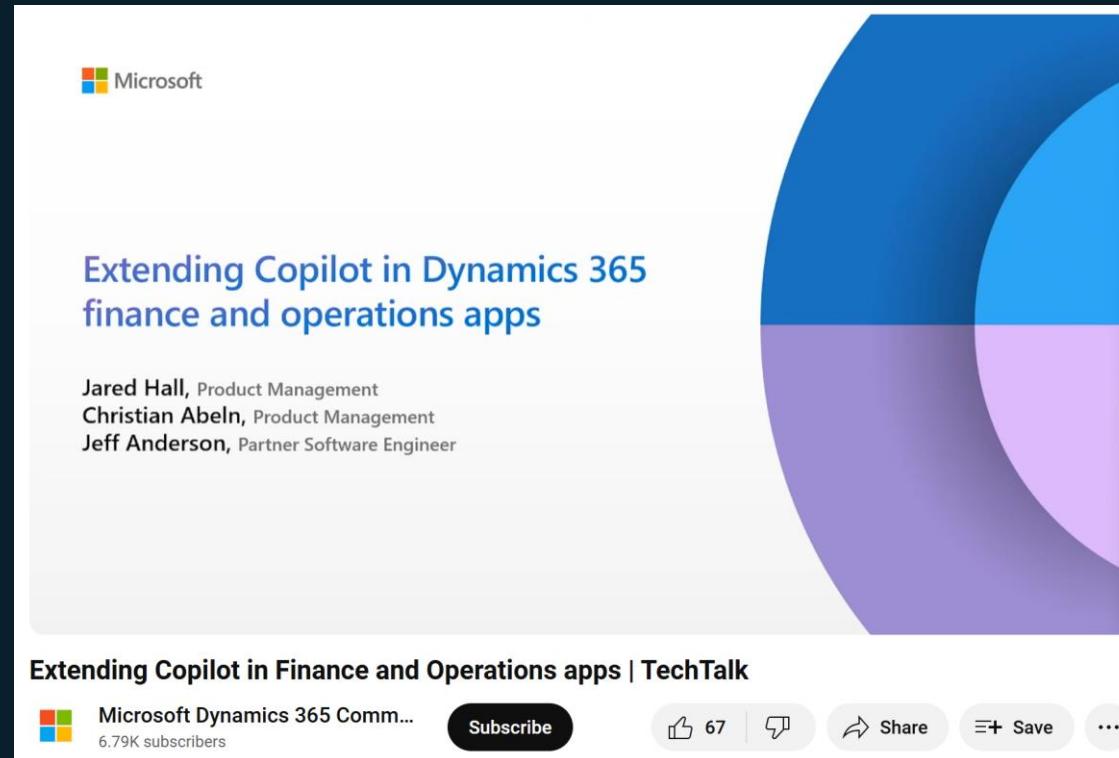
- Chatbots on your website with Microsoft Copilot GA
- Enhanced search with Microsoft Copilot Public Preview
- Security workspace Public Preview
- Built-in web application firewall GA
- Web scan GA
- Use Salesforce, Oracle, and Fabric with Power Pages Public Preview
- SAP supplier self service template Public preview
- Power Fx support in Power Pages Public Preview
- New UX controls Public Preview

# Copilot Studio



# Copilot extensions

Extending Copilot in Finance and Operations apps | TechTalk (youtube.com)



Limited Private  
Preview

# Build copilots with agent capabilities

Complete tasks independently with a copilot tailored to a specific roles or functions

Automates long-running processes

Reasons over actions and user inputs

Leverages memory and knowledge

Learns and improves

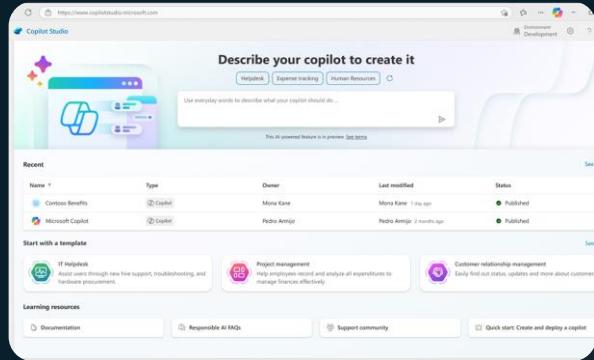
Follows human guardrails and asks for help

The screenshot shows the Copilot Studio interface for the 'IT Helpdesk' copilot. On the left, there's a sidebar with 'Copilot Studio' navigation and a 'Copilots' section. The main area has tabs for 'Overview', 'Knowledge', 'History', and 'Analytics'. A central callout box says 'Your copilot is ready! Here's what's next:' with steps: 'Add instructions to refine and improve how it responds', 'Add actions so your copilot can perform specific tasks', 'Test your copilot to ensure its comprehension', and 'Publish your copilot so others can use it'. Below this is an 'Overview' section with a 'Goal' field containing 'Assist users through new hire support, troubleshooting, and hardware procurement.' To the right, there are sections for 'Instructions' (listing actions like sending email requests and creating records), 'Actions' (listing specific tasks like 'Send email request' and 'Create record'), and a 'Test your copilot' section with a message from AI about a learned response and a button to 'Run a simulated test'. At the bottom, there are 'Teach' and 'Test' buttons.

# Copilots with agent capabilities - The new 'app' in a gen AI world

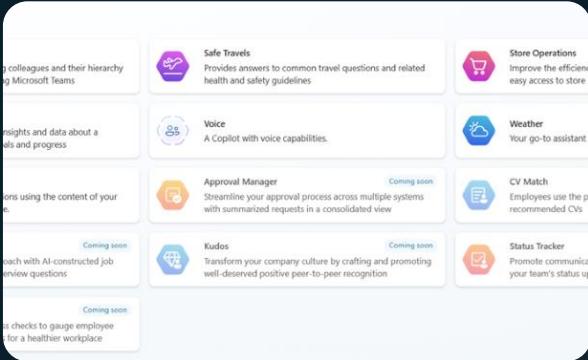
Problem	Unique business processes, industries, circumstances, and customers		Complex processes, form-heavy apps, labor intensive systems, and high setup and maintenance costs			
Build a copilot to handle this variability and complexity at infinite scale						
Solution						
Use cases	Automates long-running processes	Reasons over actions and user inputs	Leverages memory and knowledge	Learns and improves		Follows human guardrails and asks for help
How	Two technology advancements	<ol style="list-style-type: none"><li>1. New agent architectures to control and manage LLM-based orchestration with clear guardrails</li><li>2. Ability to fine tune Action models for specific tasks.</li></ol>				
GTM	What does it mean to me as a seller?	<ul style="list-style-type: none"><li>Agent capabilities are only for standalone copilots that low code sellers pitch.</li><li>The features are limited private preview and in public preview later in the calendar year.</li><li>Continue to sell the existing copilots you know today.</li><li>Customers can try agent-like chaining capabilities in generative actions now.</li></ul>				

# How it works



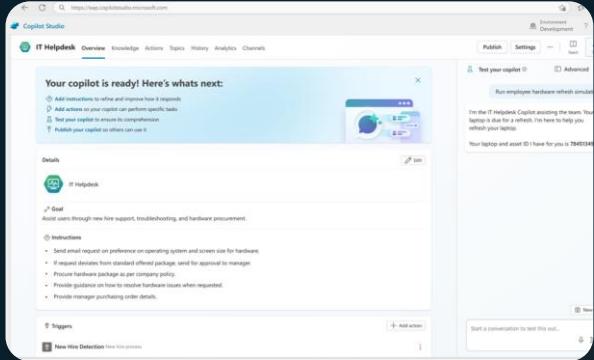
## Create

Build conversationally and see how the copilot independently orchestrates



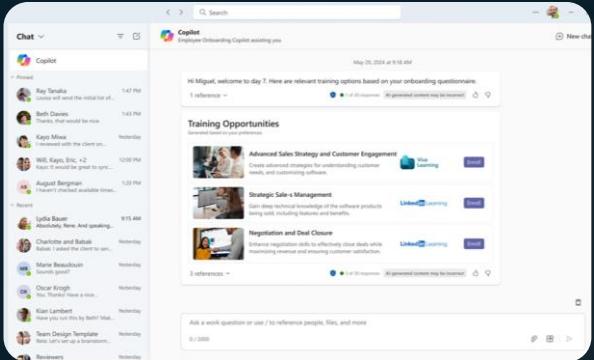
## Templates

Start quickly for prebuilt scenarios



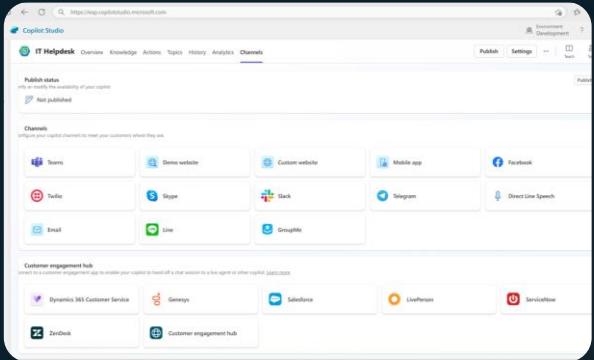
## Test

Test your copilot directly in the platform



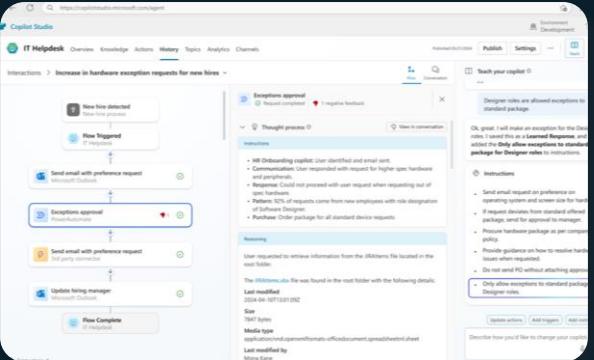
## Memory

Your copilot will remember and recall context across sessions



## Publish

Publish across several channels including Copilot for M365, Outlook, and Teams

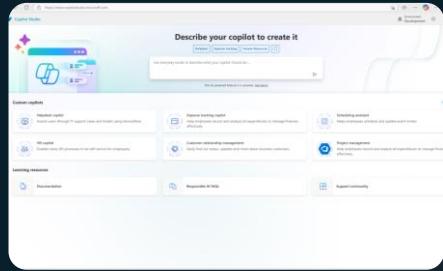


## Train

Identify learning opportunities and fine-tune behavior

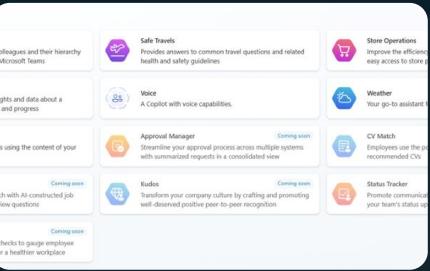
# Check out what else is coming to Copilot Studio

## Generally available



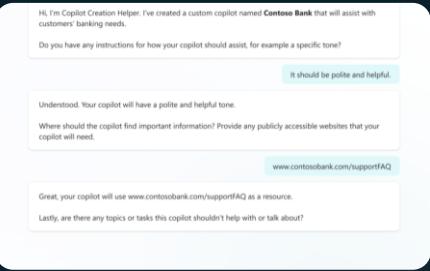
UX refresh  
Modern and fluent maker experience

## Public preview



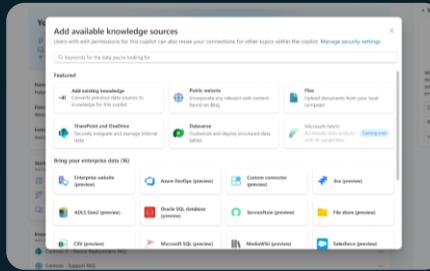
Templates  
Start quickly for prebuilt scenarios

## Public preview



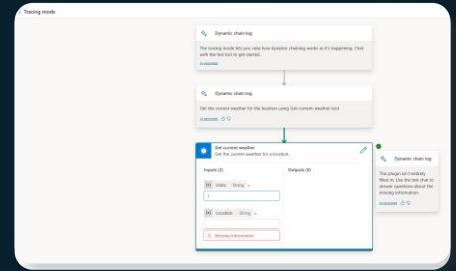
Build conversationally  
Configure your copilot with a chat experience

## Generally available



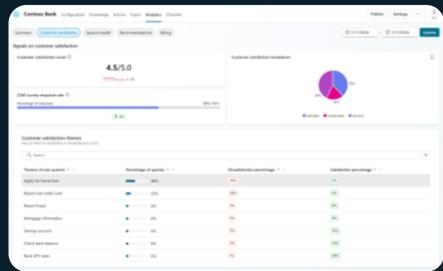
Copilot connectors  
Ground your copilot in your data sources

## Public preview



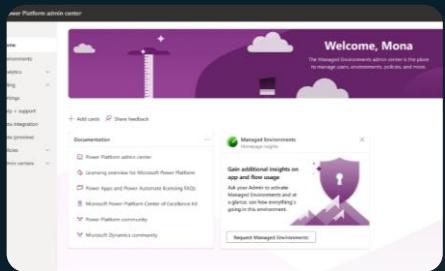
Generative Actions enhancements  
Improvement in behavior and accuracy using Azure OpenAI GPT4 support

## Public preview



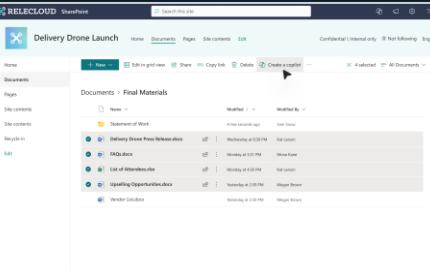
Conversational analytics  
Richer KPIs for copilot performance

## Generally available



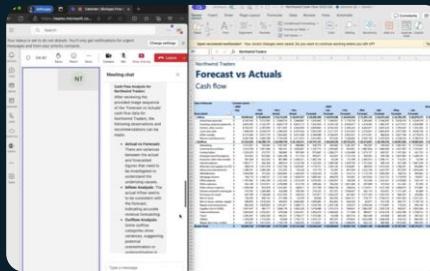
Security enhancements  
Admin defaults, Audit logs, and sensitivity labels

## Coming soon



Extend copilots in SharePoint  
Customize with additional knowledge and actions with Copilot Studio

## Coming soon



Multi-modal support  
Add a copilot to a meeting and get an analysis of the content in a shared screen

## Coming soon

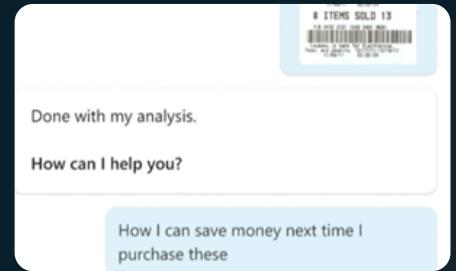
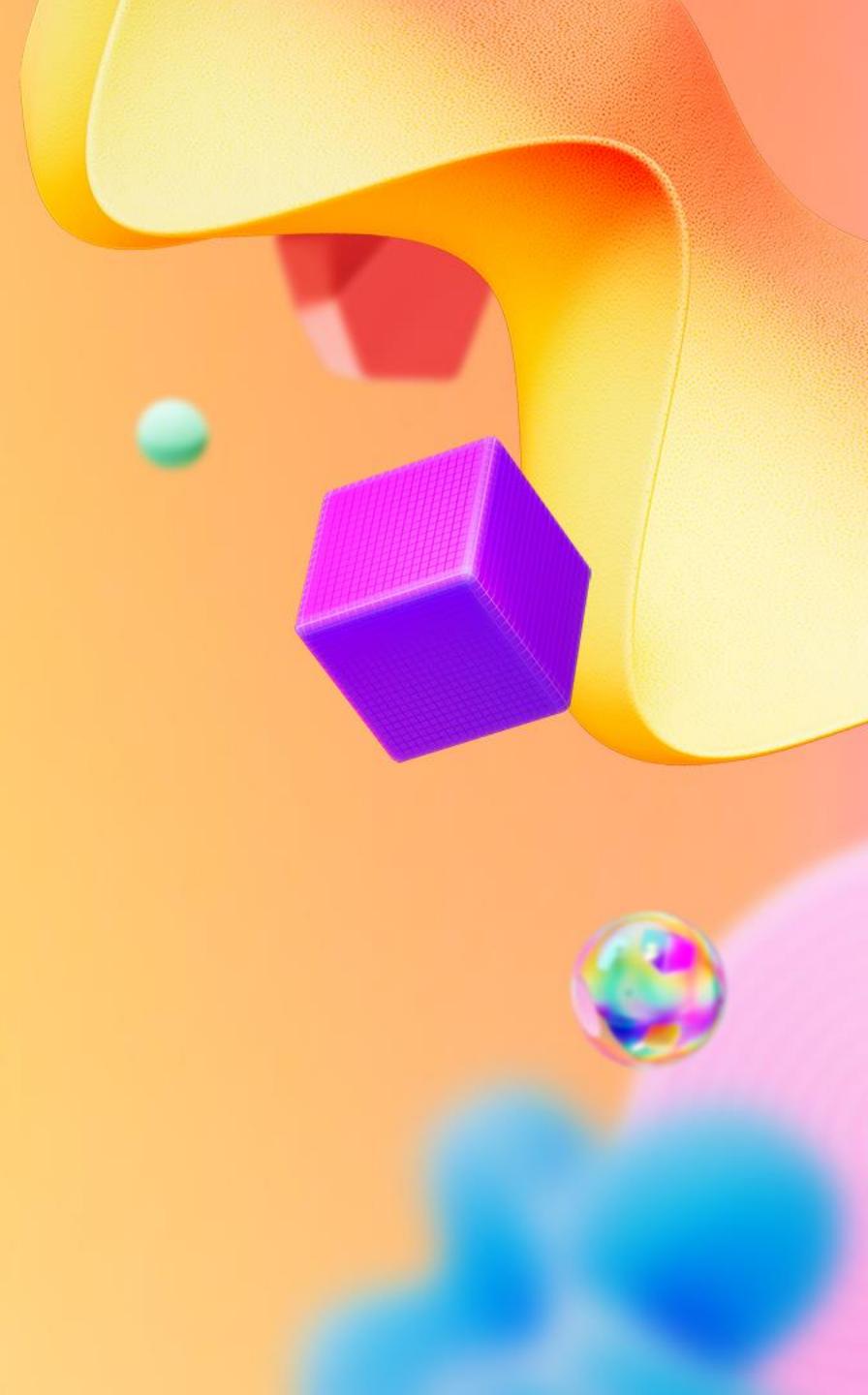


Image input support  
End users can add images into the chat

# Microsoft Dataverse



## First-class Integration



Microsoft Dataverse



Microsoft Copilot Studio

Work with **structured and unstructured data** including file uploads, all **secured** in Dataverse

**Include your Dataverse tables** as knowledge for richer, grounded data experiences

**APIs can become actions** to enable copilots actionable over any business process

### Step 1 of 3: Choose an action

Create an action or browse through our list of actions you want to use to get information from external sources.

[Learn more](#)

#### Discover an action

Search for flows, skill actions, and commonly used connector actions



#### Popular in your org

Connectors Custom Connectors Flows Skills Dataverse

**Step 1 of 3: Select Dataverse tables**  
Select up to 15 tables.

All Selected (0)

Keywords for the data you're looking for

Recommended by Copilot

Travel Feedback  
TravelFeedback

Available tables

Account  
Action Card  
Activity

Back Next Cancel

AI-generated content may be incorrect



# Knowledge in Microsoft Copilot Studio

**Get started** with enterprise knowledge in Microsoft Copilot Studio for contextually relevant responses.

**Improve reliability and accuracy** of copilot responses with specific knowledge sources.

Support a **wide range of data types**, including Dynamics 365 and public websites, Graph, and Microsoft Fabric

**Enhance** business processes and productivity with grounded enterprise data.

Provide **rich customer experiences** with added enterprise data.

Public Preview

Add available knowledge sources (Powered by Copilot connectors)

Users with edit permissions for this copilot can also reuse your connections for other topics within the copilot. [Manage security settings](#)

Keywords for the data you're looking for

Featured

- Add existing knowledge → Converts previous data sources to knowledge for this copilot
- Public website Incorporate any relevant web content found on Bing
- Files Upload documents from your local computer
- SharePoint and OneDrive Securely integrate and manage internal data
- Dataverse Customize and deploy structured data tables
- Microsoft Fabric Accelerate data analysis with AI capabilities

Bring your enterprise data (16)

- Enterprise website (preview)
- Azure DevOps (preview)
- Custom connector (preview)
- Jira (preview)
- ADLS Gen2 (preview)
- Oracle SQL database (preview)
- ServiceNow (preview)
- File share (preview)
- CSV (preview)
- Microsoft SQL (preview)
- MediaWiki (preview)
- Salesforce (preview)
- Confluence (preview)
- Azure SQL (preview)
- Zendesk (preview)
- Power Platform connector

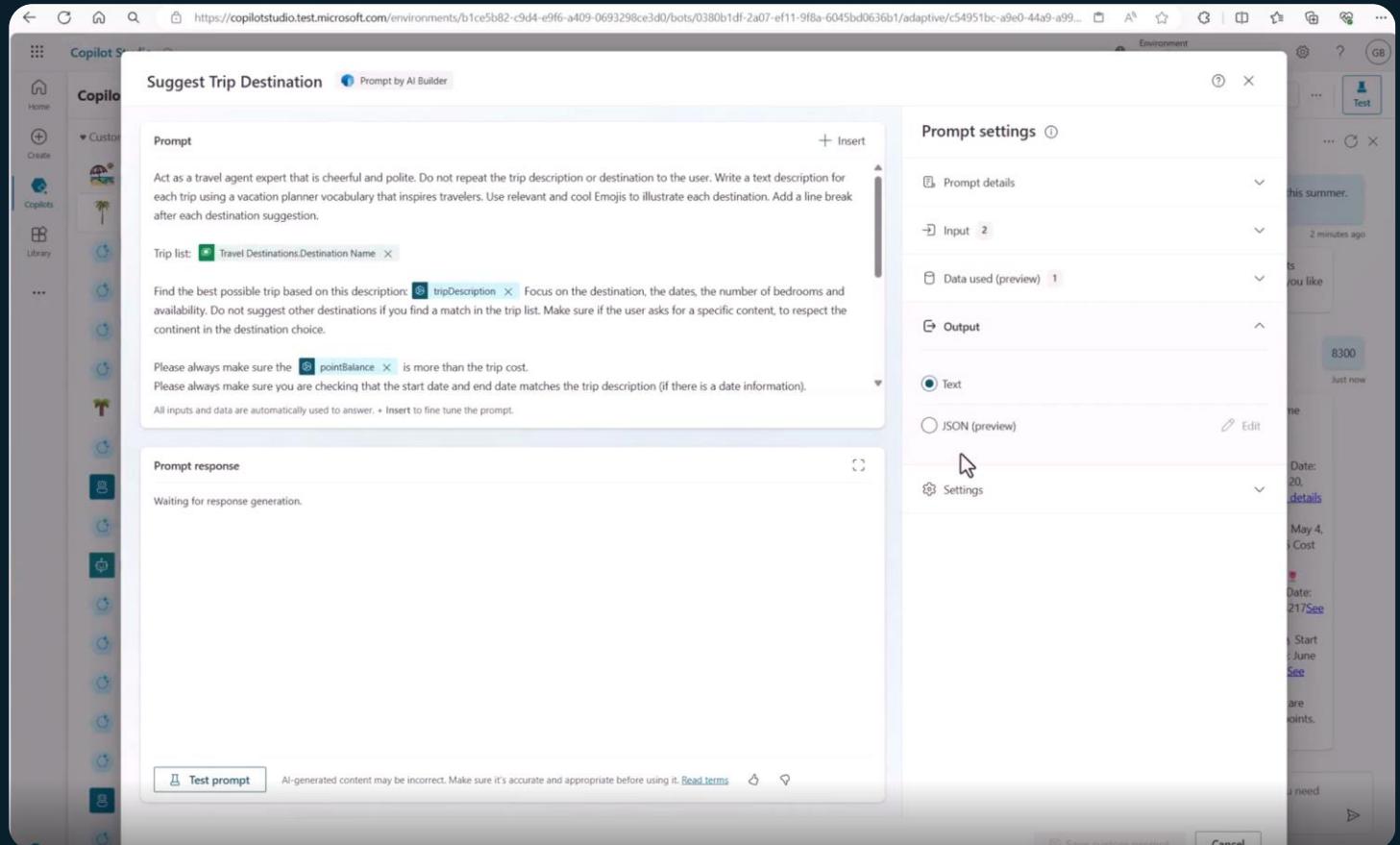
# New Prompt Building Experience

Public Preview

**Create** prompts easily using data, prompts, and topics

**Customize** with dynamic inputs variables, craft custom instructions, and link to custom data sources

**Share** and reuse components across Power Platform with data grounded in **Microsoft Dataverse**



# Power Apps



# Data Workspace in Power Apps

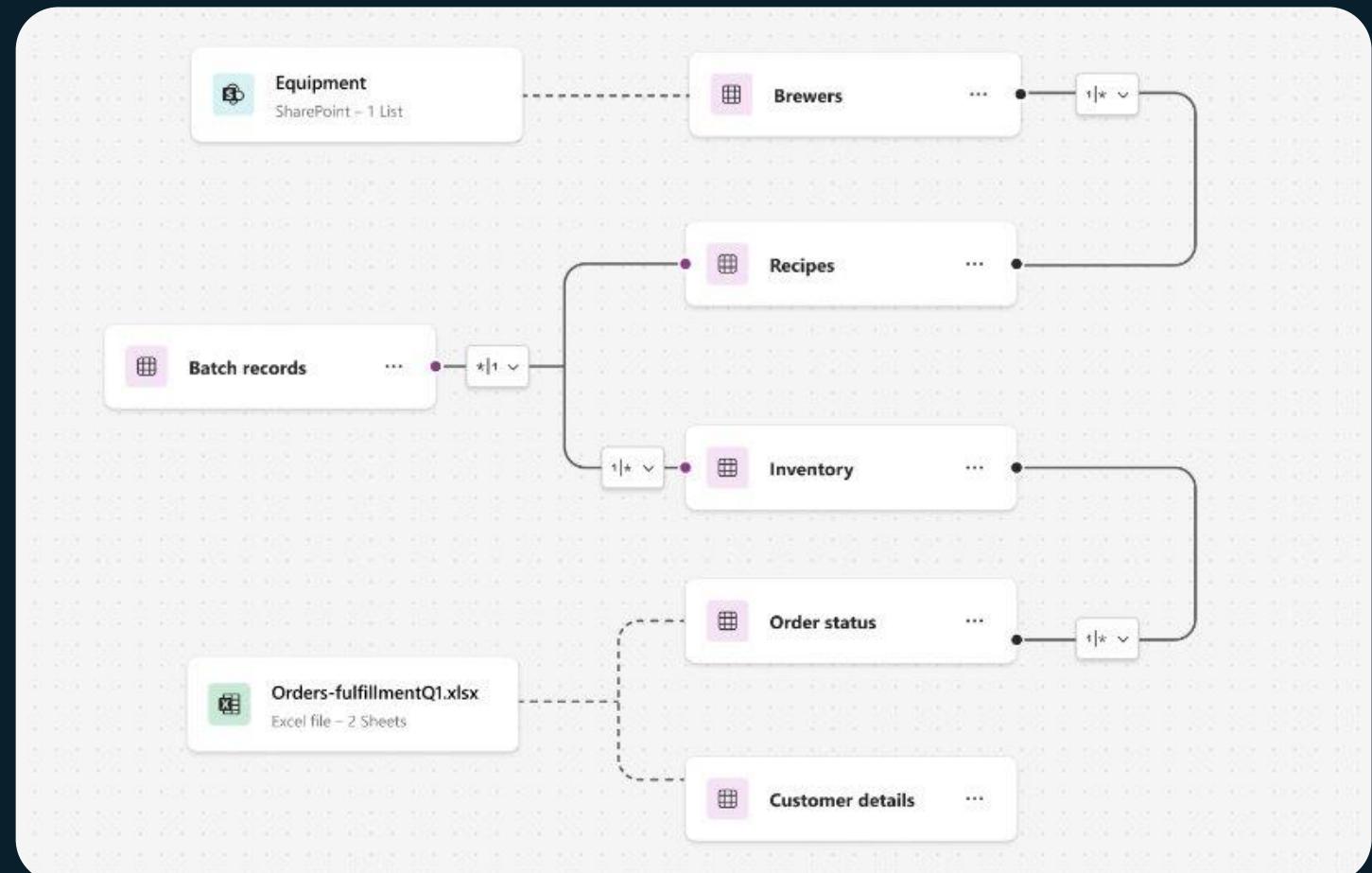
Public Preview

Leverage **Copilot** for multi-table and relationship creation

See schema view and tabular view side by side

Create multi-table app using modern app template

Create tables from scratch, conversation, Excel and SharePoint in one place



# Coauthoring in Canvas Apps



Edit apps **together** in real-time

Edit apps with up to **10 coauthors simultaneously**

Public Preview

A screenshot of the Microsoft Power Apps canvas editor interface. The title bar says "Power Apps | Kudos App (Editing)". The left sidebar shows a "Tree view" with "Screens" selected, listing "App", "Landing Screen", "Tutorial Screen", "Badge Selection Screen", "Main Page Screen", "View Kudos Screen", "Details Screen", "Message Screen", "Submission Screen", "Error Screen", "Success Screen", and "Settings Screen". The main workspace displays a "Landing Screen" card with a pink border. Inside the card, there is a Microsoft logo at the top, followed by the text "Employee Kudos" and three buttons labeled "Get Started", "Text", and "Text". On the right side, there is a "Properties" panel for the "Landing Screen". The "Display" tab is selected, showing settings for "Fill" (set to "ColorValue(varThemeColors.background)"), "Background image" (set to "None"), and "Image position" (set to "Fit"). The status bar at the bottom indicates "Editing" mode.



# View and edit source code in canvas apps

**Open source code** in Power Apps Studio in a YAML + Power Fx format

**Copy and paste code** to and from Power Apps Studio

**Share and reuse** code snippets

Public Preview

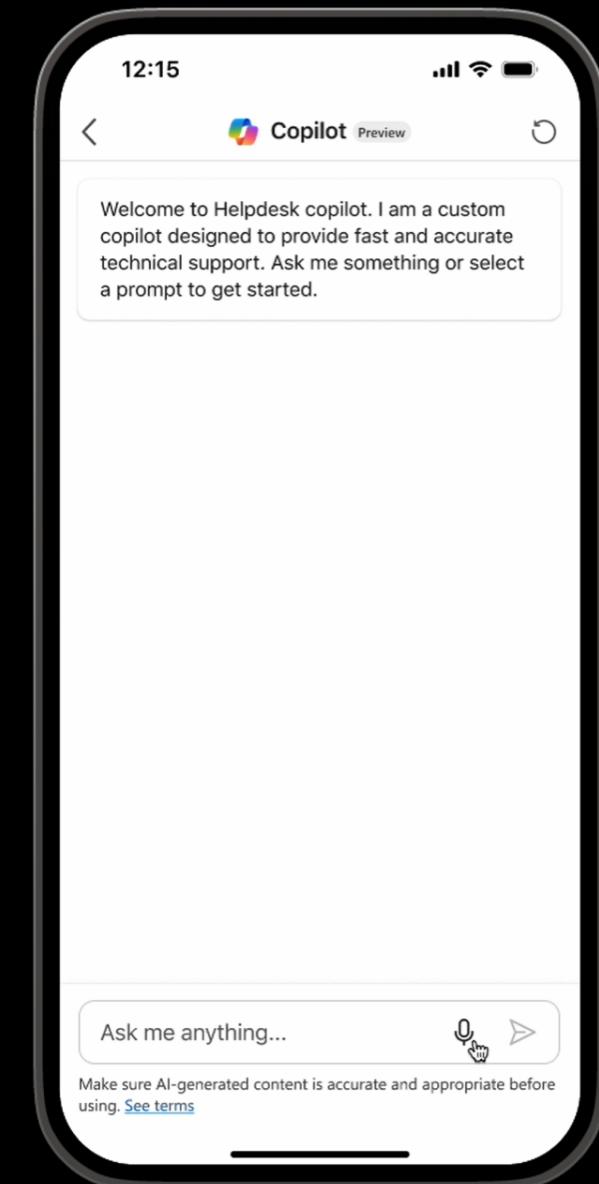
The screenshot displays the Power Apps Studio interface for editing a canvas app named "CanvasCode". The left side features a tree view of the app's structure, including screens and containers. The central area shows a "Main Screen" with a "Request approval" form. This form includes fields for Last Name, Full Name, Created By IP Address, Created By Username, Modified By IP Address, Modified By Username, Organization Name, and Time Zone. To the right of the form is a "Reviewers" section listing four individuals (A, p, N, N) with their names and titles. The right side of the screen contains the "Properties" pane, which is currently set to the "Header1" component under the "Display" tab, showing options for title, logo, and visibility. The overall interface is clean and modern, designed for easy navigation and modification of app components.

# Copilot in Mobile – Voice Dictation

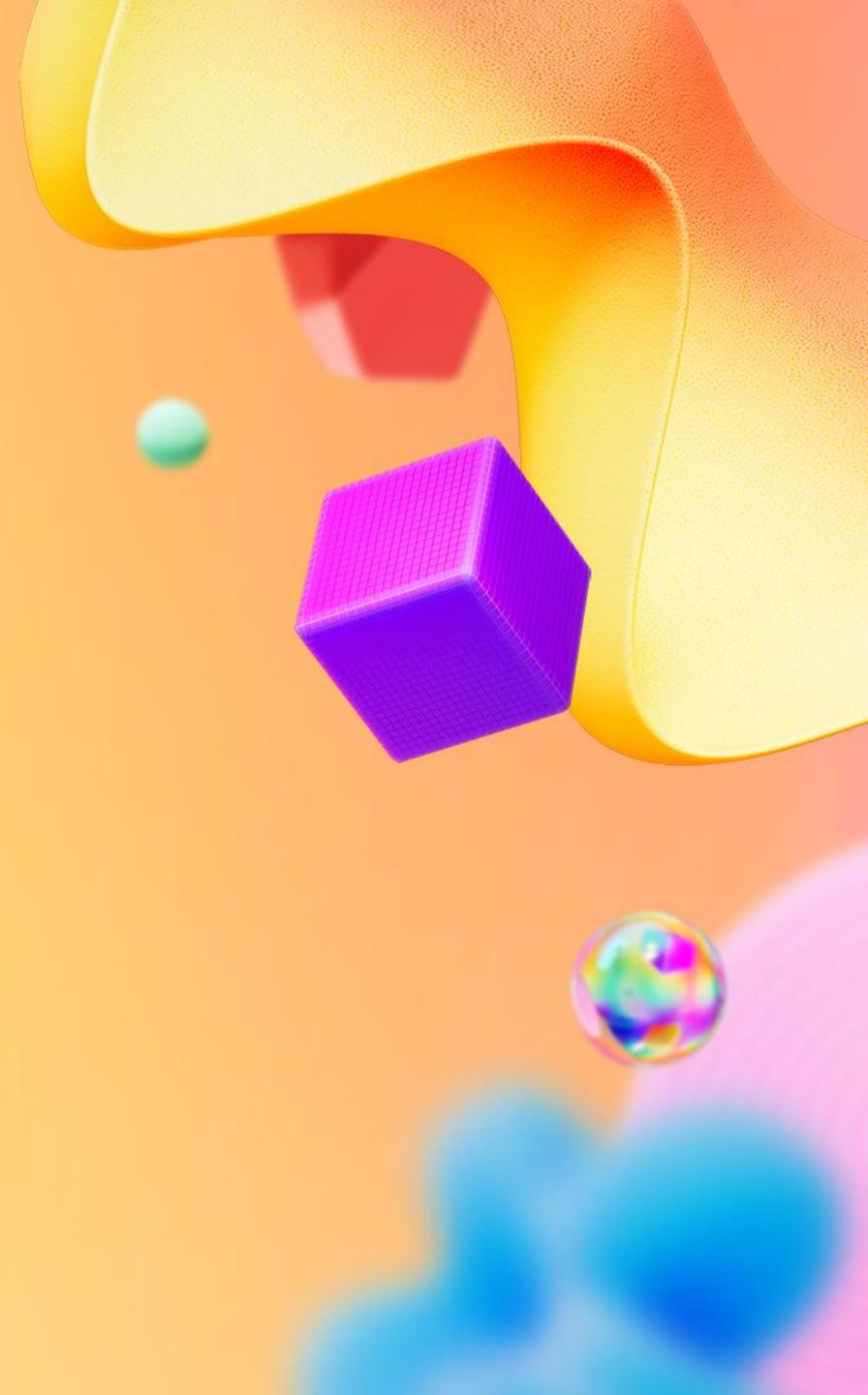
**Use voice input** to ask questions and get answers from Copilot

**Use Copilot on mobile to** complete tasks, collect data or submit information on the go

Public Preview: July



# Power Automate



NEW

Early Access Program



# AI flows

## Automation for the age of AI

Cloud flows

**Rules-based**  
cloud automation

AI flows

**Outcome based**  
generative AI powered  
automation

Desktop flows

**Rules based**  
UI automation

FOR

Modern systems with APIs

Simplifying complex processes,  
esp. with conditions, exceptions  
and loops

Inform flow based on  
unstructured data

Legacy systems without APIs

UI automation

NEW

Early Access Program



# AI flows

## Automation for the age of AI

intelligent | flexible | dynamic

A new approach to automation that is dynamic and **outcome-driven**.

User and AI collaboratively design **automation plan**, including parameters and guidelines.

**LLM reasoning, orchestration and execution** of automation.

Monitor flow success with run history, analytics, and individual flow run details.

The screenshot shows the Power Automate 'Design' interface for a 'Copilot workflow - Device returns approval'. On the left, there are three sections: 'Input' (Device name, Customer notes), 'Variables' (Customer returns team, Returns channel), and 'Output' (Approval request ID, Approval recommendation, Recommendation reasoning, Teams message ID, Approval decision). In the center, the 'Automation plan' section provides a summary of the workflow: 'Start an approval for a device return request and assign it to the Customer returns team. Recommend whether to approve the request by consulting Troubleshooting guide for device insights based on the Customer notes and Device name. If approved add a task to the Customer Returns plan with the device and approval request details. Summarize the request details, approval and the Planner task ID created and post a message to the Customer returns team in the Returns channel.' On the right, there are sections for 'Connectors' (Approvals, Teams), 'Guidelines' (Provide guidance that can influence Copilot Workflows as it executes your automation), 'Knowledge' (Troubleshooting guide), and 'Triggers' (None).

# When to use AI flows or Copilot Studio copilots with agent capabilities

## AI flows

A new approach to automate a workflow that enables users and AI to collaboratively design an automation plan, and let AI execute it.

### For single complex process automation

#### When to use it?

- Automate a single, defined end-to-end process
- Use unstructured data to inform the process and make it flexible
- Simplify a complex process, esp. those with complex conditions, process loops, and exception management.
- EG IDP for financial reporting process

#### Top capabilities that enable this

- Develop automation plan to achieve a process objective
- Generative AI dynamically selects steps when automation plan is executed to achieve a specific goal
- Manage and monitor single workflow with org guardrails

## Copilot with agent capabilities

These copilots can engage, reason and automate across a myriad of workflows, data sources, and events to achieve role level outcomes.

### For role-based orchestration requiring multiple workflows

#### When to use it?

- Need your copilot to take on a role / function that can act independently at times
- Can serve B2E, B2C and B2B scenarios
- Be triggered by conversation AND other events
- Copilot can work independently, and asks for user inputs to complete the workflow when required
- Processes that dynamically change and can be trained
- Repetitive use cases that can store memory and context from users.

#### Top capabilities that enable this

- Plan and orchestrate across multiple workflows
- Teach to improve accuracy and complexity
- Manage and monitor across workflows with org guardrails
- Memory for deep context and personalization

The future:

Better together: Trigger AI flows from within your copilots



NEW

Early Access Program



# AI recording for desktop flows

## A new way to automate

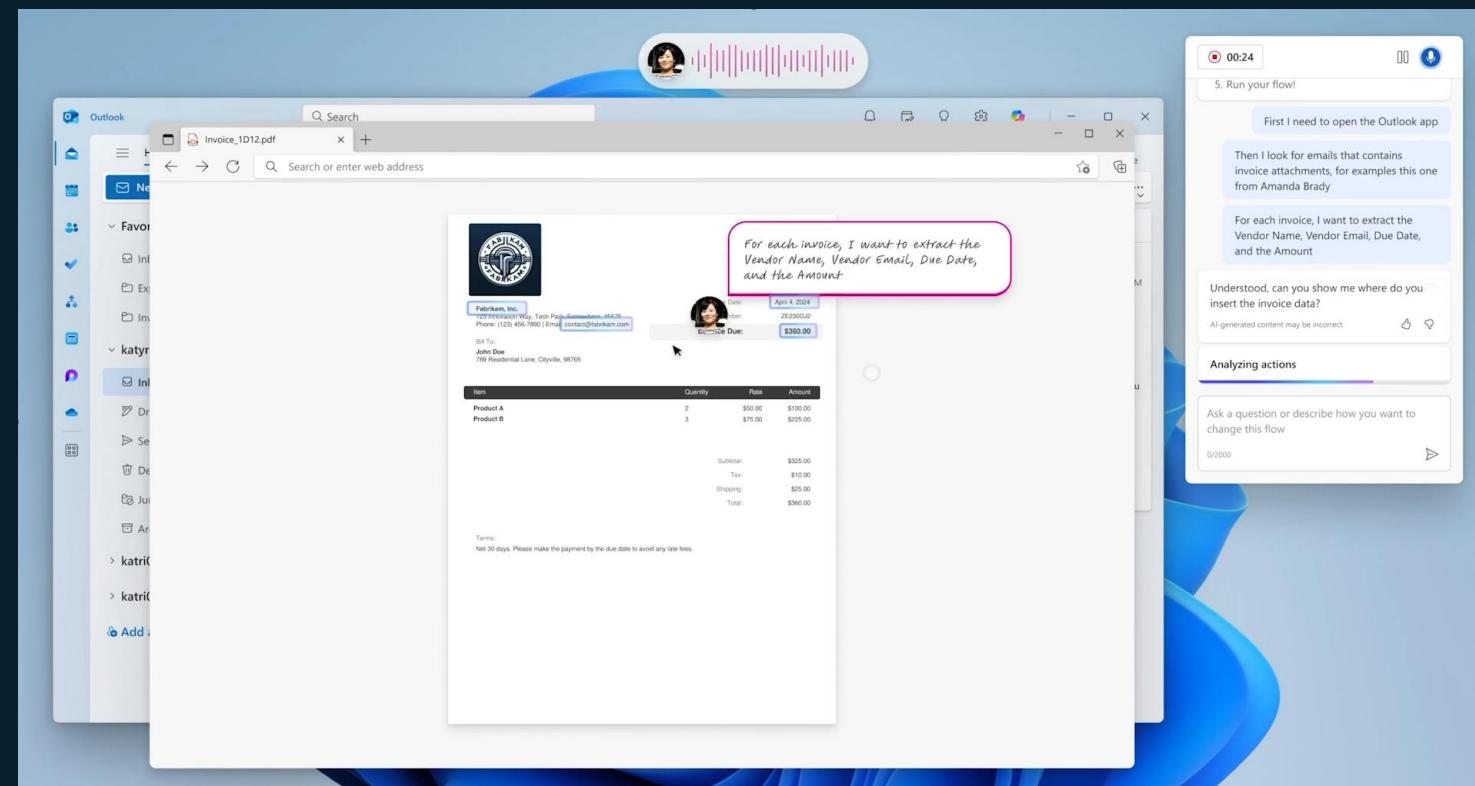
natural | accessible | resilient

Multi-modal AI recording with screen understanding and speech processing

"Show and tell" user experience is like training a new team member

Generative AI builds the flow when recording is complete

Self-healing with GPT-V when UI changes



# Power Pages

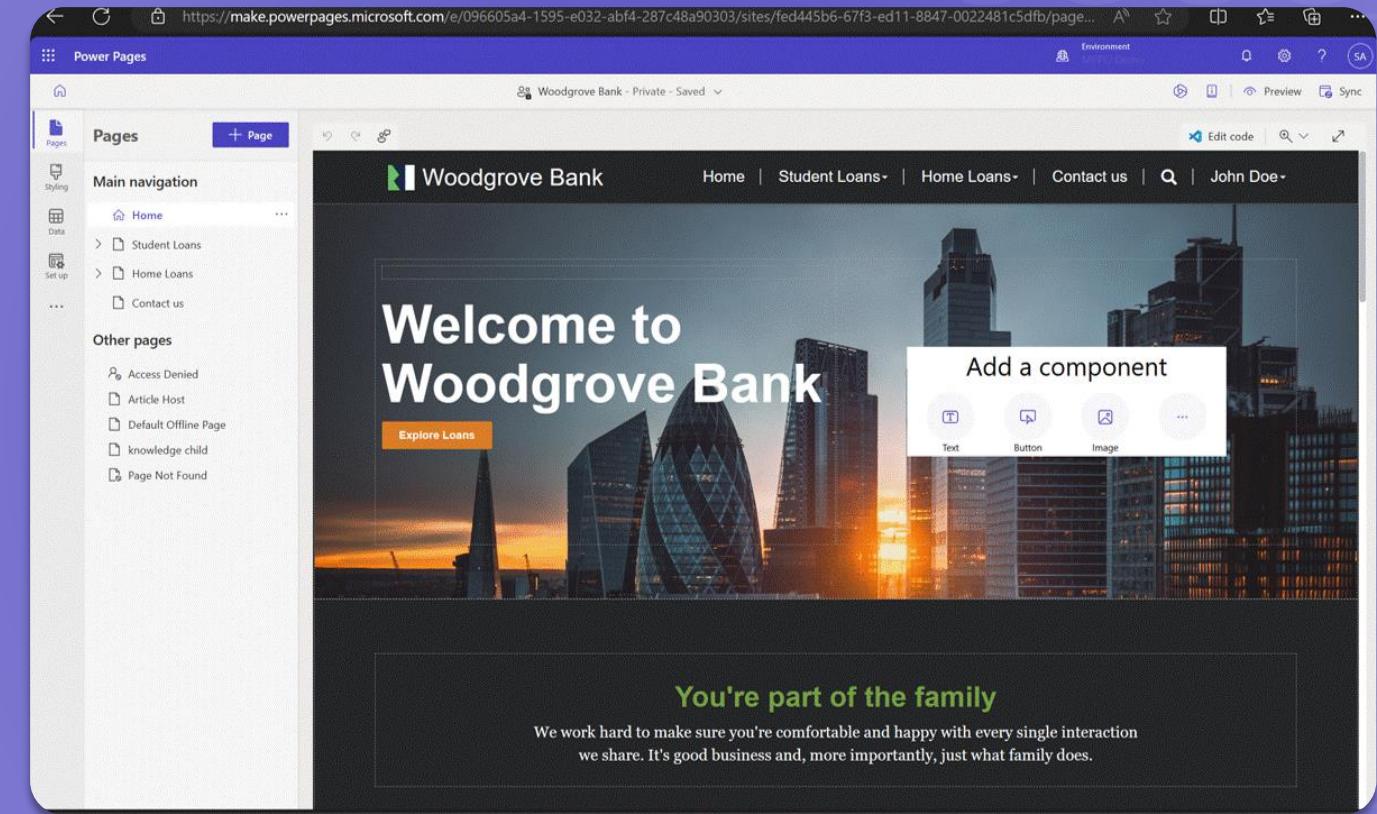


## General Availability

# Use chatbots on your website with Microsoft Copilot

Microsoft Copilot in Power Pages enables website makers to configure a copilot chatbot out-of-the-box with no additional authoring.

This chatbot will understand authenticated user context and can be customized using Copilot Studio.

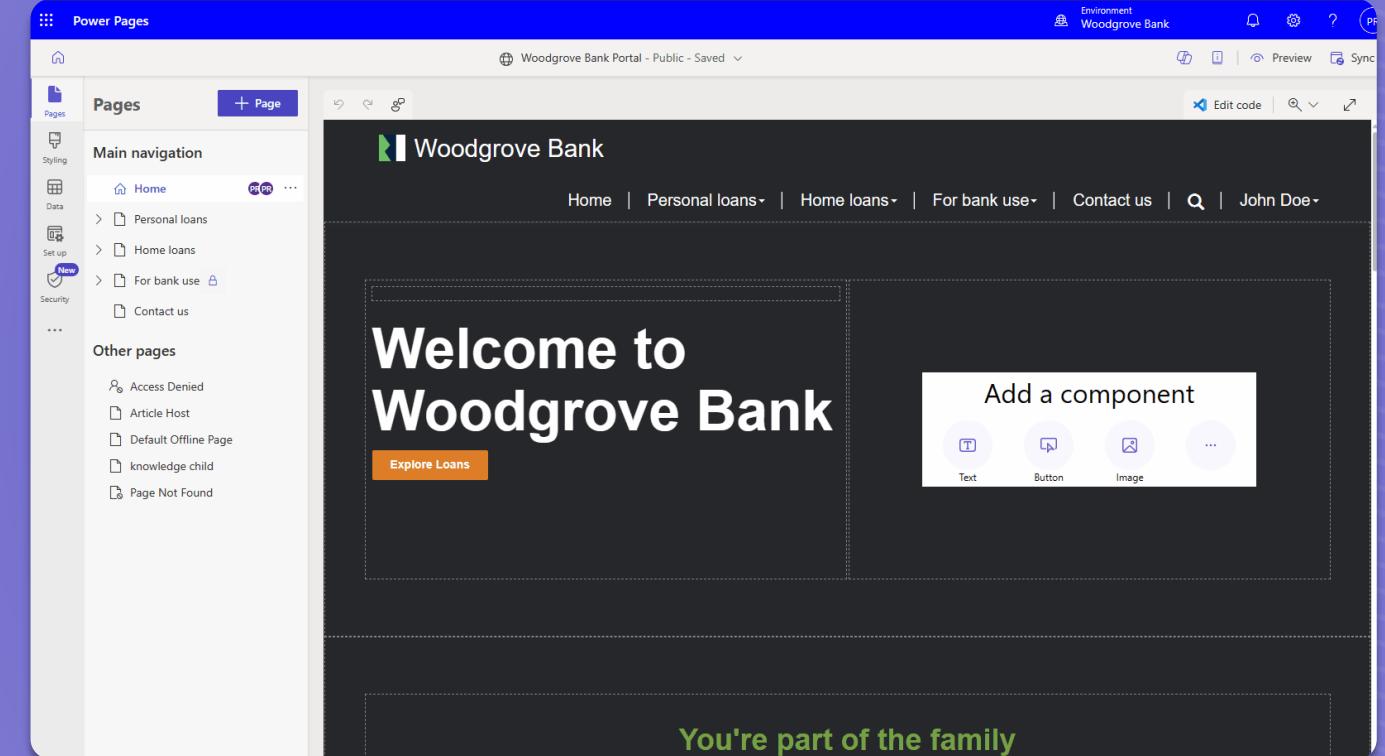


Public Preview

# Enhanced search powered by Microsoft Copilot

Microsoft Copilot in Power Pages provides AI-enabled search capabilities including summarization of results, akin to Bing copilot chat experiences.

This improves end user experience by providing information upfront.



**Public Preview**

## Security Workspace for Makers

---

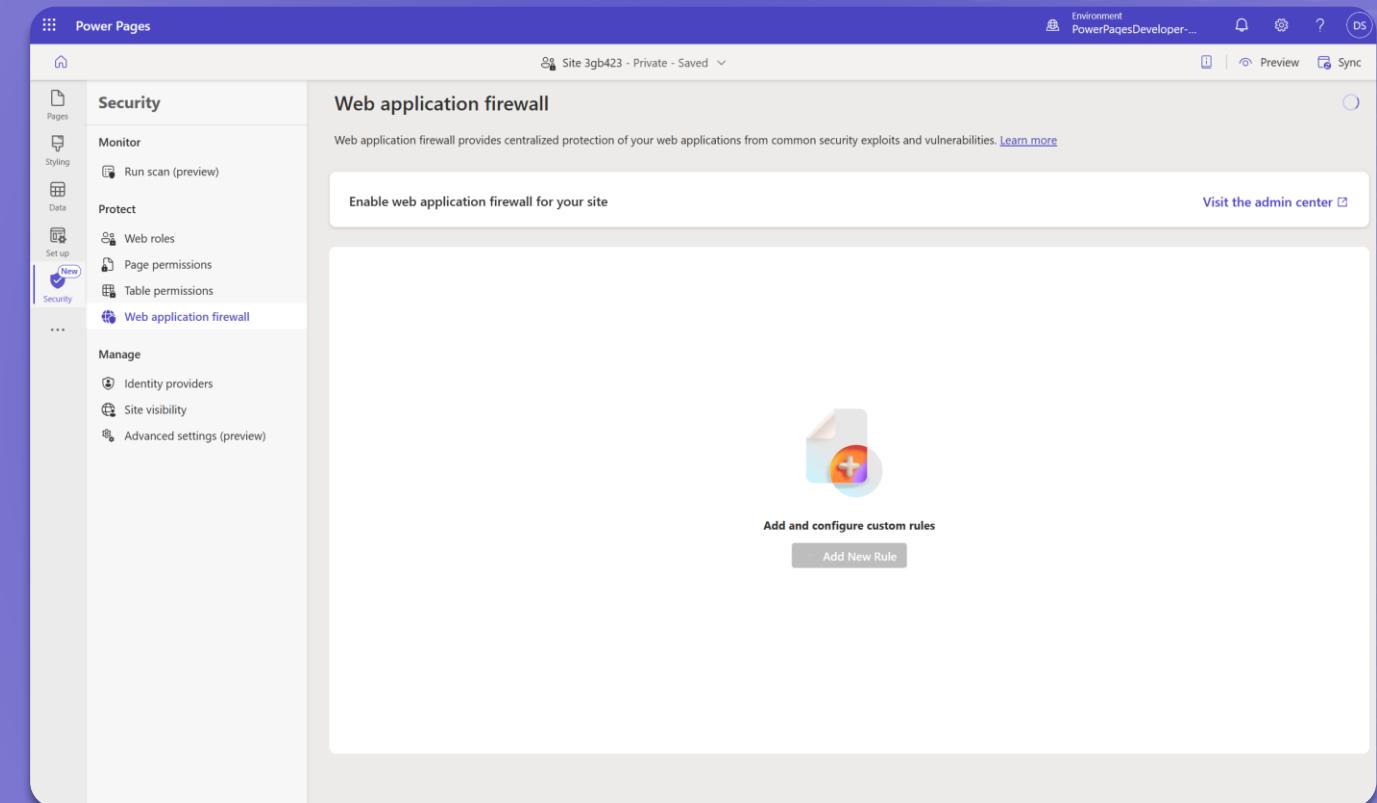
One-stop-shop for all things Security for Power Pages Makers.

Provides ability to set-up authentication providers, authorization (table, webpage based), session settings, as well as WAF/DAST setup.

## General Availability

# Built-in Web Application Firewall

WAF is going GA and will provide ability to create custom rules like geo-fencing, rate limiting, as well as WAF access logs.



## General Availability

### Web Scan

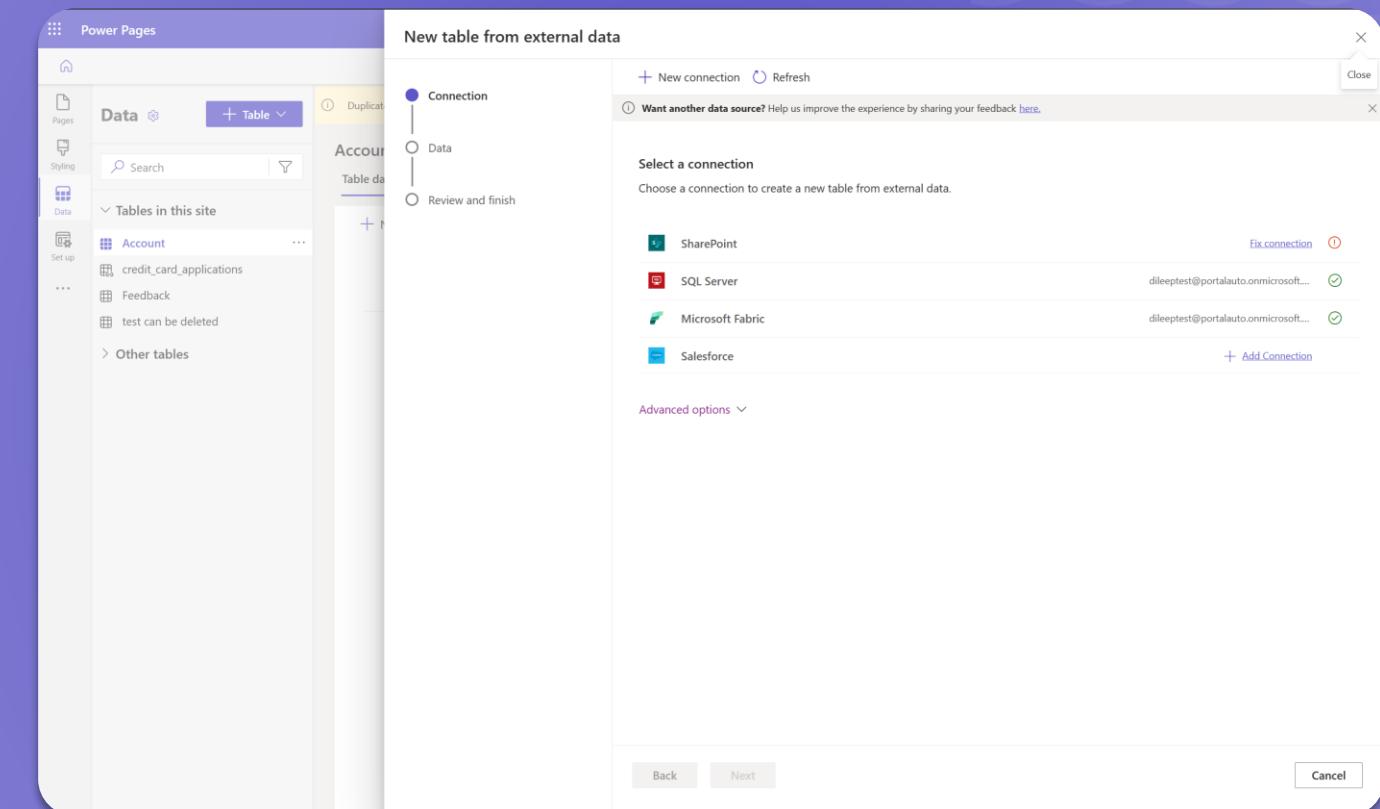
---

In-built SAST and DAST scans to ensure makers can secure their website while they are building it.

Public Preview

# Use Salesforce, Oracle and Fabric with Power Pages & Power Platform Seamlessly

Power Pages now has virtual table support for Salesforce, Microsoft Fabric and Oracle. This will support all OOB UX controls (forms, list etc), table permissions and web roles.

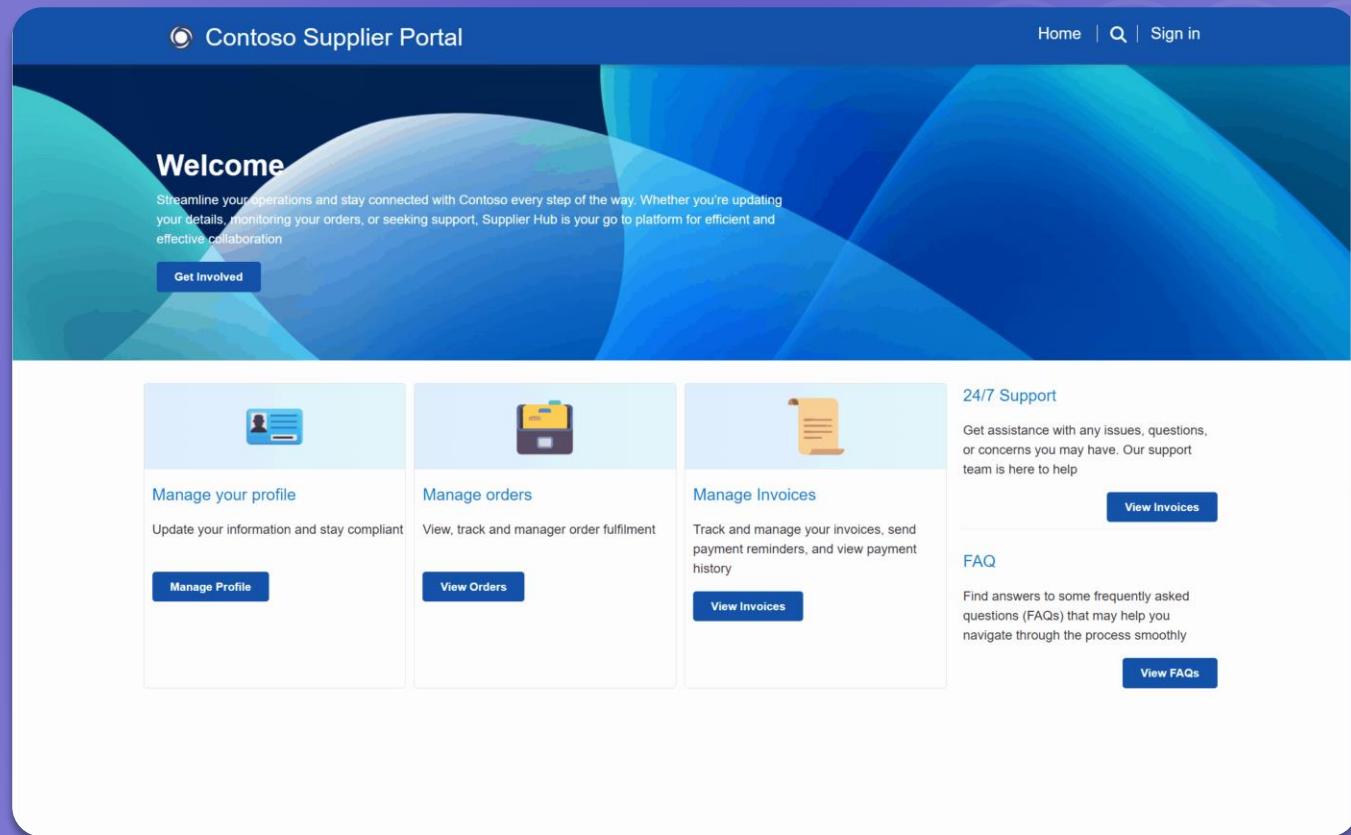


**Public Preview**

# SAP Supplier Self-Service Template

By connecting to SAP, the Power Pages supplier self-service template allows organizations to extend their SAP use cases for suppliers.

This will support OOB support for profile update, PO creation, & management experiences. Also offers full extensibility to any other SAP scenario.



# **Admin / Governance**



# Copilot hub



**Get started** with Copilot using FAQs, tech docs, and what's new

Monitor **usage** of Copilots across Power Platform

**Govern** Copilot with centralized controls and RAI

Private Preview

The screenshot displays the Microsoft Copilot hub interface, which is a central hub for managing AI features across the Power Platform. It includes sections for getting started with Copilot, monitoring usage, and governing Copilot with centralized controls and RAI.

**Readiness:** This section provides links to learn how Copilot works, prepare for Copilot, and stay updated on what's new.

**Usage:** This section shows adoption statistics for Power Apps Maker Copilot users over the last 28 days, with a total of 154 users. It also includes metrics for users editing apps with Copilot (117) and users creating apps with Express Design (2).

**Adoption:** A line chart showing the number of Power Apps Maker Copilot users from Thursday, April 14, to Wednesday, April 03. The user count increased from approximately 39 on April 14 to 154 by April 03.

**Governance:** This section provides centralized controls and RAI settings for different Power Platform components.

- Power Apps Settings:** Shows GA features that do not access your data are on by default. Features include Preview (Marker/App Copilots, App Copilot), Generally available (App descriptions, Maker Copilot), and Power Automate Settings (Desktop Flows Maker, Copilot, Process Mining Data, Ingestion).
- Power Automate Settings:** Shows GA features that do not access your data are on by default. Features include Preview (Desktop Flows Maker, Copilot, Process Mining Data, Ingestion), Generally available (Cloud Flows Maker Copilot), and Copilot Studio Settings (Preview, Generative Actions, Generally available, Generative Answers, Generative Builder).
- Copilot Studio Settings:** Shows GA features that do not access your data are on by default. Features include Preview (Generative Actions, Generally available, Generative Answers, Generative Builder).

**Microsoft Business Applications Launch Event** (April 10th) - Register now!

NEW

Public Preview



# Automation center and cloud flow trouble shooting

## End-to-end visibility for managing at scale

Dedicated work queue monitoring for SLA attainment, average processing time, and work queue throughput

Use Copilot to query desktop flow telemetry, throughput, and documentation

Recommendations for actionable insights and best practices

Identify runtime flow failures and proactively suggest guidance to fix it

The screenshot shows the Power Automate Automation center (preview) dashboard. It features a sidebar with navigation links like Home, Create, Templates, Learn, My flows, Approvals, Solutions, Process mining, AI Builder, Automation center (selected), and More. The main area has a search bar and a message about high impact recommendations. Below is the 'Automation center (preview)' section with tabs for Overview, Runs, and Work Queues. Under 'Recommendation (8)', there are cards for Frequent desktop flow errors, Automations started to fail, Work queue SLA at risk, and Improve throughput & resiliency. The 'Status' section includes cards for Top-level flows (13), Average runs per day (76), Average flow run duration (00:32), Flow runs error rate (35%), Flow runs by status (29%, 36%, 53%), and Flow runs error trend (105, 143, 184). A Copilot sidebar on the right provides suggestions for desktop flow telemetry, throughput, and documentation.

# Make security simple with Security Hub

Preview

Assess security posture for the tenant

Identify areas to improve based on strategic guidance

Set guardrails proactively

Extends to maker surfaces

The screenshot shows the 'Power Platform admin center' interface with the 'Security' tab selected. The main area displays a summary of the organization's security posture, showing an 'Enhanced' status on a gauge. Below this, sections include 'Address remaining recommendations to improve security posture' (10 recommendations for 2 of 10 environments) and '2 for securing your tenant'. To the right, there are several cards for different security components:

- Data policy:** Description: Setup environment and tenant wide guardrails for connectors to mitigate data exfiltration risks. Impact: High. Recommended action: Configure data policy for all environments. Completion status: ~50%. Refreshed: 4:34 PM, 2/6/2024.
- Customer Lockbox:** Description: Manage Microsoft attempts to access customer data for troubleshooting and diagnostics. Impact: Low. Recommended action: No recommendations. Completion status: 100%. Refreshed: 4:34 PM, 2/6/2024.
- Environment security groups:** Description: Control which users have access to environments using Entra security groups. Impact: High. Recommended action: Assign security group to environment. Completion status: ~50%. Refreshed: 4:34 PM, 2/6/2024.
- IP firewall:** Description: Restrict access to Dataverse on Managed Environments. Impact: High. Recommended action: No recommendations. Completion status: 100%. Refreshed: 4:34 PM, 2/6/2024.
- IP address-based cookie binding:** Description: Prevent cookie replay attacks on Dataverse in Managed Environments. Impact: High. Recommended action: No recommendations. Completion status: 100%. Refreshed: 4:34 PM, 2/6/2024.

Achieve the full potential of AI infused business applications  
with Enhanced Security Management

# Dynamics 365

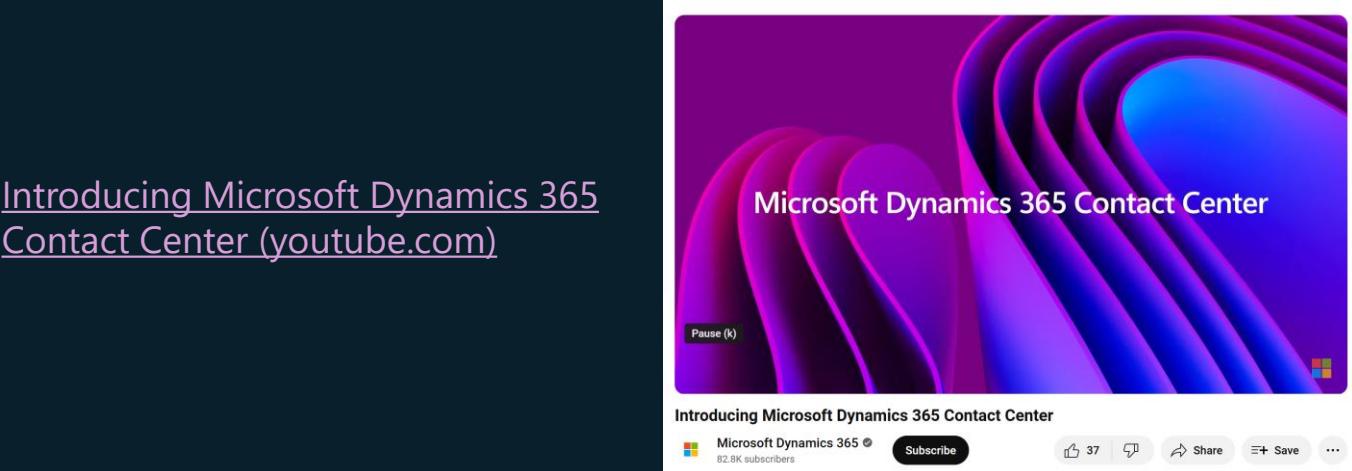


# Dynamics 365 Contact Center

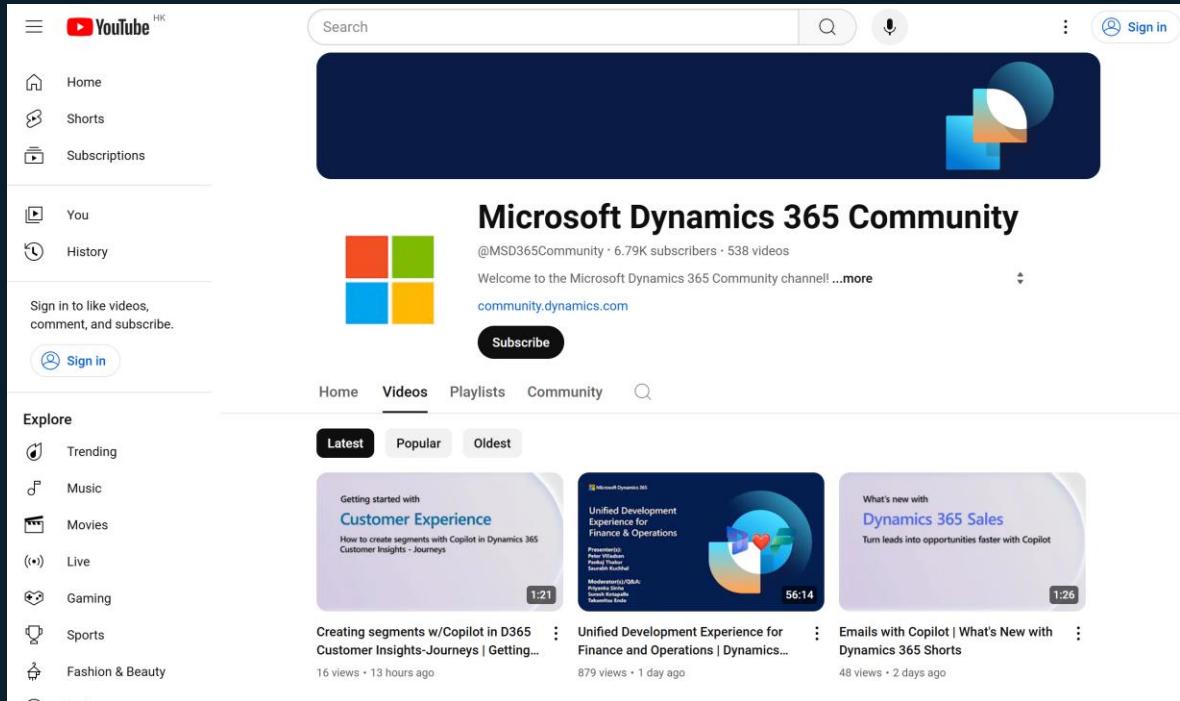
Contact Center as a Service  
(CCaaS)



[Announcing Dynamics 365 Contact Center - a Copilot-first cloud contact center to transform service experiences - The Official Microsoft Blog](#)



# Dynamics 365 Updates



## Finance & Operations

Archive with Dataverse Long-Term Retention (DTV056EXT1)

Extension options available for Data Archival and Long-Term Retention (DTV058EXT1)

Unified Development Experience for (DTV055EXT2)

Unified Admin Experience (DTV054EXT2)

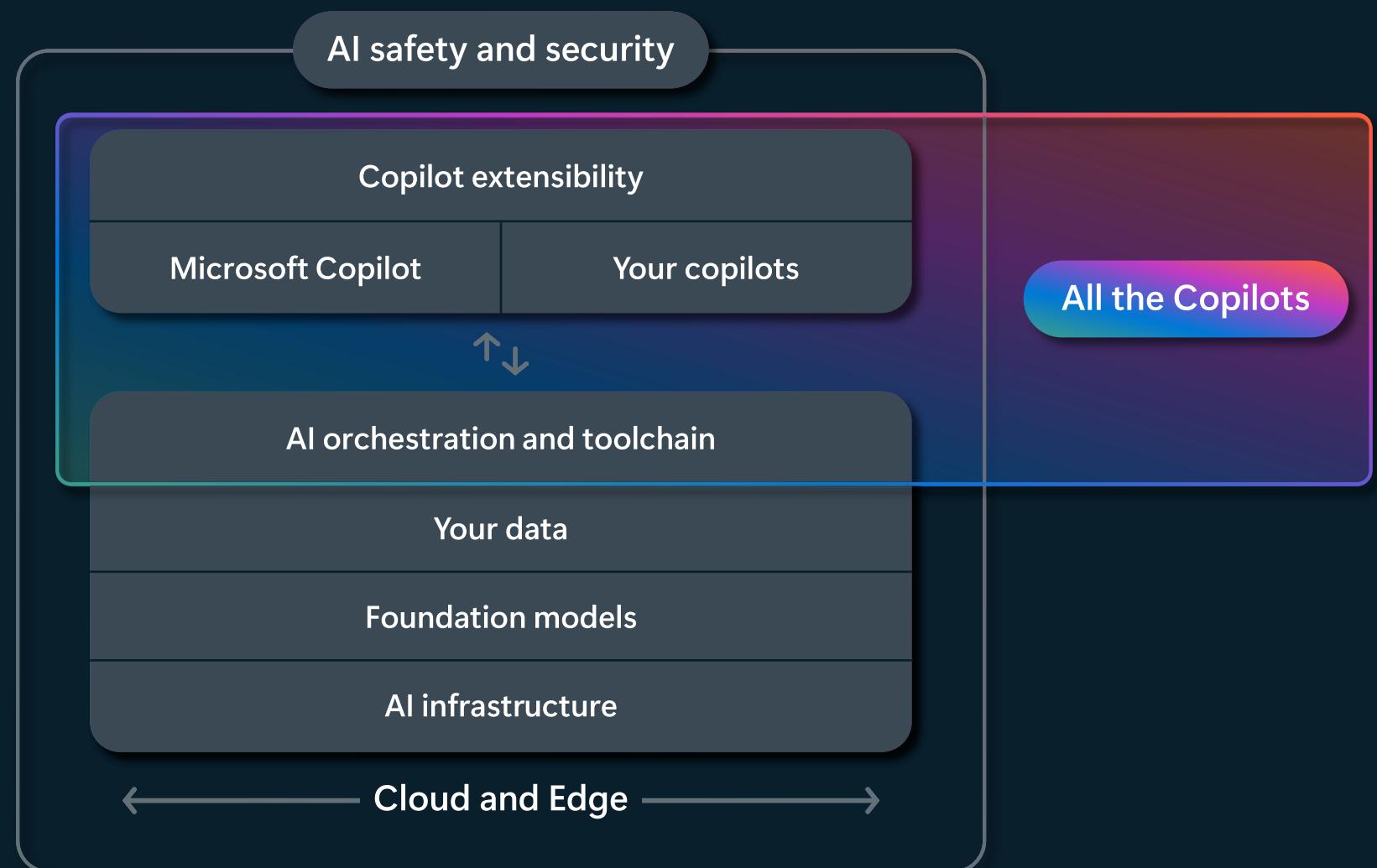
## Customer Engagement Apps

Overview of Dataverse Long-Term Retention (DTV057EXT1)

# Douglas Lam

Partner Solution Architect –  
Azure Infrastructure

# Session 1 – Copilots



# Agenda

- Copilot for Security in Microsoft Defender for Cloud (MDC)
- Copilot in Azure



# Remediating Cloud Risks with Microsoft Copilot for Security in Defender for Cloud (MDC)

# Microsoft Defender CSPM

Get in front of your most critical risks across your multicloud environments



## Gain visibility into multicloud risks

Identify critical risks and understand potential attack paths across your cloud estate with enriched code-to-cloud security signals



## Unify security standards and cloud policies

Streamline compliance and security best practices across your environment with built-in security standards and custom recommendations



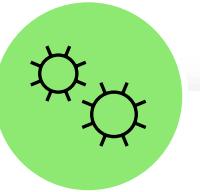
## Fortify sensitive data across clouds

Reduce risks of data breach with ongoing visibility into your cloud data estate and proactively harden at-risk resources containing sensitive data



## Mitigate code risks and “fix left”

Prevent recurring risks by understanding code to cloud context to trace and fix issues to the source code



## Prioritize remediation and collaborate at scale

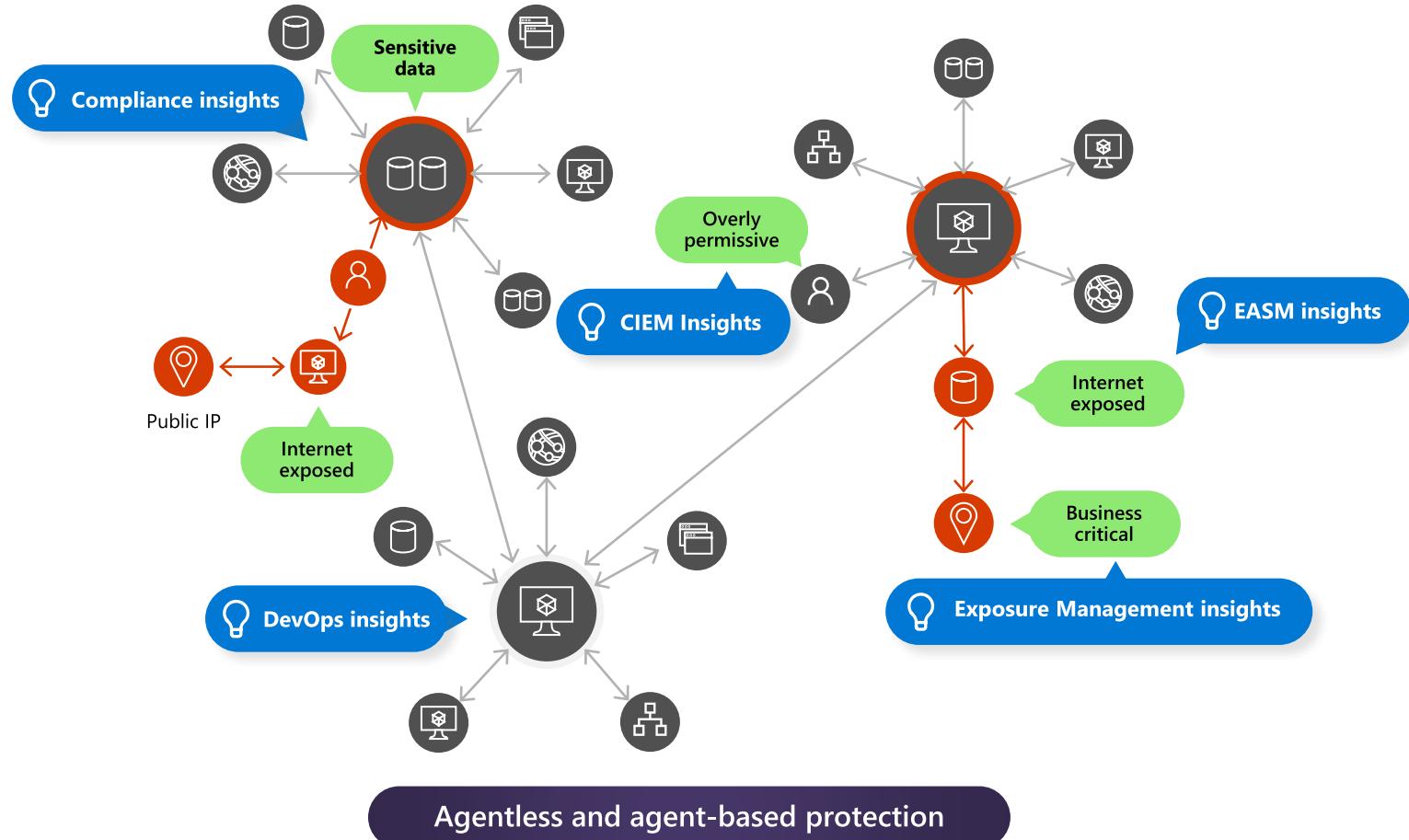
Remediate risks faster with risk-based prioritization, governance rules, and integrations with other security solutions



# Cut through the noise and get in front of your most critical multicloud risks

- › Trace risks across the development lifecycle with **code-to-cloud mapping**
- › Identify sophisticated attack paths such as **lateral movement across clouds** using contextual security insights
- › Prioritize action on **business critical resources**
- › Accelerate risk remediation further with the integrated power of **AI using Microsoft Copilot for Security**

Unify visibility with integrated CNAPP insights



# Protect at machine speed and scale with Generative AI

Accelerate risk discovery and remediation with **Microsoft Copilot** in **Defender for Cloud**

## Simplify the complexity

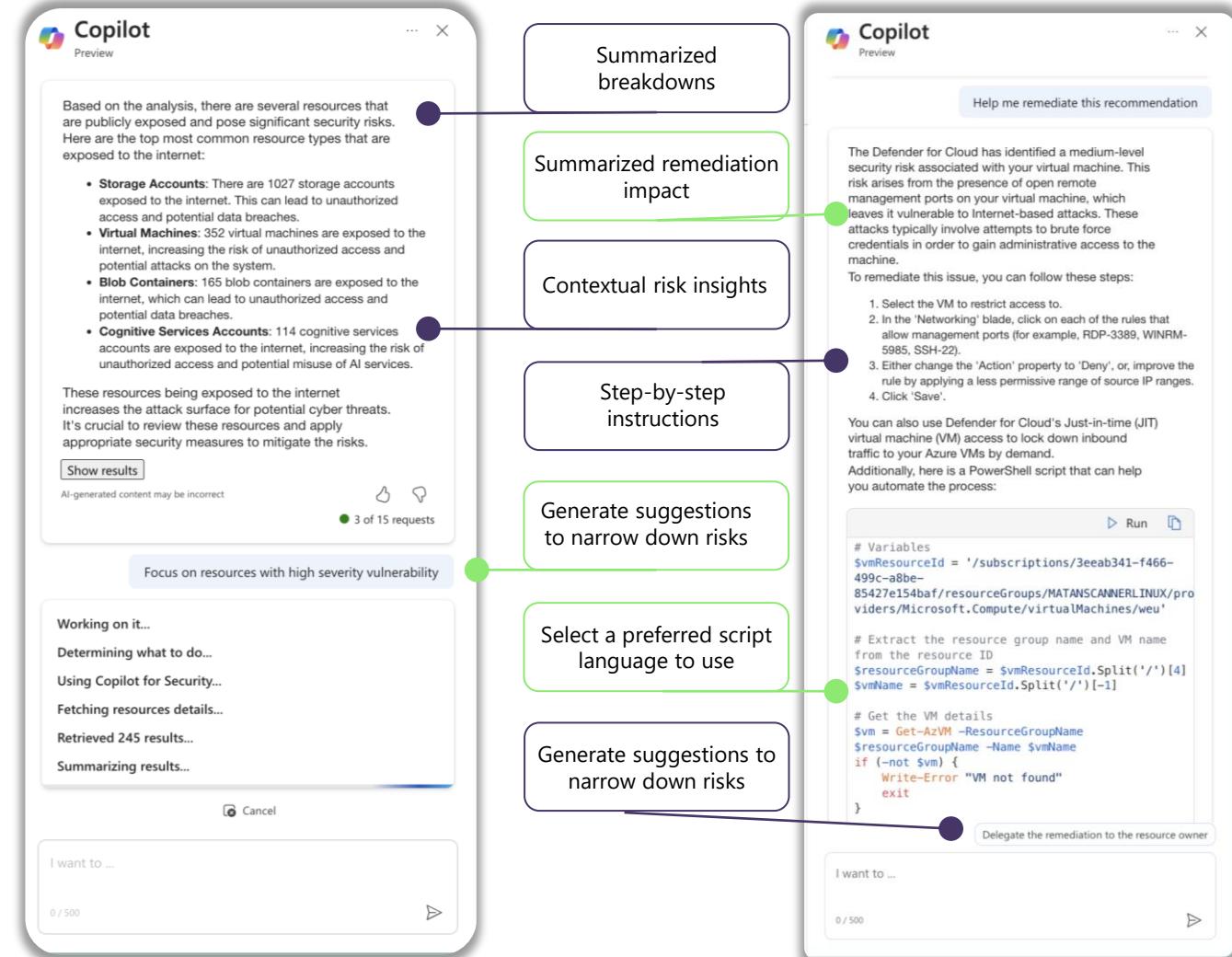
# Risk exploration with natural language queries

➤ Catch what others miss

Analyze risks at scale and get predictive guidance

## Accelerate remediation

Deploy AI-generated remediation actions and scripts or submit Pull Requests to fix the issue at the source



# Accelerate response in Defender for Cloud with Copilot for Security

The screenshot shows the Microsoft Azure Defender for Cloud Recommendations page. The main pane displays a dashboard with metrics like Active recommendations by risk (48 Critical, 55 High, 180 Medium, 1297 Low, 3448 N/A), Other metrics (216 Active attack paths, 1811 Affected resources, 342 Overdue recommendations), and Defender resource coverage (Full 105, Partial 105, Not covered 207). Below this is a table of recommendations, grouped by risk level (Critical, High, Medium, Low, N/A) and sorted by title. The table includes columns for Affected resource, Risk factors, Attack paths, and Status. The right side of the screen features a "Security Copilot" sidebar with a summary of publicly exposed resources, a detailed breakdown of specific resources at risk, and AI-generated insights for verification.

**Microsoft Defender for Cloud | Recommendations**

Showing 85 subscriptions

Search

Refresh Switch to classic Download csv report Open query Governance report Guides & Feedback Security Copilot

Scope: Azure subscriptions 205 AWS accounts 5 GCP projects 5

Active recommendations by risk

Critical	48
High	55
Medium	180
Low	1297
N/A	3448

Other metrics

Active attack paths	216
Affected resources	1811
Overdue recommendations	342

Defender resource coverage

Full 105
Partial 105
Not covered 207

Risk level ↑ Title ↑

	Affected resource	Risk factors	Attack paths ↑	Status ↑	
Critical	Management ports should be closed on your virtual machine	Exposure to the internet	4	Overdue	
Critical	All network ports should be restricted on network security group...	Exposure to the internet	+2	Overdue	
Critical	API endpoints in Azure API Management should be authenticated	modify-resource	Exposure to the internet	+3	Overdue
Critical	SQL databases should have vulnerability findings resolved	samplecrmsdemo1	Exposure to the internet	+2	Overdue
Critical	API endpoints in Azure API Management should be authenticated	submitsession	Exposure to the internet	+3	Overdue
Critical	EC2 instance should use IMDSv2	mdc-demo-w2022	Exposure to the internet	+1	Unassigned
Critical	EC2 instance should not have a public IP address	i-09e4c865a0f5675ec1	Exposure to the internet	+3	Unassigned
Critical	EC2 instance should not have a public IP address	i-04e65c8f12mz75ec1	Exposure to the internet	+3	Unassigned
Critical	EC2 instance should not have a public IP address	i-02e65438f0f5675ec1	Exposure to the internet	+2	Unassigned
Critical	EC2 instance should not have a public IP address	i-03e6123ct0f5675ec1	Exposure to the internet	+2	Unassigned
Critical	EC2 instance should not have a public IP address	i-15ev7438d15675ec1	Exposure to the internet	+3	Unassigned
Critical	EC2 instance should not have a public IP address	i-06e2438f0f5675ec1	Exposure to the internet	+1	Unassigned
Critical	EC2 instance should use IMDSv2	i-07e65ht560f5675ec1	Exposure to the internet	+3	Unassigned
Critical	Management ports of EC2 instances should be protected with...	i-01e6ty56f0f5675ec1	Exposure to the internet	+3	Unassigned
Critical	Management ports of EC2 instances should be protected with...	i-06e34w856ga5175es3	Exposure to the internet	+2	Unassigned
Critical	EC2 instance should not have a public IP address	i-01e6k3u3uerf5675ec1	Exposure to the internet	+2	Unassigned

< Previous Page 1 of 2 Next >

Give feedback

Security Copilot

Show critical risks to publicly exposed resources

Aug 01, 2023 11:14 AM

13 publicly exposed resources require attention

There are 13 publicly exposed resources, each carrying a range of potential threats such as unmanaged workloads, human errors, misconfigurations, data breaches, privacy and compliance issues, performance degradation, and security vulnerabilities.

Here's a breakdown of the specific resources at risk:

- 9 VMs with internet exposure and high severity vulnerabilities
- 1 AWS S3 bucket that's publicly accessible, housing sensitive data
- 1 SQL with excessive internet exposure, allowing basic authentication and containing sensitive data
- 2 API endpoints lacking authentication, transmitting sensitive information.

See less

Show results

AI generated. Verify for accuracy.

Focus on risks to sensitive data

Focus on risks to critical vulnerability

Focus on risks to lateral movement

<Helper text here>

[Get started with Microsoft Copilot for Security | Microsoft Learn](#)



# Microsoft Copilot in Azure



DESIGN



OPERATE

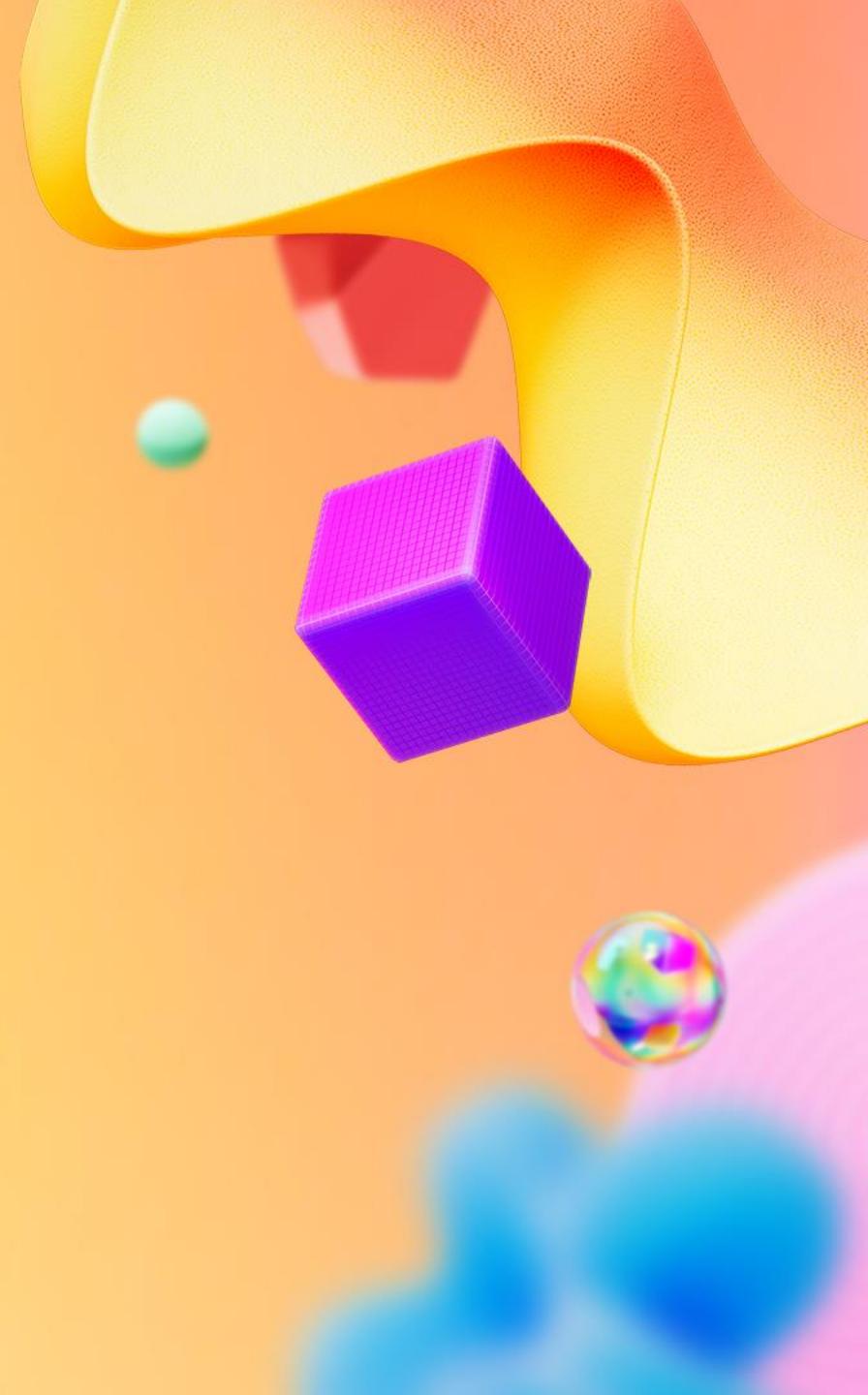


OPTIMIZE

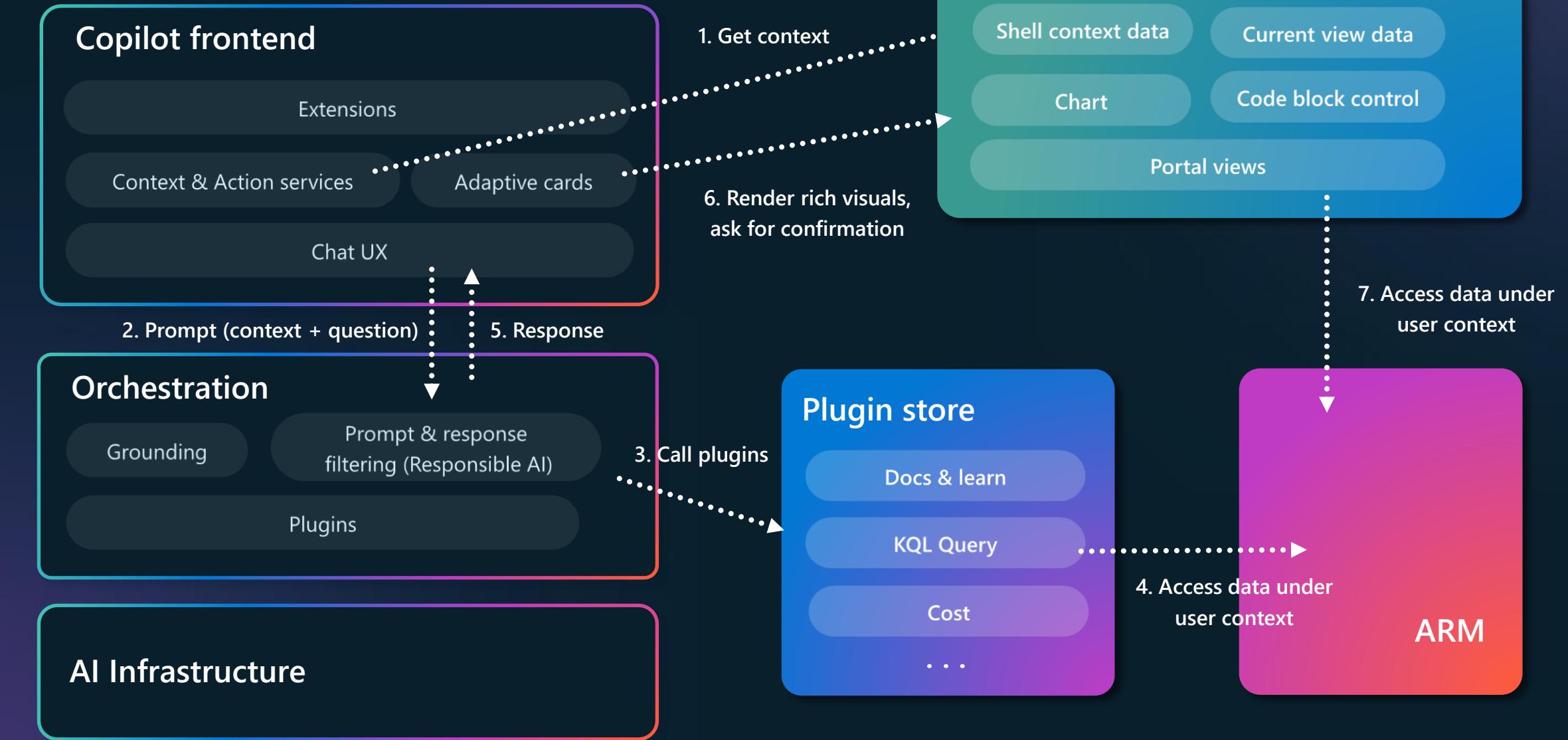


TROUBLESHOOT

# Architecture



# Architecture and data flow



# Copilot Availability

Until now:

Sign up required to join the public preview of Copilot in Azure

Starting **next week**

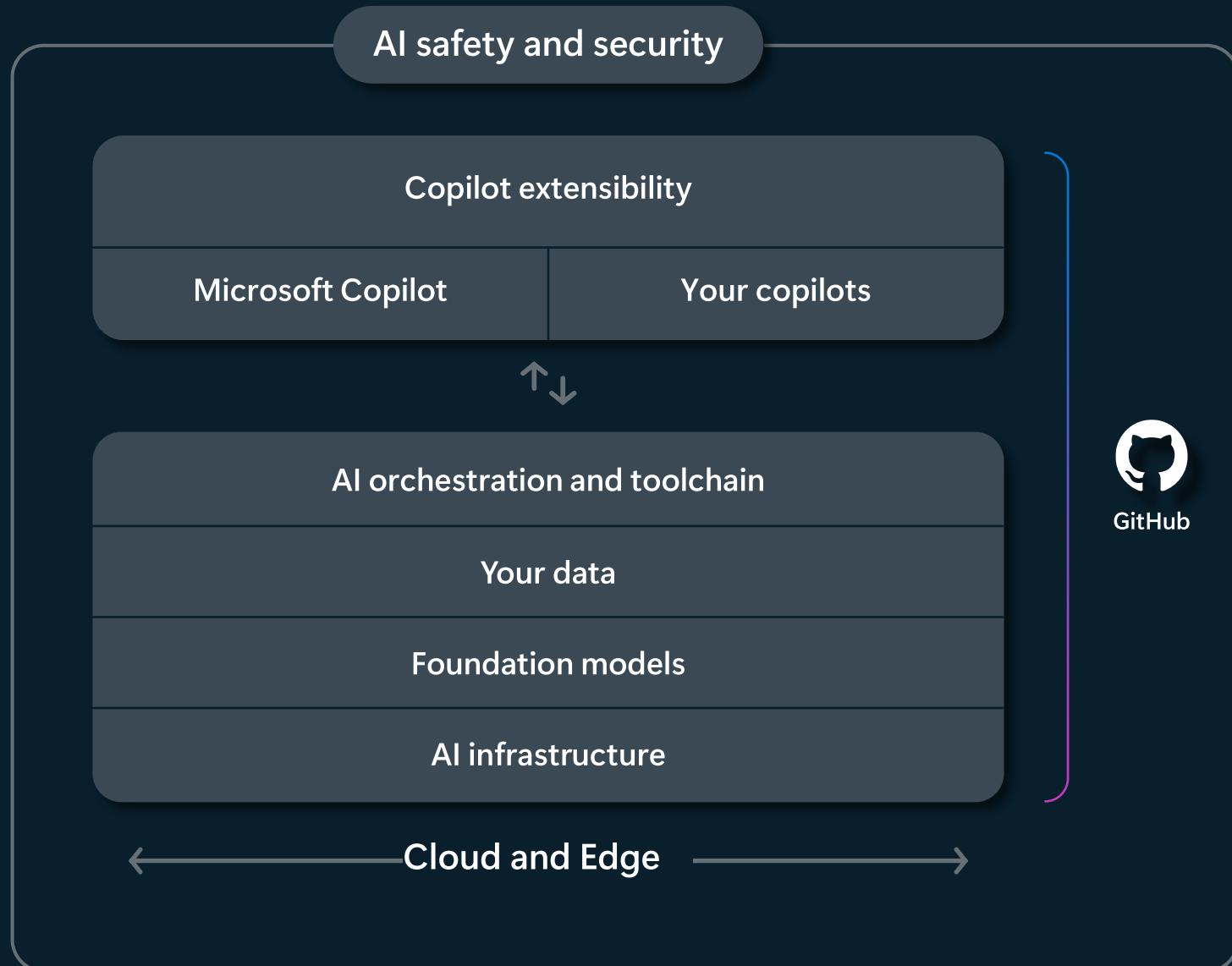
Copilot in Azure will be rolled out to all users and tenants in stages over a few weeks. Copilot will be enabled by default.

## Configuring access to Copilot

You can manage access to Copilot in your organization in the Azure Portal **next week** ahead of Copilot rollout across Azure.

Admins can now configure access management to Copilot in Azure by assigning individual users/groups permissions at a tenant level.

Existing Preview access maintained





# GitHub Copilot

Your AI pair programmer

Offers code suggestions,  
contextualized to your project

- ✓ Convert comments to code
- ✓ Autofill for repetitive code
- ✓ Autosuggest tests
- ✓ Show alternatives



```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8     const response = await fetch(`http://text-processing.com/
9         method: "POST",
10        body: `text=${text}`,
11        headers: {
12            "Content-Type": "application/x-www-form-urlencoded"
13        },
14    );
15    const json = await response.json();
16    return json.label === "pos";
17}
```



Technical Preview

# Copilot Workspace

A Copilot-native dev environment, designed for everyday tasks.

Join the waitlist >

[gh.io/copilot-workspace](https://gh.io/copilot-workspace)

Update the product page to make description a rich text field #1

Open Damovisa opened this issue 2 days ago · 5 comments

Damovisa commented 2 days ago

The description field currently only shows plain text.

CONTOSO

Products Catalog / New Product

Add new product

Upload Picture

Product name

Item name

Product Description

Short description of the item  
Be great to support italics, bold and bullets here!

Click to upload

Item category

Select category

Item sub category

Select subcategory

Assignees

mona-workspace

Labels

enhancement help wanted

Projects

None yet

Milestone

No milestone

Development

Open in Workspace

jtfeisher/eShopOnWeb: Samp | didactic-parakeet-96776gp | Update the product page to ... | +

github.com/mona-workspace/contoso-admin/issues/1

New Chrome available :

mona-workspace / contoso-admin

Type  to search

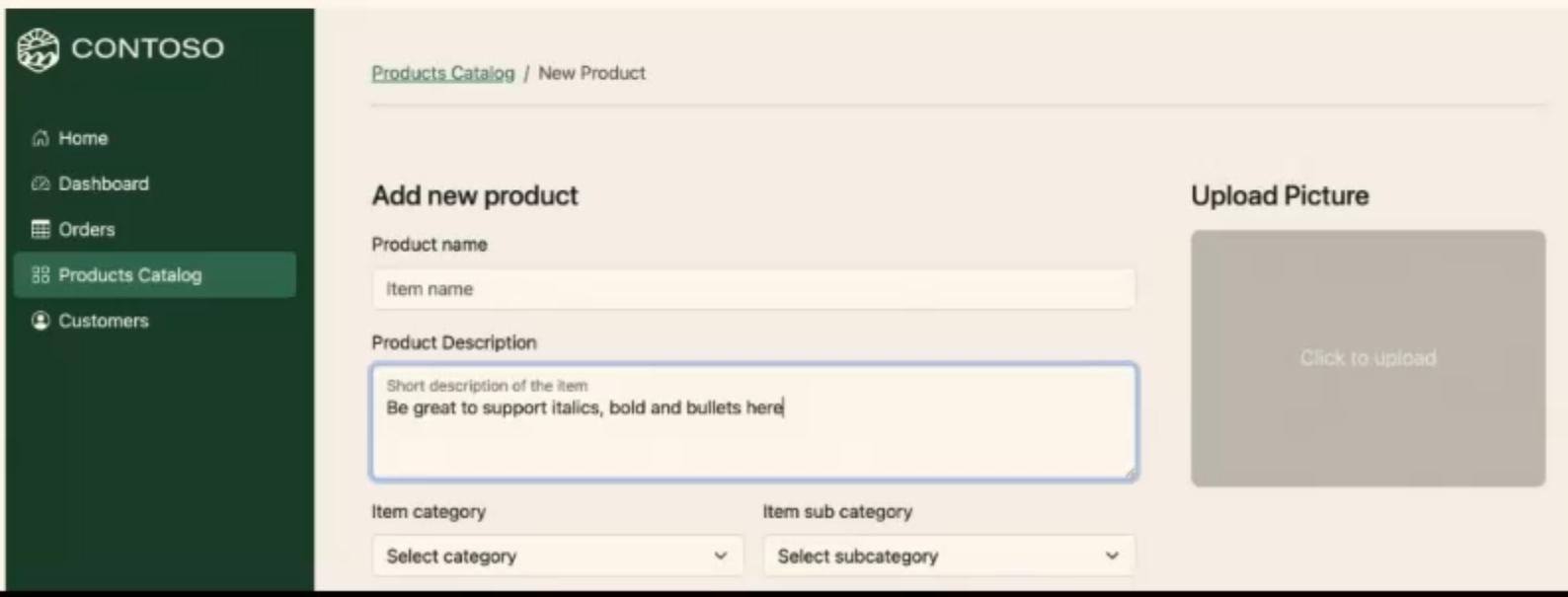
Code Issues 1 Pull requests Actions Projects Security Insights Settings

# Update the product page to make description a rich text field #1

Open Damovisa opened this issue 2 days ago · 5 comments

Damovisa commented 2 days ago

The description field currently only shows plain text.



CONTOSO

Products Catalog / New Product

Add new product

Product name

Item name

Product Description

Short description of the item  
Be great to support italics, bold and bullets here!

Upload Picture

Click to upload

Item category

Select category

Item sub category

Select subcategory

Assignees

mona-workspace

Labels

enhancement help wanted

Projects

None yet

Milestone

No milestone

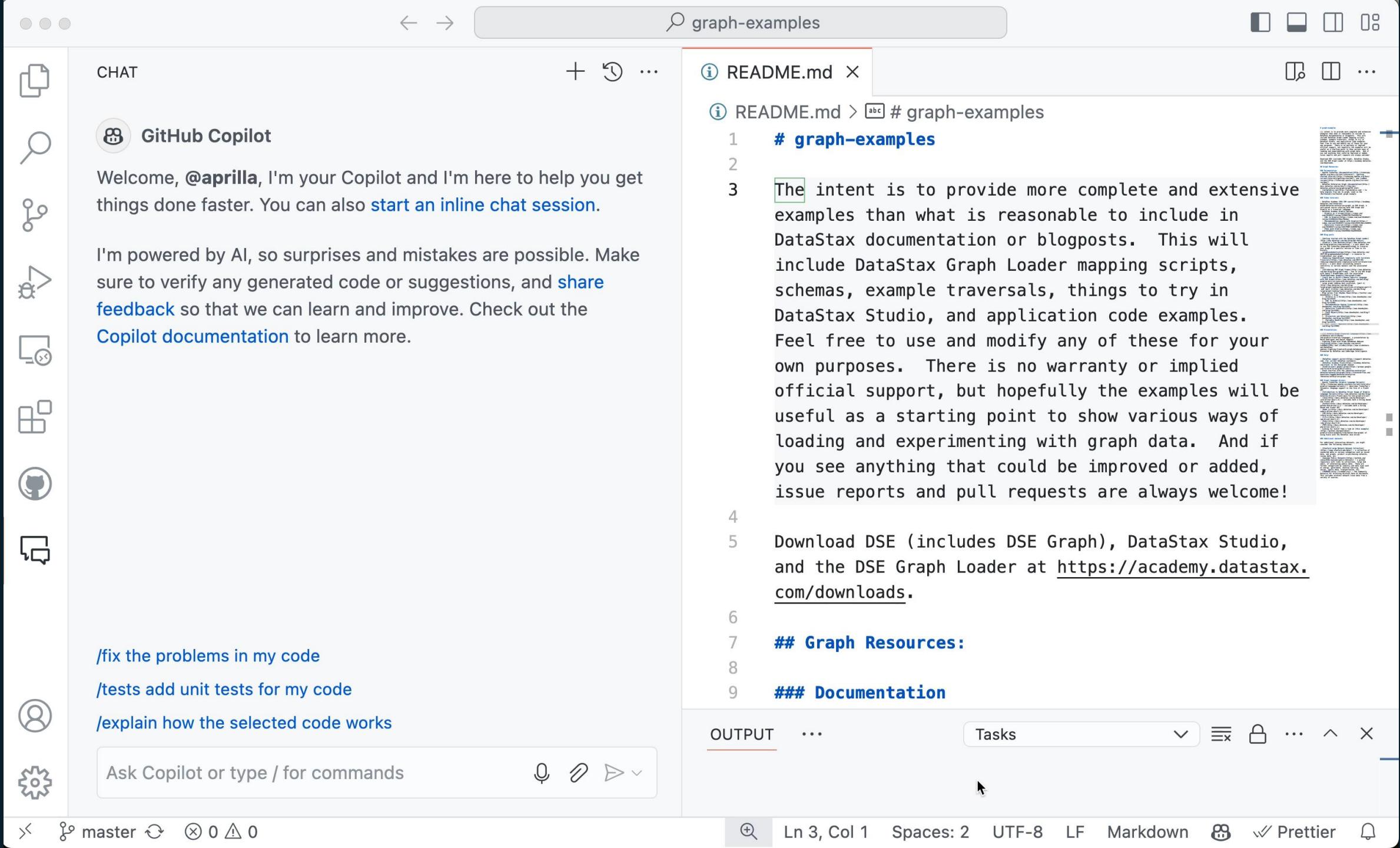
Development

Open in Workspace

Limited Beta

 GitHub Copilot  
**Extensions**





Marketplace x github/copilot-api: The engine x Bug: Seeing 500s for JSON p: +

github.com/github/copilot-api

github / copilot-api 🔒

Type ⌘ to search

Code Issues 316 Pull requests 117 Discussions Actions Projects 3 Security Insights

copilot-api Private

Watch 85 Fork 0 Star 12

main 444 Branches 0 Tags Go to file Add file <> Co

gnfisher Merge pull request #4804 from github/gf/buildfilepath-regex 7a65547 · 19 hours ago 13,550 Commits

.devcontainer Add Node.js to Codespace devcontainer 3 weeks ago

.github fix that syntax again last week

.vscode Merge pull request #4077 from github/uoodsq/overmind-... last month

cmd Do not use regex conversion to build query path for filepa... yesterday

config Adding uks gpt3.5-turbo ptum endpoint and mirror config... last month

docs Add a couple of notes about testing last month

pkg Do not use regex conversion to build query path for filepa... yesterday

proto gen files 2 weeks ago

script WIP 2 weeks ago

.air.toml copilot-platform-api -> copilot-api 10 months ago

.env.example Merge branch 'main' into ci-aip-tests 3 months ago

.env.feature\_flags\_development Update .env.feature\_flags\_development 5 months ago

All repositories ...

Chatting about github/copilot-api

Private repository

Copilot is powered by AI, so mistakes are possible. Review output carefully before use. [Learn more about GitHub Copilot Chat.](#)

Indexed for improved understanding and accuracy.

Ask about the repository:

What questions can I ask?

What does this repository do?

How should I get started exploring this repo?

Ask Copilot

Marketplace x | github/copilot-api: The engine x Bug: Seeing 500s for JSON p... x +

github.com/github/copilot-api/issues/4431?reload=1?reload=1

Code Issues 324 Pull requests 103 Discussions Actions Projects 3 Security Insights

Bug: Seeing 500s for JSON parsing error Cancel Save

**Open**

JasonEtc0 opened 3 weeks ago edited by aprilla · Edits · ...

Customers are reporting issues with the application returning 500 errors, due to a JSON parsing error. I found this exception in Sentry ID: 5088933476

+ Add tasklist 😊

JasonEtc0 added bug 3 weeks ago

howie-work changed the title [Demo]: Seeing 500s for JSON parsing error Bug: Seeing 500s for JSON parsing error 3 weeks ago

JasonEtc0 transferred this issue from [github/copilot-extensibility](#) 3 weeks ago

github-project-automation added this to Copilot API & Core Productivity 3 weeks ago

github-project-automation moved this to Triage in Copilot API & Core Productivity 3 weeks ago

JasonEtc0 changed the title Bug: Seeing 500s for JSON parsing error Bug: (FOR A DEMO) Seeing 500s for JSON parsing error 3 weeks ago

sharonlo changed the title Bug: (FOR A DEMO) Seeing 500s for JSON parsing error Bug: Seeing 500s for JSON parsing error 3 weeks ago

brannon 3 weeks ago

Assignees No one - Assign yourself

Labels bug

Projects Le Copilot Grand Memex Status Backlog ▾ Copilot API & Core Productivity Status Triage ▾

Milestone No milestone

Development Open in Workspace

Create a branch for this issue or link a pull request.

Notifications

Private Preview

# GitHub Copilot for Azure

Developers can use Copilot Chat from within their development tool to:

- Learn about Azure resources and best practices
- Explore and manage Azure resources
- Troubleshoot and optimize apps on Azure quickly and easily

Join the preview

[aka.ms/azcode/copilot/preview-signup](https://aka.ms/azcode/copilot/preview-signup)



GitHub Copilot

Welcome @oliviadev, I'm your Copilot and I'm here to help you get things done faster.

I'm powered by AI, so surprises and mistakes are possible. Make sure to verify any generated code or suggestions, and [share feedback](#) so that we can learn and improve. Check out the [Copilot docs](#) to learn more.



oliviadev

@azure what's the best way to deploy this app?



GitHub Copilot using @azure...



Analyzing project

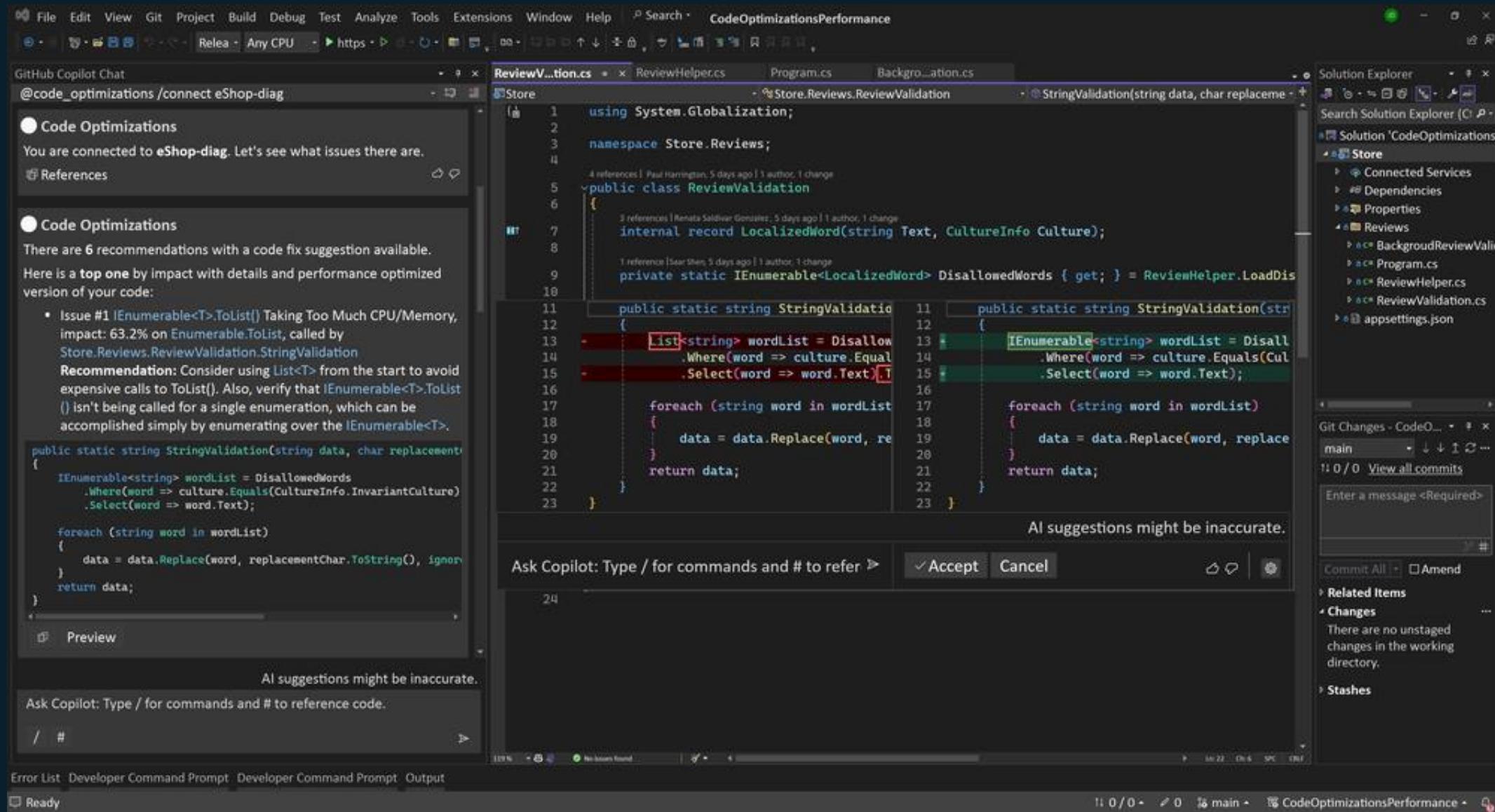


Deciding which project information to collect

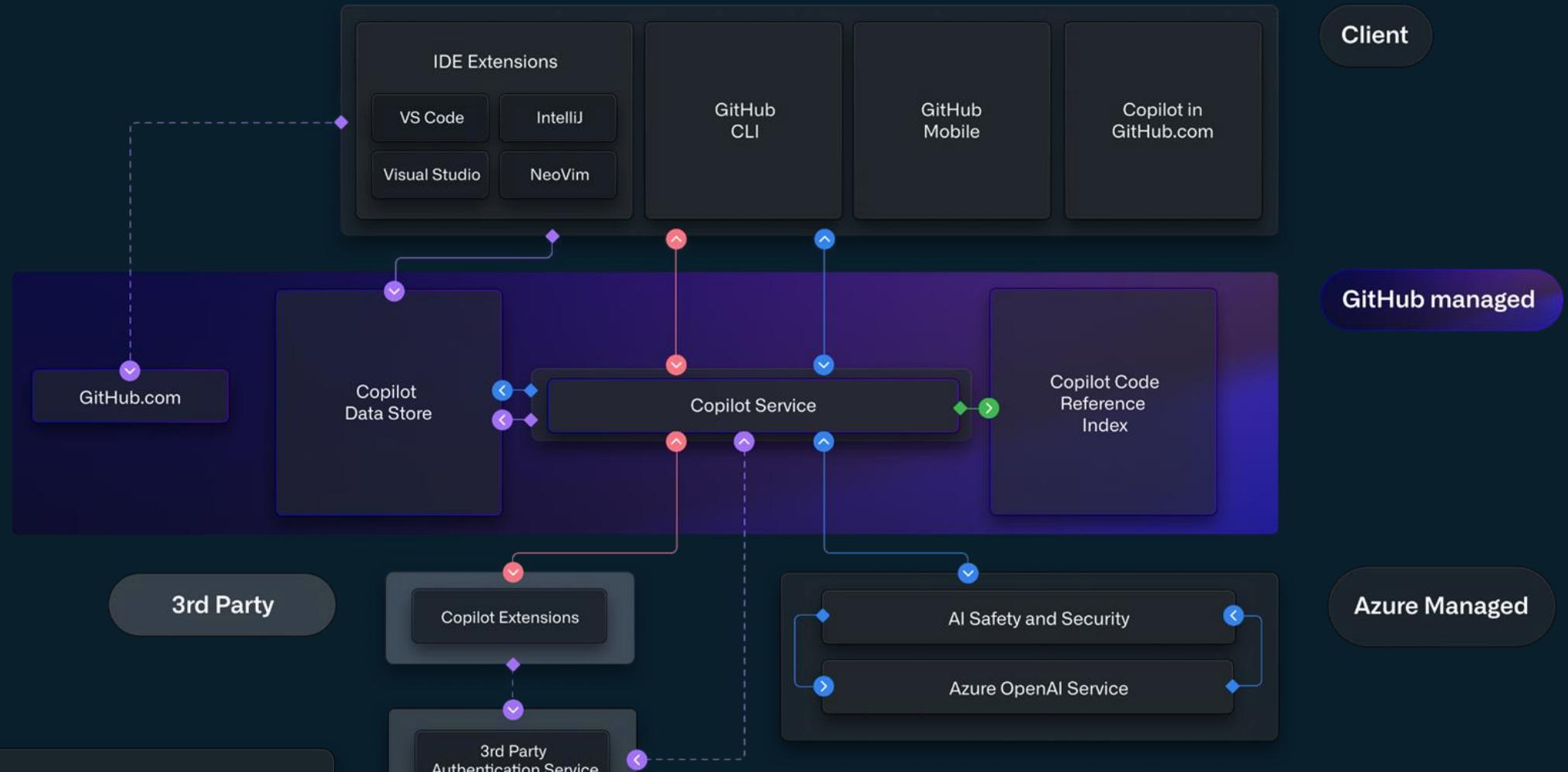


Creating specifications

# GitHub Copilot with Code Optimizations extension



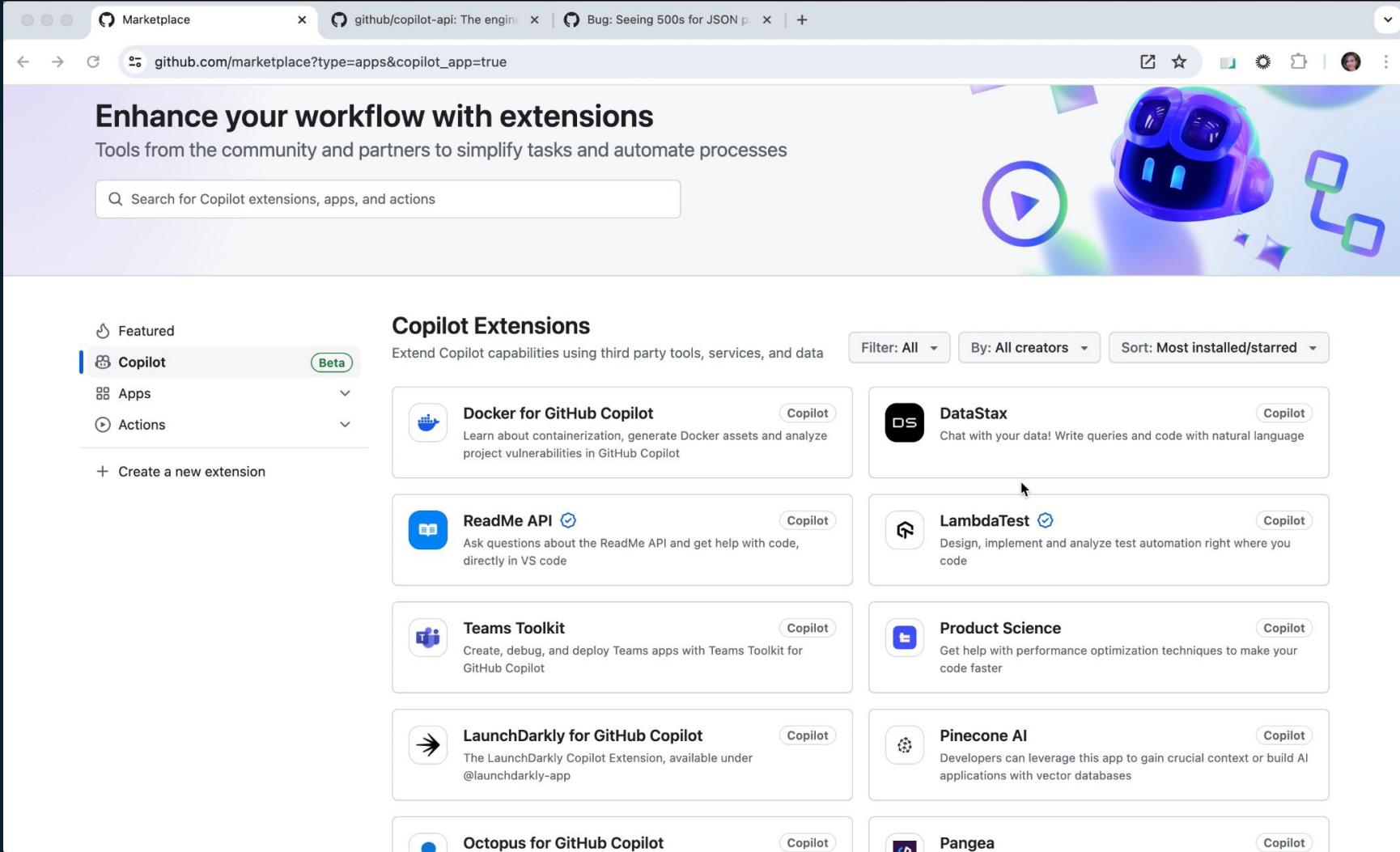
# GitHub Copilot Data Flow



## Legend: Data Pathways

- Prompts + suggestions
- User engagement data (UED)
- Authentication
- Code Matching
- 3rd party extensions

# Extension Marketplace



The screenshot shows the GitHub Extension Marketplace interface. At the top, there's a search bar with placeholder text "Search for Copilot extensions, apps, and actions". Below it, a sidebar on the left lists categories: "Featured", "Copilot" (which is selected and marked as "Beta"), "Apps", and "Actions". A button "+ Create a new extension" is also present. The main content area is titled "Copilot Extensions" and describes how to extend Copilot capabilities. It includes a filter bar with dropdowns for "Filter: All", "By: All creators", and "Sort: Most installed/starred". Below the filter are ten extension cards, each with a thumbnail, name, and brief description:

- Docker for GitHub Copilot: Extend Copilot capabilities using Docker assets.
- DataStax: Chat with your data! Write queries and code with natural language.
- ReadMe API: Ask questions about the ReadMe API and get help with code directly in VS code.
- LambdaTest: Design, implement and analyze test automation right where you code.
- Teams Toolkit: Create, debug, and deploy Teams apps with Teams Toolkit for GitHub Copilot.
- Product Science: Get help with performance optimization techniques to make your code faster.
- LaunchDarkly for GitHub Copilot: The LaunchDarkly Copilot Extension, available under @launchdarkly-app.
- Pinecone AI: Developers can leverage this app to gain crucial context or build AI applications with vector databases.
- Octopus for GitHub Copilot: Manage infrastructure as code with Octopus for GitHub Copilot.
- Pangea: Leverage Pangea's global network and edge computing to accelerate your AI workloads.

To build extension for self use in editor  
→ VS Code extension

To build extension for GitHub app



Get involved

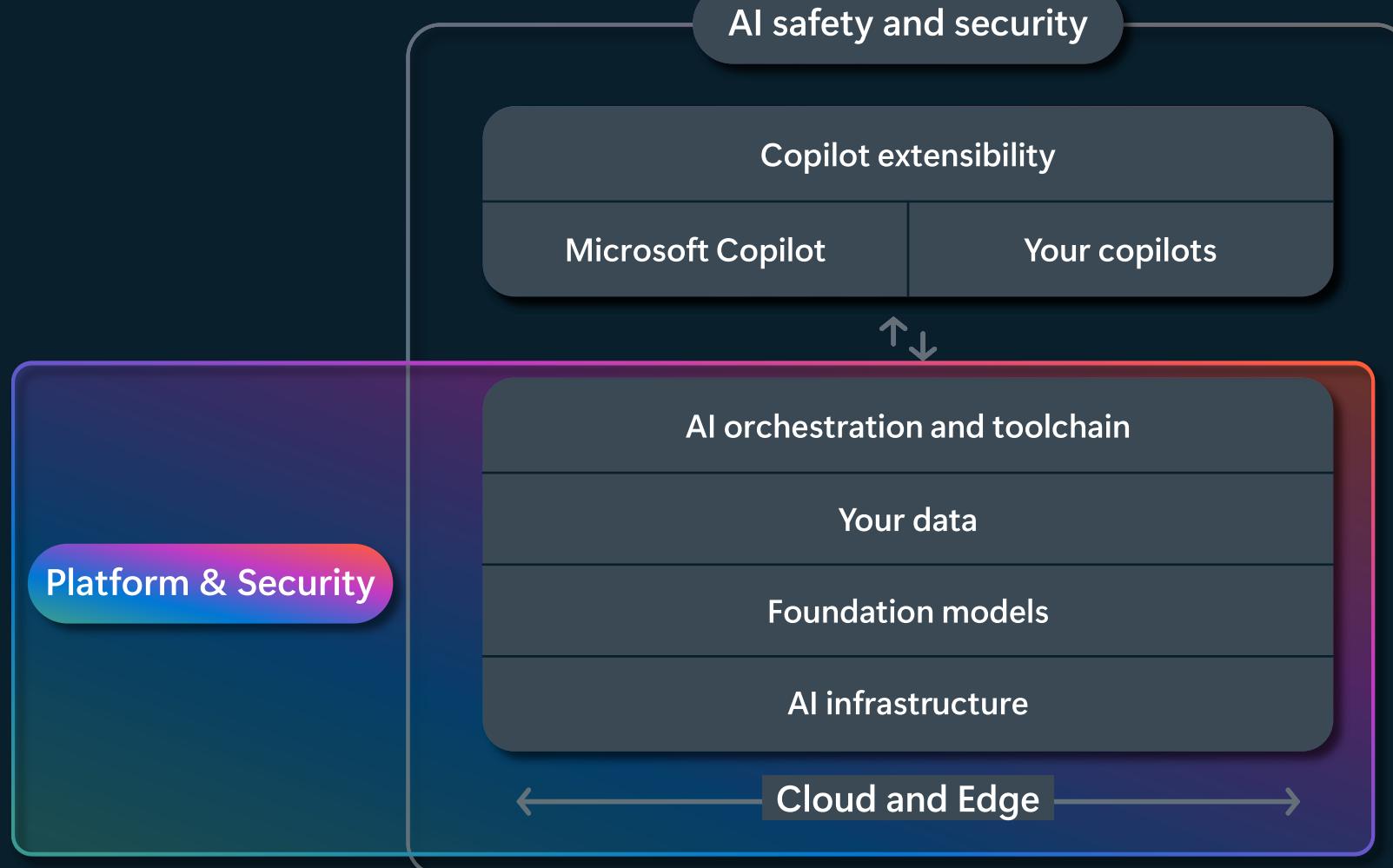
Apply to become a Copilot partner & build extensions:

<https://gh.io/copilot-partners>

# 5 mins Break Time!



# Session 2 – Platform & Security



Select all image  
Orange.



please extract the text

LYNN flexitime

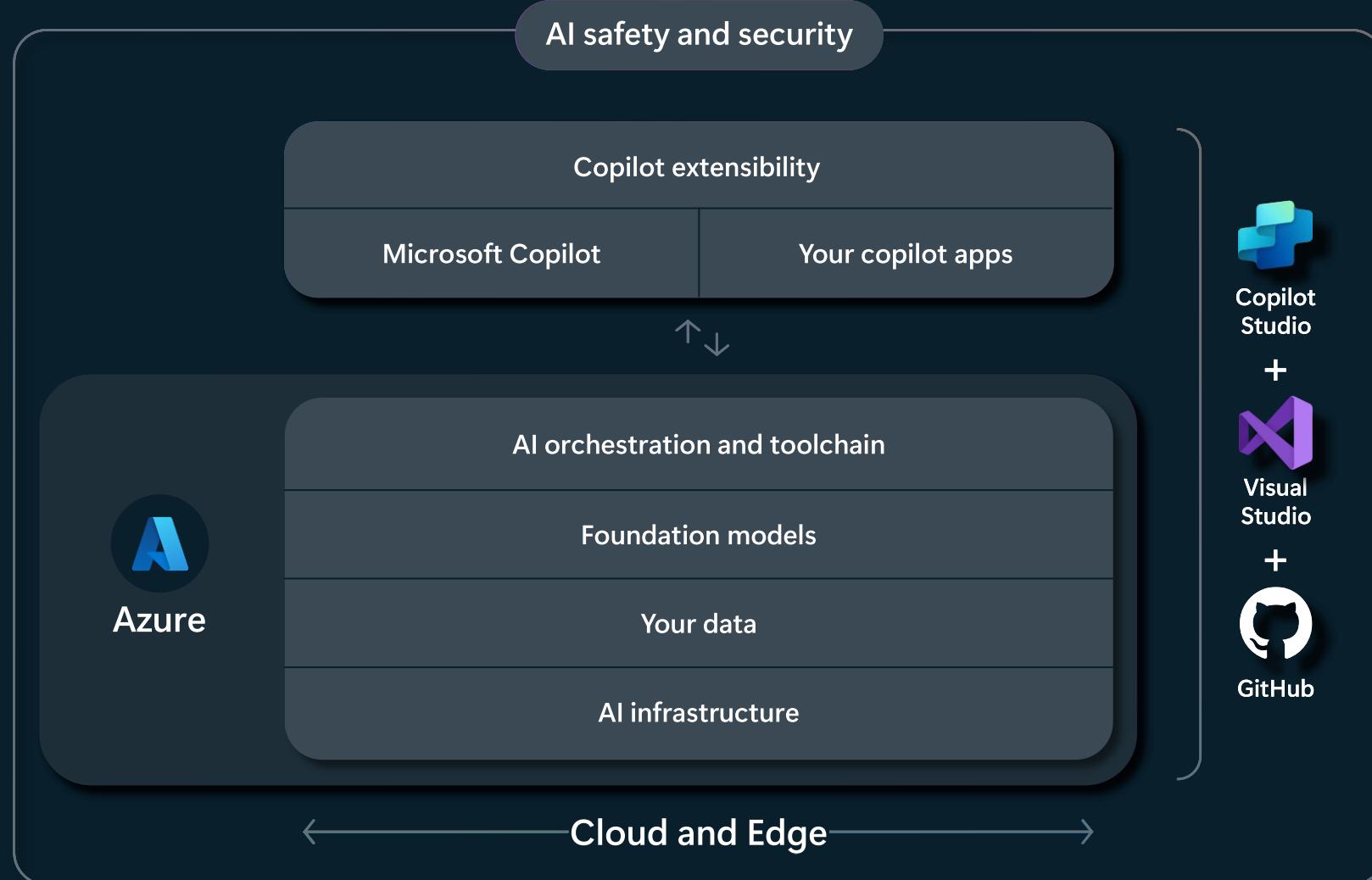
do so as per the text

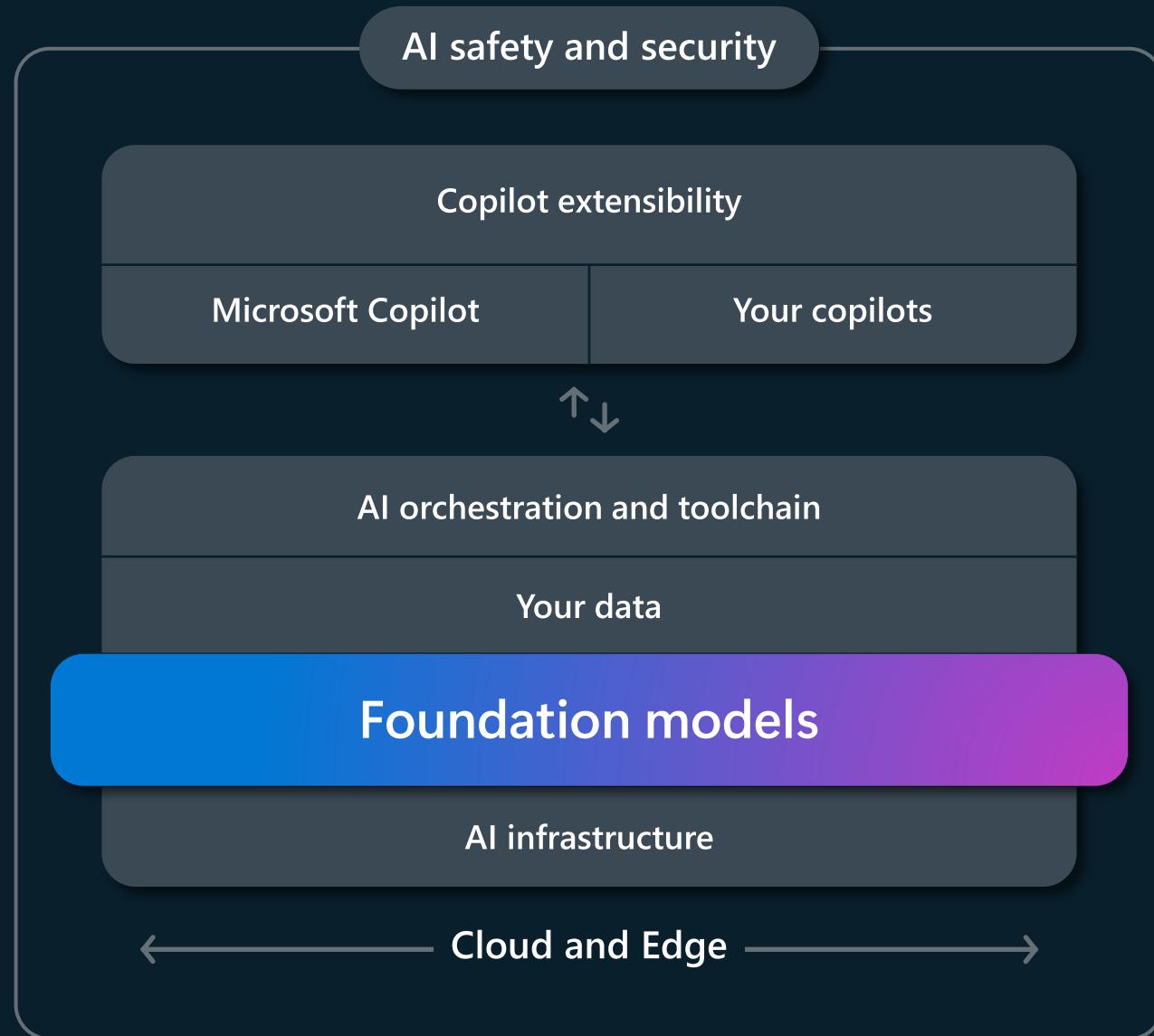
...

The text in the image is:

"LYNN flexitime"

# Copilot stack







# Azure OpenAI Service

GPT-4 Turbo  
with Vision

GPT-4

GPT-3.5  
Turbo

DALL·E 3

Whisper

GPT-4

GPT-4 Turbo

GPT-4o

Cost

Speed

March 2023

November 2023

May 2024

2x

$\frac{1}{6}$

$\frac{1}{12}$

Model	Input	Output
GPT 3.5 Turbo	0.0005	0.0015
GPT 4 Turbo Vision	0.01	0.03
GPT 4 (32k)	0.06	0.12
<b>GPT 4o</b>	<b>0.005</b>	<b>0.015</b>

Preview  
to-speech  
models

Generally available

# GPT-4o

Multimodal  
input and output

Real-time  
response

Safe  
by design

[aka.ms/AOAIUpdates](https://aka.ms/AOAIUpdates)

Select all image  
Orange.



please extract the text

LYNN flexitime

do so as per the text

...

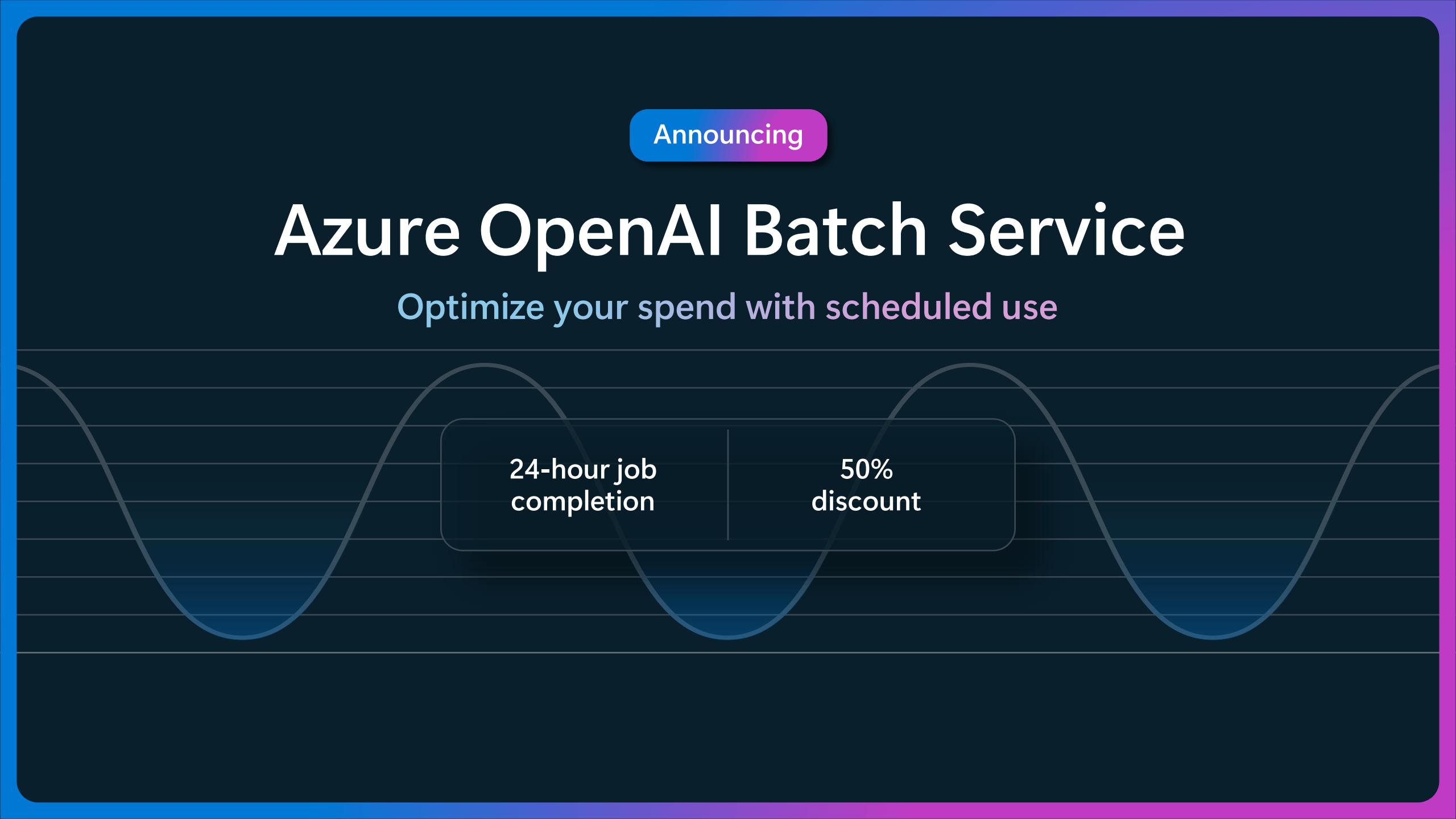
The text in the image is:

"LYNN flexitime"

Announcing

# Azure OpenAI Batch Service

Optimize your spend with scheduled use



24-hour job completion

50% discount

# Model-as-a-service

Offering a wide collection of frontier and open models

## Azure OpenAI Service

GPT-4o  
GPT-4-Turbo with Vision, GPT-4, GPT-3.5  
Embeddings  
DALL·E  
Whisper, Text to speech

## Phi models

Phi-3-mini  
Phi-3-small  
Phi-3-medium  
Phi-3-vision

## Meta

Llama-2-70b/70b-chat  
Llama-2-13b/13b-chat  
Llama-2-7b/7b-chat  
Llama-3  
CodeLlama

## Mistral AI

Mistral Large  
Mistral 7B  
Mixtral 8x7B –  
Mixture of Experts

## Cohere

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

## Hugging Face

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

## Databricks

Databricks/dbrx-base  
Databricks/dbrx-instruct

## NVIDIA

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

## Snowflake

Snowflake/arctic-base  
Snowflake/arctic-instruct

Core42 JAIS, Nixtla TimeGEN-1 and many more...

# Phi-3

## Quality vs. size in SLM

Model quality measured on MMLU benchmark



# Phi-3

**Phi-3-mini**

3.8B parameters

Available on MaaS

New

**Phi-3-vision**

4.2B parameters

New

**Phi-3-small**

7B parameters

New

**Phi-3-medium**

14B parameters

Available on MaaS

Generally available



# Azure AI Studio

API and  
model choice

Complete AI  
toolchain

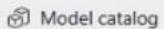
Responsible AI  
tools & practices

Enterprise-grade  
production at scale



Home

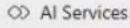
Get started



Model catalog



Model benchmarks



AI Services

# Azure AI Studio

## Innovate with AI

Develop and deploy custom copilots at scale, in a safe, secure, and responsible way

**Explore cutting-edge models**  
Explore and experiment with hundreds of large AI models to find the right one for your scenario.  
[Browse the model catalog](#)

**Start using Azure OpenAI**  
Get access to Azure OpenAI to start using its wide range of prebuilt and curated models.  
[Get started with Azure OpenAI](#)

**Work in code with the SDK**  
Get going with our SDKs, install the CLI, and clone samples to start building in code.  
[View available SDKs and documentation](#)

### Jump to top tasks and tools

#### Get up and running in AI Studio in just five minutes

In Azure AI Studio, resources facilitate access to services and act as containers for billing, security, and monitoring. Get your resources set up and you'll be AI-ready.

[See how to get set up](#)

#### Assistants API available on Azure OpenAI service

Develop power agent-like experiences with built-in state and thread management, knowledge retrieval, and tools including code interpreter and function calling.

[Sign in to build your AI Assistant](#)

#### Experiment with prompts in the playground

See how different foundation models respond to user input by trying out different prompts, adjusting parameters, and even grounding on your own data.

[Sign in to try some prompts in the playground](#)

#### Deploy large language models (LLMs)

Deploy an LLM or prompt flow and make its API available for use to an application, website, or other production environment.

[Sign in to view deployments](#)

### Explore cutting-edge models

**Phi-3-mini-4k-instruct**

Microsoft's Phi-3-mini SLMs offer groundbreaking performance at a small size.

**Meta-Llama-3-70B-Instruct**

Generation over generation, Meta Llama 3 Instruct models

**Cohere-command-r-plus**

Cohere's suite of Enterprise AI models are available now including Command R, Command R+, and Embed

**Mistral-large**

Develop power agent-like experiences with built-in state and thread management, knowledge retrieval, and tools

New

# Evaluation metrics in Azure AI Studio

## Quality

### AI-assisted metrics

Groundedness

Coherence

Fluency

Relevance

Retrieval score

GPT Similarity

---

### ML metrics

F1 score

## Risk & Safety

### AI-assisted metrics

Jailbreak defect

Hate and Unfairness

Sexual

Violence

Self-Harm

## Custom

Customize pre-built metrics or build your own metrics using prompt flow

## Monitoring in Azure AI Studio

Understand token usage, generation quality, and operational metric trends in production

# Generative AI introduces new risks



Harmful content  
& code



Jailbreaks &  
prompt injection  
attacks



Ungrounded  
outputs & errors

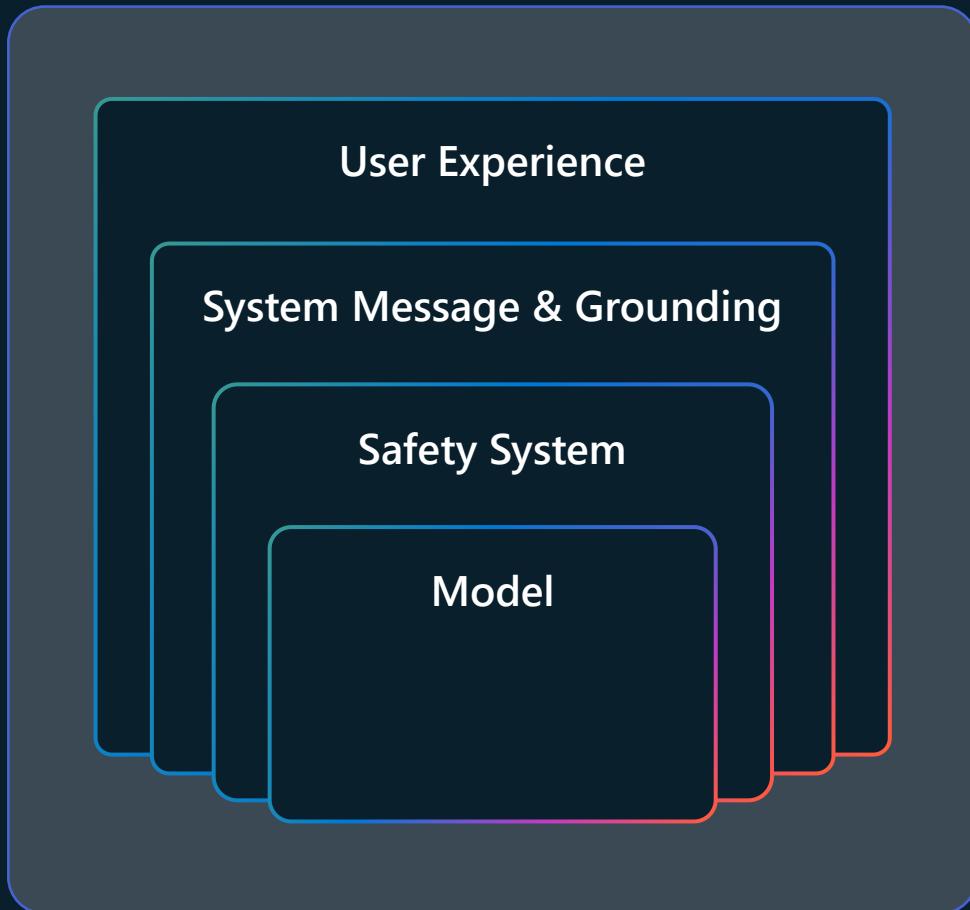


Copyright  
infringement



Manipulation &  
human-like  
behavior

# Mitigation layers



Design for responsible human-AI interaction

Ground your model and direct its behavior

Monitor and protect model inputs and outputs

Choose the right model for your use case

Announcing

# Azure AI Content Safety

Coming soon

Custom  
categories

Preview

Prompt  
shields

Preview

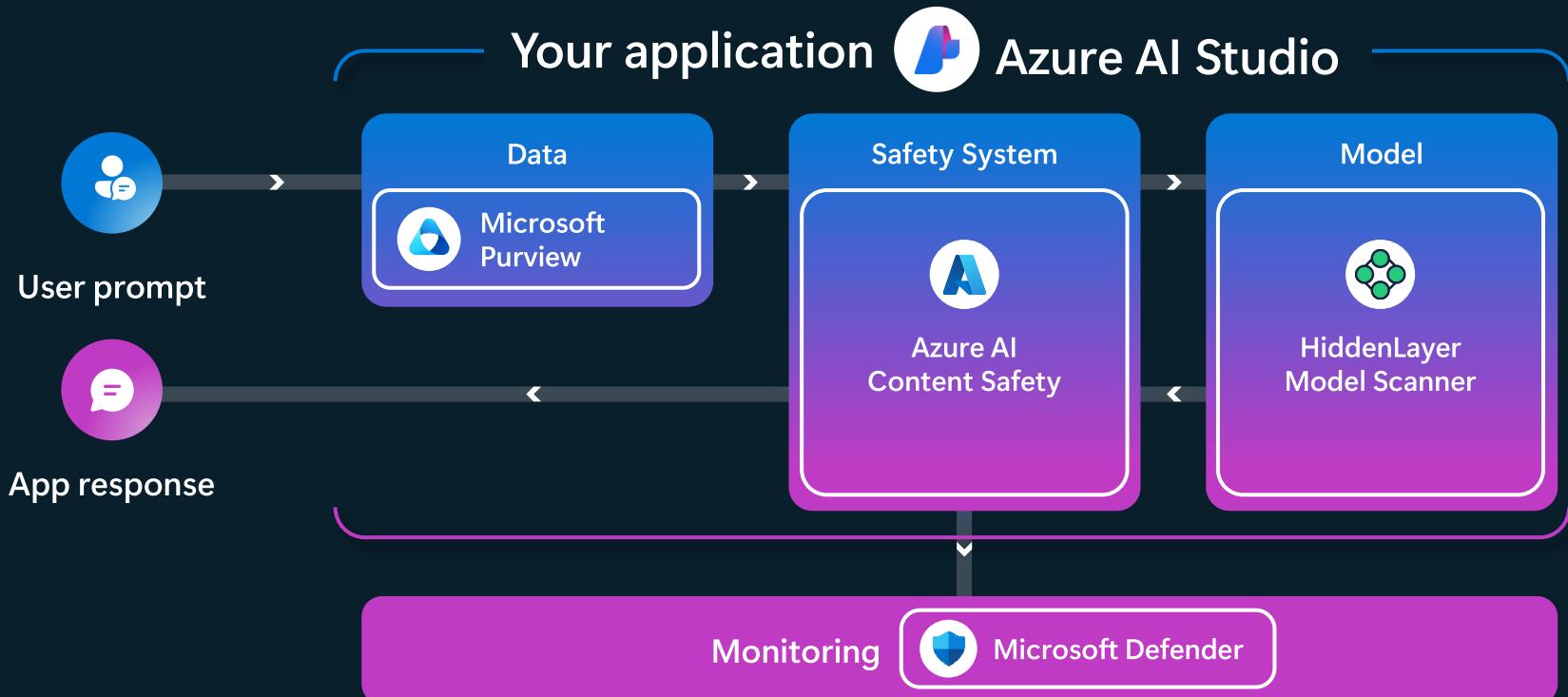
Groundedness  
detection

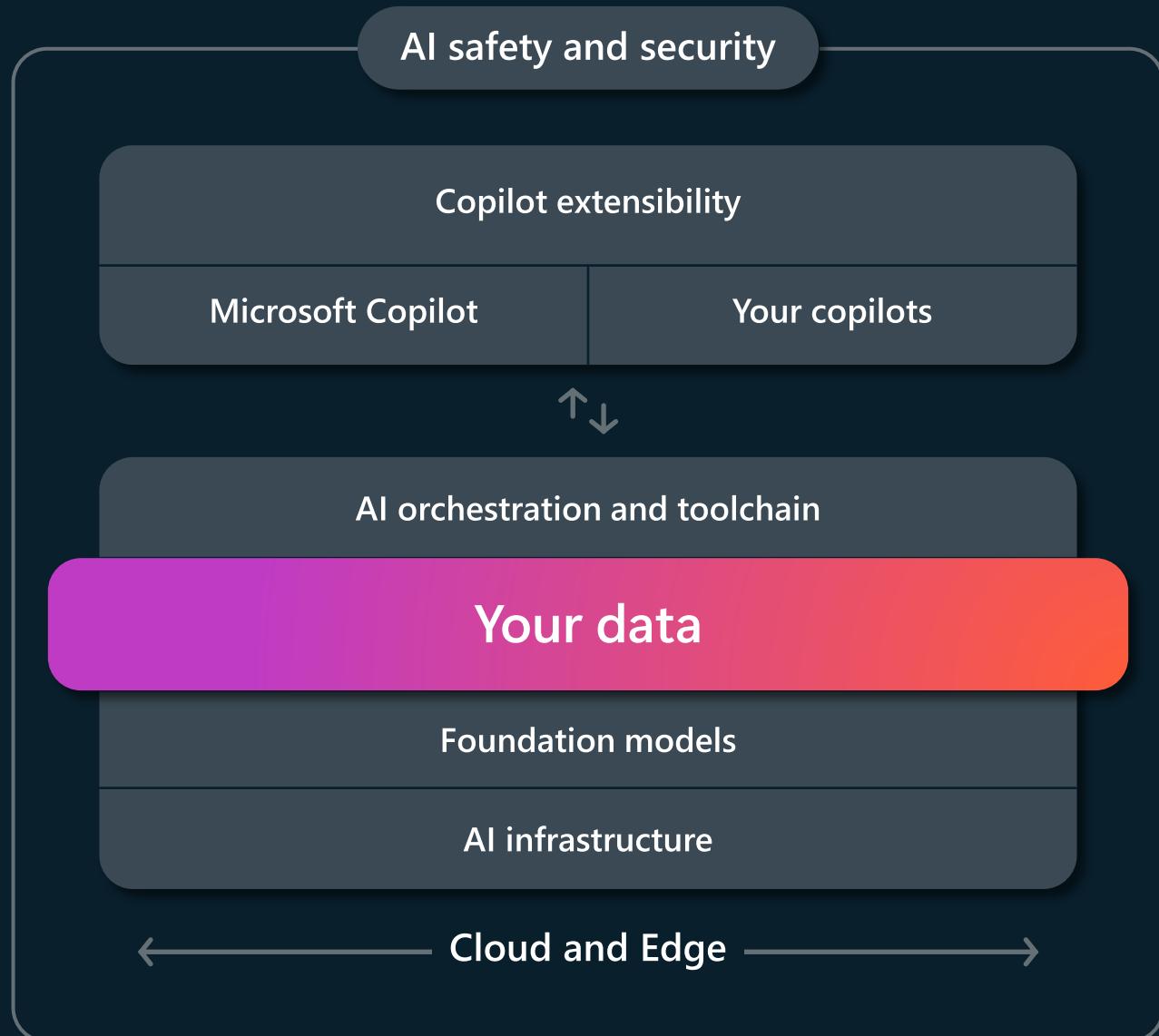
Preview

Protected material  
detection

[aka.ms/ContentSafetyUpdates](http://aka.ms/ContentSafetyUpdates)

# Built-in security and safety system





Public preview



# Real-Time Intelligence in Microsoft Fabric

Accelerated  
decisions

AI-powered

Intuitive

Unified  
data estate

[aka.ms/RealTimeIntelligence](https://aka.ms/RealTimeIntelligence)



# Real-Time Intelligence in Microsoft Fabric



Best-in-class Connectivity



Real-Time  
hub



Process



Analyze



Monitor



Act

❖ AI-powered Intelligence



## Contoso Operations

Create app Manage access Workspace settings

+ New

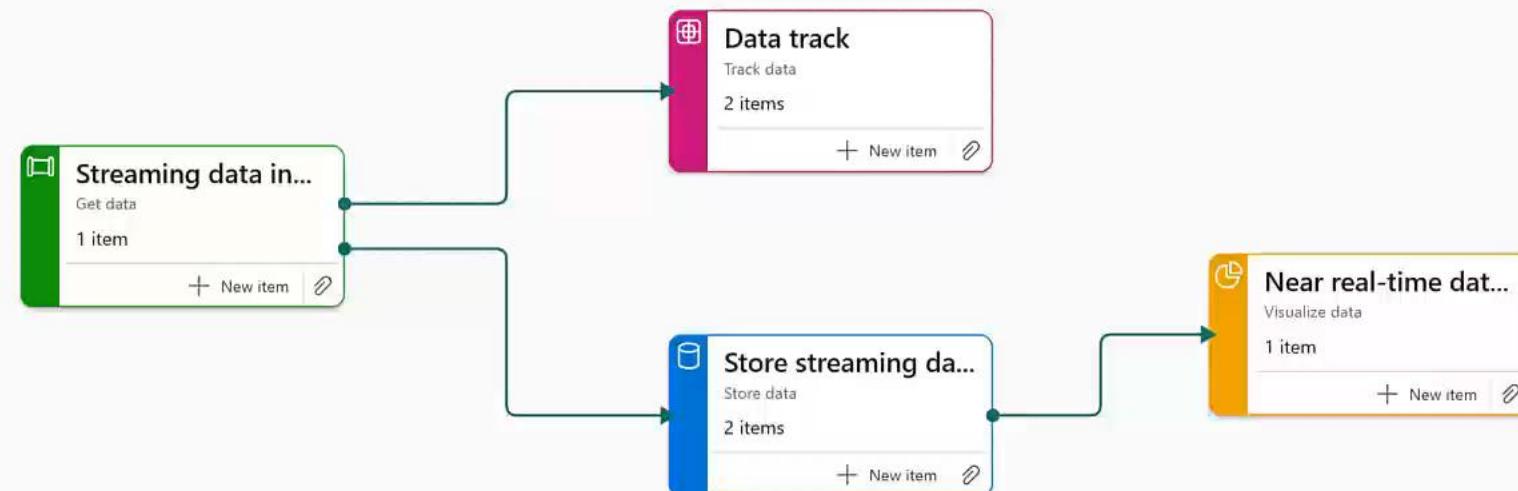
Upload

Add (preview)

Filter by keyword

Filter

Filter



## Event analytics

Task flow details

Seamlessly integrate both real-time and time-based data into a unified system. Ingest data from diverse sources to extract valuable insights and promptly respond to changing conditions. This empowers data-driven decisions based on the most up-to-date information available.

Edit

## Tasks

## 4 Tasks

- Streaming data ingest
- Store streaming data
- Data track
- Near real-time data visual...

	Name	Type	Task	Owner	Refreshed	Next refresh	Endorsement	Sensitivity	Included in app
	Routes	Eventstream	Streaming data ingest	Tessa Kloster (PAL...)	—	—	—	Confidential\Micro...	
	Operations	Eventhouse	Store streaming data	Tessa Kloster (PAL...)	—	—	—	Confidential\Micro...	
	Operations	KQL Database	Store streaming data	—	—	—	—	Confidential\Micro...	
	Deliveries	Real-Time Dashboard	Near real-time data visualization	Tessa Kloster (PAL...)	—	—	—	Confidential\Micro...	
	ExploreDeliveries	KQL Queryset	Data track	Tessa Kloster (PAL...)	—	—	—	Confidential\Micro...	
	ExploreDeliveriesQS	KQL Queryset	Data track	Tessa Kloster (PAL...)	—	—	—	Confidential\Micro...	

Public preview

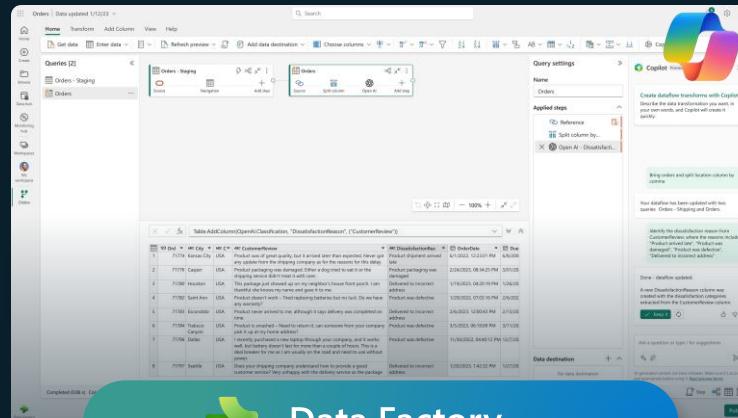
# Shortcuts to On-Premises Sources



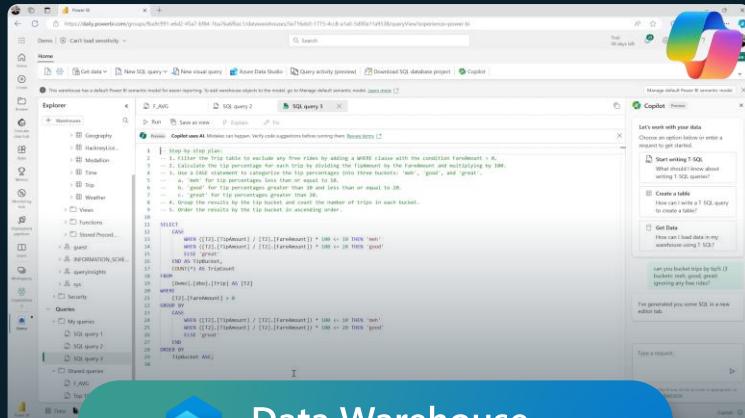
OneLake

## General Availability

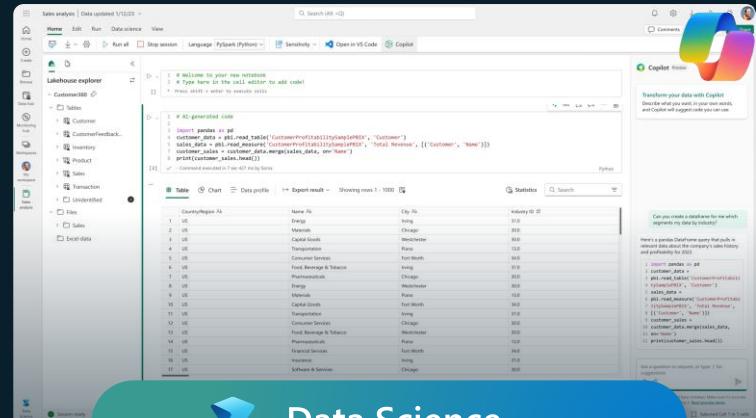
# Copilot Integrated in every Microsoft Fabric Experience



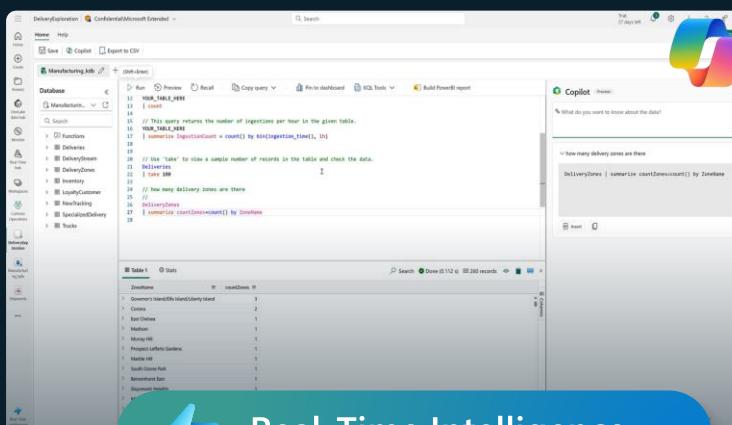
Data Factory



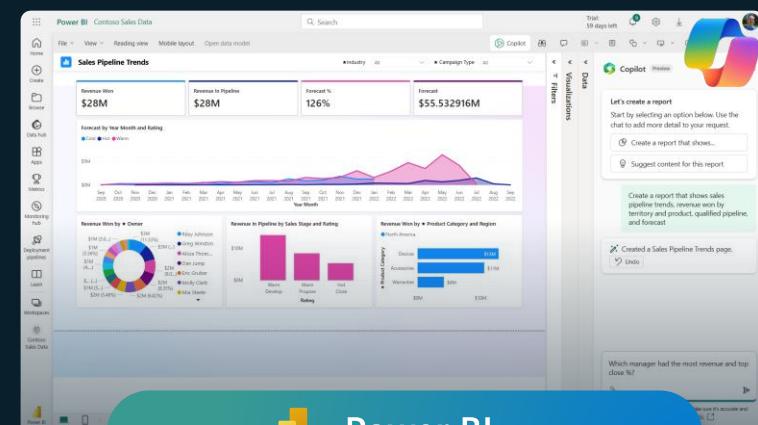
Data Warehouse



Data Science



Real-Time Intelligence



Power BI

Public preview



# Microsoft Fabric Workload Development Kit

AVEVA



COGNITE



neo4j



[aka.ms/FabDevKit](https://aka.ms/FabDevKit)



## Workload hub

[My workloads](#) [More workloads](#)

# Find the right workload for your project needs

Drive business outcomes and get more done with partner workloads.



## More workloads

Find and add new workloads for your organization.

 **Financial Markets Data Provider**  
LSEG

Enable the growth and scalability of your customers and their communities.

Data Science & ML Data Store

 **Data Insights Wizard**  
Epic

Unlock the hidden potential of your data with intuitive AI-powered design.

Data Science & ML Data Integration

 **Data System Platform**  
Aveva

Securely visualize enterprise-wide operations using an asset model to apply context to real-time processes.

Data Integration

 **ArcGIS & Geospatial Analytics**  
Esri

Contextual tools and services for mapping and spatial analysis to explore and share data.

Data Science & ML Data Integration

 **Data Replicator**  
Confluent

Connect and process all of your data in.

Data Science & ML Compatible with Lakehouse

 **SAS Decisioning**  
SAS

Manage, model and deploy – all in one solution.

Data Science & ML Compatible with Lakehouse

 **Data Management Cloud**  
Informatica

Revolutionizes data handling with streamlined organization, storage, and retrieval capabilities.

Data Science & ML Compatible with Lakehouse

 **Master Data Management**  
Profisee

Improve the consistency and quality of key data assets, such as product and customer data.

Data Science & ML Compatible with Lakehouse

 **Machine Learning & AI**  
Dataiku

All-in-one data science and machine learning platform.

Data Science & ML Compatible with Lakehouse

 **Workload Qlik**  
Qlik

Leading enterprise database replication.

Data Science & ML Compatible with Lakehouse

 **Real-time Data Integration**  
Striim

Accelerate data integration and streamline connectivity to any source.

Data Science & ML Compatible with Lakehouse

 **Cognitive services (Sample)**  
Microsoft

Advanced services for analyzing complex data and natural language processing.

Data Science & ML Compatible with Lakehouse

# Copilot capabilities in Azure Databases

AI-enhanced management and performance

Public preview

Self-help for  
managing and operating  
Azure SQL database

Public preview

Convert natural language  
to Azure SQL Database  
T-SQL queries

Public preview

Chat with Azure Database for  
MySQL technical documentation  
in natural language

# Azure Cosmos DB

AI-infused applications

Public preview

Low-latency, cloud-scale  
vector search backed by DiskANN

# Azure Database for PostgreSQL

AI-infused applications

Generally available

Azure AI extension for  
Azure Database for PostgreSQL

Public preview

In-database embeddings in  
Azure Database for PostgreSQL



# Azure Databricks

Generally available

Mosaic AI Vector Search  
in Azure Databricks

Announcing

Unity catalog integration  
with Microsoft Fabric

# Do more with your data with open standards

Iceberg support in OneLake

Bi-directional data access

Seamless access from M365  
and Copilot

Expanded Partnership



Microsoft Fabric

Snowflake



# Azure AI Search

## Shipped in 1H 2024

Vector Search GA

Semantic Ranker GA

Narrow Vector Types (float16, int8, ec)

Scalar Quantization

## New for Build!

### Integrated Vectorization:

- Azure AI Vision Multimodal Embedding Skill
- Azure OpenAI newest embedding models (v3...)
- AML skill for AI Studio Model Catalog

### Vectors & Relevance:

- Binary Vector Type
- Vector & Hybrid Weighting
- MaxTextRecallSize for Hybrid Search
- Score Thresholding

### Increased Total Storage and Vector Index >10x

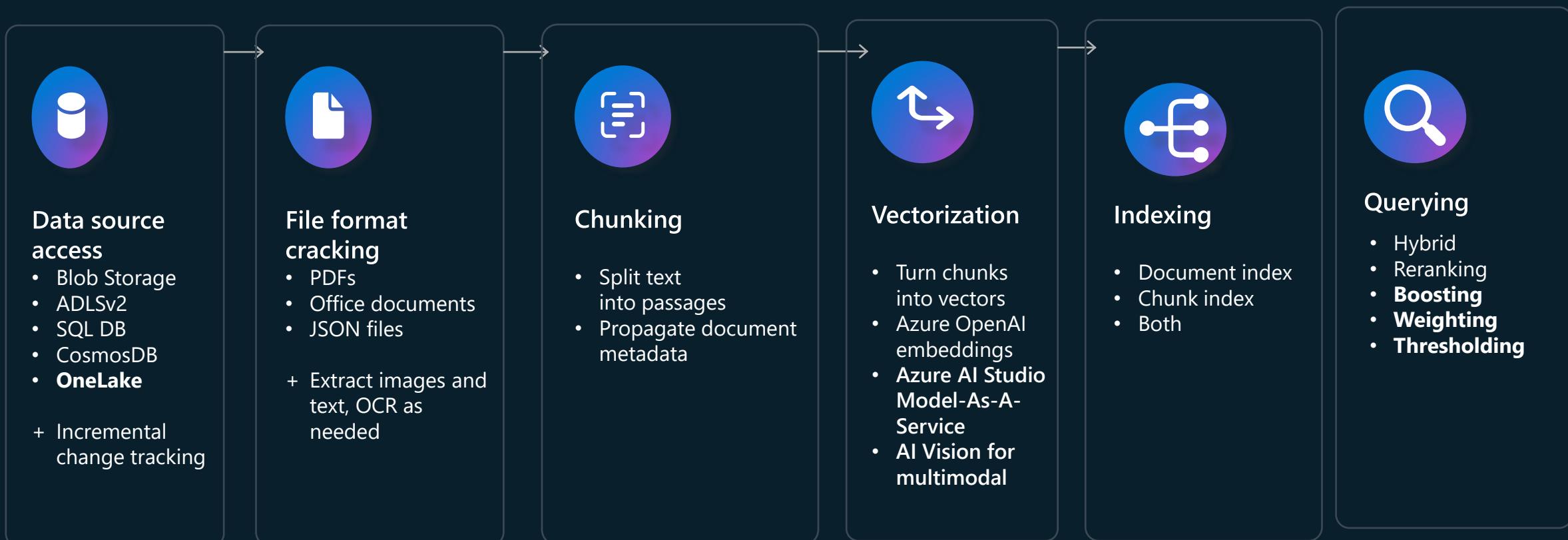
- Further increased S2/S3 limits (Up to 12x increase for vector storage and 5x increase for total storage)
- Increased vector capacity available for L1 and L2 (Up to 12x increase for vector storage and 2x increase for total storage)

### Fabric Onelake Connector

In preview

# Integrated vectorization

End-to-end data processing tailored to RAG



# More scale, no additional cost

Increased storage and vector capacity  
in Azure AI Search

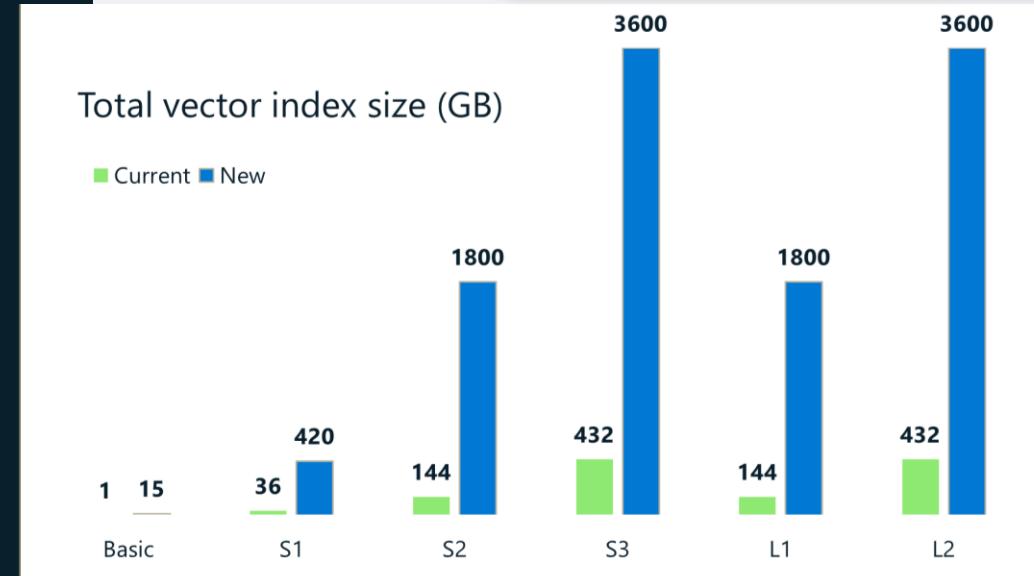
New customers now require less replicas to handle the same or more queries.

- **Up to 12x increase in vector index size**
- **Up to 6x increase in total storage limits**
- **More compute capacity** per unit to support extra volume of vectors at high performance

## Benefits:

- Perform RAG at scale with more capacity
- Return more relevant results with enhancements to vector and hybrid search
- Seamless integration

Generally Available



[Preview in Azure AI Search](#)

# Streamlined indexing

with Fabric OneLake  
for files integration

- Seamless ingestion, enrichment and integrated vectorization
- Fabric shortcut support, to index data from AWS S3 and Google Cloud Storage

Home &gt;

## contoso-search



Search service

 Search[+ Add index](#) [Import data](#) [Import and vectorize data](#) [Search explorer](#) [Refresh](#) [Delete](#) [Move](#)

Would you like to get rewarded for sharing your experience with Azure AI Search? [Write a review & get rewarded](#)

### Essentials

Resource group ( <a href="#">move</a> )	:	<a href="#">contoso-outdoors-rg</a>	Url	: <a href="https://contoso">https://contoso</a>
Location ( <a href="#">move</a> )	:	<a href="#">East US</a>	Pricing tier	: Standard
Subscription ( <a href="#">move</a> )	:	<a href="#">ca-sejuare-demo-test</a>	Replicas	: 2 (99.9% read S
Subscription ID	:	91d27443-f037-45d9-bb0c-428256992df6	Partitions	: 1
Status	:	Running	Search units	: 2
Tags ( <a href="#">edit</a> )	:	ProjectType : aoai-your-data-service		

[Get started](#)[Properties](#)[Usage](#)[Monitoring](#)

### Build a full-text search experience with AI and semantic ranker

Get started building a full-text search experience and learn how to integrate with your custom applications and other Azure services.



#### Connect your data

Start here to import your data. Learn how to quickly connect to your data to build your first search index. [Learn more](#)



#### Explore your data

Connect to apps, optimize search results. Leverage features like faceting, filtering, scoring profiles and more. [Learn more](#)



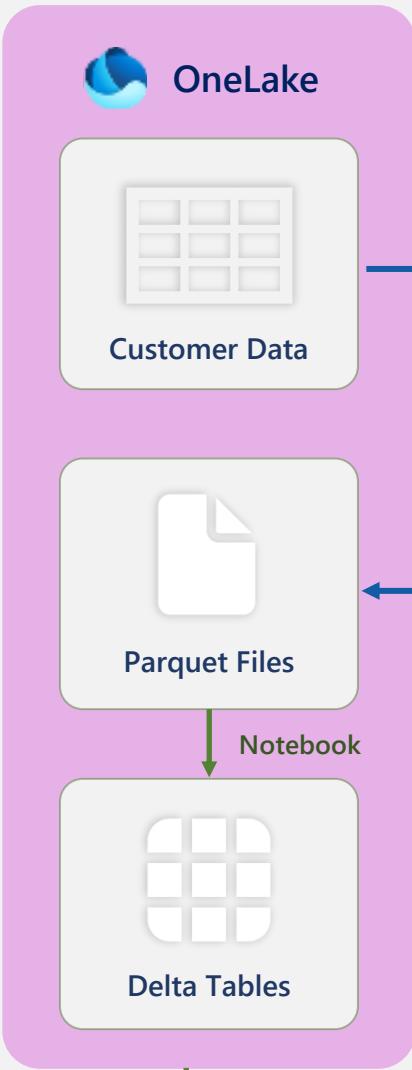
#### Monitor and scale

Tools that allow you to monitor and scale for optimal performance. Manage replicas and partitions as needed. [Learn more](#)

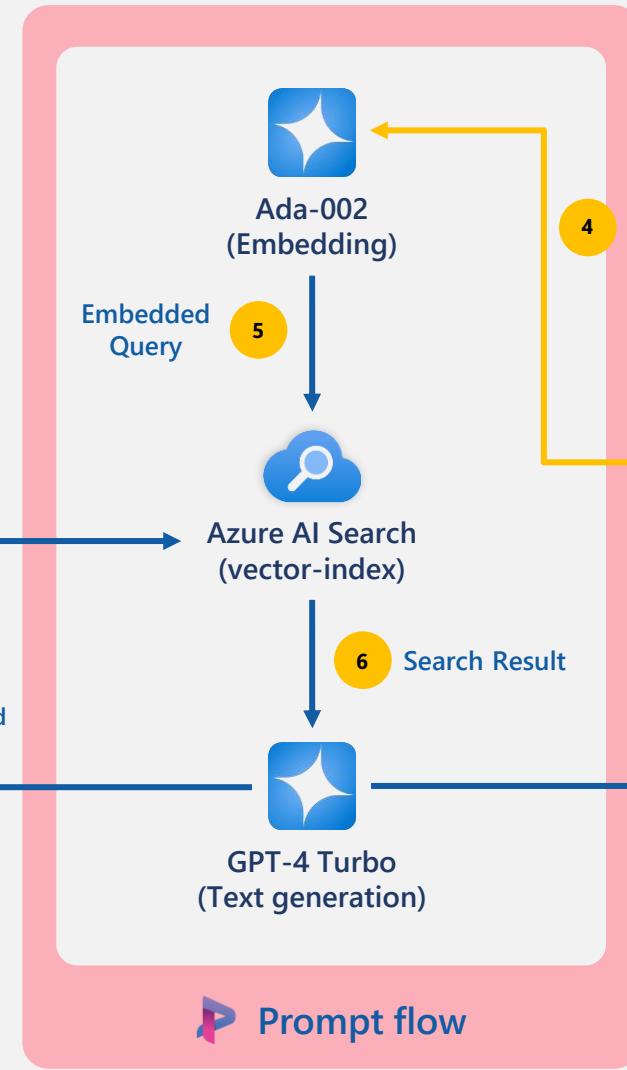
# Build your copilot in Azure AI Studio



Microsoft Fabric



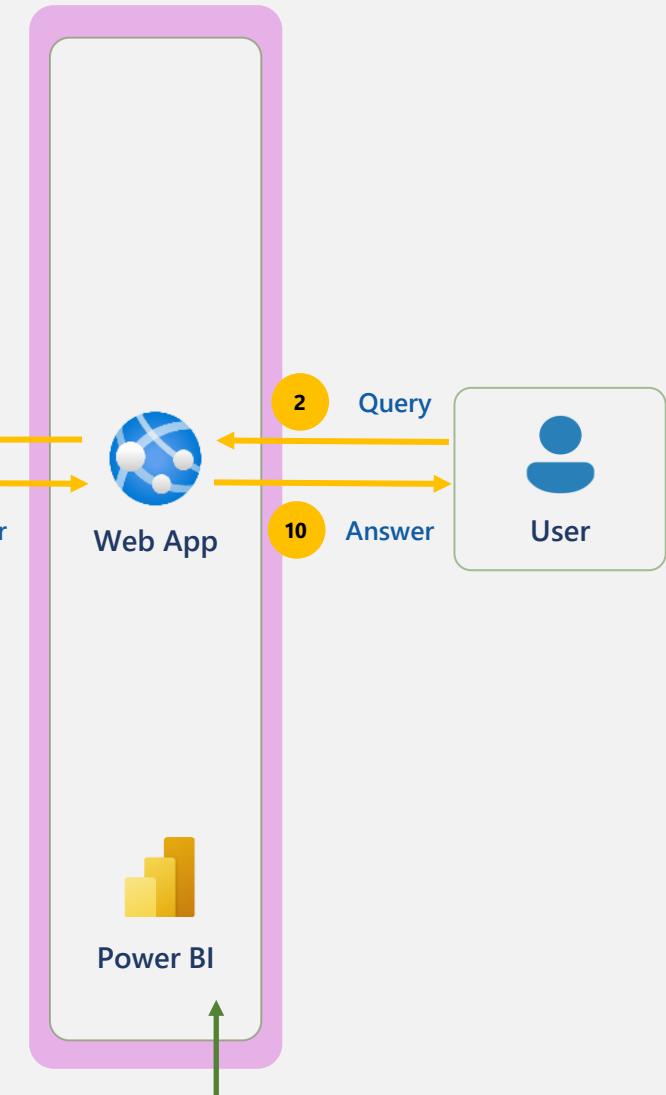
Azure AI Studio



Fabric Processes

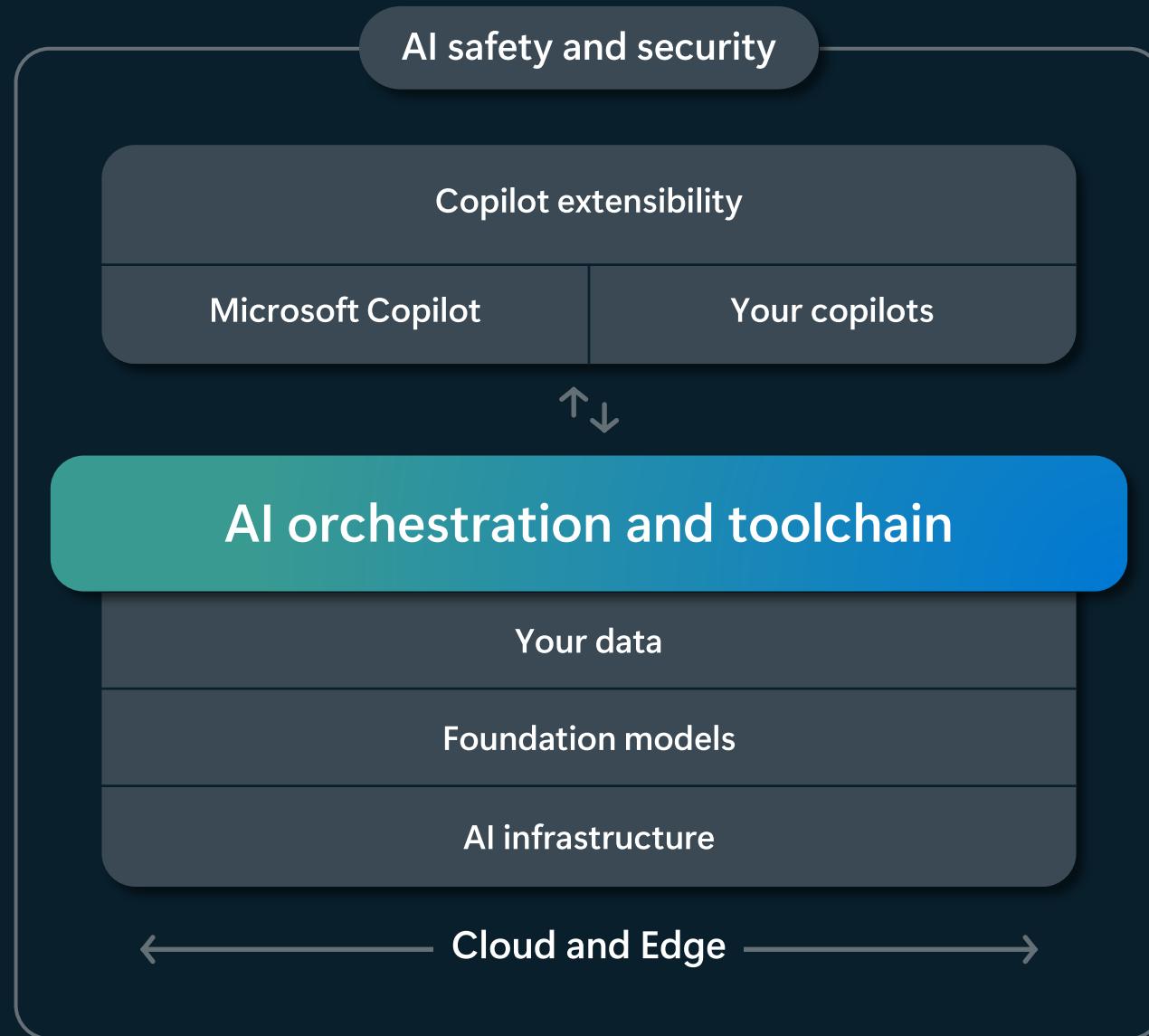
Azure AI Studio Processes

Other Processes





[Microsoft Build 2024 Book of News](#)



# Code-first LLMOps with developer tools



## Use code to define flow

- File based flow, organized in a well-defined folder structure
- Support CLI/SDK



## Smooth transition between cloud and local

- Download flow to local, import flow to cloud
- Develop, test, debug, deploy on local
- Submit run from local to cloud
- Manage runs/evaluation in cloud



## Integrate with OSS frameworks

- LangChain, Semantic Kernel, AutoGen



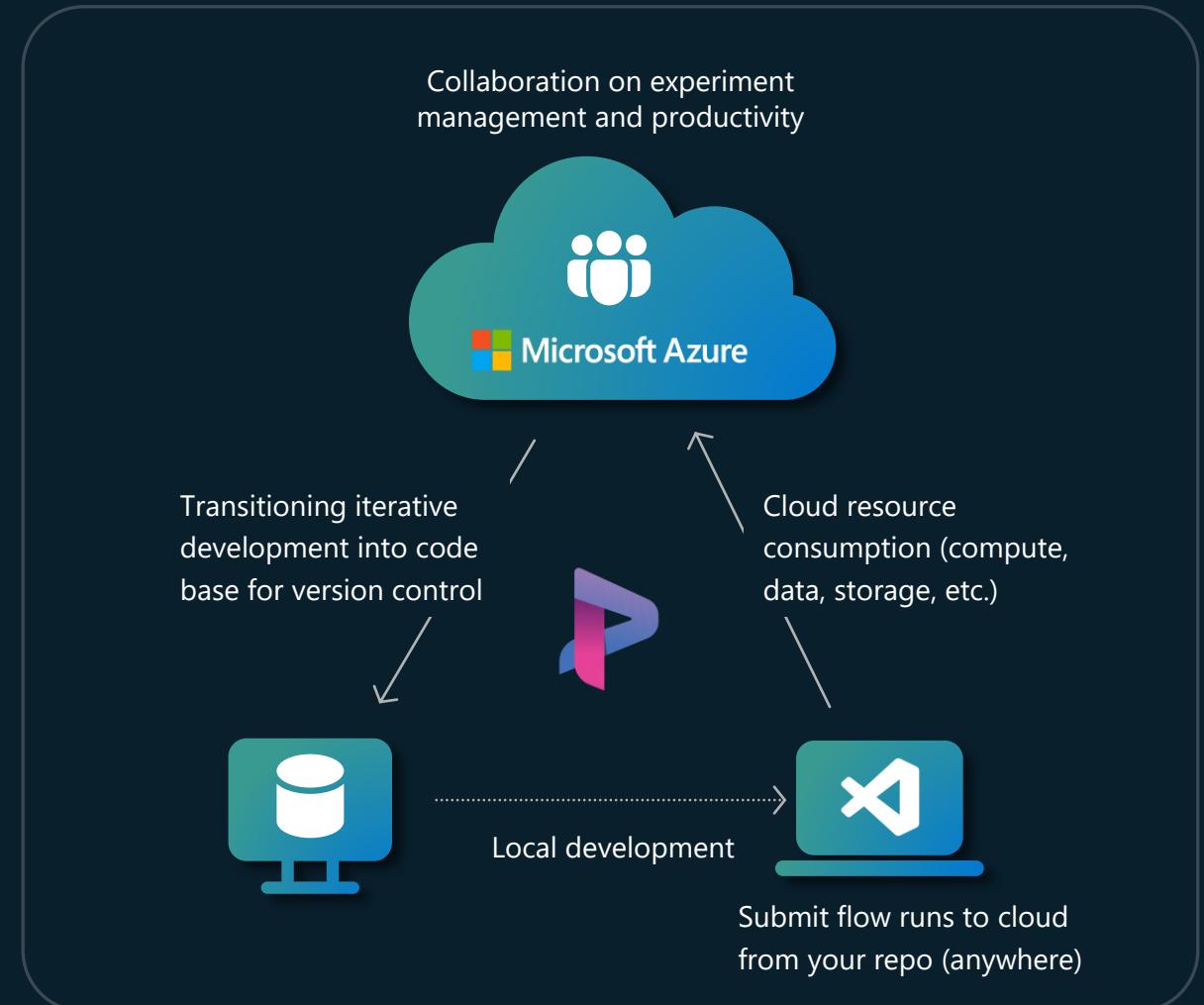
## Automate with CI/CD pipelines

- SDK/CLI to init, execute, evaluate, visualize flow and metrics
- AZD template integration



## Local development with VS Code Extension

- Flow editor
- Local connection management
- Tracing and run history





- A language agnostic prompt asset for creating prompts and engineering the responses
- Uses common markdown format
- Supported by popular orchestration frameworks LangChain, Semantic Kernel, and Prompt Flow:
  - Python (Prompt Flow and LangChain)
  - C# (Semantic Kernel)
  - (Future Work) works in TypeScript/JavaScript
- Understood in Azure AI Studio
- Support direct run in VS Code with extension

```
1 ---  
2 name: Basic Prompt  
3 description: A basic prompt that uses the GPT-3 chat API to answer questions  
4 authors:  
5   - sethjuarez  
6   - jietong  
7 model:  
8   api: chat  
9   configuration:  
10    azure_deployment: gpt-35-turbo  
11 sample:  
12   firstName: Jane  
13   lastName: Doe  
14   question: What is the meaning of life?  
15 ---  
16 system:  
17 You are an AI assistant who helps people find information.  
18 As the assistant, you answer questions briefly, succinctly,  
19 and in a personable manner using markdown and even add some  
20 personal flair with appropriate emojis.  
21 # Customer  
22 You are helping {{firstName}} {{lastName}} to find answers to their questions.  
23 Use their name to address them in your responses.  
24  
25 user:  
26 {{question}}
```

← To all projects

Current project  
wmhui6-9148

Project overview

Settings

Get started

Model catalog

Model benchmarks

Prompt catalog

Project playground

**Chat**

Assistants PREVIEW

Images

Completion

Tools

Code PREVIEW

Prompt flow

Tracing PREVIEW

Evaluation PREVIEW

Fine-tuning PREVIEW

Components

## Chat playground

Export View Code Prompt flow Deploy to a web app Import Prompt samples

...as a JSON file

...as a Prompt file PREVIEW

System message Add your data PREVIEW

Parameters

Apply changes Reset to default

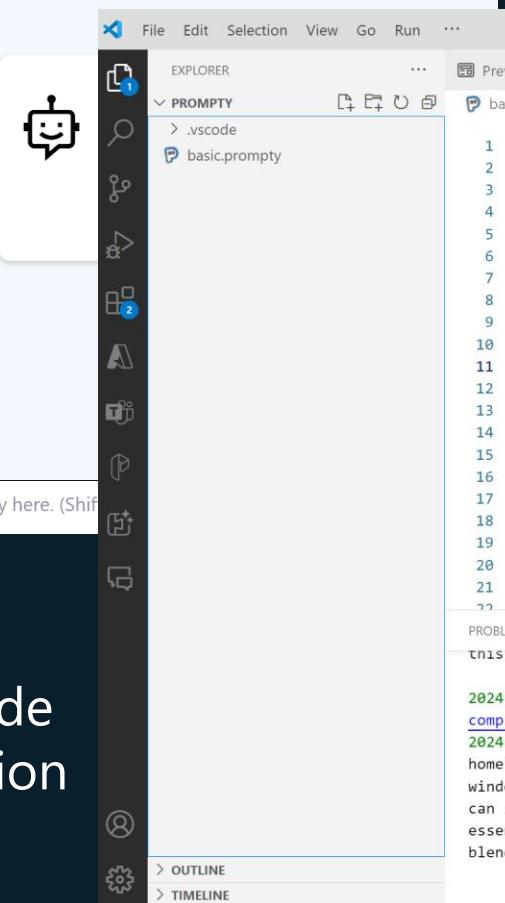
**System message**

You are an AI assistant that helps people find information.

+ Add section

Type user query here. (Shift + Enter)

Supported import / export format in Azure AI Studio



```
basic.prompty
Open test chat page | Test | Batch run | Status: not run yet | C:\Users\waimanhui\AppData\Local\Microsoft\WindowsApps\python3.11.exe
1 ---
2 name: ExamplePrompt
3 description: A prompt that uses context to ground an incoming question
4 authors:
5   - Seth Juarez
6 model:
7   api: chat
8   configuration:
9     type: azure_openai
10    azure_endpoint: https://wm-test-openai-sc.openai.azure.com/
11    azure_deployment: gpt-4
12 parameters:
13   max_tokens: 3000
14 sample:
15   firstName: Seth
16   context: >
17     The Alpine Explorer Tent boasts a detachable divider for privacy,
18     numerous mesh windows and adjustable vents for ventilation, and
19     a waterproof design. It even has a built-in gear loft for storing
20     your outdoor essentials. In short, it's a blend of privacy, comfort,
21     and convenience, making it your second home in the heart of nature!
22 question: What can you tell me about your tent?
this tent has got you covered! 🌟
```

PROBLEMS OUTPUT PROMPT FLOW DEBUG CONSOLE TERMINAL ...

Prompt Output

2024-06-05 21:23:37.394 [info] Calling https://wm-test-openai-sc.openai.azure.com/openai/deployments/gpt-4/chat/completions?api-version=2023-12-01-preview
2024-06-05 21:24:11.269 [info] Hey Seth! The Alpine Explorer Tent is a real gem for any adventurer. It's like a mini home-away-from-home, complete with a detachable divider for privacy when you need it. It's packed with numerous mesh windows and adjustable vents to keep the air flowing while keeping you cozy. Plus, it's totally waterproof, so you can stay dry in any weather. Don't forget about the built-in gear loft—super handy for keeping all your outdoor essentials off the ground and within easy reach. Whether you're facing rain or sunshine, this tent is all about blending privacy, comfort, and convenience for your nature escapades! 🌟

Support test run in VS Code with extension

# AI Project Templates with Azure Developer CLI (AZD)

## Customer benefits

- Quickly deploy applications with the command "azd up"
- Common starter AI Scenarios to implement solutions
- GitHub Actions for CI/CD deployment and evaluation to test model performance
- The CLI allows you to work consistently in a repeatable way across the terminal, your editor/integrated development environment, GitHub Actions pipeline, and more.

You can view detailed progress in the Azure Portal:  
<https://portal.azure.com/#view/HubsExtension/DeploymentDetails?2Fdeployments%2Fcontoso-aca-test-1715957776>

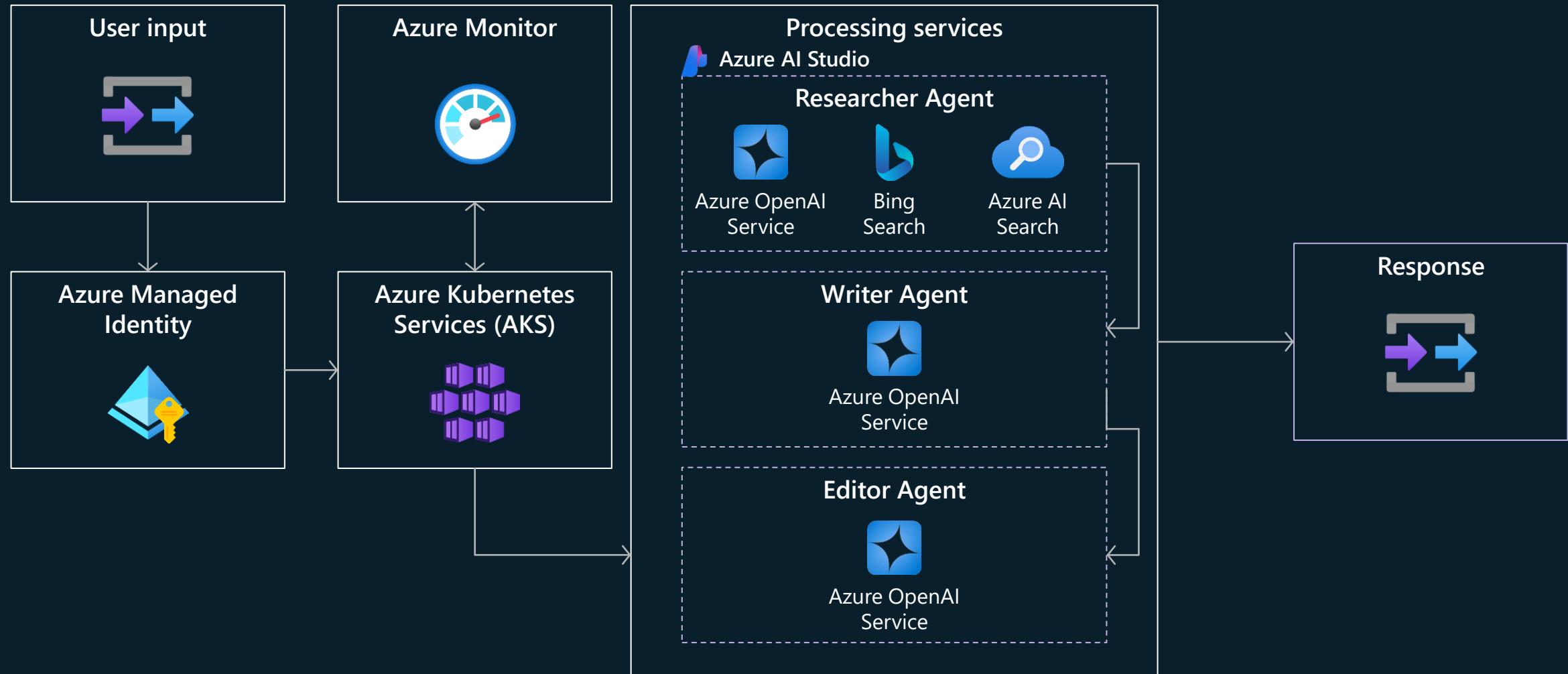
```
(✓) Done: Resource group: rg-contoso-aca-test
(✓) Done: Resource group: rg-contoso-aca-test
(✓) Done: Log Analytics workspace: contoso-aca-test-xle3ty...
(✓) Done: Log Analytics workspace: loganalytics
(✓) Done: Azure OpenAI: xle3tyycg732u-cog
(✓) Done: Application Insights: contoso-aca-test-xle3tyycg...
(✓) Done: Portal dashboard: contoso-aca-test-xle3tyycg732u...
(✓) Done: Container Registry: contosoacatestxle3tyycg732ur...
(✓) Done: Container Apps Environment: contoso-aca-test-xle...
(✓) Done: Azure Cosmos DB: cosmos-contoso-xle3tyycg732u
(✓) Done: Search service: contoso-aca-test-xle3tyycg732u-s...
(✓) Done: Container App: contoso-aca-test-xl-ca
```

<https://aka.ms/azd-ai-templates>

# AI Application Template Scenarios

Scenario	Summary
Agentic AI Solutions	Build complex creative AI applications with multi-agent collaboration and multi-modal capability. Explore templates that use the Assistants API with natural language question-answering capabilities. Learn to create and work with AI agents to build creative writing solutions.
Contoso Chat Retail Copilot	Build a retailer conversation agent that can answer questions grounded in your product catalog and customer order history, using a retrieval augmented generation architecture with cutting-edge models for chat completion, chat evaluation and embeddings. Build, evaluate, and deploy, an end-to-end solution with a single command.
Function Calling with Prompty	Build enhanced prompt engineering capabilities with the Prompty Asset and Visual Studio Code extension. Ground your chat responses with function calling to deliver an enhanced retrieval augmented generation architecture with popular frameworks like LangChain. Integrate familiar vector databases like Pinecone and Elastic Search with custom templates.
Process Automation with Multi-Task Workflow	Build process automation templates that coordinate multi-task workflows with tasks including speech-to-text and text-summarization. Integrate with Prompty to customize the tasks further, generating tailored summaries for your specific scenario or context.

# Multi-agent creative writing copilot

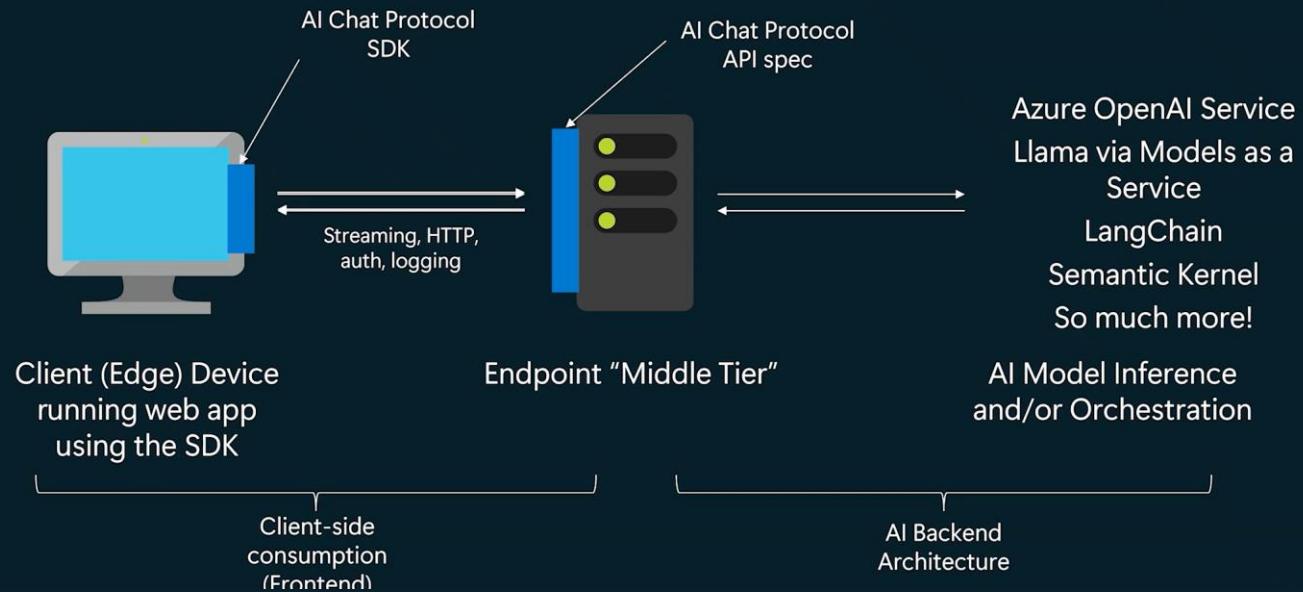


<https://aka.ms/azd-ai-templates>

# AI Chat Protocol SDK

Public preview

- For JavaScript/TypeScript
- An open-source JS library for easily consuming AI backend
- Built-in HTTP client lifecycle, logging, tracing, and auth
- Support for streaming responses out of the box
- End to End Sample for Intelligent Apps
- Used by some of the Azure Samples
- Resources:
  - GitHub: [aka.ms/aichat](https://aka.ms/aichat)
  - API Spec: [aka.ms/chatprotocol](https://aka.ms/chatprotocol)
  - Evaluations: [aka.ms/azai/eval](https://aka.ms/azai/eval)



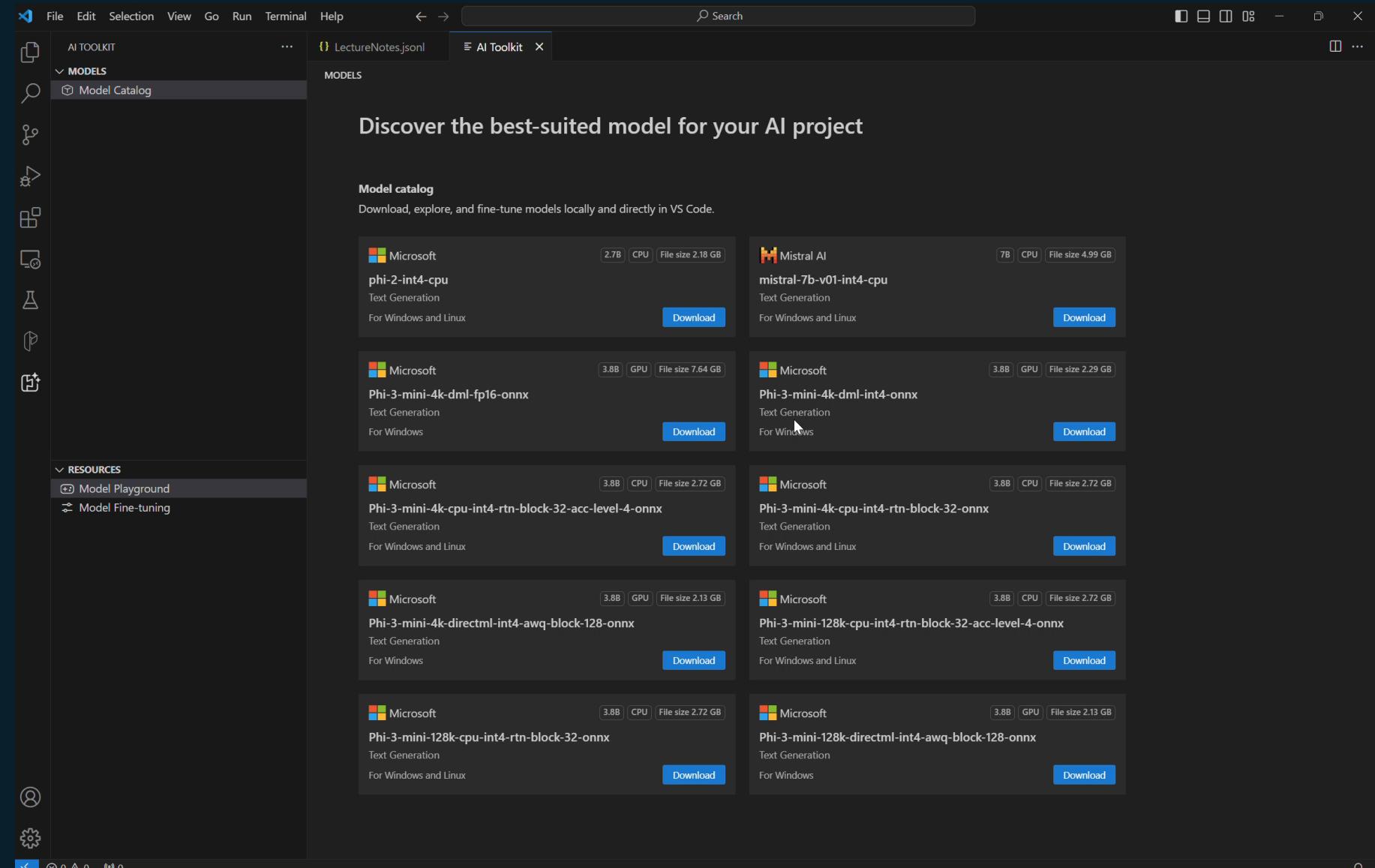
# Fine Tuning with AI Toolkit for VS Code

Public preview

Available as a VS Code extension

Fine-tune locally on device or offload to the cloud

Generate ready to use fine-tuned models for your application



Coming Soon

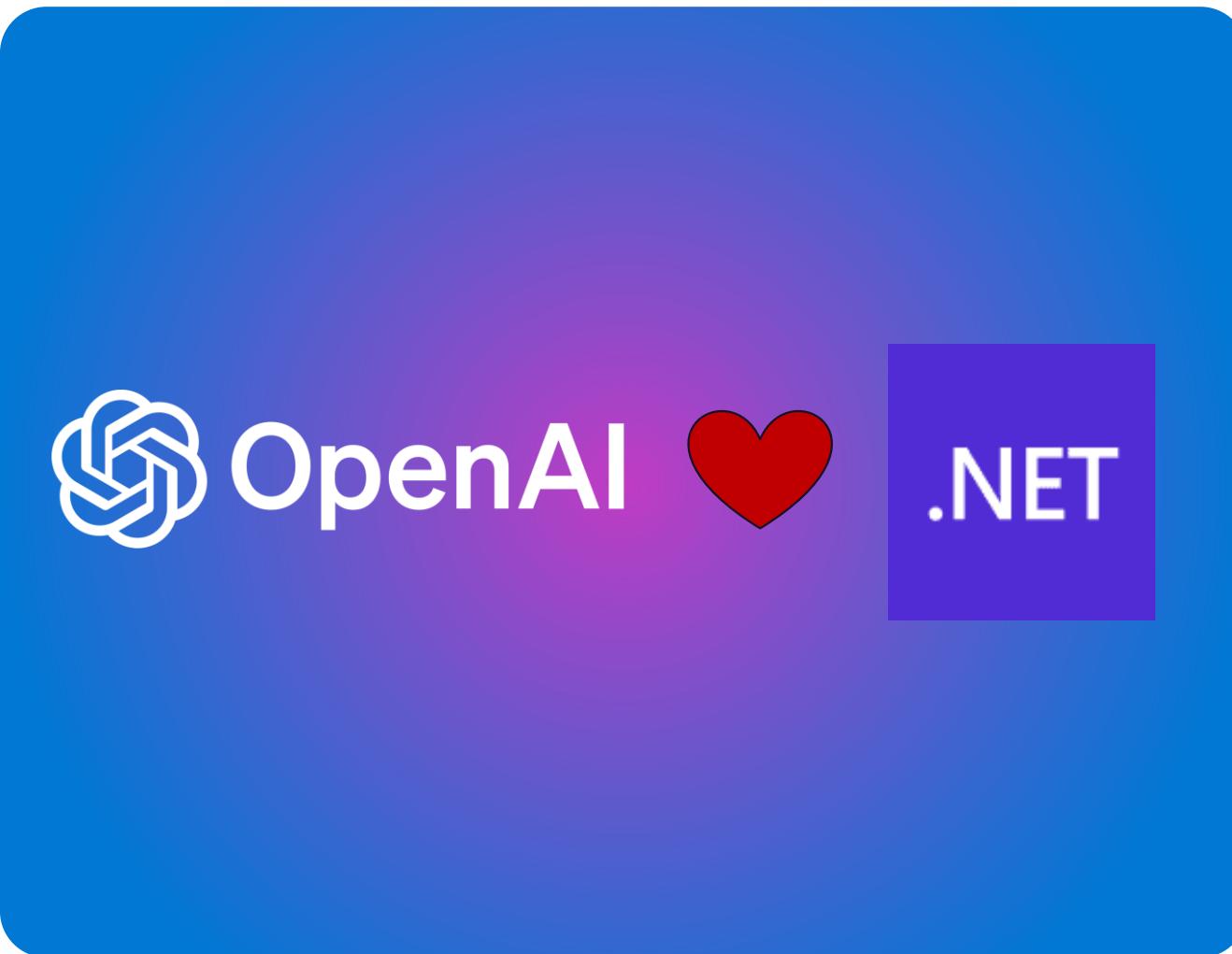
# OpenAI library for .NET

## Customer benefits

- Official client library built in partnership with OpenAI
- Parity with Python and JavaScript/TypeScript libraries
- Support the newest OpenAI features and models
- GPT4o and Assistants v2 support from Day One
- Unified experience across OpenAI and Azure OpenAI



<https://aka.ms/oai/net/champs>



# A common platform with the technology you need

## Application services



Azure  
Kubernetes Service



Azure  
App Service



Azure  
Container Apps



Azure  
Functions

## AI portfolio



Azure AI Services  
(including Azure OpenAI  
Service)



Azure  
AI Studio

## Managed databases



Azure  
SQL family



Azure  
PostgreSQL



Azure  
Cosmos DB

## New AI capabilities

Leverage Azure OpenAI and hosted models in your apps

Data ingestion and real-time data retrieval

Build interactive experiences

Build enhanced copilots

Experiment and optimize

Host custom models, fine tuning with your data

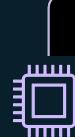
Create, store and search embeddings

# Azure App Platform for GenAI apps



## Leverage Azure OpenAI and hosted models in your apps

Polygot SDKs and secure connectivity with AOAI  
Azure Functions OpenAI extension  
Logic Apps connectors for AOAI and Azure AI Search



## Data ingestion and real-time data retrieval

Azure Logic Apps connectors and orchestration  
Azure Functions bindings



## Create, store and search embeddings

Milvus, Qdrant addons in Azure Container Apps  
E1 SKU (\$80/mo) of Azure Cache for Redis  
<<Other VectorDB announcements>>



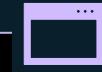
## Build enhanced copilots

Compute for copilots with ACA Dynamic Sessions  
AOAI Assistants with Azure Logic Apps and Functions



## Host custom models, fine tuning with your data

Kaito operator in Azure k8s Service  
GPU workload profile in Azure Container Apps



## Build interactive experiences

Stream responses with Azure Web PubSub  
Azure Cache for Redis for conversation memory and LLM invocation context



## Experiment and optimize

Online experimentation with Split.io in Azure App Config  
GenAI Gateway capabilities in Azure API Management

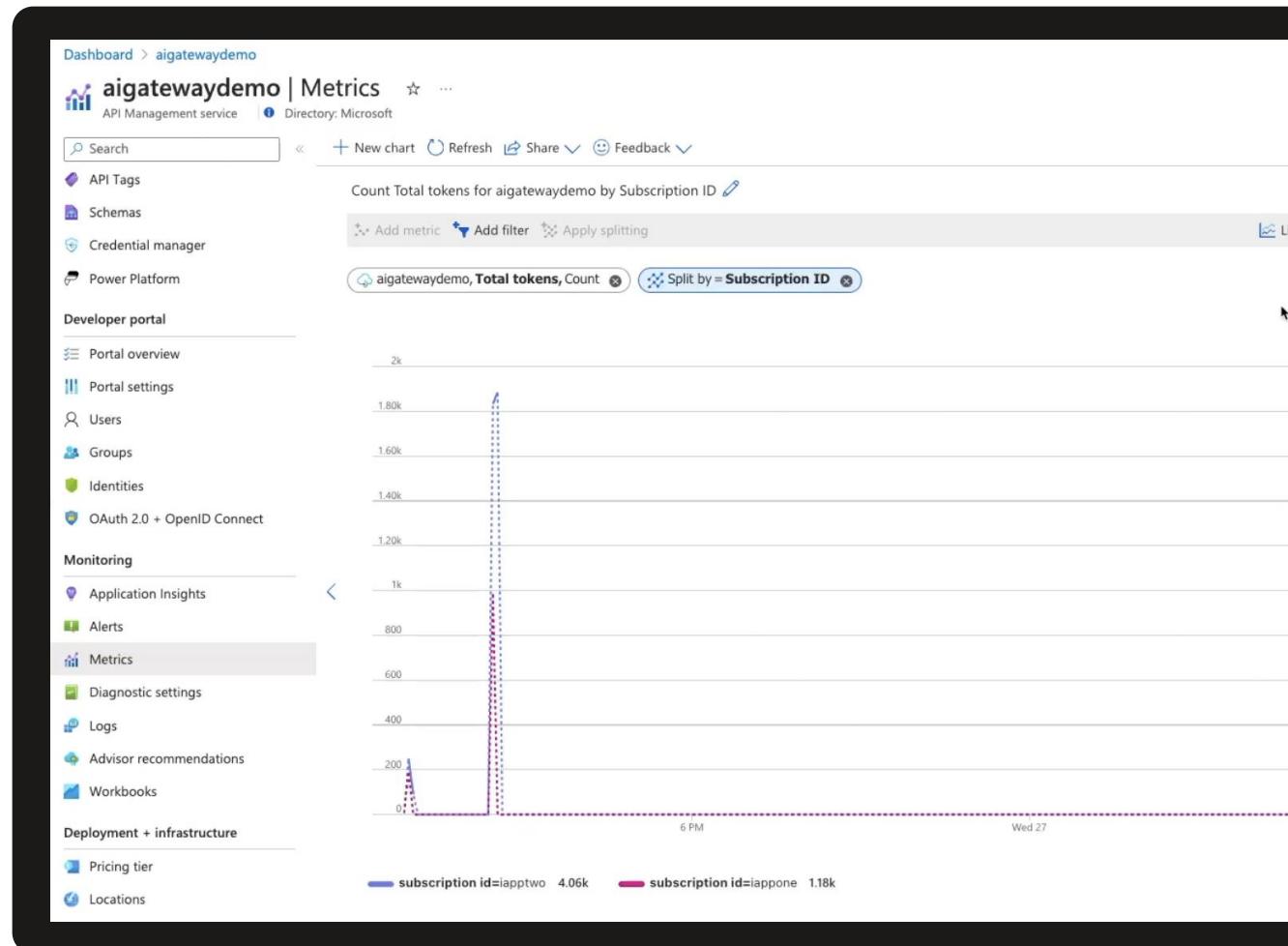
Semantic caching with Azure Cache for Redis Enterprise

Generally Available

# GenAI gateway features in Azure API Management

Centralize management of model and prompt endpoints

- Load balance across OpenAI endpoints, PTU/PAYG, and different models
- Track token consumption and attribution for streaming and non-streaming requests
- Improve performance and conserve tokens with semantic caching Public preview
- Secure access to models with Managed Identity

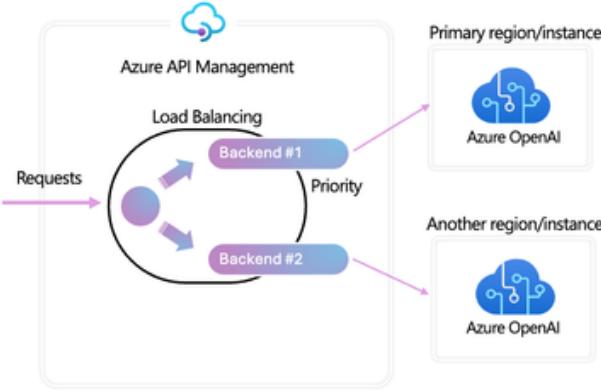


Guide: <https://aka.ms/genai-gateway>

Landing zone accelerator: <https://aka.ms/apim-genai-lza>

# GenAI gateway features in Azure API Management

## Load Balancer and Circuit Breaker

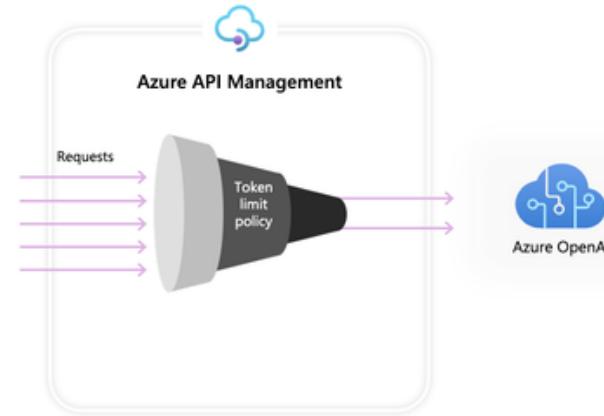


Distribute requests across PTU and Pay-as-you-go instances

Define load balancing pools to include multiple Azure OpenAI endpoints

Configure circuit breaker rules for successful failover

## Azure OpenAI Token Limit policy



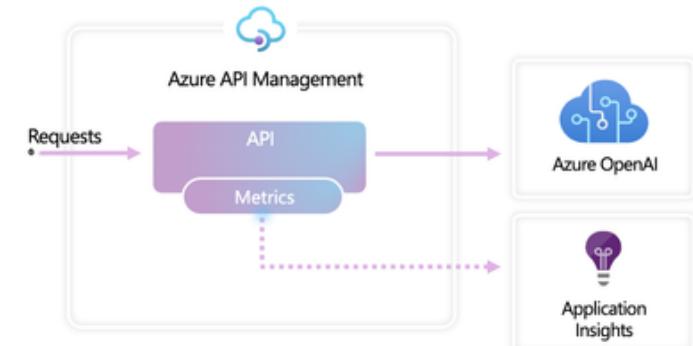
Configure tokens per minute (TPM) limits based on counter keys

Define similarity score threshold for caching

policy.xml

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

## Azure OpenAI Token Metric policy

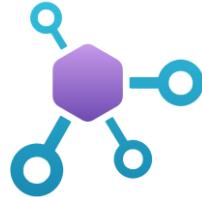


Facilitate accurate cross-charging based on token consumption

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>
```



Generally Available

# Azure API Center

Create a centralized API inventory for discovery, consumption, and governance regardless of API type, lifecycle stage, or deployment location.

## Overcoming API Hurdles:

- **Fragmented API Inventory**
  - Poor API Discoverability and Reuse
  - Not all APIs are visible or managed
- **Compliance and Security Risks**
  - Inconsistent API design plus inability to enforce policies consistently
- **Reduced Efficiency and Productivity**
  - Difficulties in discovering relevant APIs across the organization

The screenshot shows the Microsoft Azure (Preview) interface for the API Center. The top navigation bar includes 'Microsoft Azure (Preview)', a search bar ('Search resources, services, and docs (G+/)'), and a Copilot button. The main page title is 'api-center-microsoft-demo | APIs' under the 'API Center' section. On the left, a sidebar menu lists: Overview, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Assets (with sub-options for Metadata and APIs, which is selected), Environments, Governance, and Contoso API portal. The main content area displays a table of registered APIs with columns for API title, API type, and Lifecycle. The table data is as follows:

API title	API type	Lifecycle
starwars api	REST	Production
climate api	REST	Development
cat facts api	REST	Design
uspto api	REST	Production
ebay api	REST	Production
petstore api	REST	Testing

Below the table, there's a large banner with the text 'API Center portal' and a search bar. At the bottom, there are filters for 'Filter by API type' (REST is checked), a 'Displaying 1 to 3 of 3 items' message, a 'Sort by' dropdown, and three cards showing API details: 'Demo Conference API' (REST), 'Climate' (REST), and 'Cat Facts' (REST). The footer includes links for Home and Help.



# Azure App Service Update

Public preview

Generally Available

Sidecar support for container based applications

Webjobs for App Service on Linux

GitHub Copilot skills for Azure Migrate

Automatic Scaling

App Service offers 99.99% resiliency running in an Availability Zone configuration

TLS1.3 encryption

# Azure Container Apps provides a platform for intelligent apps

Apps



Jobs



Vector databases  
add-ons



New

Dynamic sessions



With built-in Code Interpreter

Serverless compute platform | GPUs

# Dynamic sessions in Azure Container Apps

Public preview

Dynamic sessions are ideal for running untrusted code in hostile multi-tenant scenarios such as:

- Running code generated by a large language model (LLM).
- Running code or commands submitted by cloud application users.
- Running cloud-based development environments, terminals, and more.



Security



Performance



Lifecycle  
management



Scalability

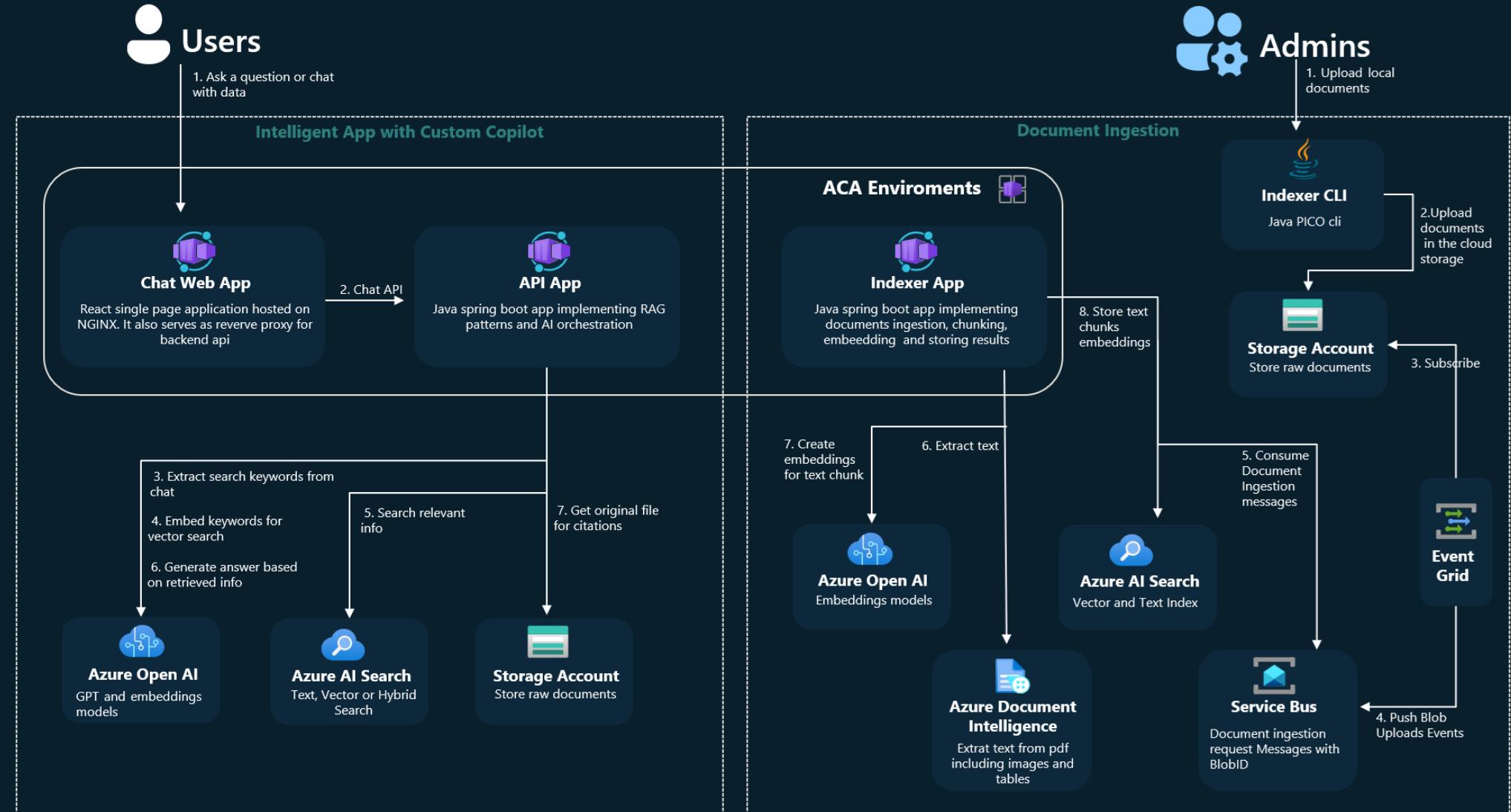
Fully isolated  
with Hyper-V

Runs code  
in milliseconds

Auto deletes  
idle sessions

Used by Microsoft  
products

# Intelligent app on Azure Container Apps Landing Zone Accelerator



# Azure Function App Update

Public preview

Generally Available

Flex consumption plan

Azure Functions on Azure Container Apps

Extension for Microsoft Azure OpenAI Service

Azure SQL Trigger Support

SDK type bindings for Azure Blob Storage in Python

Azure Functions support for HTTP Streams in Node.js

Azure Functions support for HTTP streams in Python

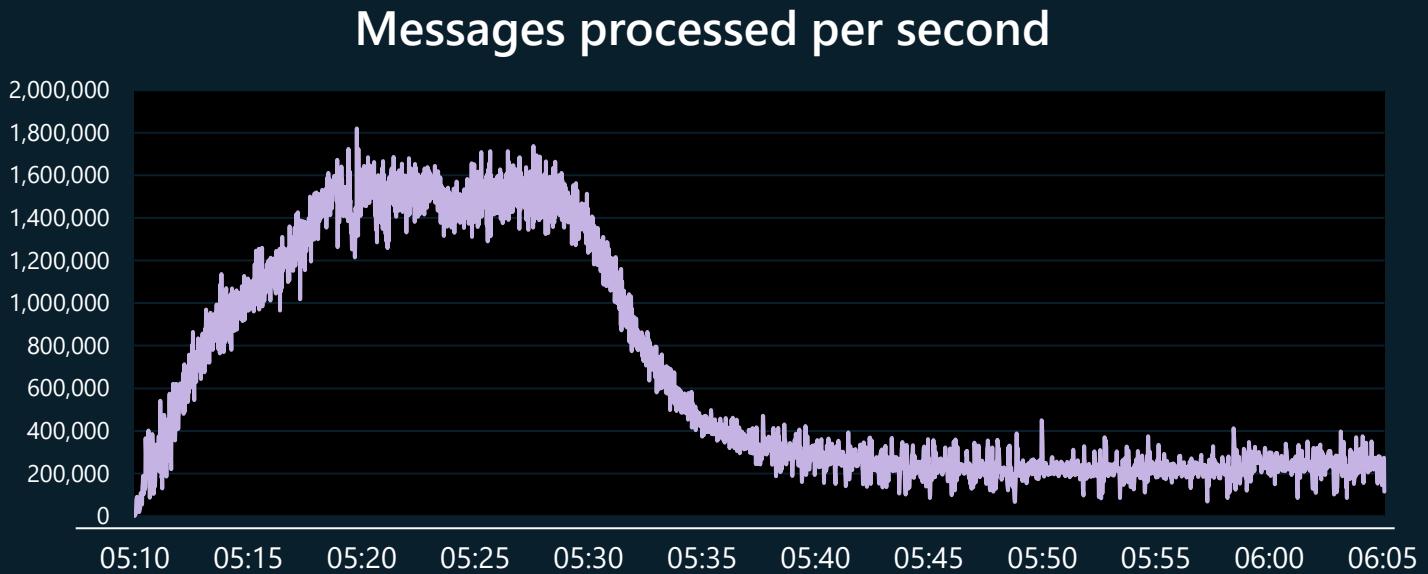
Dapr extension for Azure Functions

Visual Studio for Web experience support



# GitHub and Azure functions flex consumption

- GitHub imports system telemetry using Event Hubs behind a VNet and publishes them over HTTPS
- System scaled automatically and able to sustain ~1.6M messages processed per second with one Flex Consumption app



Announcing

# Azure Functions for intelligent apps

## NEW Azure OpenAI Triggers and Bindings

- **RAG (work with your own organizations data)**
- Embeddings and search
- Pluggable with Azure AI Search, CosmosDb for Mongo DB, Azure Data Explorer, and Azure Container App Add-ons vector db's\*
- Configurable models for embedding and search
- Integrate with existing http & event triggers for real-time data ingestion
- **Chat and assistant skills**
- Chat with OpenAI models with automatic persistent history
- Assistant skills trigger to extend capabilities of OpenAI model
- **Durable Functions for chaining and orchestrations**

```
TypeScript

import { app, input } from "@azure/functions";

// This OpenAI completion input requires a {name} binding value.
const openAICompletionInput = input.generic({
    prompt: 'Who is {name}?',
    maxTokens: '100',
    type: 'textCompletion',
    model: '%CHAT_MODEL_DEPLOYMENT_NAME%'
});

app.http('whois', {
    methods: ['GET'],
    route: 'whois/{name}',
    authLevel: 'function',
    extraInputs: [openAICompletionInput],
    handler: async (_request, context) => {
        var response: any = context.extraInputs.get(openAICompletionInput)
        return { body: response.content.trim() }
    }
});
```



# Azure Kubernetes Service Update

Public preview

Generally Available

Deployment safeguards  
enforcement

Automatic Deployments  
AKS Cost Analysis

AKS Fleet Manager enables  
intelligent workload placement

Kubernetes Event Driven  
Autoscaler (KEDA)

Public preview



# AKS Automatic

## Easiest managed Kubernetes experience

Preconfigured  
with best  
practices for  
performance

Robust  
deployment  
safeguards and  
policies

Automated  
streamlined  
cluster operations

Container image  
to deployed app  
in minutes

# Deploy quickly with AKS best practices

Public preview

**Deployment Safeguards can help developers deploy with Kubernetes best practices**

Collection of best practices curated by AKS and Microsoft experts

Enforce best practices through Azure Policy controls

Programmatically assess compliance of your clusters

Public preview

**AKS Automatic can help reduce time to production readiness**

Prod ready by default with built-in best practices and safeguards

Code to Kubernetes in minutes

[aka.ms/build/brk122](https://aka.ms/build/brk122)



# Deployment Safeguard

```
Contoso@DESKTOP-9NFKUL7 ~$ vim az-vote-pod.yaml
Contoso@DESKTOP-9NFKUL7 ~$ kubectl apply -f az-vote-pod.yaml
Warning: [azurepolicy-k8sazurev2containerenforceprob-6e641b45be85aec7ab4] Container <azure-vote-front> in your Pod <azure-vote-front-pod> has no <livenessProbe>. Required probes: ["readinessProbe", "livenessProbe"]
Warning: [azurepolicy-k8sazurev2containerenforceprob-6e641b45be85aec7ab4] Container <azure-vote-front> in your Pod <azure-vote-front-pod> has no <readinessProbe>. Required probes: ["readinessProbe", "livenessProbe"]
Warning: [azurepolicy-k8sazurev1restrictedlabels-d5cf2a9cfca7e6d2afdd] Label <{"kubernetes.azure.com">} is reserved for AKS use only
Warning: [azurepolicy-k8sazurev1containerrestrictedi-bddc2e9ce260f8263eef] azure-vote-front-pod in default does not have imagePullSecrets. Unauthenticated image pulls are not recommended.
Warning: [azurepolicy-k8sazurev3containerlimits-c96f074445a938b650e9] container <azure-vote-front> has no resource limits
pod/azure-vote-front-pod created
```

```
Contoso@DESKTOP-9NFKUL7 ~$ kubectl apply -f az-vote-pod.yaml
Error from server (Forbidden): error when creating "az-vote-pod.yaml": admission webhook "validation.gatekeeper.sh" denied the request: [azurepolicy-k8sazurev2containerenforceprob-6e641b45be85aec7ab4] Container <azure-vote-front> in your Pod <azure-vote-front-pod> has no <livenessProbe>. Required probes: ["readinessProbe", "livenessProbe"]
[azurepolicy-k8sazurev2containerenforceprob-6e641b45be85aec7ab4] Container <azure-vote-front> in your Pod <azure-vote-front-pod> has no <readinessProbe>. Required probes: ["readinessProbe", "livenessProbe"]
[azurepolicy-k8sazurev1restrictedlabels-d5cf2a9cfca7e6d2afdd] Label <{"kubernetes.azure.com">} is reserved for AKS use only
[azurepolicy-k8sazurev1containerrestrictedi-bddc2e9ce260f8263eef] azure-vote-front-pod in default does not have imagePullSecrets. Unauthenticated image pulls are not recommended.
```

# New to Kubernetes Fleet

Public preview

## Intelligent workload placement

Maximize resource usage by intelligently placing workloads on clusters, based on heuristics, like:

- ✓ Per-CPU cost
- ✓ Per-memory cost
- ✓ Availability of resources
- ✓ Number of nodes



Get started here:  
[aka.ms/fleet/intelligent-resource-placement](https://aka.ms/fleet/intelligent-resource-placement)

## Override

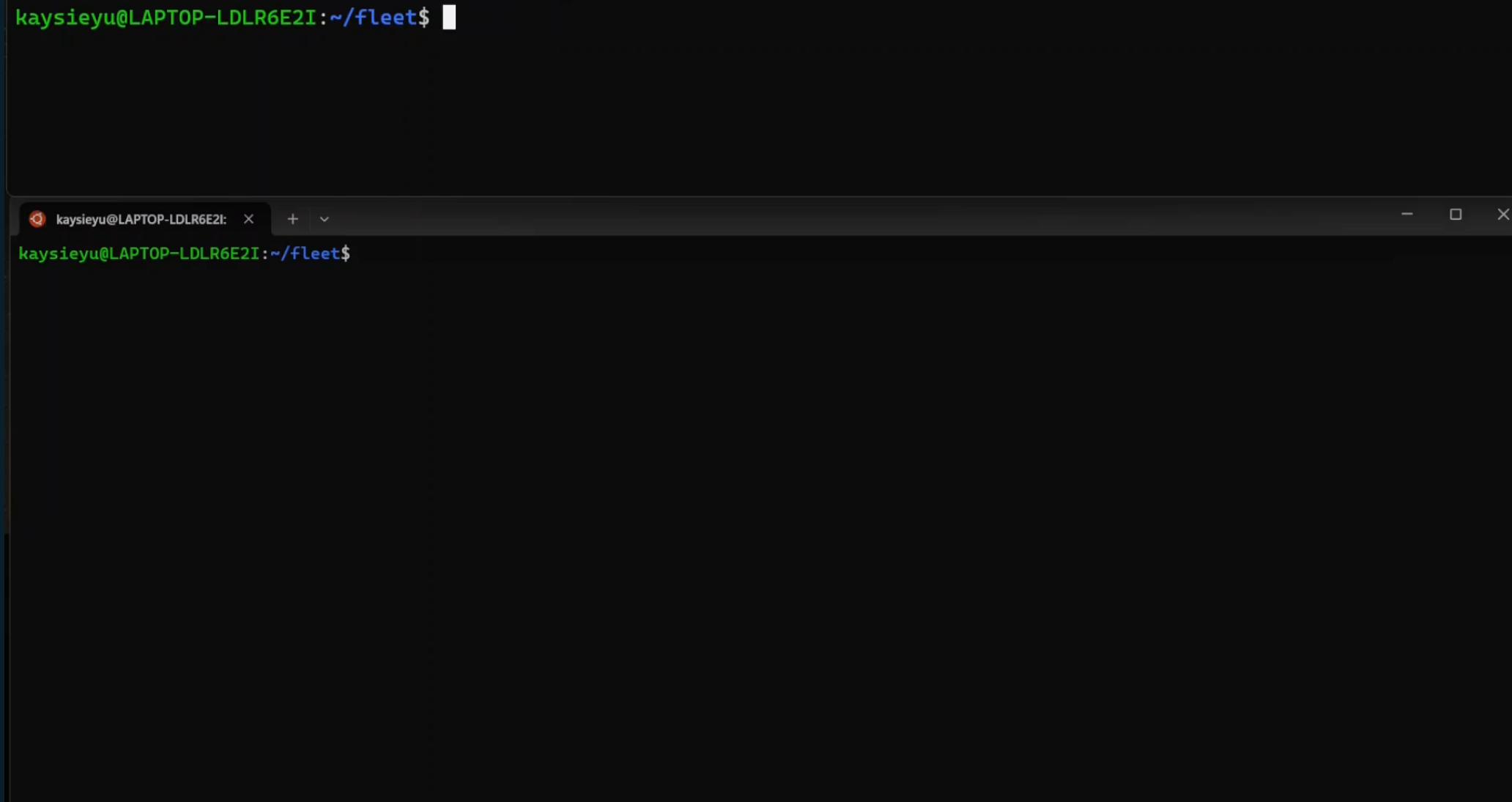
Continue to tailor your workload placement on cluster groups by changing the resources propagated from hub to member clusters

- ✓ As a **Fleet admin**, use `ClusterResourceOverride` for cluster resources
- ✓ As an **app developer**, use `ResourceOverride` to target namespace-scoped resources



Learn more:  
[aka.ms/aks/fleet-override](https://aka.ms/aks/fleet-override)

# AKS Fleet Manager workload placement



The image shows a screenshot of a terminal window with a dark background. It contains two identical command-line sessions. The top session is a new instance, and the bottom session is a previous instance of the same command. Both sessions show the prompt "kaysieu@LAPTOP-LDLR6E2I:~/fleet\$". The terminal has a standard window title bar at the top.

```
kaysieu@LAPTOP-LDLR6E2I:~/fleet$  
kaysieu@LAPTOP-LDLR6E2I:~/fleet$
```

# What is KAITO?

**Kubernetes** **AI** **toolchain** **operator**

Leverage the ecosystem

Deployment tools other than  
a “Platform/Service”

Streamline workflows

Check out the KAITO  
project on GitHub

[github.com/Azure/KAITO](https://github.com/Azure/KAITO)



# AI toolchain operator add-on for AKS

Public preview



## AI toolchain operator add-on for AKS

Workspace and infra  
setup to model  
inferencing



in a matter of  
minutes

Load model  
weights



Model  
Containerization



Host image



Provision GPU  
infrastructure



Try the KAITO add-  
on for AKS



[aka.ms/aks/  
ai-toolchain-operator](https://aka.ms/aks/ai-toolchain-operator)

# { } Logic Apps Update

Support for Zero Downtime  
Deployment Scenarios

.NET 8 custom code support

Logic Apps connectors for IBM  
Mainframe and Midranges

Improved on boarding experience in  
VS Code

Logic Apps standard deployment  
scripting tools in VS Code

Logic Apps monitoring dashboard

Public Preview



# Azure OpenAI Assistants with Azure Logic Apps

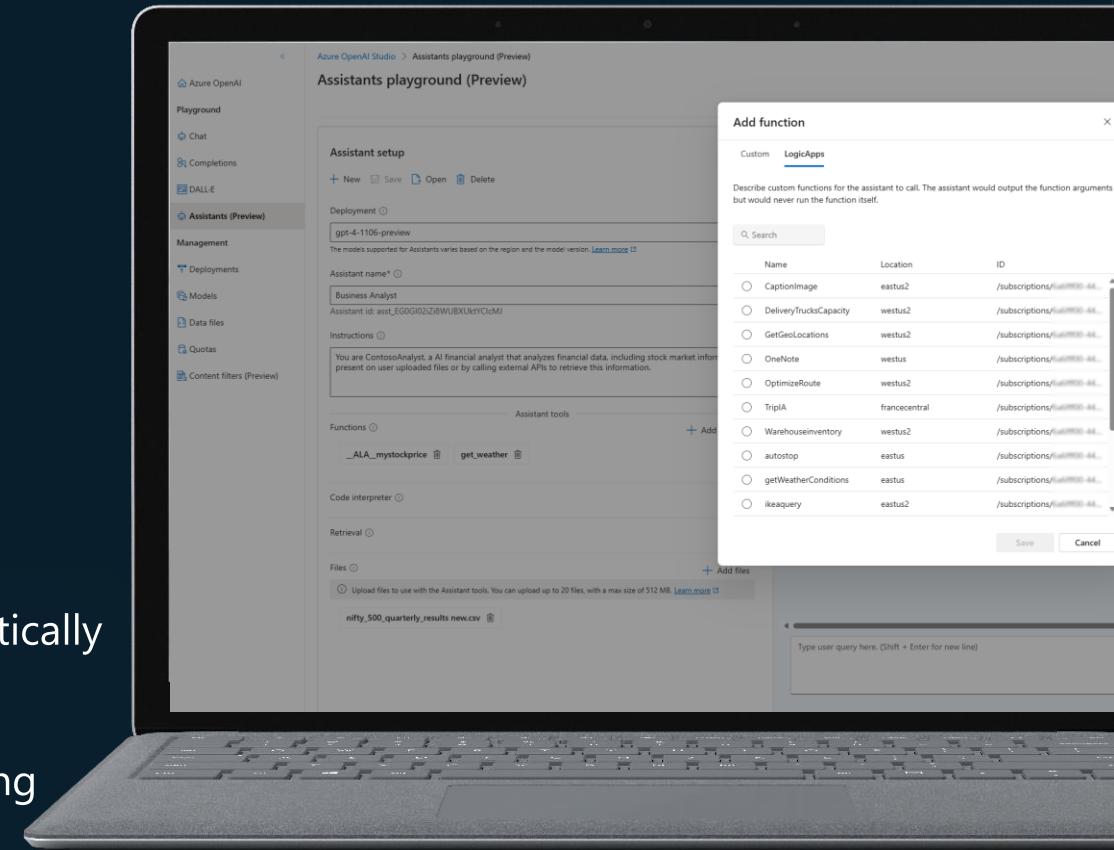
Leverage Azure Logic Apps to build AOAI assistants with function calling

## Customer benefits

Import your REST APIs implemented in Logic Apps as *AI Functions*

AI Studio invokes the Logic Apps workflow (as AI function) automatically based on the user prompt

Build new Logic Apps workflows as AI functions in AI Studio (Coming soon)



# AI RAG patterns with Logic Apps

Add an action

[Return to search](#)



Azure OpenAI

Connects to Azure Open AI to perform operations on large language models.

Gets multiple embeddings

[Preview](#) [In App](#)

Gets a single embedding

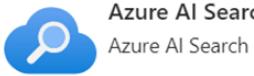
[Preview](#) [In App](#)

Gets the chat completions

[Preview](#) [In App](#)

Add an action

[Return to search](#)



Azure AI Search

Azure AI Search operations.

Index documents

[Preview](#) [In App](#)

Index a document

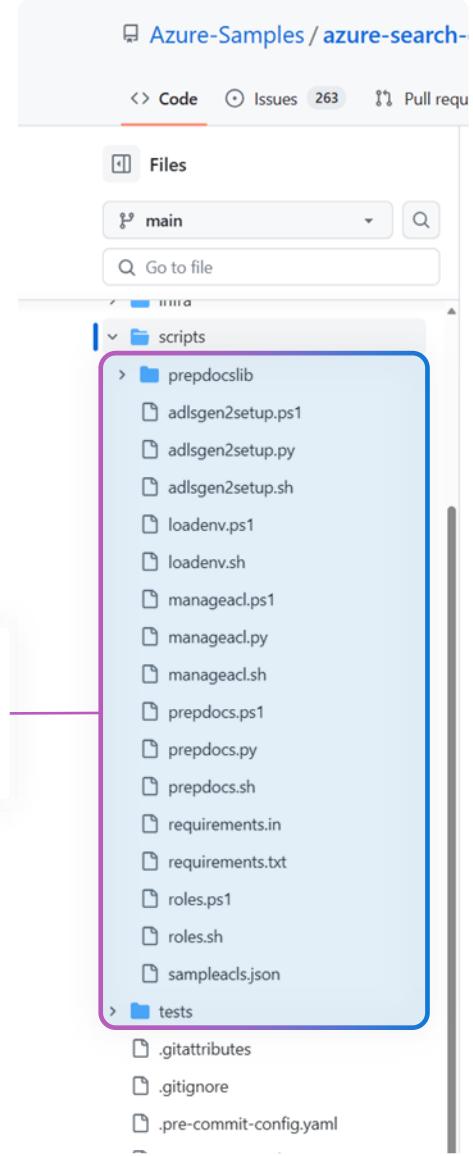
[Preview](#) [In App](#)

Vector search

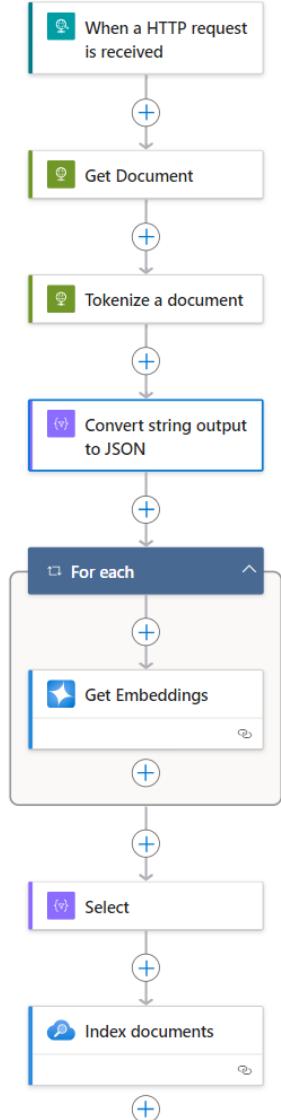
[Preview](#) [In App](#)

Lines of code to  
write/manage with  
code-first approach

Ingestion flow  
with code



Ingestion flow  
with Logic Apps connector



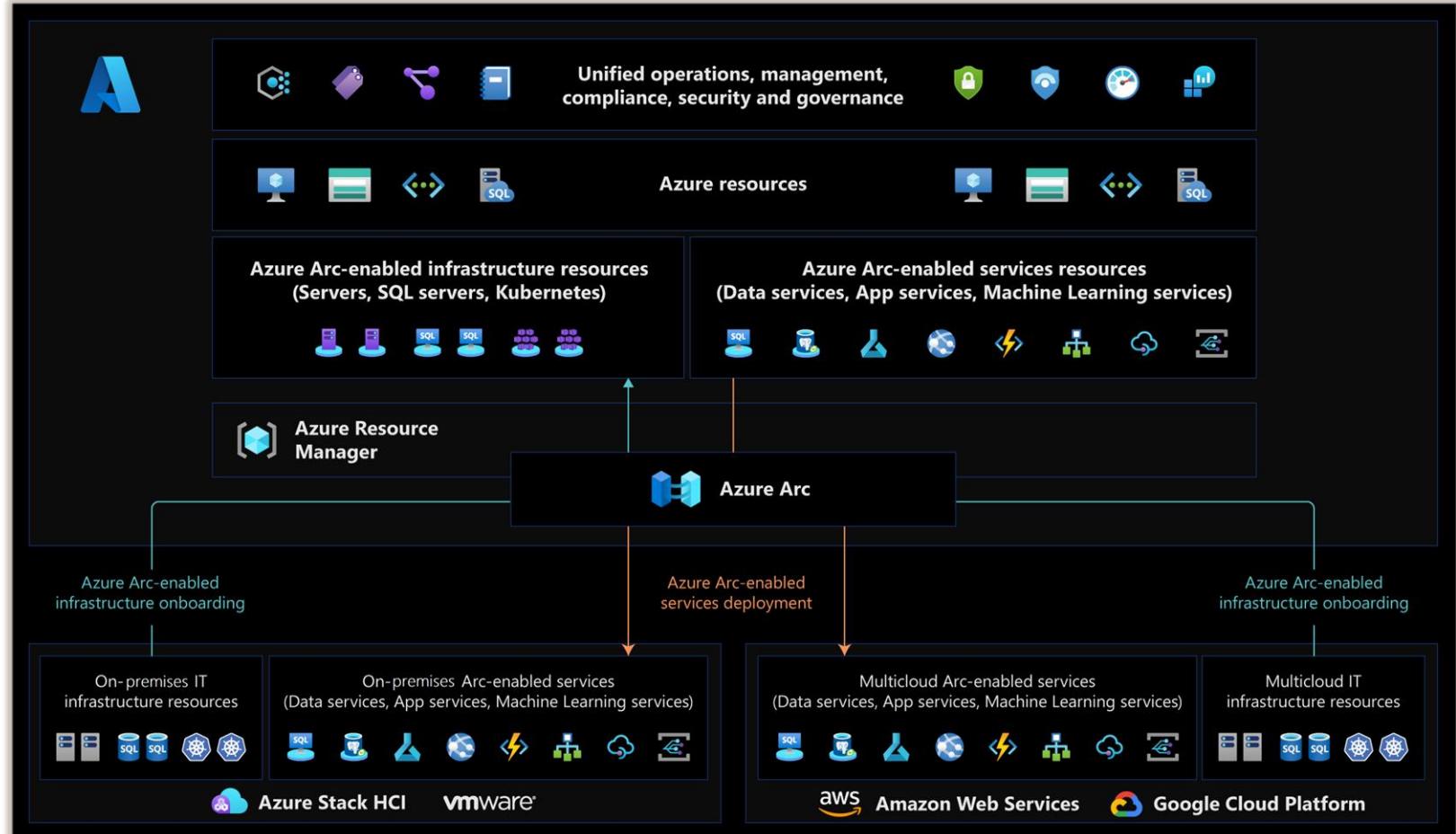
# Logic Apps Hybrid

## Background:

- Customers looking for local storage/execution:
  - BizTalk Migration
  - Regulatory/Compliance
- Initial focus on semi-connected scenarios
- Enables multi-cloud deployment opportunities for customers

## Investments:

- Public Preview Summer '24
- Build Workflows in VS Code
- Deploy to customer-managed Kubernetes using custom location
- Use the Azure Portal for management experience
- .NET 8 Custom Code extensibility
- Target-based scalability





# Azure Communication Service Update

Microsoft PowerPoint Live for my inclusive content sharing

Live Reactions: Collect reactions from the audiences

File sharing during a meeting from custom apps

Real-time transcription using Azure AI Speech (Preview)

Closed captions

Picture-in-picture for iOS and Android

Noise suppression during a video call

Stream live audio via the audio streaming API

WhatsApp Message Analysis with Azure OpenAI (Preview)



# Static Web App Update

Public preview

	Free	Standard	Dedicated
<b>Web hosting</b>	✓	✓	✓
<b>CI/CD</b>	✓	✓	✓
<b>Global Distribution</b>	✓	✓	
<b>Custom Domain</b>	2 per app	5 per app	10 per app
<b>Max app size</b>	250 MB per app	500 MB per app	2 GB
<b>Private Link</b>		✓	✓
<b>Managed Backends</b>		✓	✓
<b>SLA</b>		✓	✓
<b>Data Residency</b>			✓ Currently support Central US East US 2 East Asia West US 2
<b>Always On Managed Functions Support</b>			✓

- Select the **Dedicated** plan when:
  - Your application requires regional data residency.
- Select Standard or Dedicated when:
  - Expected traffic volumes exceed bandwidth maximums.
  - The existing Azure Functions app you want to use either has triggers and bindings beyond HTTP endpoints, or can't be converted to a managed Functions app.
  - Security requirements that require a [custom provider registration](#).
  - The site's web assets total file size exceed the storage maximums.
  - You require formal customer support.
  - You require more than three [staging environments](#).

# Azure Messaging and Streaming Services Update



## [Batch Delete \(Public Preview\)](#)

Azure Service Bus now supports Batch Delete providing a more convenient way to manage your queues, including dead letter queues.



## Event Hub

### [Event Hubs Emulator \(Public Preview\)](#)

The Event Hubs Emulator is a containerized instance of Azure Event Hubs that can run on Windows or Linux for development or test purposes. In this first delivery of the emulator, it only will support AMQP traffic.

### [Large Message size \(Public Preview\)](#)

support messages as large as 20 MB.

### [Kafka Compression \(GA\)](#)

### [Schema Registry updates](#)

New schema types support:  
[JSON Schema \(GA\)](#)

### [Protobuf in Schema Registry \(Public Preview\)](#)



## Event Grid

### [MQTT features:](#)

Enable Event Grid to be MQTT pub sub broker

### [MQTT Last Will and Testament \(LWT\) \(GA\)](#)

### [OAuth 2.0 authentication for MQTT clients \(Public Preview\)](#)

### [Namespace features:](#)

#### [Custom domain names support \(Public Preview\)](#)

#### [Push delivery to Azure Event Hubs \(GA\)](#)

#### [CloudEvents 1.0 Binary Content Mode \(GA\)](#)

#### [Shared Access Signature \(SAS\) tokens authentication \(Public Preview\)](#)

#### [Namespace Topic as a destination \(GA\)](#)

### [Event sources:](#)

#### [Microsoft Graph API events \(GA\)](#)

#### [Azure Resource Notifications health resources events to Azure Monitor alerts \(Public Preview\)](#)

#### [API Center system topic \(Public Preview\)](#)



# Azure Load Testing Update

Public preview

Azure Functions integration

Azure Functions optimizer

Generally Available

Multi-region load tests

App service integration

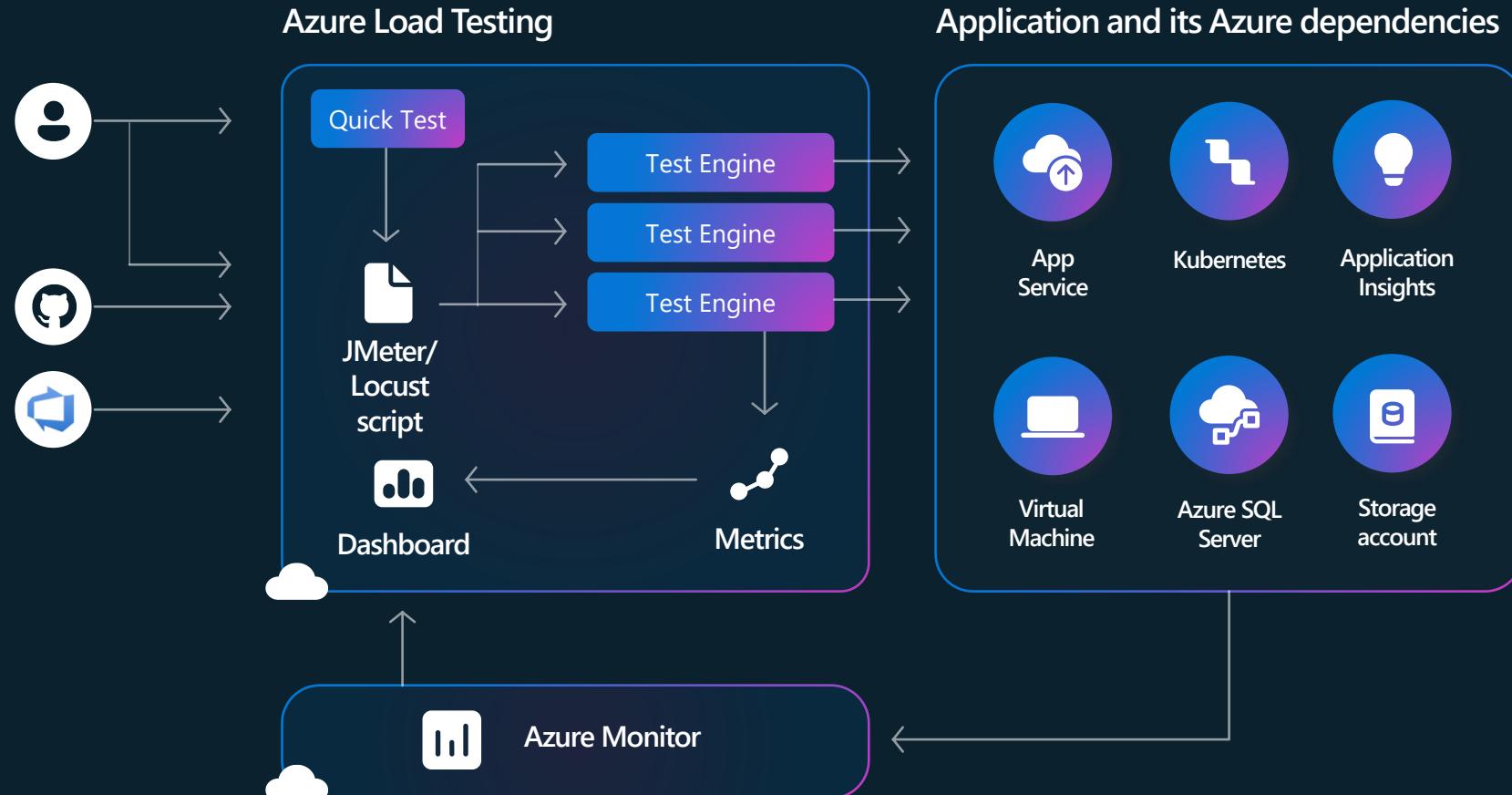
Preview by invite

Support for Locust



[aka.ms/ALT-Locust-SignUp](https://aka.ms/ALT-Locust-SignUp)

# What's happening behind the scenes in Azure Load Testing





# Microsoft Playwright Testing



- Open-source framework developed and maintained by Microsoft
- Tests web apps across all modern web browsers
- Multiple language support: JS/TS/Python/.NET.JAVA
- Parallel browser testing
- Powerful, built-in tooling like codegen, Trace Viewer, and more



Speed up  
CI validation



Improve test  
coverage



Test with  
confidence



Resolve issues  
faster

Preview by invite

# Rich Reporting and Troubleshooting

Quickly identify failed and flaky tests.

Troubleshoot tests easily using error messages, logs & other test artifacts

In-context trace viewer for exploring Playwright traces for faster debugging.

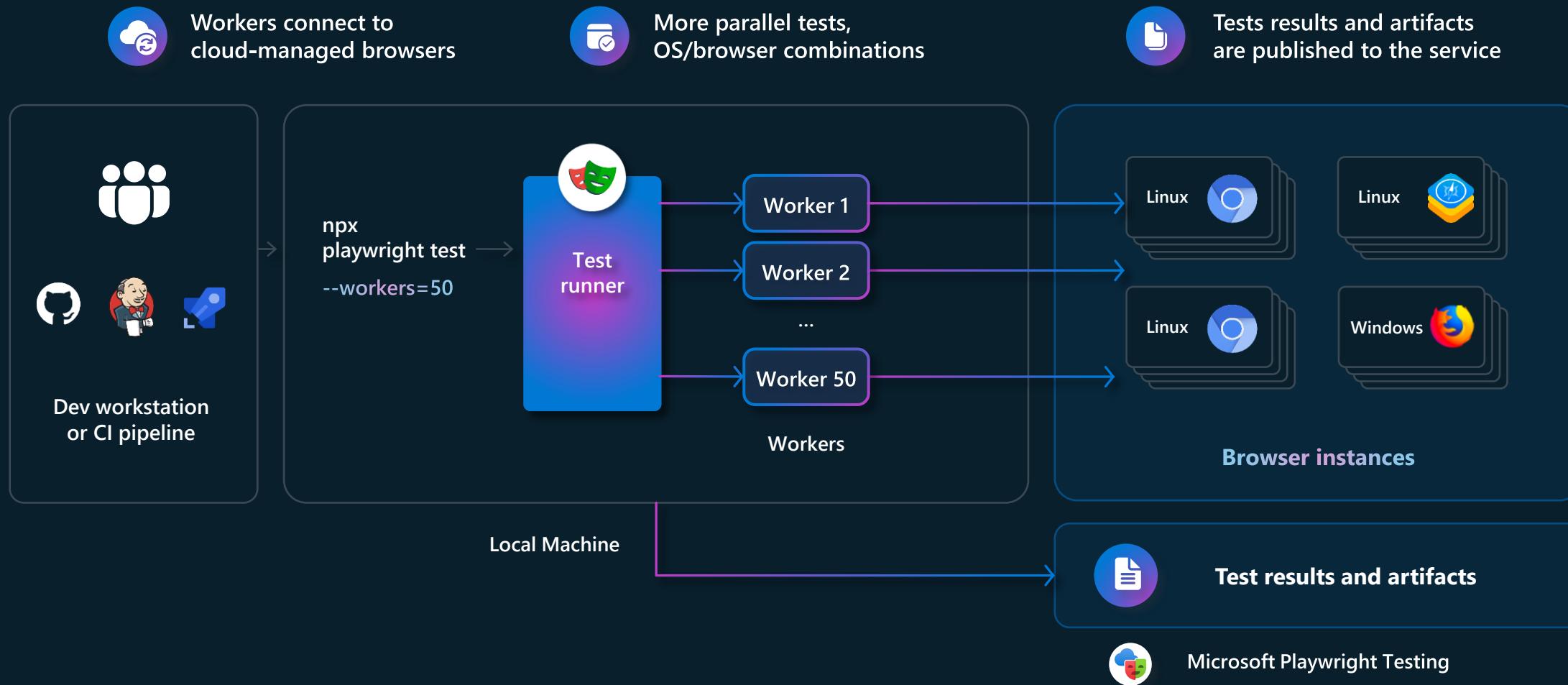
Available for both dev inner loop and CI pipelines.



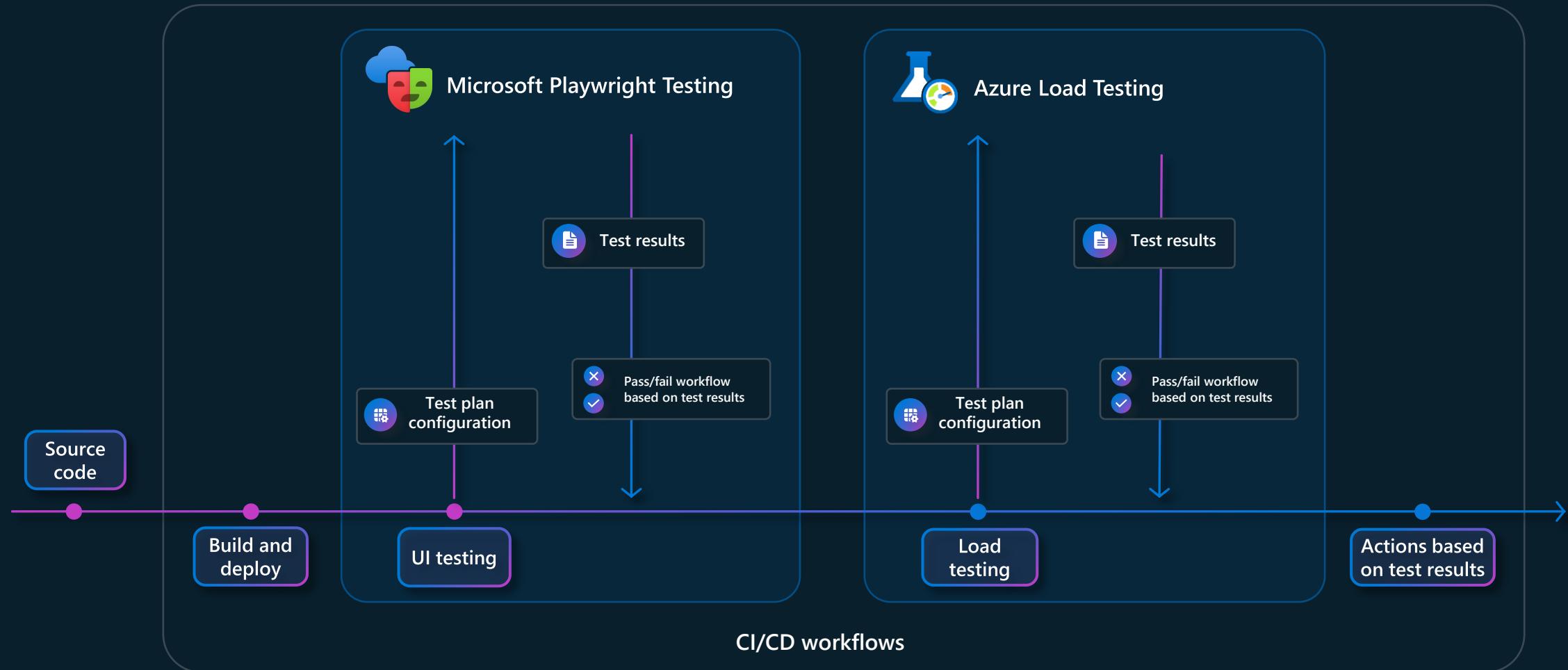
[aka.ms/mpt/reporting-signup](https://aka.ms/mpt/reporting-signup)

The screenshot shows the Microsoft Playwright Test interface. At the top, it displays a summary of a failed test run titled "mptdemo-reporting" with a status of "Test run Failed". Below this is a "Test Summary" section with a search bar and a "Failed" filter applied. The main content area is titled "Test case" and lists two expanded sections: "Sample.spec.ts" (3) and "offers.app.spec" (9), each containing multiple failed test steps. A detailed error message for one step is shown: "Error: Timed out 5000ms waiting for expect(locator).toHaveTitle(expected)". The error message includes the locator used ("locator(':root')"), the expected pattern ("/Playwright55/"), and the received string ("Fast and reliable end-to-end testing for modern web apps | Playwright"). At the bottom, there is a "Call log" section listing a series of locator resolutions. The interface also includes sections for "Annotations", "Test steps", and "Attachments", along with a "View trace" button.

# Run parallel tests with cloud browsers using Microsoft Playwright Testing



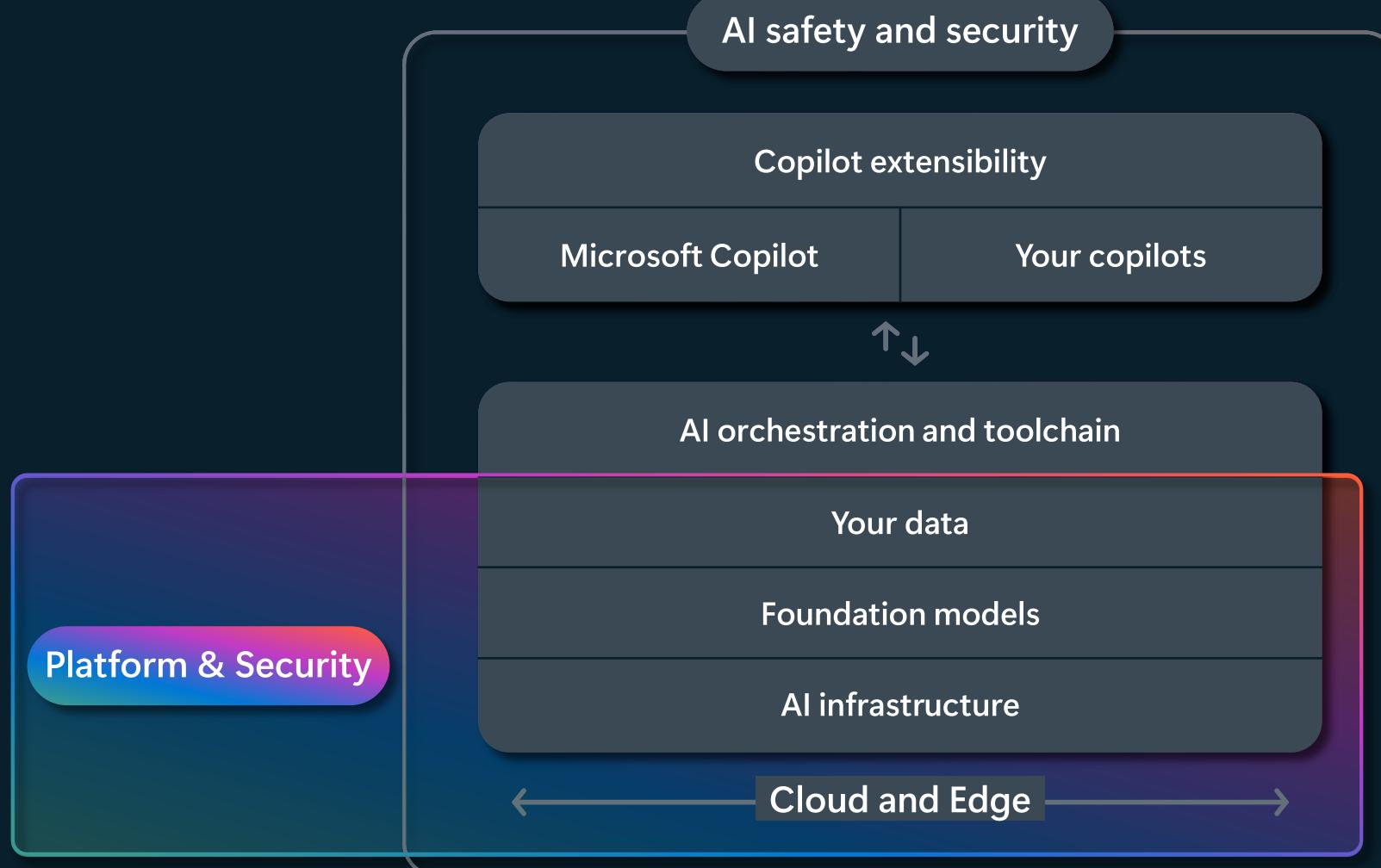
# Using Azure testing services in your DevOps workflows



# Douglas Lam

Partner Solution Architect –  
Azure Infrastructure

# Session 2 – Platform & Security



# Agenda

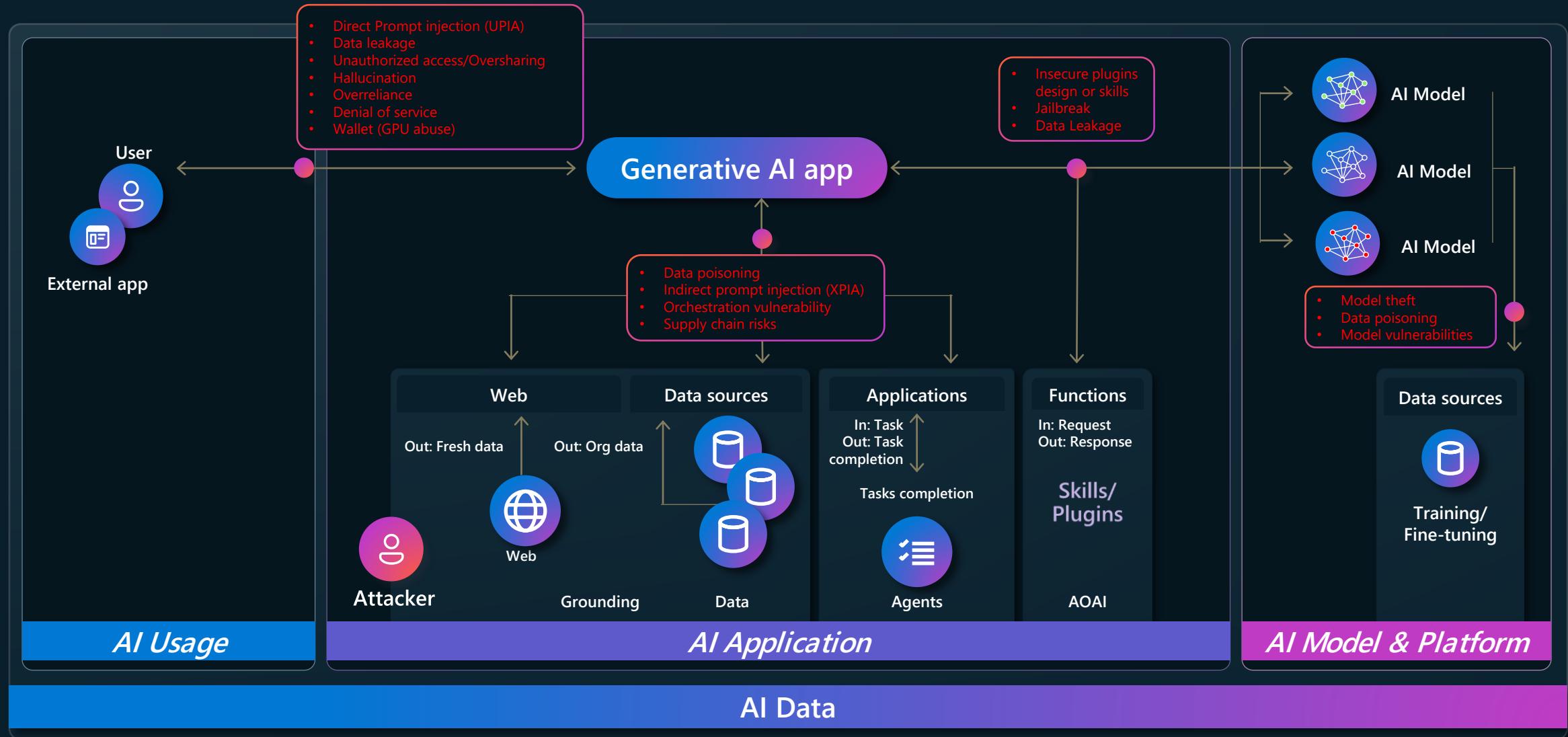
- Securing AI deployments with Microsoft Security
- Hybrid cloud observability with Azure Monitor
- Azure cloud platform innovation for AI apps
- Building AI apps with Azure Blob Storage
- Azure Arc Site Manager for edge site monitoring



Secure your AI application  
transformation with  
Microsoft Security



# Generative-AI threat landscape



# How secure is your custom-built AI applications?

## *AI Usage*

Do you have safety controls in place to detect and block adversarial prompt attacks or risky content?

Do you have security and privacy controls in place to protect sensitive or personal data?

## *AI Application*

Do you have visibility into all the AI applications deployed in your environment and their vulnerabilities?

Does your security team have the ability to detect and respond to threats impacting your AI applications?

## *AI Model & Platform*

Are the models you use for your AI applications secured and governed?

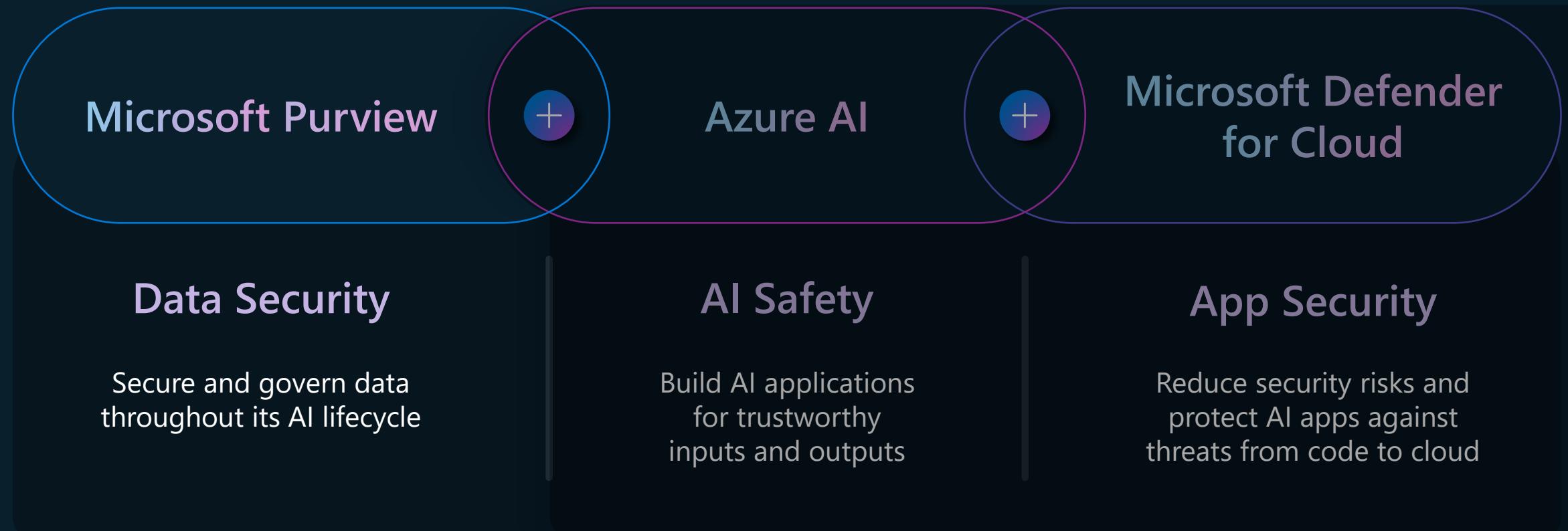
Can you detect and assess model vulnerabilities before deploying them to your environment?

## *AI Data*

Do your app customers have visibility into the sensitive data flowing in and out of your AI apps?

Do your AI apps honor permission controls on data and prevent data overexposure?

# An integrated approach to build secure, safe, and trustworthy AI



Home

Create

My Content

Feed

Apps

M365 Chat

Outlook

Teams

Word

Excel

PowerPoint

 M365 Chat  
What are the important documents for Project Obsidian?

November 9, 2023 at 07:58 PM

What are the important documents for Project Obsidian?

Here are some important documents for Project Obsidian:

- "Project plan for project obsidian" is a confidential document authored by Alex Wilber about next generation chip design. It provides overview of the project objectives, scope, deliverables, timelines, and resources. [①](#)
- "Next generation chip" is a confidential document authored by Joe Fromm discusses the design of this next generation chip. It discusses the need for building dedicated chips to power future cars and compares the advantages and disadvantages of ARM and x86 chip architectures. [②](#)
- "Obsidian" is a presentation that is authored by and last modified Tuesday evening by Edison Gil. It was used to kick off Project Obsidian. [③](#)
- [2023\\_Obsidian\\_Resources](#) is also related, but organization policy restricts me from discussing its content in this chat.

 Copy  Share

**Confidential \ Project Obsidian**  
Confidential data that requires protection, blocks users from copy and share of content.

3 references ^  1/30 AI-generated content may be incorrect  

  **Project plan for project obsidian**

Confidential \ Project Obsidian  
Word | Alex Wilber modified on 11/07/23  
"Project Plan for Project Obsidian. A brief overview of the project objectives and scope. Project Objectives: The main objective of Project Obsidian is to develop..."

  **Next generation chip**

Confidential \ Project Obsidian  
Word | Joe Fromm modified on 11/15/23  
"A next generation chip that is more capable than any current generation in harnessing raw compute power while being able to efficiently..."

  **Obsidian**

General  
Powerpoint | Edison Gil modified on 10/12/23  
"Kick off for a new project..."

[Who is Alex Wilber?](#) [What is the project objective of Project Obsidian?](#) [Timelines for Project Obsidian](#) 

Ask a work question or use / to reference people, files and more  

# Secure and govern data, natively integrated into your apps

- Label awareness
- Label inheritance
- Honor access rights
- Log maker activities

Announcing:  
**Microsoft**  
**Purview integration**  
**with Copilot Studio**

Public Preview

...and coming soon for Azure AI Studio

**Secure and govern  
data in your apps  
with just a few  
lines of code**

- Discover AI risks
- Protect sensitive data
- Govern AI usage

Announcing:  
**Microsoft  
Purview SDK**

Coming soon

# Microsoft Purview enables data security and compliance

## Discover AI risks

Total prompts with sensitive data

Google Bard



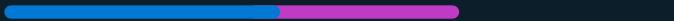
OpenAI ChatGPT-4



Contoso Built App



Microsoft Copilot

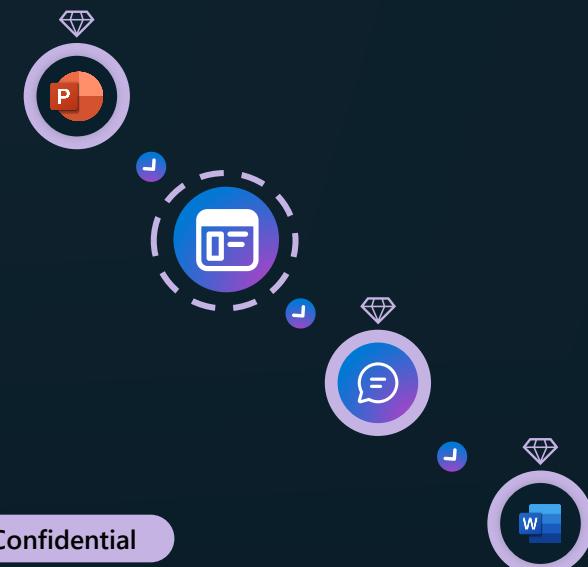


Credit card  
numbers

US SSN

Physical  
address

## Protect sensitive data



## Govern AI usage



⚠ Confidential projects

⚠ Corporate sabotage

⚠ Regulatory collusion

[Microsoft Purview data security and compliance protections for Microsoft Copilot and other generative AI apps | Microsoft Learn](#)

# Build secure, safe, and trustworthy AI

**Microsoft Purview**

**Azure AI**

**Microsoft Defender  
for Cloud**

## Data Security

Secure and govern data  
throughout its AI lifecycle

## AI Safety

Build AI applications  
for trustworthy  
inputs and outputs

## App Security

Reduce security risks and  
protect AI apps against  
threats from code to cloud

# Continuously reduce risk

Contextual and prioritized security posture management across the entire cloud application lifecycle

AppSec + CI/CD security

## Enable secure development

Prevent vulnerabilities, misconfiguration, secrets in code and secure your software supply chain



CWP + CDR

## Remediate threats faster

Near real-time detection and response for compute and AI workloads, data and APIs in a unified XDR experience



# Protect AI apps from code to runtime

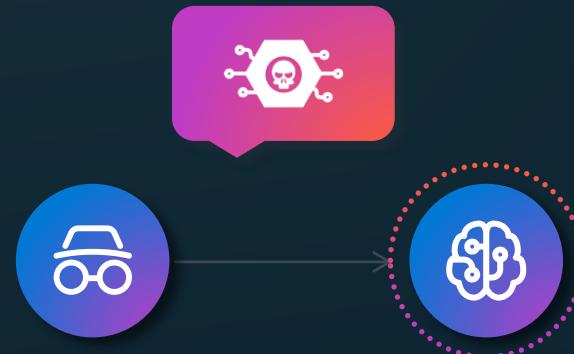
Start secure

AI security posture management (AI-SPM)



Stay secure

Threat protection for AI workloads



Microsoft Defender  
for Cloud

# Discover new AI attack surfaces from code to cloud

- Discover AI models and workloads
- Identify GenAI vulnerabilities and risks
- Map exploitable attack paths
- Delegate and remediate risks

Now Available:  
**AI security posture management (AI-SPM)**

in Microsoft Defender for Cloud

Public Preview

Azure OpenAI Service | Azure Machine Learning | Amazon Bedrock

# Monitor and protect GenAI apps from emerging threats

- Identify threats posed to apps using Azure OpenAI Service
- Evidence-based security alerts enriched with Microsoft threat intelligence
- Native integration with Azure AI Content Safety prompt shields
- Correlate alerts into incidents in Defender XDR

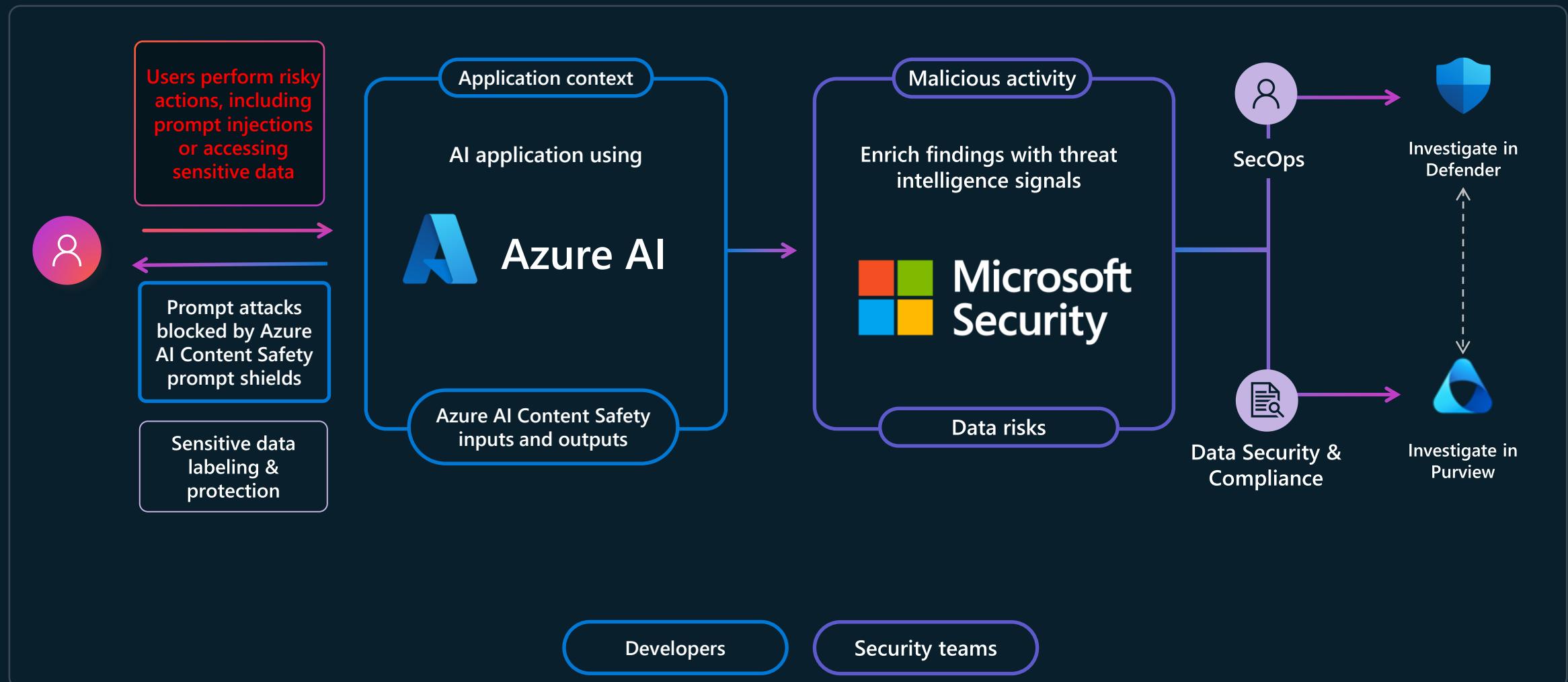
Now Available:  
**Threat protection  
for AI workloads**

in Microsoft Defender for Cloud

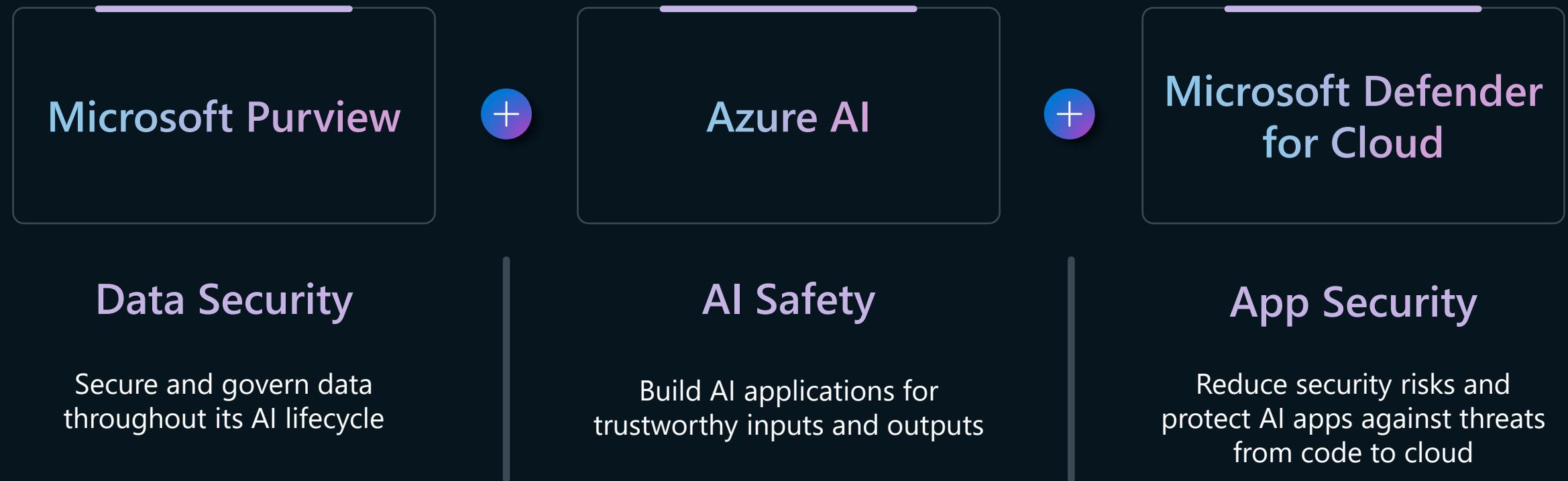
Preview

Azure OpenAI Service

# Secure and govern custom-built AI applications

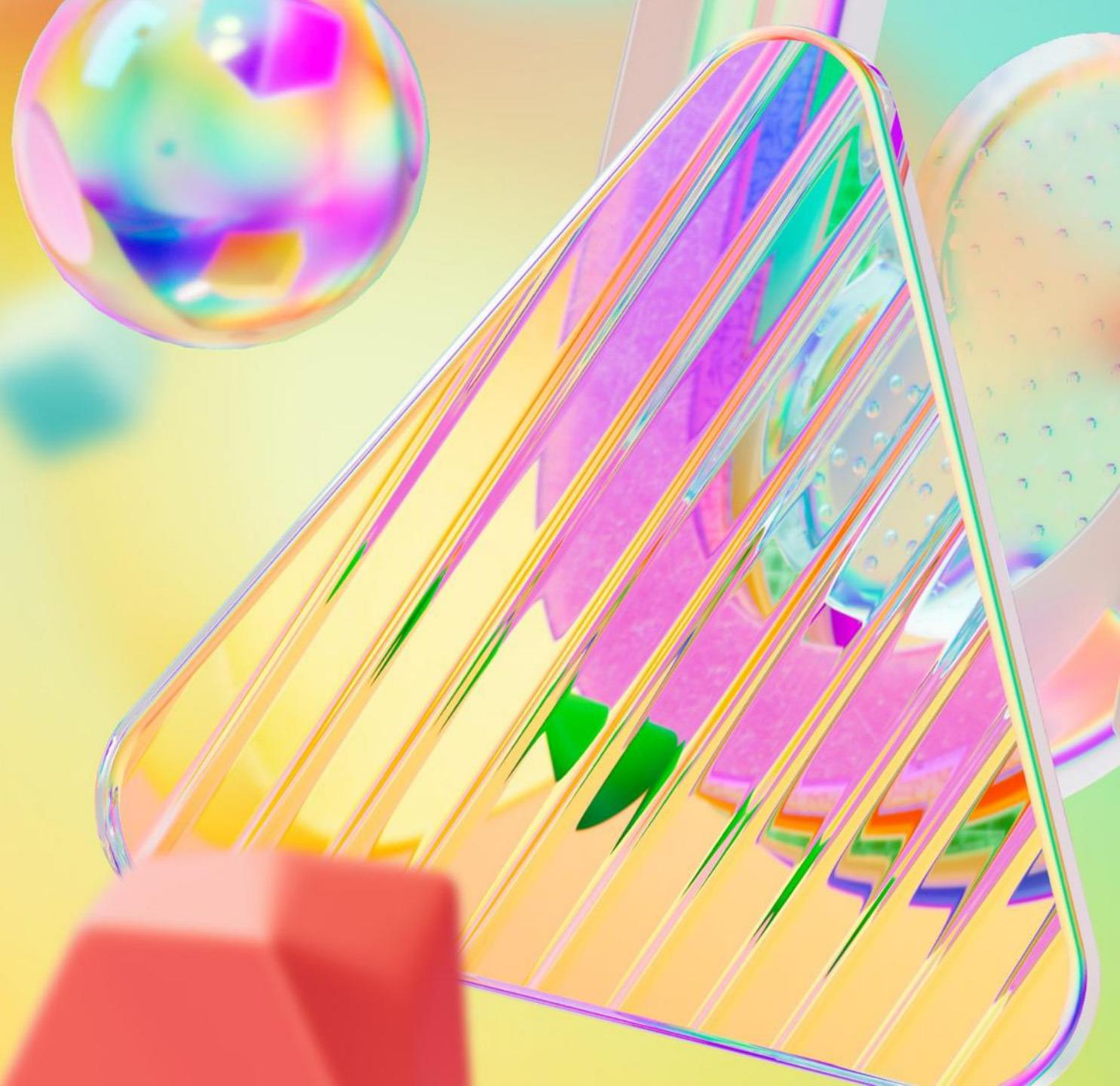


# An integrated approach to build secure, safe, and trustworthy AI





# Azure Monitor: Observability from Code to Cloud



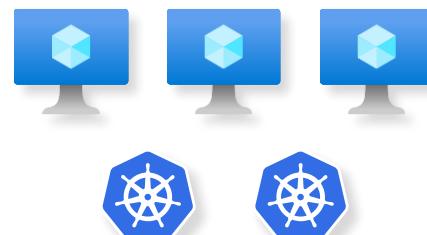
# Azure Monitor anywhere enabled by Azure Arc

Consistent monitoring across your hybrid and multi-cloud compute

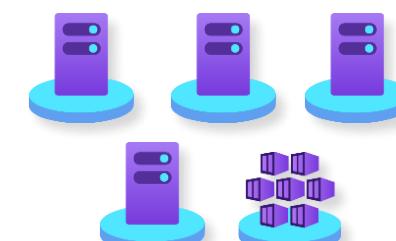


Azure Monitor

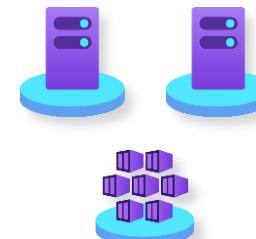
Monitoring across your infrastructure



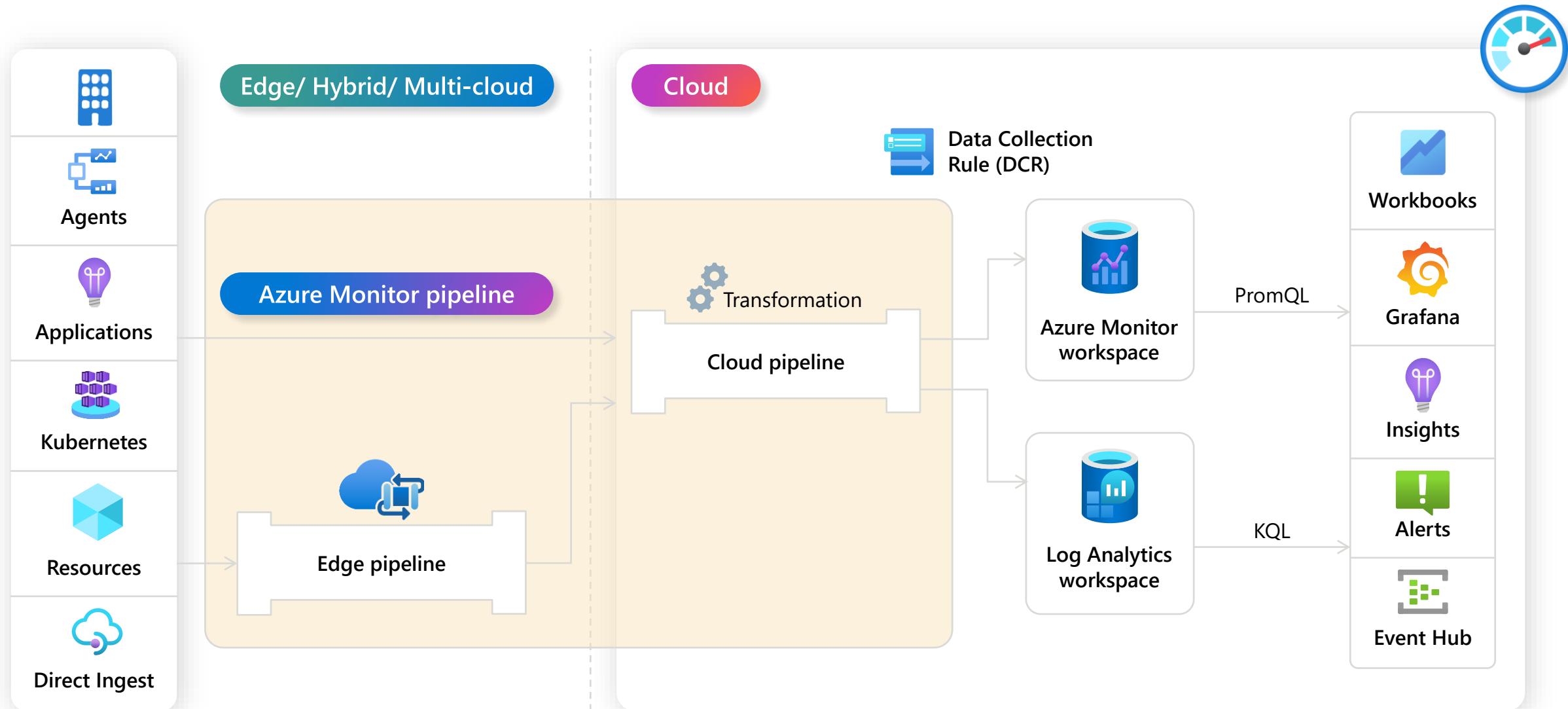
Google Cloud



On-premises



# Azure Monitor pipeline at edge (preview)



# Azure Monitor pipeline (preview)

Powerful new solution designed to facilitate **high-scale data ingestion & routing** from edge environments to seamlessly enable observability across Azure and the adaptive cloud approach

## High scale data ingestion

Ingest high volume of telemetry (in the range of million events per second) from IT/OT devices and resources

Easy to scale horizontally by **running multiple replica sets of Azure Monitor pipeline instance** for **routing high volume telemetry to cloud**.

## Observing resources in isolated environments

Facilitate a connection between Azure & K8s clusters in isolated networks (ISA-95/ Purdue network) with Azure IoT Layered Network Management and Azure Monitor pipeline

Collect telemetry from resources in segmented networks and route it to Azure Monitor.

## Reliable data ingestion and data loss prevention

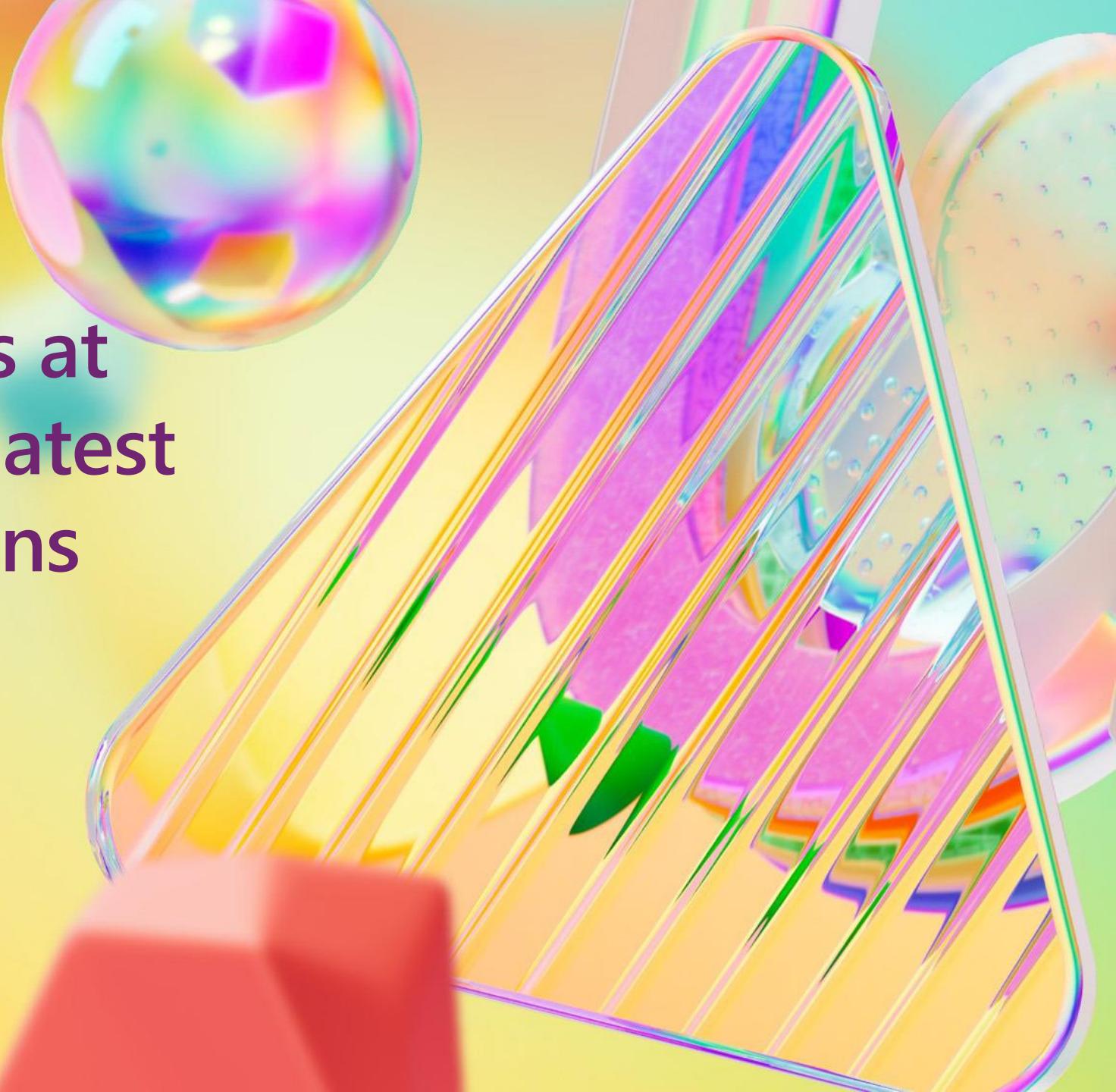
Prevent data loss by **caching telemetry** during periods of intermittent connectivity and **backfill to Azure Monitor** once connection is restored

<http://aka.ms/AzMonPipelineEdgeBlog>

<http://aka.ms/AzMonPipelineOverview>



**Building applications at  
hyperscale with the latest  
Azure IaaS innovations**



GPT-4

GPT-4 Turbo

GPT-4o

Cost

Speed

March 2023

November 2023

May 2024

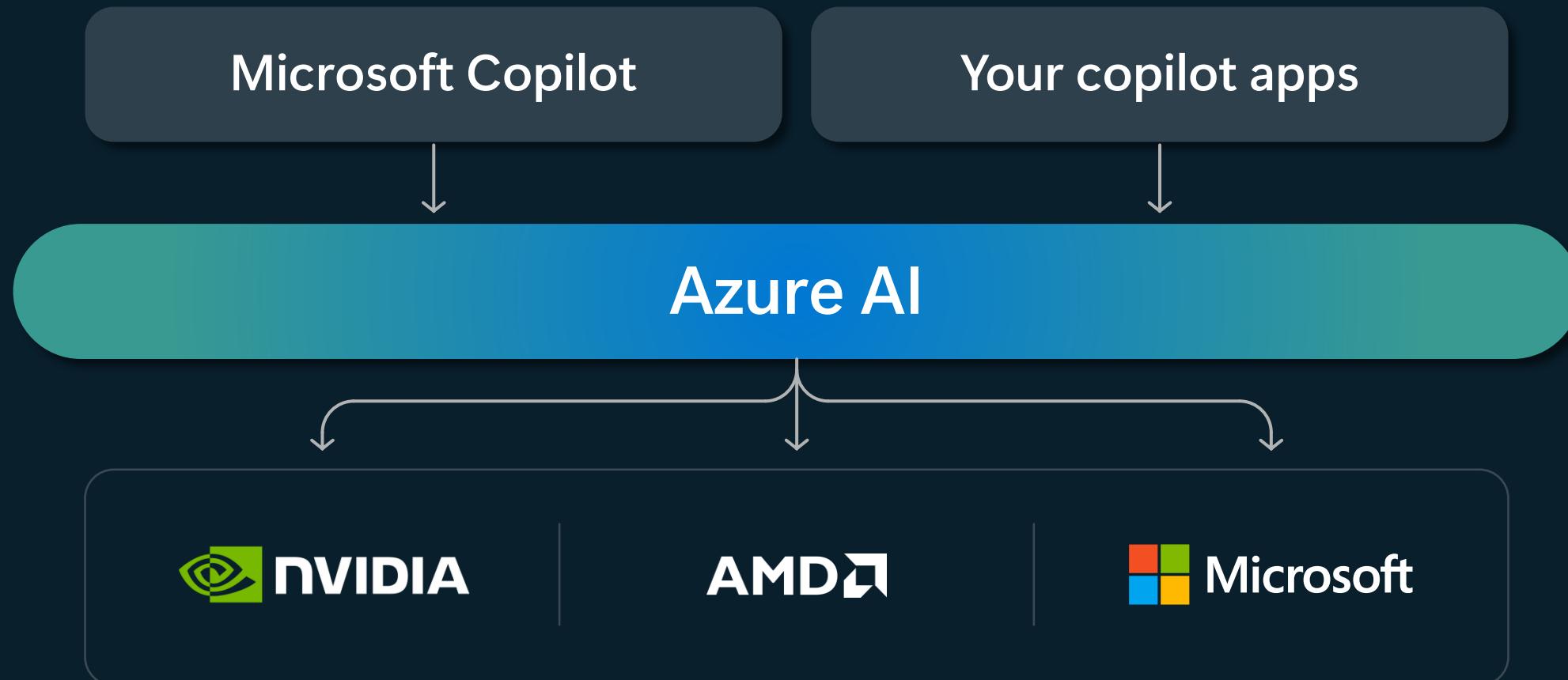
2x

$\frac{1}{6}$

6x

$\frac{1}{12}$

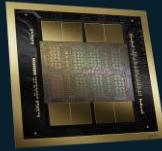
# Best AI performance. Lowest cost.



Full-stack  
partnership



Microsoft  
Azure



NVIDIA  
Blackwell  
Platform



NVIDIA  
NIM-optimized  
models



NVIDIA on  
Azure

Microsoft  
Apps



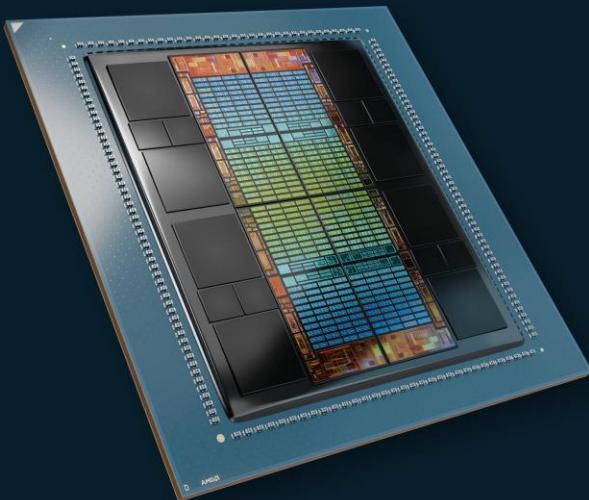
Windows 365  
and Copilot  
GPU acceleration

[aka.ms/NVIDIAPartnership](https://aka.ms/NVIDIAPartnership)

General  
availability

# ND MI300X V5 virtual machine

Leading  
price/performance  
on GPT-4



Optimized for  
Microsoft Azure  
workloads

Introducing the new Azure AI infrastructure VM series ND MI300X v5 - Microsoft Community Hub

[aka.ms/AMDPartnership](https://aka.ms/AMDPartnership)

**AMD**

Preview

# Azure Cobalt 100 Arm-based VMs in Preview now

- General purpose (Dpsv6-series and Dplsv6-series)
- Memory optimized (Epsv6-series)
- Suitable for various scale out workloads
- Available in seven Azure regions



Request access to the preview  
[aka.ms/Cobalt100-VM-Preview-Signup](https://aka.ms/Cobalt100-VM-Preview-Signup)



New Cobalt VMs compared to our previous generation Arm-based VMs offer:

- Up to 1.4x CPU performance
- Up to 1.5x performance on Java-based workloads
- Up to 2x performance on web servers, .NET applications and in-memory cache applications
- 4x local storage IOPS with NVMe
- More power efficient

In preview

# New Azure Cobalt 100 Arm-based VMs

General Purpose (lower memory to vCPU ratio)	General Purpose	Memory Optimized
 <b>Dpls and Dplds v6</b> 2- 96 vCPU VM size Up to 192 GiB of memory (2 GiB of memory per vCPU)	 <b>Dps and Dpds v6</b> 2- 96 vCPU VM size Up to 384 GiB of memory (4 GiB of memory per vCPU)	 <b>Eps and Epds v6</b> 2- 96 vCPU VM size Up to 672 GiB of memory (up to 8 GiB of memory per vCPU)
<p>Up to 5280 GiB of local storage Up to 60Gbps of network bandwidth Up to 40Gbps of remote storage throughput</p>		
Workloads		
Media encoding, gaming servers, small databases, microservices	Application servers, web servers, small to medium databases	Large databases, in-memory caches, and analytics workloads

Ability to attach Standard, Premium and Premium SSD v2, and Ultra Disks (based on regional availability)

# Azure Boost advancements enabling cutting edge performance

With newest Intel VMs

**12.5 GB/s**

Industry leading storage throughput

**200G**

Network throughput

**<3s**

Target for reduced maintenance impact

**2x**

Dual TORs 2x100G for resiliency

**100%**

New Azure VMs\* ship with Boost

**7+**

million customer VMs

- Now supports native Linux/Windows for MANA SR-IOV and DPDK drivers, with broad availability for the largest Network Virtual Appliances
- MANA support is available with the newest Intel Dv6 & Ev6 VMs coming in preview (<https://aka.ms/mana>)

99

This is impressive to see:  
Driving 100Gbps with pretty  
much no CPU utilization on  
the VMs"

– Sr. Staff Performance Engineer,  
[Nutanix](#)

99

The latest Azure Boost instances powered by MANA NIC have enabled Keysight Elastic Network Generator to transmit up to 200 Gbps of unidirectional traffic while utilizing only 4 vCPU cores"

– Sam Alsous, Lead Cloud Solutions Architect, [Keysight Technologies](#)

\*Does not include HPC or GPU VMs

Updates

# Azure Maia

Integrated system  
of silicon, racks,  
and software



Designed for  
Azure OpenAI  
workloads

[aka.ms/AzureMaia](http://aka.ms/AzureMaia)

# VM Scale Sets Standby Pools – Accelerate deployment speed

Standby pools for Virtual Machine Scale Sets enables you to increase scaling performance by creating a pool of pre-provisioned VMs that the scale set can use when scaling out. Standby pools is the perfect addition to **bursty workloads** or workloads that have **time-consuming post provisioning steps**.



- Reduce scale out latency
- Complete initialization steps
- Maintain a ready capacity
- Reduce costs



- Scale 50 running VMs in <1 second
- Scale 50 deallocated VMs in 40 seconds
- Scale 1000 running VMs in 20 seconds
- Scale 1000 deallocated VMs in 200 seconds

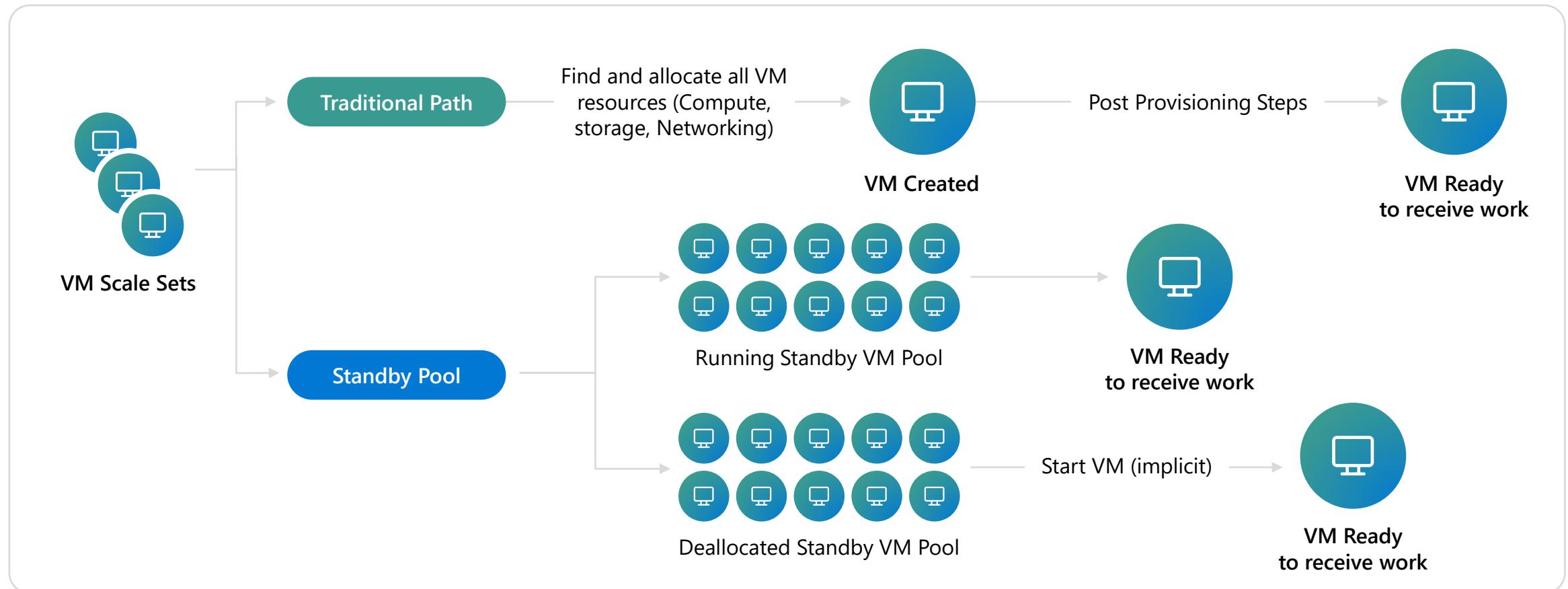
[Standby pools for Virtual Machine Scale Sets - Azure Virtual Machine Scale Sets | Microsoft Learn](#)

Learn more:  
[aka.ms/StandbyPoolsVMSSDocs](https://aka.ms/StandbyPoolsVMSSDocs)



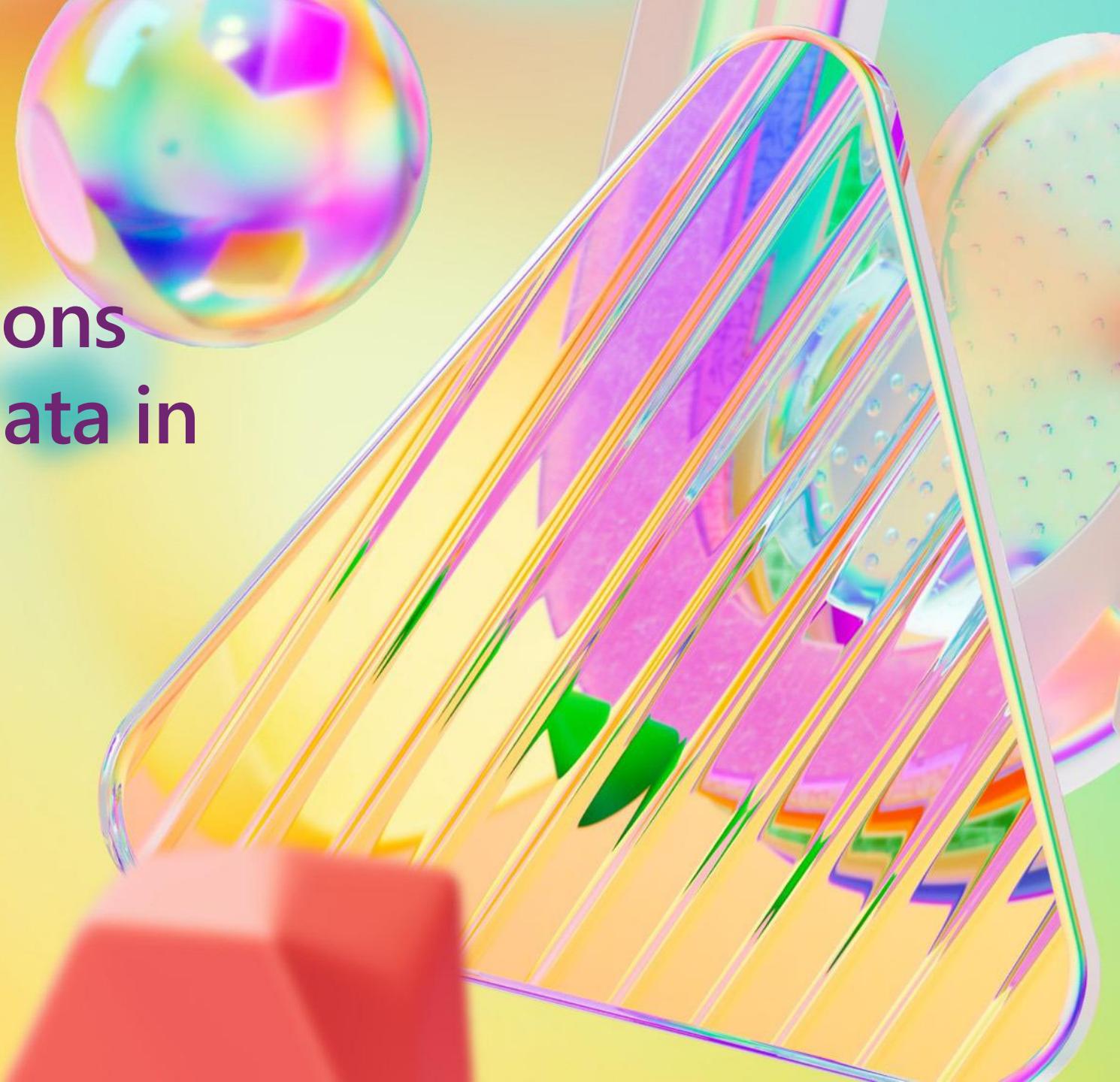
# Complete long post provisioning steps ahead of time

The VMs in the pool complete various initialization steps such as installing applications/software or loading large amounts of data prior to being put into the scale set. Use running VMs when latency requirements are strict and use deallocated VMs to reduce costs while maintaining fast scale out times.





# Building AI applications that leverage your data in Object Storage



# Bringing domain knowledge to LLMs



## Prompt engineering

In-context learning



## Fine-tuning

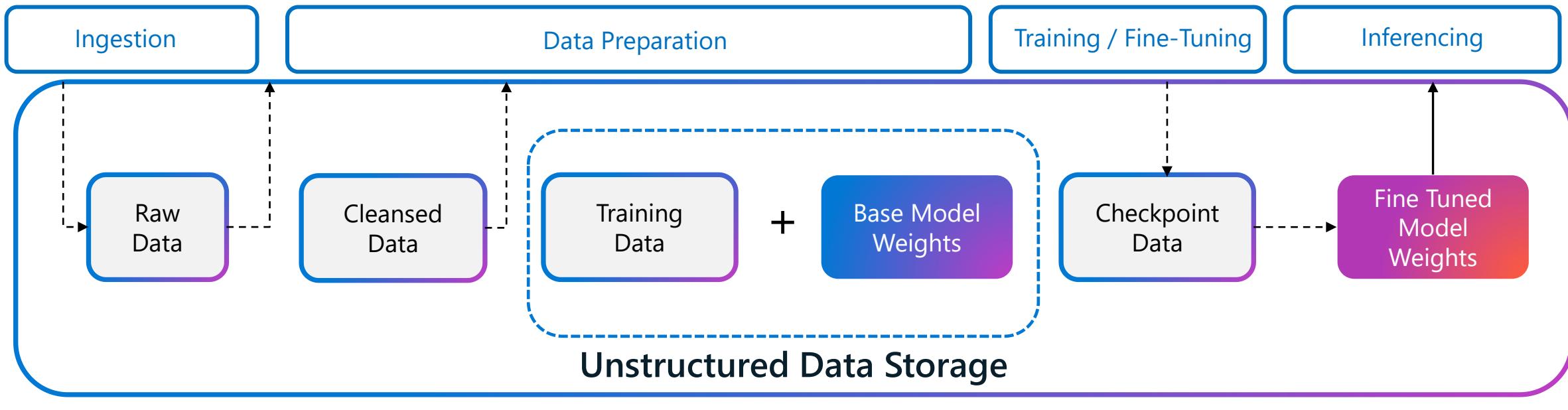
Learn new skills



## Retrieval augmentation (RAG)

Learn new facts

# AI Pipeline - Storage Requirements



## Requirements

### Training / Fine-Tuning

- **Ingestion:** Bring raw training data to Azure
- **Data Preparation:** Integration with Spark, MosaicML, etc.
- **Training/Fine-Tuning:** Data to GPU nodes, checkpoints to storage. Integration with PyTorch and other ML frameworks
- **Data Management:** Secure & cost-efficient retention

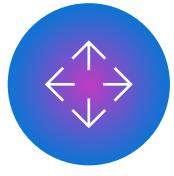
### Deployment/Inference

- **Deployment:** Model distribution and load times
- **Data Management:** Model versioning, retention of inference inputs and outputs

# Training & Fine-tuning with Azure Blob Storage



# Training and Fine-Tuning with Blob Storage



## Scalable

PBs of data, Tbps of R/W throughput



## Cost Effective

Tiered storage for long-term retention



## Multi-Protocol

Integrates well with analytics engines



## Optimized for node-local access

BlobFuse2 with caching



## Data Management

Blob index tags, automated lifecycle management

# Blobfuse2

GA +  
update



High Throughput access  
to Blob Storage



Easy to install and work  
with PiB scale data



Open-sourced &  
supported by Microsoft

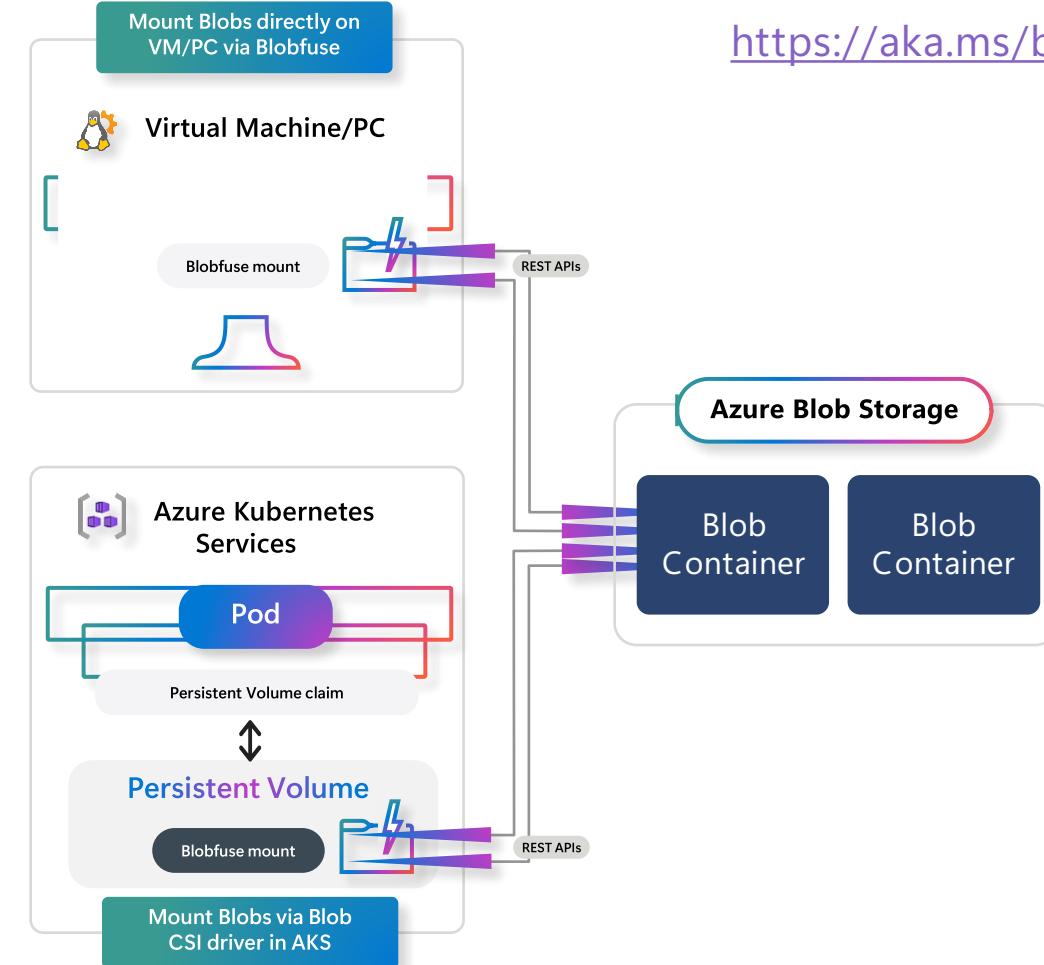


Secure access to data



Virtual File System Driver to  
mount blob storage in your  
computing environment

<https://aka.ms/blobfuse>



# Blobfuse2: Caching options

## File caching

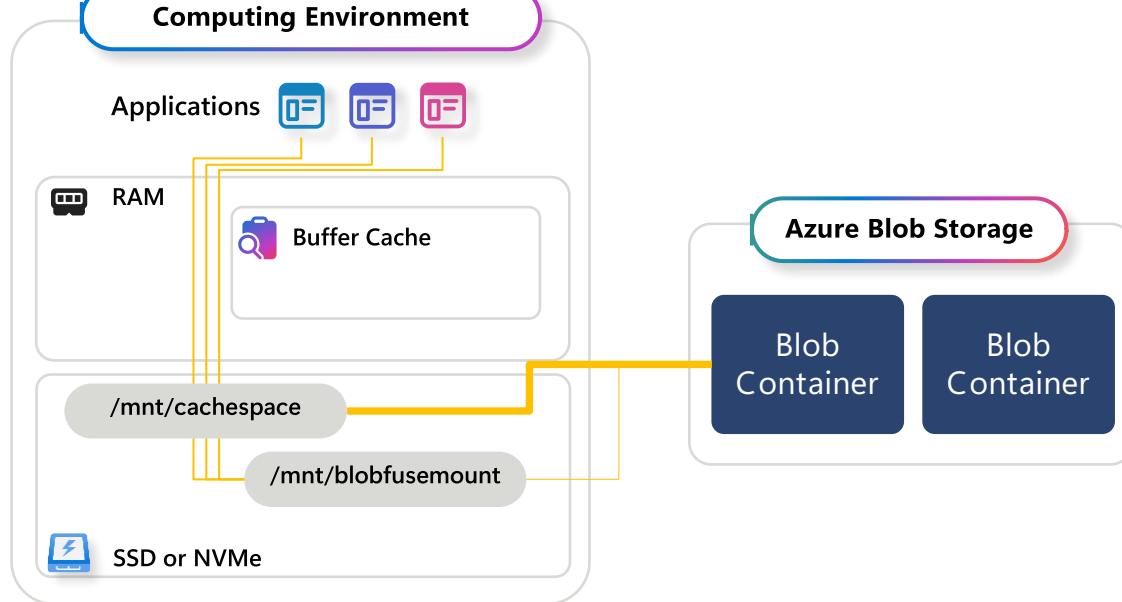
- **Downloads the files completely in chosen caching folder before serving to applications**
- Can only read files that can fit into chosen cache space
- Configurable timeouts for predictable eviction policy
- **Useful for repeated reads** that do not require frequent updates from Storage account: AI Training datasets

## Streaming with Block-cache

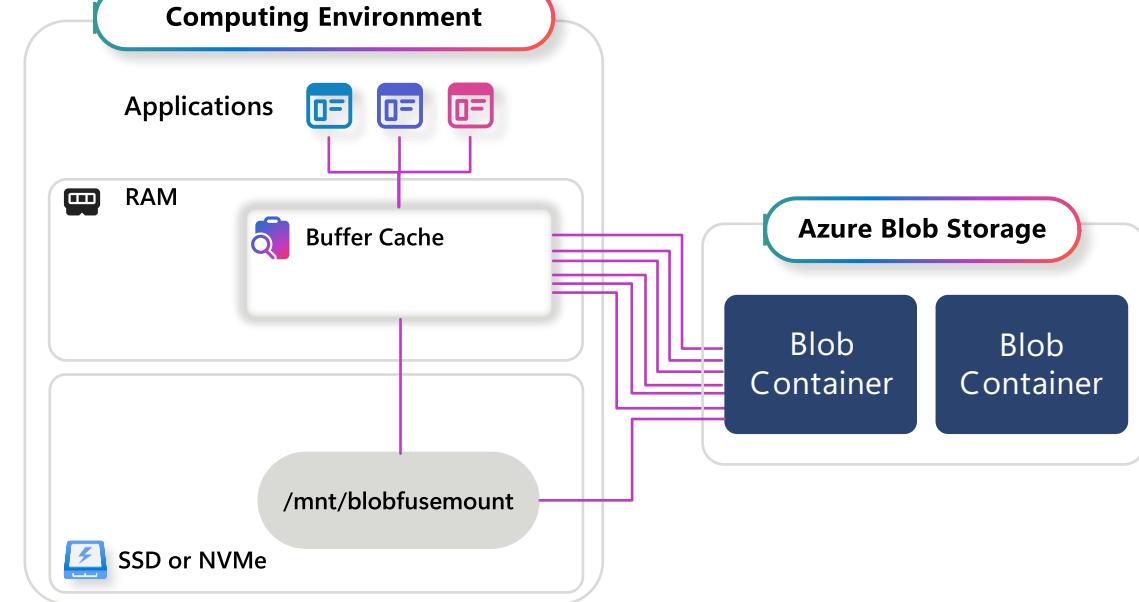
New

- Serves the file for reads while downloading
- **Can read files larger than local disk/RAM space**
- Configurable pre-fetch, parallelism and block sizes
- Uses the Buffer cache to deliver accelerated access
- **Useful for large file reads and writes**
  - Checkpointing for AI training/fine tuning
  - HPC: ADAS Simulations, Genomic Sequencing

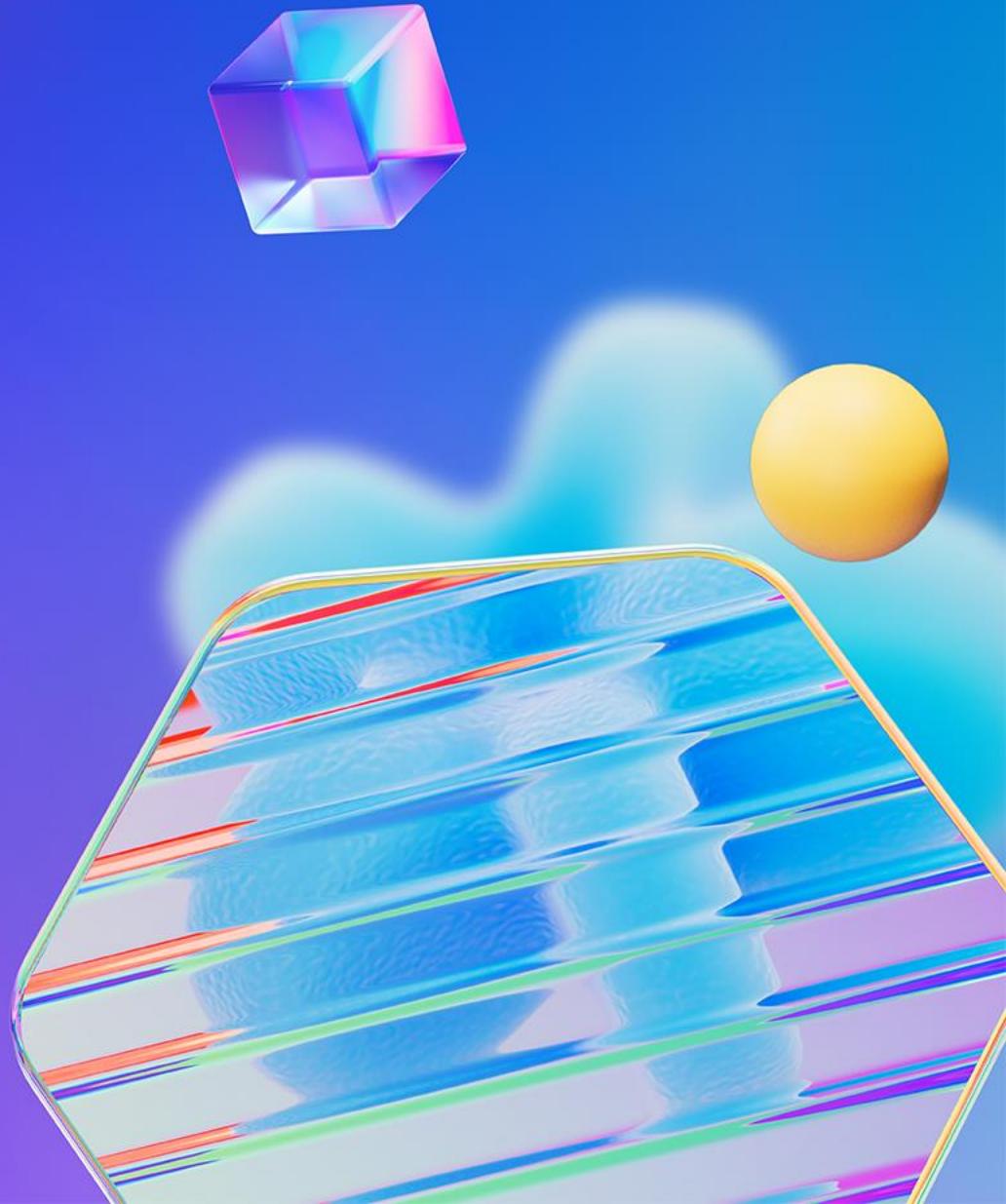
## Computing Environment



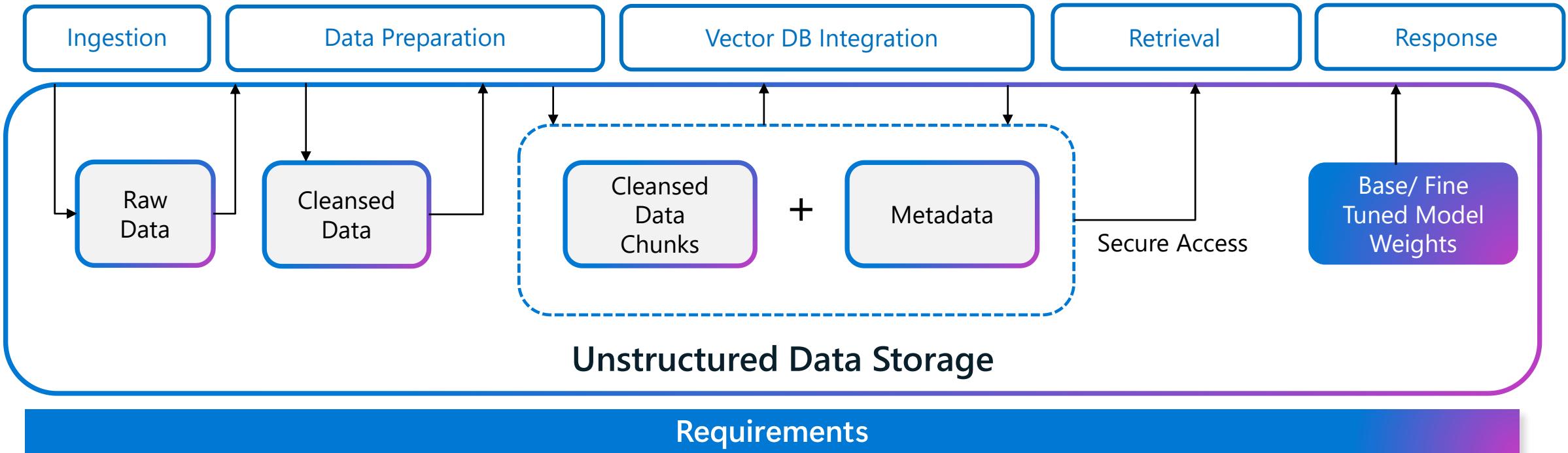
## Computing Environment



# RAG with Azure Blob Storage



# RAG Pipeline – Storage Requirements



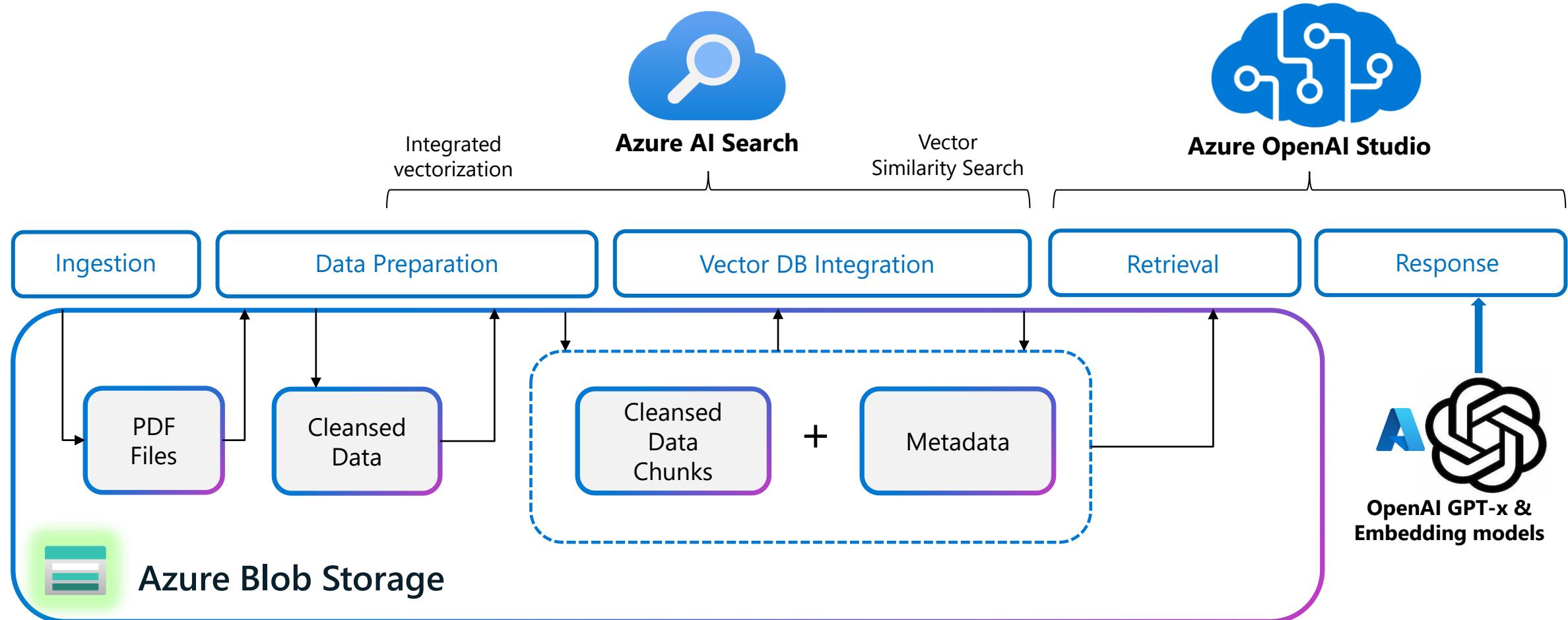
-  Bring enterprise and domain-specific data to foundation models
-  Low latency access to models and data
-  Integration with Vector DB functionality
-  Timely data and index updates
-  Security and access control to core enterprise data

# Premium Blob Storage for RAG

Latency (400KiB PDFs)	
Operation	Premium Blob vs Standard (Hot) Blob
PutBlob (Data Ingest + Chunk Writes)	Premium ~2x faster
GetBlob (Vectorization)	Premium ~3x faster
GetBlob (Retrieval)	Premium ~3x faster

Premium delivers ~3x faster RAG performance with 65% savings on Transactions!

# Azure Blob integration with Azure AI services



# Key Takeaways

## Leveraging Azure Blob for building AI Apps ...

### Ideal for AI Training and Fine Tuning

#### Scalable

to Exabytes of data and many Tbps of throughput

#### Cost-effective

with storage tiers and automated lifecycle management

#### Integrated

with analytics engines for data preparation

#### Interoperable

BlobFuse2 for seamless use in GPU clusters

### Accelerates building RAG based LLM Apps

#### Integrated

with Azure AI services

#### Interoperable

with vector DBs and orchestrators with SDK/tools

#### Secure

with Entra ID integration, RBAC and ABAC

#### Low-Latency Access

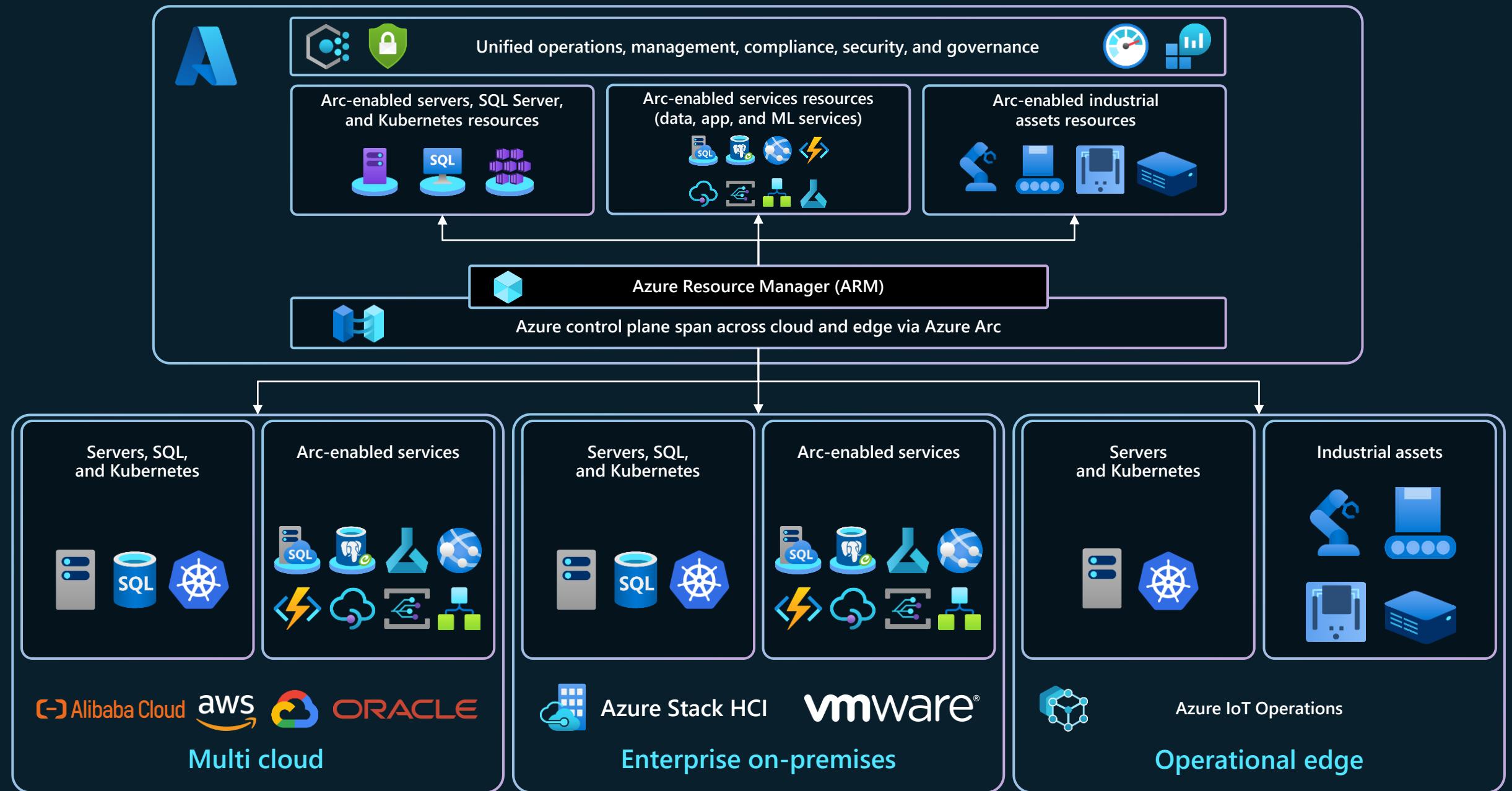
with Premium Blob Storage

#### Freshness

with Blob change feed and change notifications

Public Preview

# Azure Arc Site Manager



# Edge sites monitoring

Home > Azure Arc

## Azure Arc | Site manager (preview) ⚡ ...

Microsoft



Get started Overview Sites



Create



Manage view



Refresh



Export to CSV



Open query

Filter for any field...

Connection : all

Alerts : all

Updates : all

Resource group : All

Scope : all

+ Add filter

Showing 4 to 4 of 4 records.

<input type="checkbox"/>	Site name ↑	Site resources ↑↓	Scope ↑↓	Connection	Alerts	Updates	
<input type="checkbox"/>	United States	81	US_10001	✖ 1 needs attention	✖ 2 critical, +3 more	✖ 1 needs attention	...
<input type="checkbox"/>	Los Angeles	12	LA_10001	✓ Connected	✓ No alerts	✓ Up to date	...
<input type="checkbox"/>	San Francisco	27	SF_10001	✓ Connected	✓ No alerts	✓ Up to date	...
<input type="checkbox"/>	New York	27	NY_10001	✖ 1 needs attention	✖ 2 critical	1 update available	...

[What is Azure Arc site manager \(preview\) - Azure Arc | Microsoft Learn](#)

Tailored for on-prem scenarios such as a store, restaurant, factory



# Microsoft Build





Thank you!

