



Azure AI Foundry & AI Agent Service (Pro-Code)

[Name]
Azure Partner Solutions Architect
Microsoft Americas



AI Alignment Guide



I want a generative AI solution

I want out-of-the-box solutions that work with my existing data estate for my employees		I want to build a solution with custom data and UI, and deploy internally or externally		
Persona-Based Copilots	Microsoft 365 Copilot	Copilot Studio	Azure AI Foundry	
 				
Enhance	Extend	Integrate		
Licensing	Per User	Per User*	PAYG or Capacity Pack	Azure Services Meters
Stories	Link	Link	Link	Link
Persona	Line-of-Business Owner	Knowledge Worker	Power User	Developer
Out-of-box Value	★★★	★★★	★★	★
Customization	★	★	★★	★★★

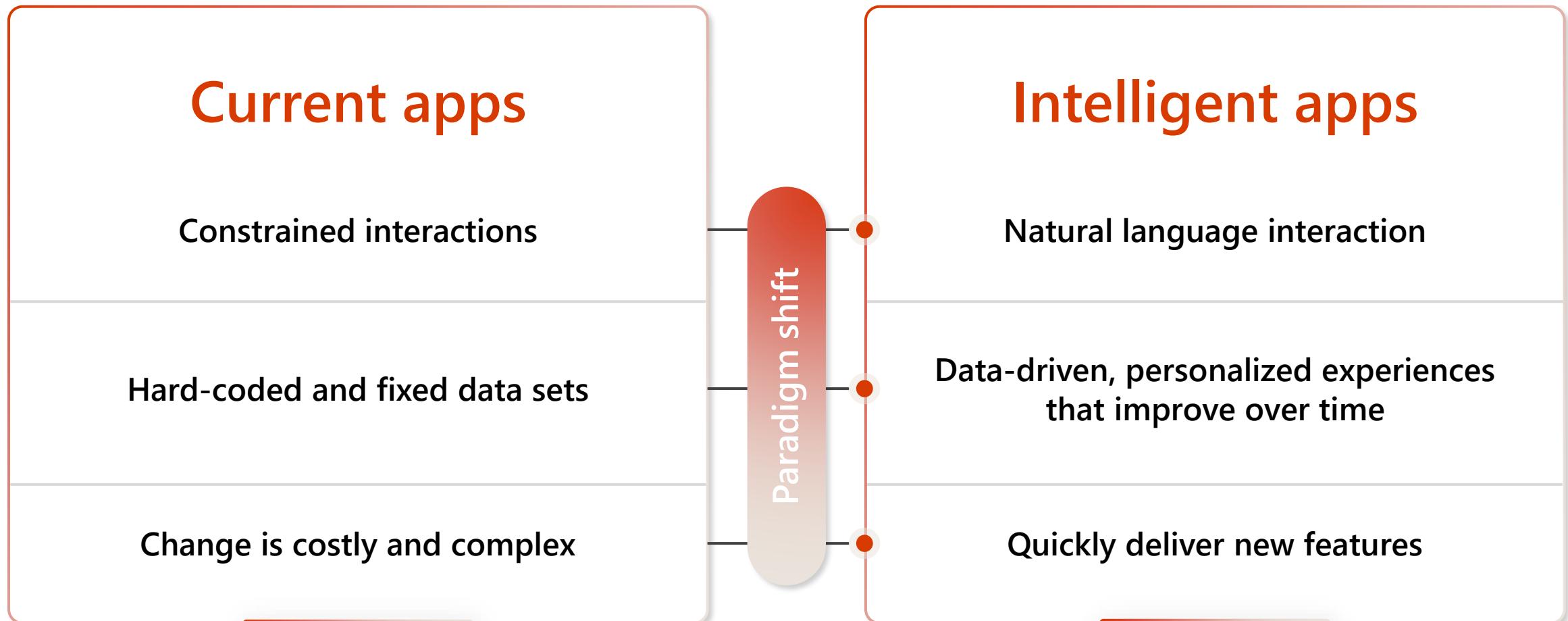


Integrate with Fabric + Purview



* [Microsoft 365 Copilot Chat](#) advanced Agent Capabilities requires Copilot Studio Messages

There is a paradigm shift over current app development



Top use cases for

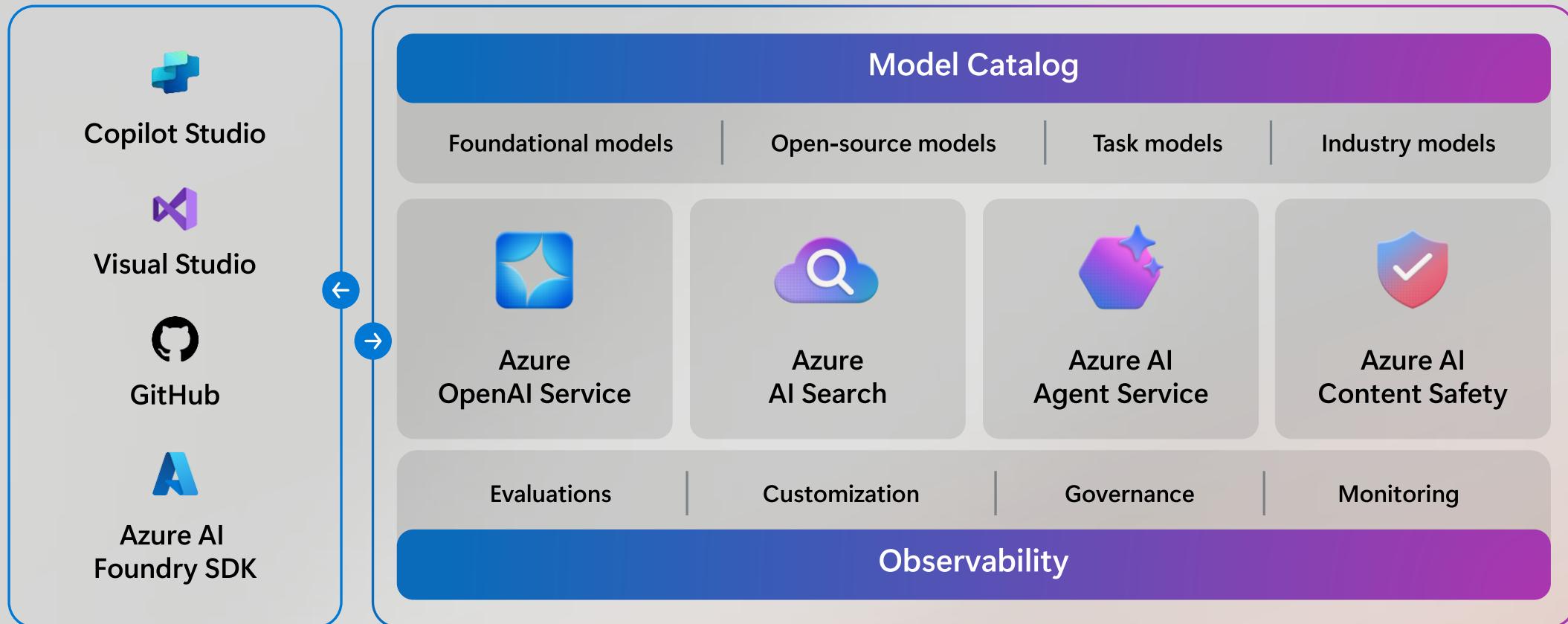
Azure AI Foundry



-  Build your own AI assistant
-  Chat with your data
-  Analyze and summarize documents
-  Generate code and documentation
-  Deliver personalized recommendations
-  Improve information discovery
-  Generate engaging content
-  Process transactions and detect fraud
-  Improve forecasting



Azure AI Foundry



Azure AI Foundry is designed for developers to:

- Build generative AI applications on an enterprise-grade platform.
- Explore, build, test, and deploy using cutting-edge AI tools and ML models, grounded in responsible AI practices.
- Collaborate with a team for the full life-cycle of application development

Build intelligent apps with Azure AI services

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

Azure OpenAI Service

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

Azure AI Search

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

Azure AI Speech

- Speech to text (including the Whisper model on Azure OpenAI Service)
- Text to speech
- Speech translation
- Speaker recognition

Azure AI Vision

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

Azure AI Content Safety

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

Azure AI Document Intelligence

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

Azure AI Language

- Task-optimized AI models for text analytics
- Custom industry-specific AI for healthcare
- Custom, industry-specific models

Azure AI Translator

- Multilingual text and speech translation
- Synchronous and asynchronous translation request support
- Native translation of documents and manuals

Management layers in Azure AI Foundry

AI Foundry targets three different management needs:

Provide **AI developers and business stakeholders** with a SaaS-like self-serve experience, to allow for rapid AI experimentation

Provide **team leads** with central configuration and governance for managing capacity, spend, shareable assets for their team

Provide a compliant, yet non-repetitive or duplicate setup by **IT security** using templates

AI Foundry: projects

Customize in projects

AI Foundry: hub

Share connections, compute, base models

Govern quota and usage

Azure portal

Platform setup

Govern security

Azure AI Model Catalog

Empowering
you to find the
best model for
every use case.

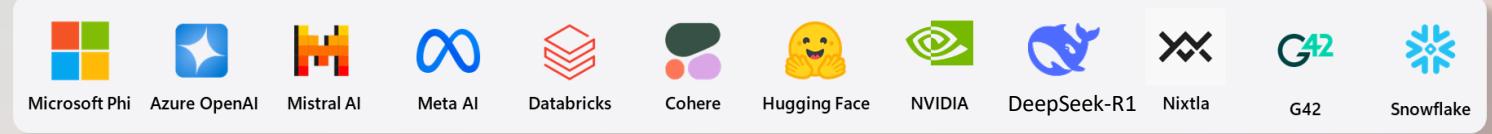
Avoid model lock-in

Flexible model selection

Curated for enterprise

Explore, compare, and
swap models quickly

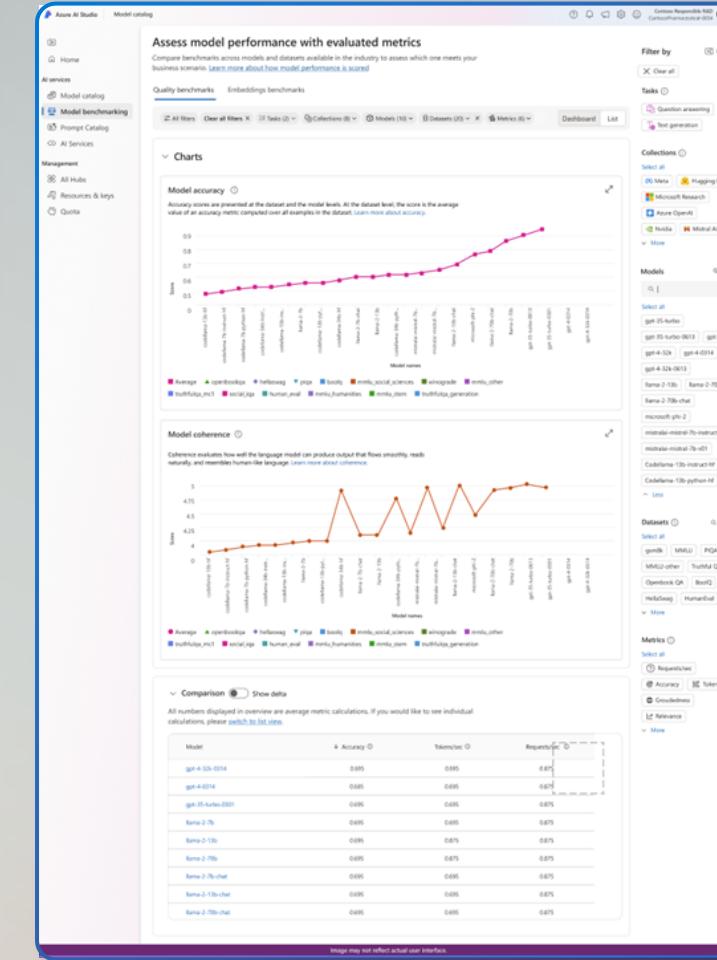
LLMs and SLMs	Regional and domain specific	Open and proprietary	Modalities, tasks, and tools
<p>Flagship LLMs</p> <ul style="list-style-type: none">GPT-4Mistral LargeLlama3 70bLlama 405BDeepSeek-R1 <p>Small language models</p> <ul style="list-style-type: none">Phi3Mistral OSS modelsLlama3 8bMinstral	<ul style="list-style-type: none">Core42 JAIS Arabic language LLMNixtla TimeGEN-1 Timeseries forecastingNTT DATA tsuzumiIndustry models from Bayer, Fidelity, Sight Machine, Rockwell Automation, Cerence, HF	<p>100s of open models from Hugging Face</p> <p>Open models from Meta, Databricks and Snowflake, NVIDIA</p>	<p>Multi-modal</p> <ul style="list-style-type: none">GPT-4o, Phi3-vision, Bria AI <p>Image generation</p> <ul style="list-style-type: none">DALL-E 3, Stability AI <p>Embedding models</p> <ul style="list-style-type: none">Ada, Cohere



Model benchmarking

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

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



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

API and model choice



Models as a Platform (MaaP)

Offering within the model catalog

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

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

API and model choice

The screenshot shows the Azure AI Studio Model catalog interface. At the top, there's a header with the title "Model catalog". Below the header, there's a search bar and a section titled "Find the right model to build your custom AI solution". This section includes announcements about "Phi-3 is now available" and "Build the future of AI with Meta Llama 3". There are also sections for "Cohere models are here" and "All filters". The main area displays a grid of model cards. Each card contains the model name, a small icon, and a brief description. The models listed include "gpt-4-32k", "databricks-dolly", "llama-12b", "gpt-4", "gpt-4-32k", "databricks-dolly", "llama-12b", and "gpt-4". Other visible models include "llama-7b", "Deci-DeciLM-7B", "microsoft-phi-2", "tiiuae-falcon-40b", "llama-7b", "Deci-DeciLM-7B", "microsoft-phi-2", and "tiiuae-falcon-40b". The interface is clean and modern, designed for easy navigation and discovery of AI models.

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



aka.ms/modelcatalog



Models as a Service (MaaS)

Offering within the model catalog

Models as a Service (MaaS) lets you fine-tune models using serverless APIs, making it easier for generative AI developers to build custom copilots.

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

API and model choice

The screenshot shows the Azure AI Studio Model catalog. At the top, there's a search bar and filter options for 'All filters', 'Collections', 'Deployment options: Serverless API', 'Inference tasks', 'Fine-tuning tasks', and 'Licenses'. Below this, a section titled 'Find the right model to build your custom AI solution' displays a grid of 30 models. Each model entry includes a thumbnail, the model name, its purpose (e.g., Chat completion, Embeddings), and a small circular icon with a question mark. To the right of the grid, there's a sidebar titled 'Filter by' with sections for 'Collections' (Curated by Azure AI, Azure OpenAI, Meta, Hugging Face, NVIDIA, Microsoft, Mistral AI, Deci AI, Nixtla, JAIS, Cohere, Databricks, Snowflake), 'Deployment options' (Serverless API, Managed compute), 'Inference tasks' (Conversational, Fill mask, Question answering, Summarization), 'Fine-tuning tasks' (Image classification, Image segmentation, Object detection, Question answering), and 'Licenses'. At the bottom of the catalog area, a purple banner says 'Image may not reflect actual user interface.'



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

aka.ms/modelcatalog



AI Foundry SDK for streamlined development



Create an
AI project to
connect services
and data



Use a common
inferencing API
for our most
popular models



Use agents
to unlock another
level of intelligence

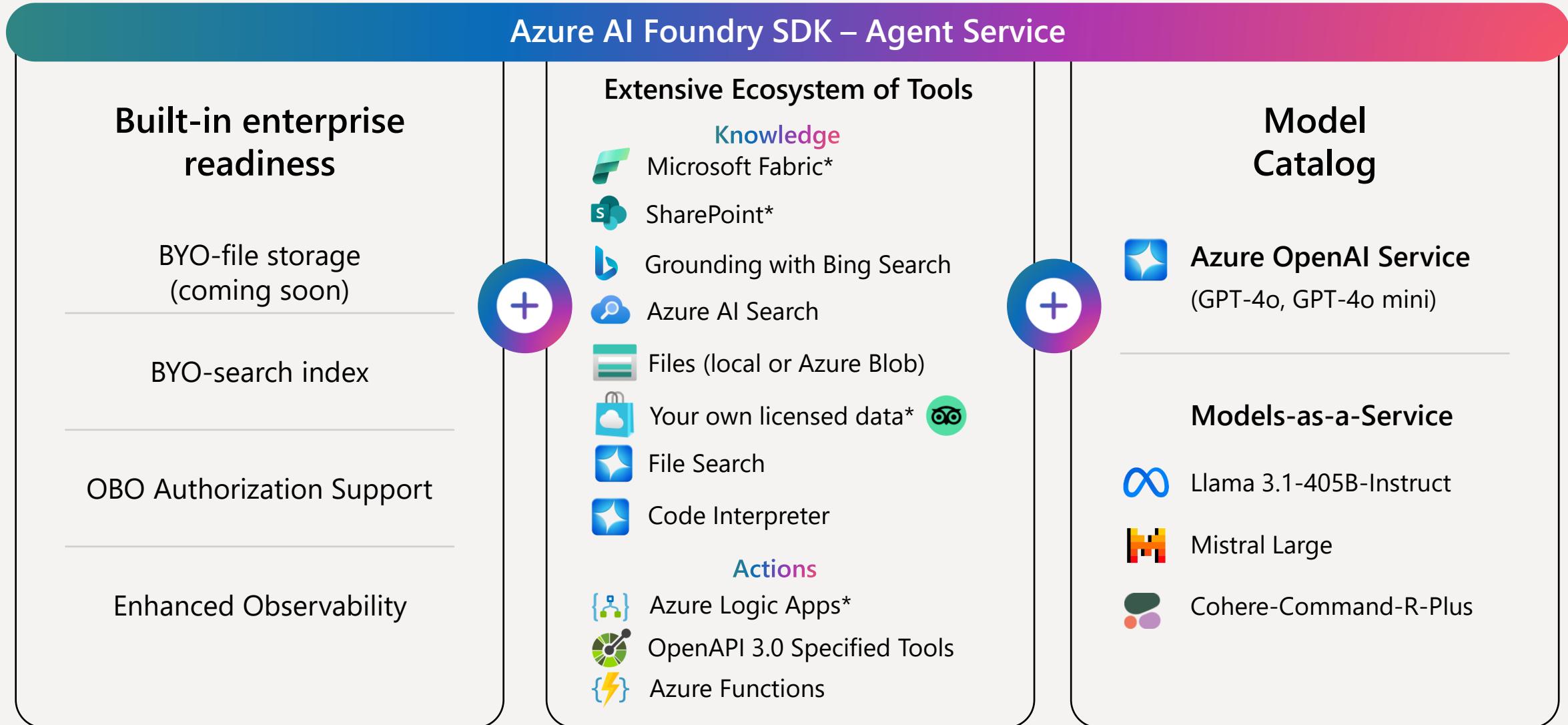
Automatically take actions,
including self-correct

Integrate your existing data
and services (Bing Search,
Azure AI Search, Microsoft
Sharepoint, Microsoft Fabric)



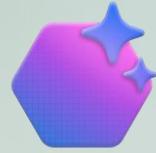
Deploy apps
to Azure using
pre-built
templates

A complete package



*Indicates feature is coming soon

Public Preview



Introducing Azure AI Agent Service

Empower developers to securely build, deploy, and scale AI agents with ease

Rapid Development
and Automation

Extensive Data
Connections

Flexible Model
Selection

Enterprise-grade
Security

AI.Azure.com



When do I use AI agents?

Automate Complex or repetitive tasks that requires human intelligence like **data analysis, natural language processing**.

Enhance Human Capabilities or Augment Human Decision Making such as providing **recommendations, feedback or guidance based on data or preference**.

Create engaging and interactive experiences, such as games, simulations or virtual AI assistants that can **adapt to user behavior and preferences**.

Explore and Discover new knowledge or solutions, such as finding optimal strategies, generating novel designs or solving hard problems that are beyond human reach.

Improve Social and Environmental outcomes, such as supporting education, health or sustainability initiatives that can benefit from AI agents **scalability, efficiency or creativity**.



AI Agents for many Business Use Cases

PRIVACY-COMPLIANT DATA COLLECTION

- Legal Agent: Ensures privacy regulations.
- Marketing Agent: Collects customer data.

VENDOR EVALUATION AND COST OPTIMIZATION

- Procurement Agent: Selects suppliers.
- Product Agent: Assesses quality.

EMPLOYEE ONBOARDING AUTOMATION

- HR Agent: Verifies documents, generates contracts.
- Legal Agent: Ensures compliance with employment laws.

CONTRACTOR INVOICE VERIFICATION

- Procurement Agent: Manages contractor payments.
- Invoice Reconciliation Agent: Validates invoices.

SOFTWARE COMPLIANCE MANAGEMENT

- Tech Agent: Ensures licensing compliance.
- Legal Agent: Reviews software contracts.

MARKETING CAMPAIGN COST ANALYSIS

- Invoice Reconciliation Agent: Analyzes costs.
- Marketing Agent: Evaluates ROI.

IT SUPPORT AUTOMATION

- HR Agent: Handles technical issues.
- Tech Agent: Resolves support requests.

EMPLOYEE DEVELOPMENT RECOMMENDATIONS

- HR Agent: Recommends training programs.
- Product Agent: Suggests career growth opportunities.

LEGAL COMPLIANCE IN MARKETING CONTENT

- Legal Agent: Reviews marketing materials.
- Marketing Agent: Ensures compliance.

PRODUCT COMPLIANCE ASSURANCE

- Legal Agent: Validates industry standards.
- Product Agent: Monitors compliance.

VENDOR CONTRACT NEGOTIATION

- Legal Agent: Evaluates contracts.
- Procurement Agent: Negotiates terms.

PERSONALIZED PRODUCT RECOMMENDATIONS

- Product Agent: Analyzes customer behavior.
- Marketing Agent: Tailors recommendations for campaigns.

PRODUCT QUALITY MONITORING

- Product Agent: Identifies defects.
- Tech Agent: Suggests improvements.

INTELLECTUAL PROPERTY COMPLIANCE

- Marketing Agent: Ensures IP adherence.
- Legal Agent: Reviews content.

TAX-COMPLIANT EXPENSE REPORTING

- Expense Billing Agent: Ensures tax compliance.
- Legal Agent: Reviews expenses.

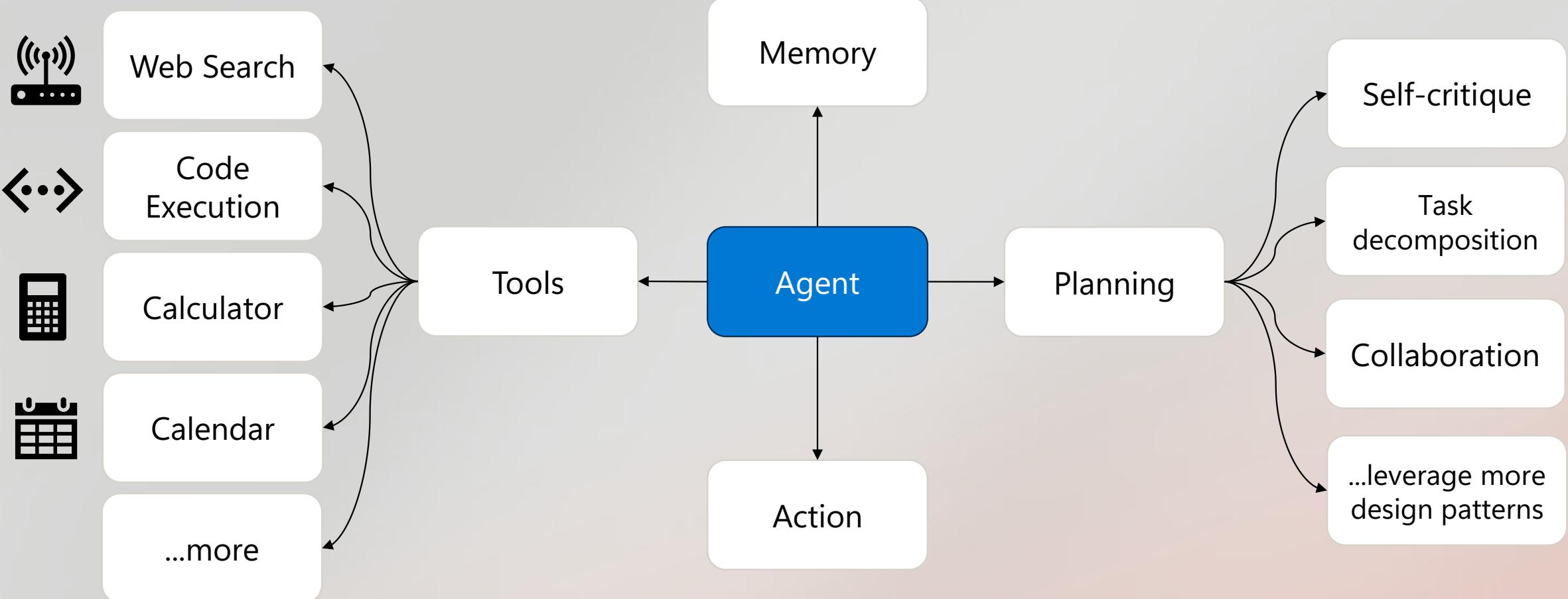
SUPPLY CHAIN OPTIMIZATION

- Tech Agent: Predicts demand, manages inventory.
- Procurement Agent: Automates purchasing decisions.

EXPENSE REPORTING AND INVOICE RECONCILIATION

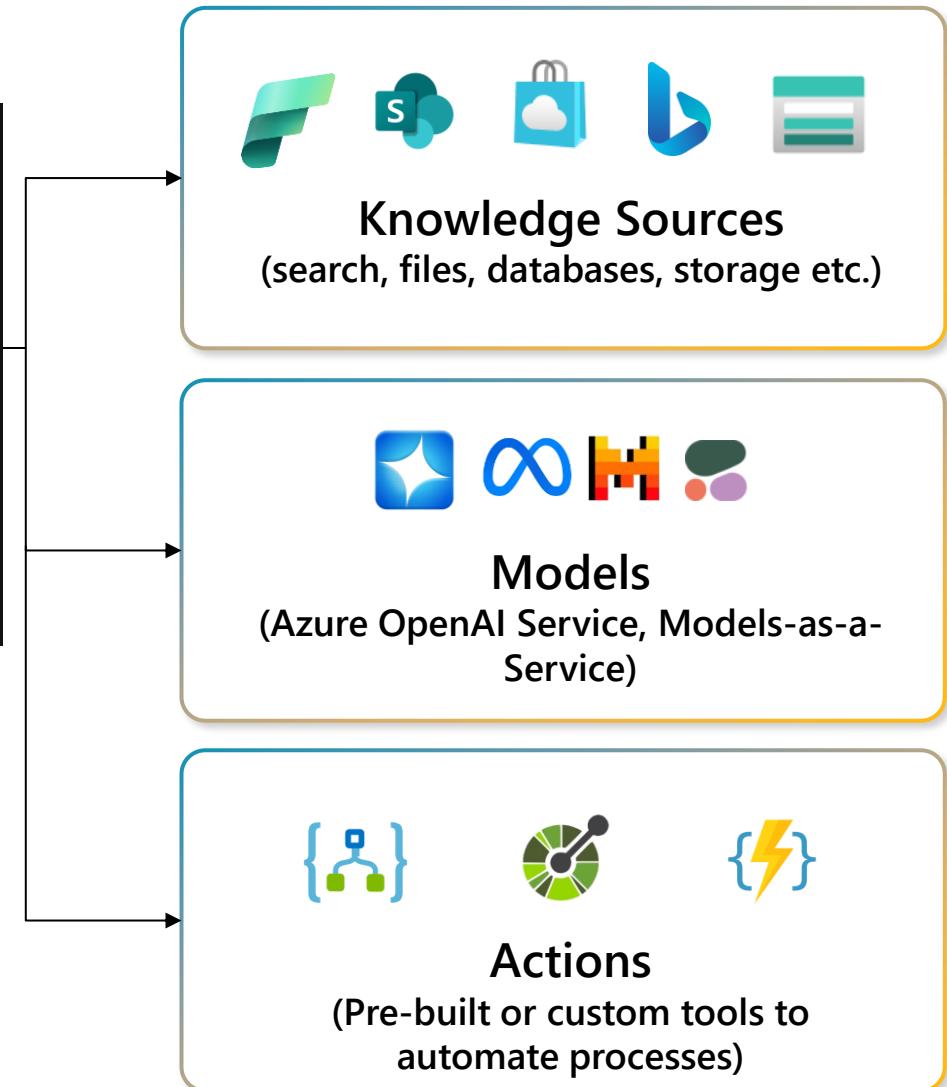
- Expense Billing Agent: Verifies receipts.
- Invoice Reconciliation Agent: Matches invoices with expenses.

Agentic AI capabilities



How does your agent work?

```
Python                                         Copy  
  
agent = project_client.agents.create_agent(  
    model="gpt-4o-mini",  
    name="my-agent",  
    instructions="You are helpful agent",  
    tools=code_interpreter.definitions,  
    tool_resources=code_interpreter.resources,  
)
```



How does your agent work?

Step 1:
Create an Agent

Step 2:
Create a Thread

Step 3:
Run the Agent

Step 5:
Check the Run status

Step 6:
Display the Agent's Response

Agent
Travel Planning Agent

Instructions

You are a travel booking and expense management assistant designed to help employees plan, book, and manage business travel.

Model



Your data (optional)

Azure AI Search

Files (local or Azure Blob)

Tools (optional)

File Search
Code Interpreter
Function Calling
Bing Search
Microsoft SharePoint (coming soon)
Microsoft Fabric (coming soon)
Azure Logic Apps (coming soon)
Azure Functions
OpenAPI 3.0 specified tools

Thread
Travel Planning

User's message

I need to book a hotel in New York for 2 stays.

Agent's message

Here are some suggestions:

User's message

What's the daily meal allowance for the business trip?

Agent's message

The daily allowance for your business trip is \$75, as per company policy.

Run 1

1 Use Tripadvisor API to search the nearest hotel

2 Create message

Run 2

1 Use Microsoft SharePoint to query the company travel policy

2 Create message

Chat Completions API vs Assistants API vs Agents

Chat Completions API (Chat Playground)

- Lightweight and powerful
- Inherently stateless

VS

Assistants API (Assistants Playground)

- Build with OpenAI models
- Stateful (inbuilt conversation state management)
- Access persistent threads
- Automatic management of the model's context window
- Access files in several formats
- Utilized Microsoft-managed storage
- File Search (API handles chunking, embeddings storage and creation, and implementing vector search)
- Code Interpreter
- Function Calling

VS

Azure AI Agent Service (Agents' Playground)

- **ASSISTANTS API FEATURES, plus**
- Build with a model of your choice (OpenAI, Llama, Mistral, Cohere...)
- Real time web-grounding with Bing
- Secure grounding on enterprise data in SharePoint and Fabric
- Bring Your Own Licensed data (Tripadvisor)
- Connect to 1400+ data sources and services with Azure Logic Apps
- Long running, event driven actions with Azure Functions
- Standardized OpenAPI 3.0 tools
- Bring Your Own Storage
- Bring Your Own Private Network
- Bring Your Own AI Search Resource
- Limitless scaling with PTUs
- Open Telemetry based tracing

Comparison Chart

Feature/Aspect	Persona-based Copilots	Microsoft 365 Copilot	Copilot Studio	Azure AI Foundry
Purpose	Specialized copilots tailored for user personas.	Enhancing productivity within Microsoft 365 apps.	Authoring and customizing copilots for specific use cases.	Enterprise-grade AI adoption with customizable solutions.
Target Users	End-users needing tailored copilots for specific roles.	Knowledge workers and professionals using Office apps.	Developers and knowledge workers creating copilots for apps or workflows.	Developers, data scientists, and enterprises.
Capabilities	Focused assistance for specialized tasks or industries.	Integrates generative AI in Word, Excel, Teams, etc.	Building, testing, and deploying custom copilots.	Large-scale AI model training, deployment, and optimization.
Integration	Can integrate across platforms based on persona needs.	Embedded directly into Microsoft 365 ecosystem.	Tightly integrates with Azure AI Foundry and APIs.	Integrates with Azure AI services, data platforms, and custom ML models.
Customization	Moderate: Persona-based customization predefines scope.	Limited: Customization through settings or plugins.	High: Enables bespoke copilots for specific workflows.	High: Supports tailored AI models and pipelines.
Pricing	Varies based on implementation and platform usage.	Microsoft 365 subscription plus Copilot add-on fee.	Subscription (\$200/month for 25K messages) or pay-as-you-go.	Metered, capacity packs, or subscription-based.
Key Differentiator	Persona-driven AI designed for specific roles/tasks.	Seamless integration within Microsoft Office apps.	Flexibility in creating custom copilots across apps or tools.	Advanced enterprise tools for scalable, bespoke AI projects.
Example Use Cases	Virtual assistants for healthcare, education, or HR tasks.	Generating reports in Excel or drafting emails in Outlook.	Building a customer service chatbot or workflow assistant.	Developing an AI system for predictive analytics or fraud detection.
Business Outcome	Provide highly personalized, role-specific assistance to boost employee performance and job satisfaction.	Increase productivity, automate routine tasks, and enhance collaboration within Microsoft 365 apps.	Improve customer engagement and optimize workflow automation across apps.	Drive efficiency and innovation in large-scale enterprise applications (e.g., predictive analytics, automation).
Deployment	Cloud-based or on-device, depending on persona need.	Cloud-based within Microsoft 365 services.	Cloud-based with Azure dependencies.	Cloud-based, with on-prem options for enterprise solutions.
Gaps/Limitations	Less flexible for tasks requiring general-purpose AI tools.	Not suitable for workflows outside the Microsoft ecosystem.	Limited to AI-based copilots; lacks broader productivity tools.	Complex setup may not suit smaller businesses or non-enterprise users.

Capabilities Chart

Feature/Aspect	Microsoft 365 Copilot	Persona-based Copilots (D365 Copilot for Sales example)	Copilot Studio	Azure AI Foundry
Models	Fine-Tuned M365 App Specific	Fine-Tuned Persona Specific	Fine-Tuned Best-in-Class	+ 3000 model catalog Custom models
Deployment	License Allocation	License Allocation	Tenant Capacity or Meter	Dedicated Serverless
Integration	Microsoft Graph Power Platform	Copilot Studio	Microsoft Graph Power Platform	Multiple Options Azure AI Services
Development	Extend Microsoft 365 Copilot	Deploy for D365 or Salesforce	Studio Web UI	Studio Web UI AI Foundry SDK
Knowledgebase (RAG)	Built-In	Built-In	Built-In BYOM and Custom Engine	Azure AI Search
Action	Built-In M365 App Specific	Built-in	Built-In Power Platform	Built-in Pro-Code Actions
Orchestration	Proprietary Generative Orchestration	Proprietary Generative Orchestration	Built-In Classic or Generative Orchestration	Prompt Flow Semantic Kernel LangChain LlamaIndex AutoGen
Agents	Extend	Built-in	Built-In	Azure AI Agent Service
Channels	Built-in M365 App Specific	Built-in M365 App Specific	Internal/External Channels	REST Endpoint Serving
Responsible AI	Built-in	Built-in	Built-in	Azure AI Content Safety
Monitoring & Reporting	Copilot Analytics and Dashboards	Copilot Analytics and Dashboards	Power Platform Admin Center	Multiple Monitoring Options

Resources

Agents Principles

[AI agents — what they are, and how they'll change the way we work](#)

[Microsoft's Agentic AI Frameworks: AutoGen and Semantic Kernel](#)

Agents with AI Foundry

[Introducing Azure AI Foundry Agent Service to scale your AI agents \(video\)](#)

[Azure AI Agent Service documentation](#)



Thank you