Microsoft Azure

Estado del Arte Agentic AI

Generative AI trends

93%
organizations are experimenting with multiple models[1]

61%
people are wary about trusting AI systems[3]

50%
generative AI will launch agentic AI pilots or POC by 2027[2]

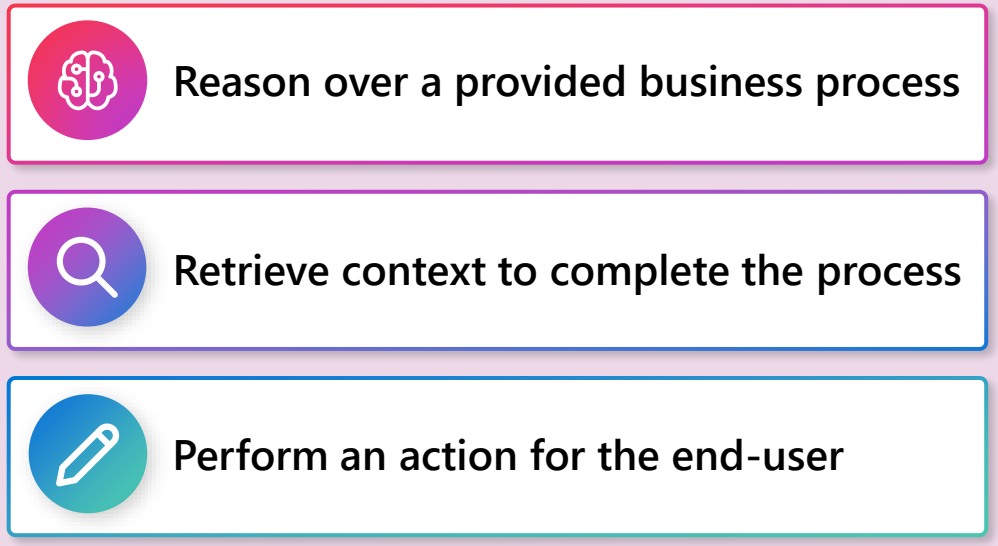
30%
or fewer generative AI experiments moved to production by most respondents[4]

1. 16 Changes to the Way Enterprises Are Building and Buying Generative AI | Andreessen Horowitz
2. Autonomous generative AI agents | Deloitte Insights
3. Trust in artificial intelligence – 2023 Global study on the shifting public perceptions of AI, KPMG
4. GenAI and the future enterprise | Deloitte Insights

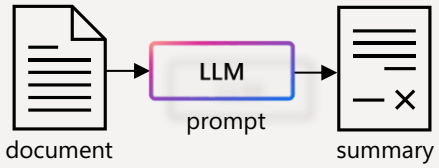
Agentic AI

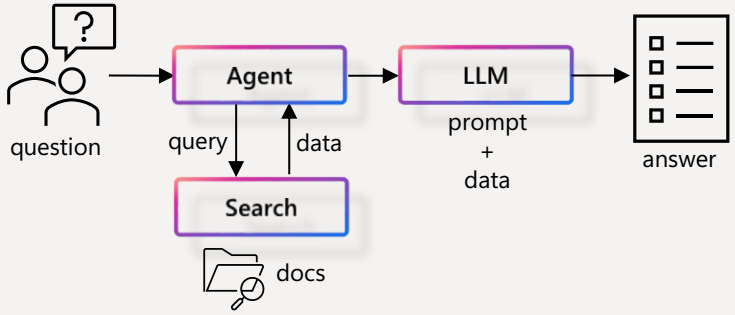
use AI to automate and execute business processes, nearly autonomously

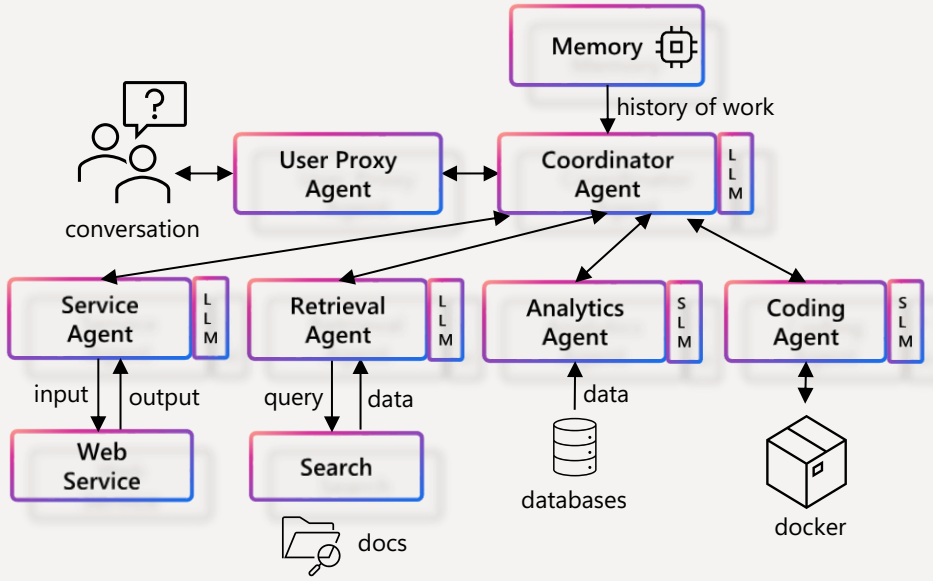
AI agents can plan, adapt to new information, and execute tasks independently, making them capable of handling complex, dynamic environments, decision-making and autonomous action. They can learn from feedback, adjust their strategies, and operate with minimal human oversight.



Evolution of LLM-based Solutions







No Agent

Very narrow one shot task

Ex: log to JSON

Single Agent

Very clearly scoped iterative task

Ex: providing an answer with supporting evidence to a complex question

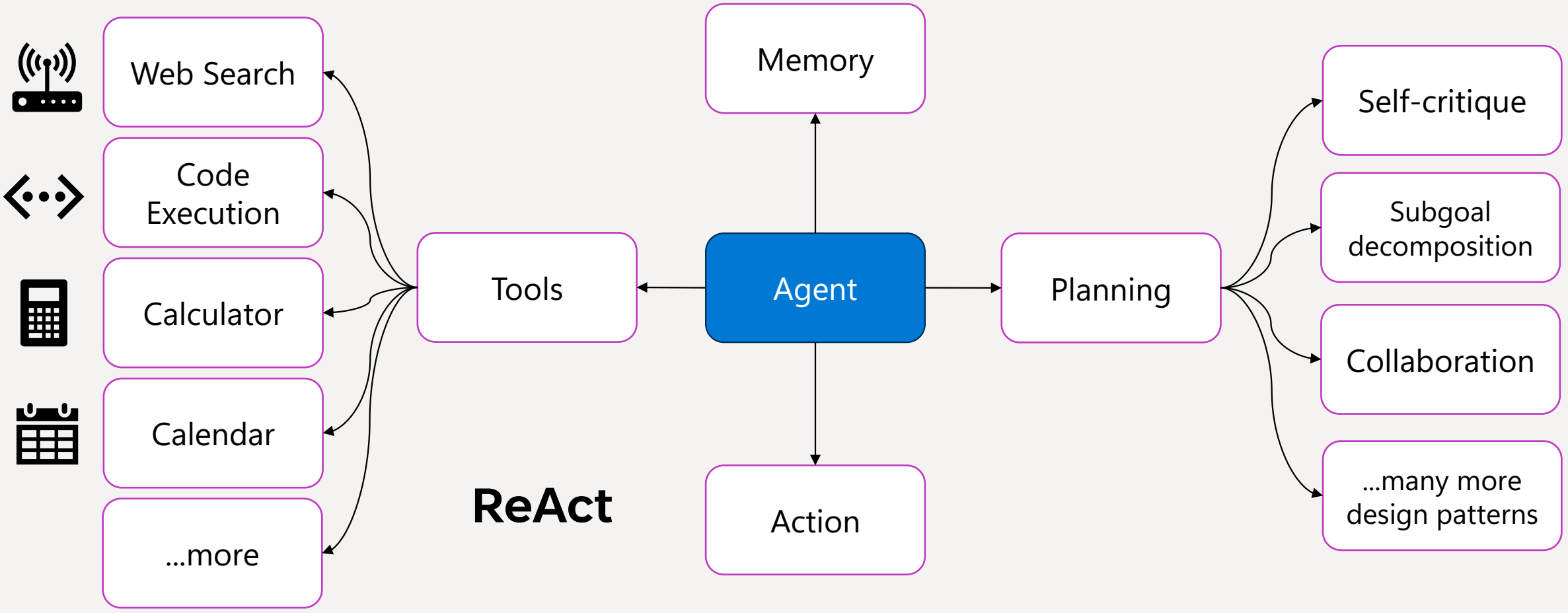
Multi-agent Systems

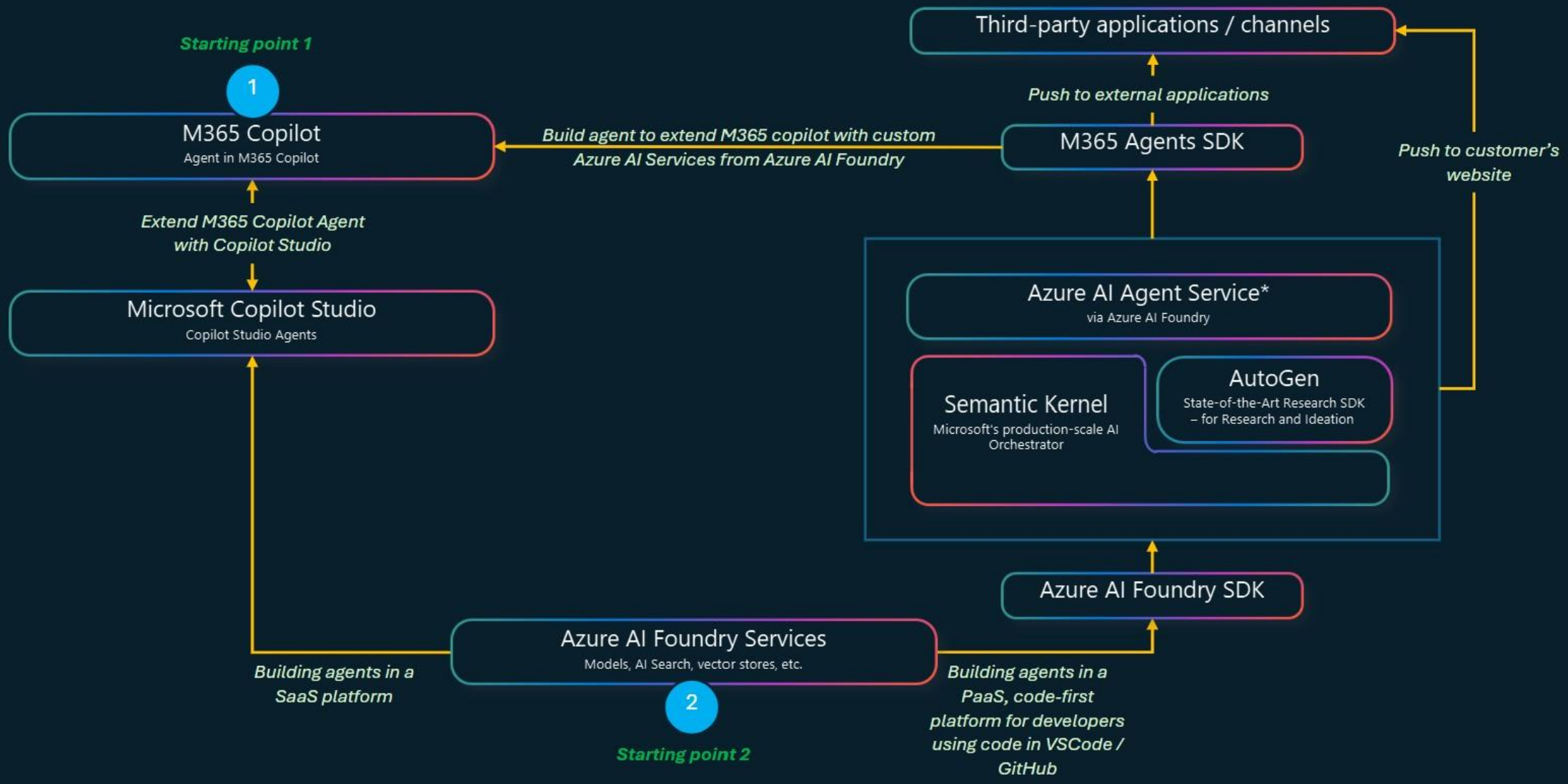
Wide scope complex use case requiring diverse skills

Ex: Propose 2 Instagram marketing campaigns including assets that would leverage the top 2 recent trends in our past quarter US Sales to boost our mailing list user base and predict the impact of each campaign

VALUE

Agentic AI capabilities





** Azure AI Agent Service is also part of the Azure AI Foundry*

Common Design Patterns

RAG Agent

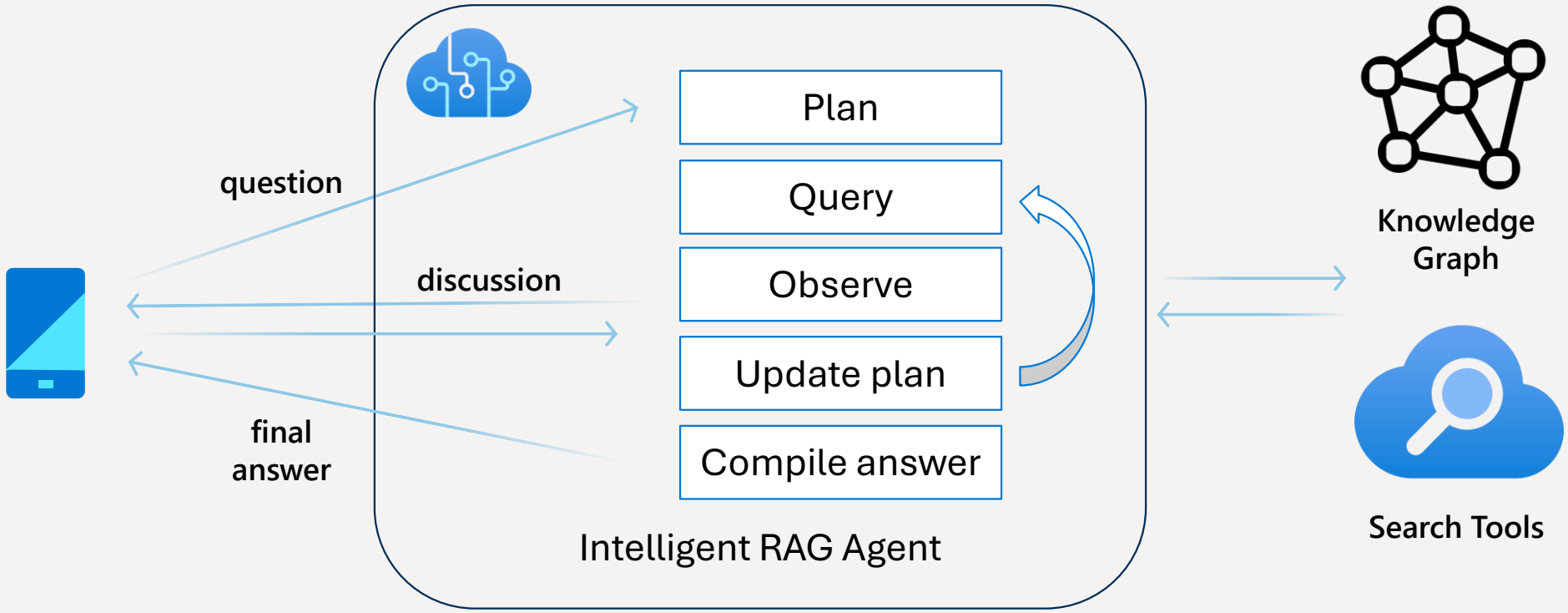
Code Generation Agent

Multi-agent Systems

Multi-domain Agent Systems

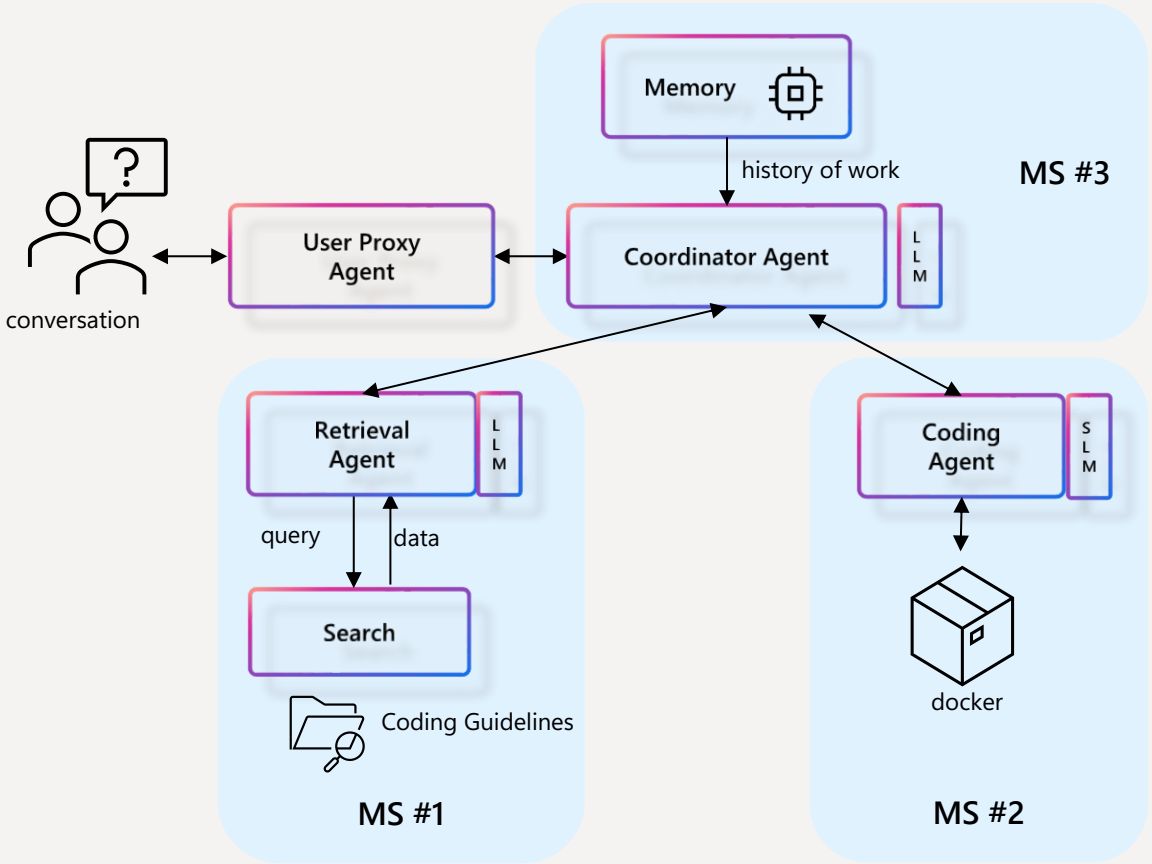
Retrieval Augmented Generation Agent

Translates questions into a research problem with human in the loop to produce high quality answers to complex questions within the scope of its domain



Multi-Agent System

A complex problem is decomposed into smaller, manageable parts, each addressed by specialized agents, effectively a micro-service (MS). These agents work together in a coordinated manner within a workflow to efficiently solve the overall problem.

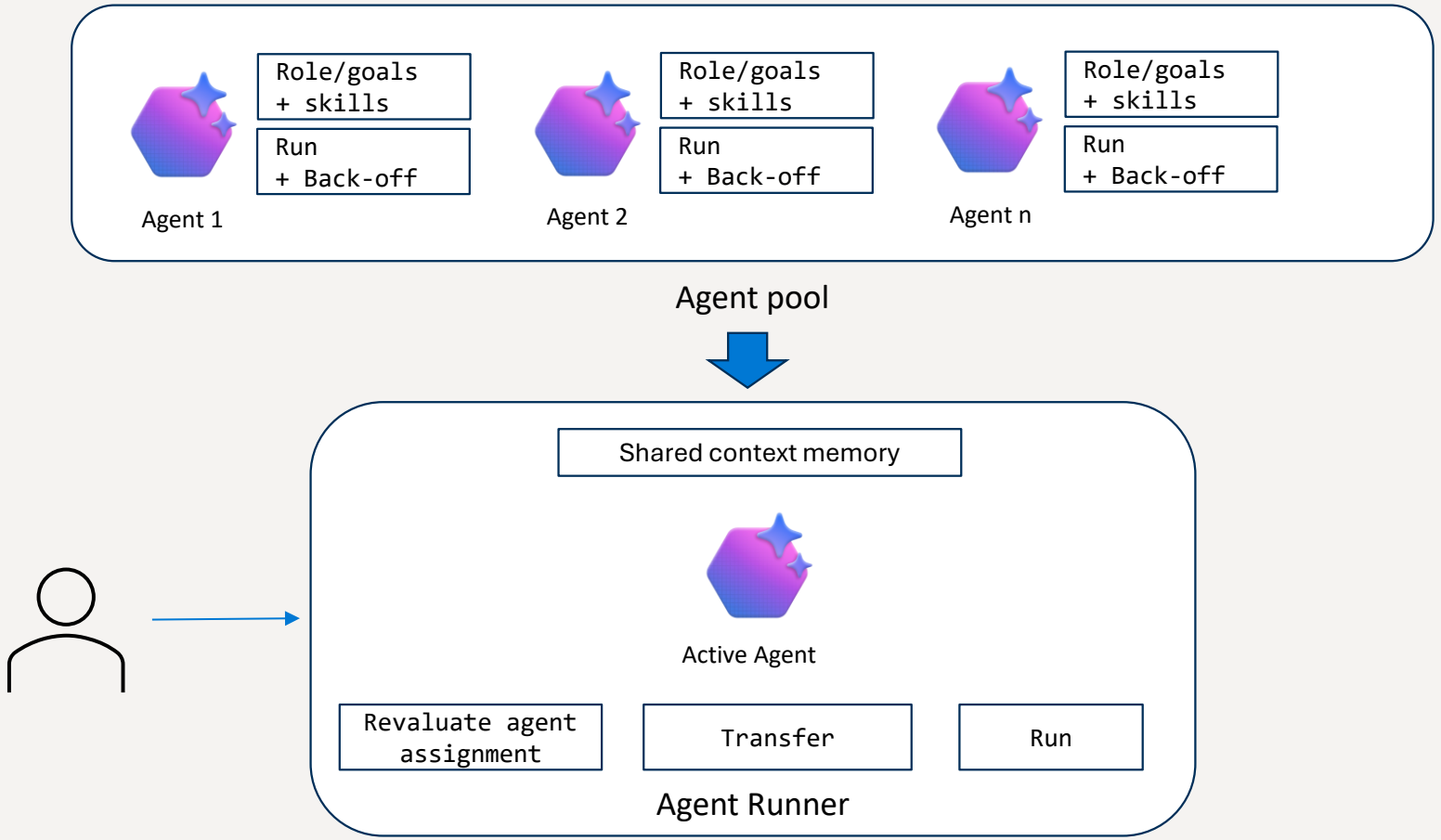


Critical Design Elements

- Adaptive planning within scope of existing tightly scoped skills (agents)
- Handles ambiguity by discussing and refining requirements with human
- Memory to handle complex long running execution of a plan
- Effective inter agent communications
- Test, monitor, release & maintain each agent independently to quickly handle quality & safety issues

Multi-Domain Agents System

Multiple domain-specific agents are orchestrated by an Agent Runner to scale across multiple domains while appearing as a single agent to users.



Critical Design Elements

- Agents capability descriptors
- Scalable Agent Runner able to manage 10s to 100s of agents
- Ability to manage domain switching with proper memory management
- Avoid single interceptor problem as individual agents maintain direct communication with user and can hand off when needed

The full enterprise package

Azure AI Agent Service

Trust

Customer control over data, networking, and security

- BYO-file storage
- BYO-search index
- BYO-virtual network*
- BYO-thread storage*

Choice

Model choice and flexibility with the model catalog

Azure OpenAI Service

GPT-4o, GPT-4o mini, etc.

Models-as-a-Service

- Llama 3.1-405B-Instruct
- Mistral Large
- Cohere-Command-R-Plus

Skills

Richest set of enterprise connectivity

Knowledge

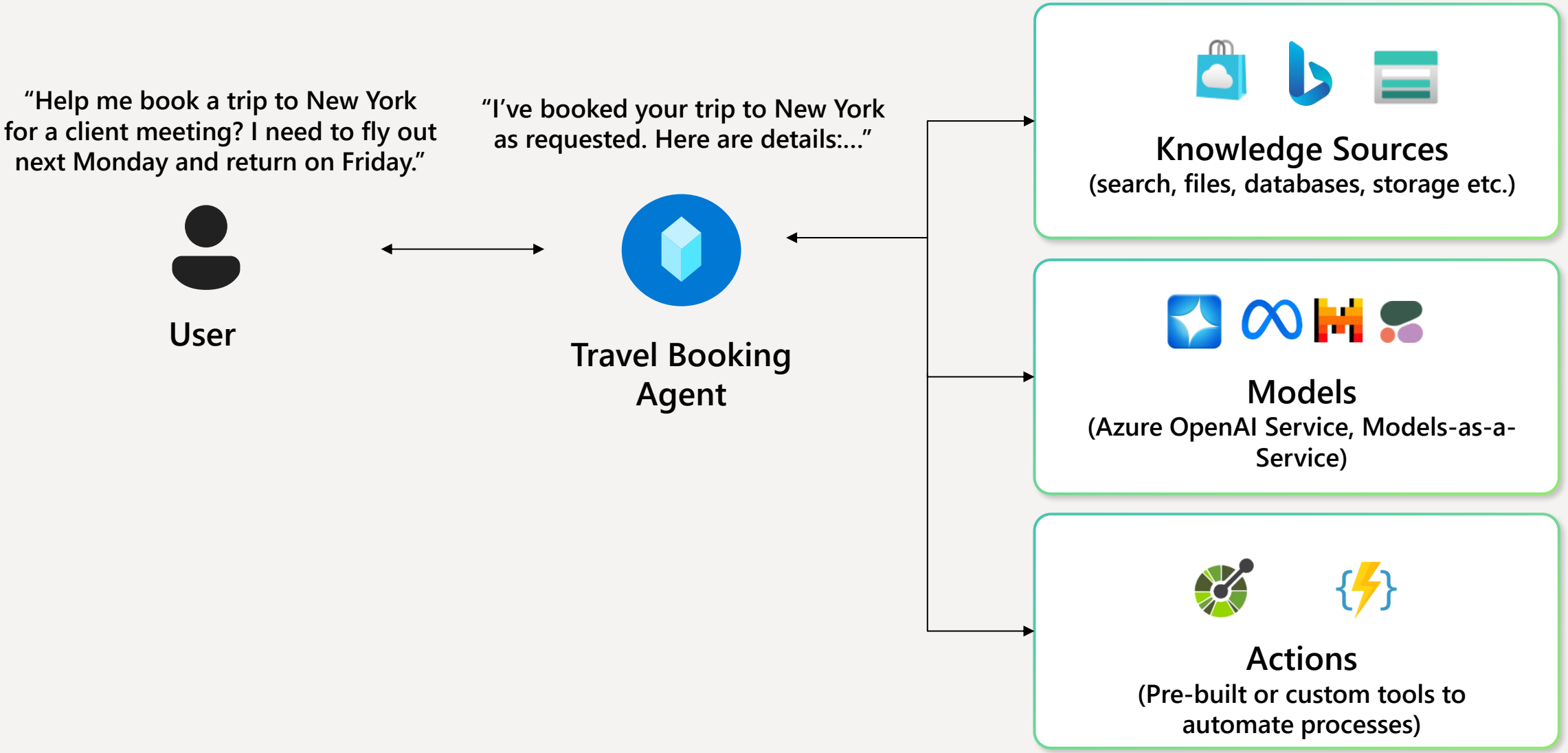
Actions

Logic Apps* Azure functions OpenAPI

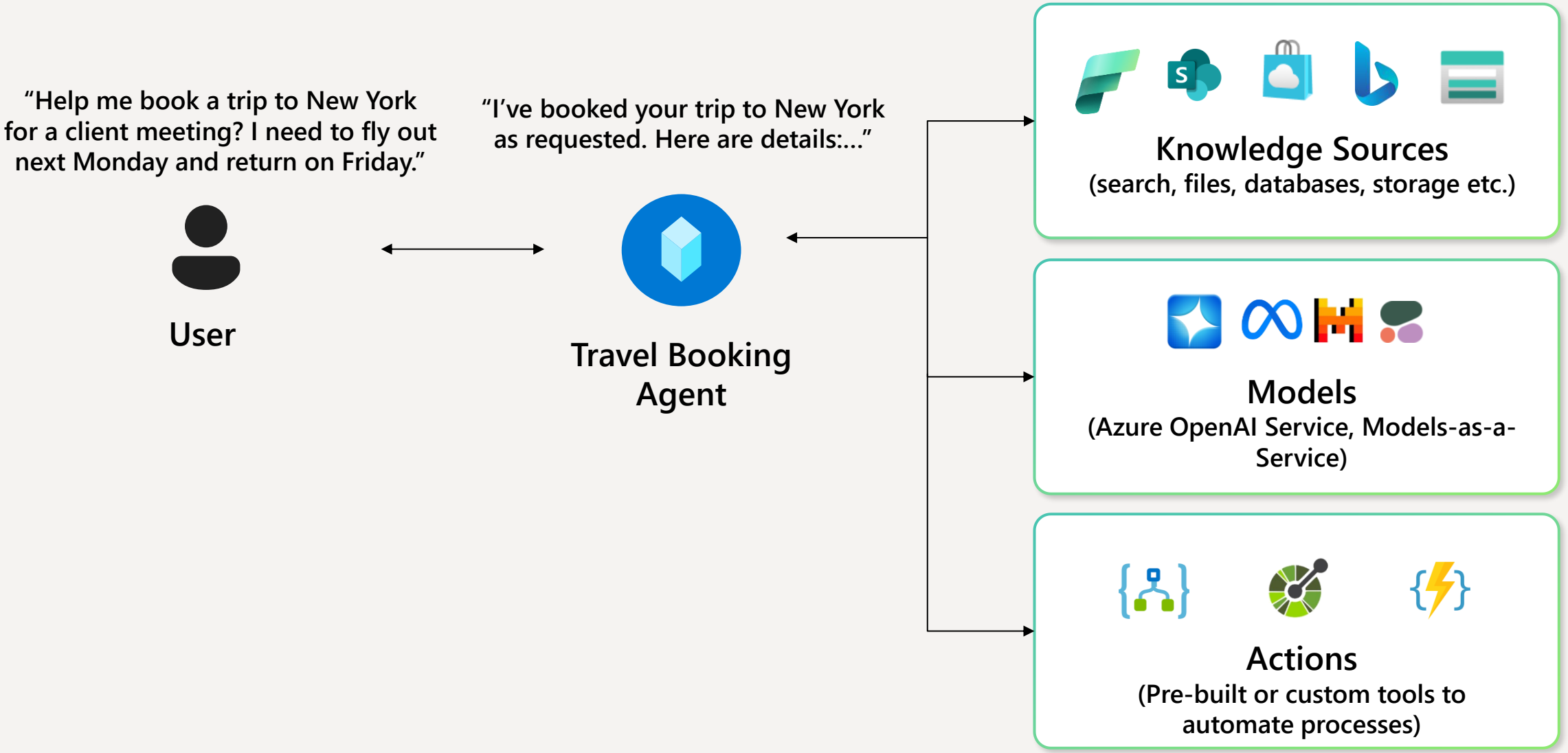
Azure AI Foundry SDK

Azure AI Foundry portal

How does your agent work?



How does your agent work?



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

Extensive Data Connections

Extensive Knowledge Connections

Public web data

Corporate data

Structured data

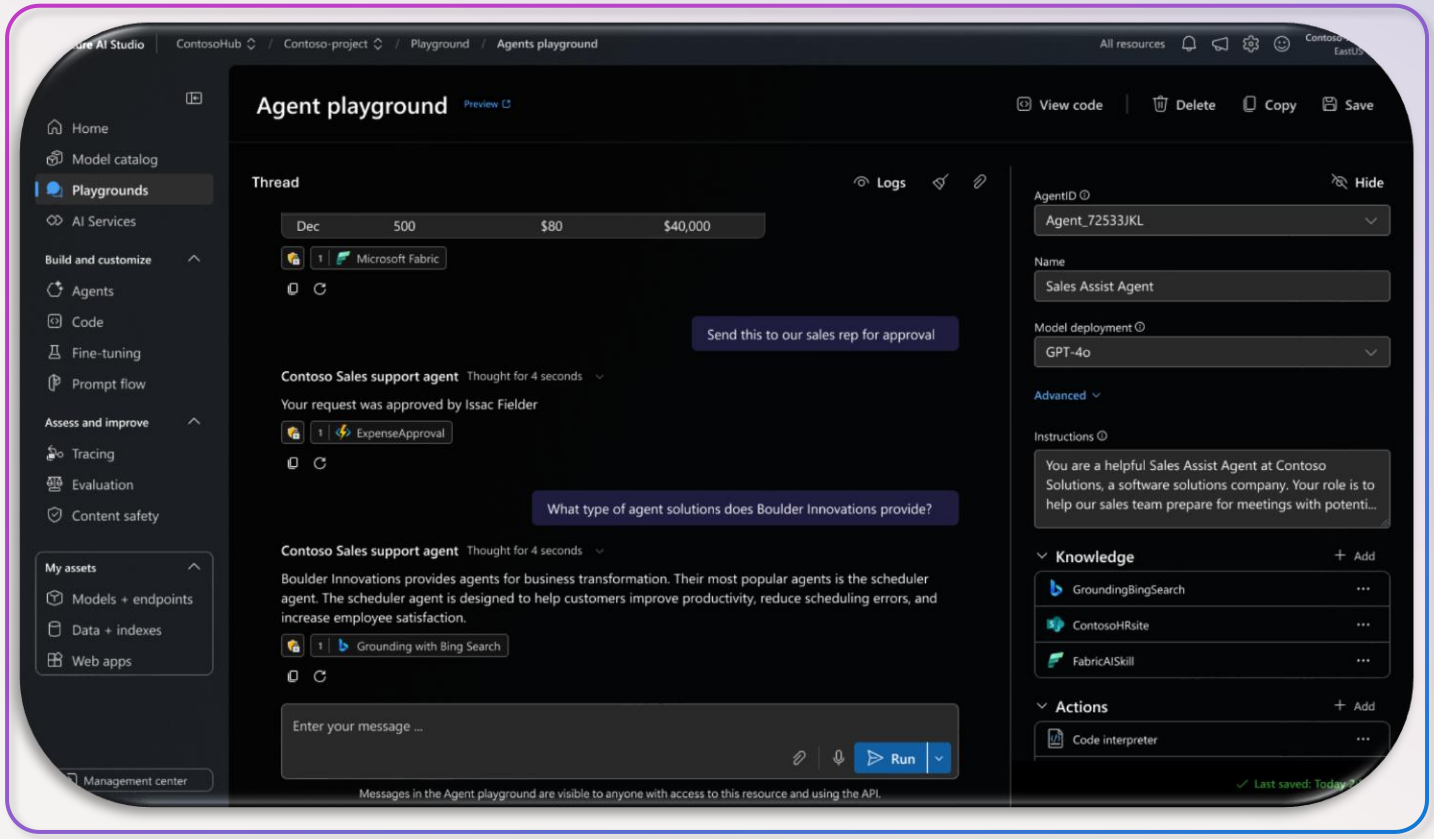
Unstructured data

External data

Real-time, web data grounding with Bing Search

- Create more reliable and trustworthy applications with real-time public information from Bing
- Grounding with Bing Search allows your agents to integrate real-time public web data, ensuring their response is accurate and up to date.
- By including supporting URLs and search query links, Grounding with Bing Search enhances trust and transparency, empowering the users to verify responses with the original sources.

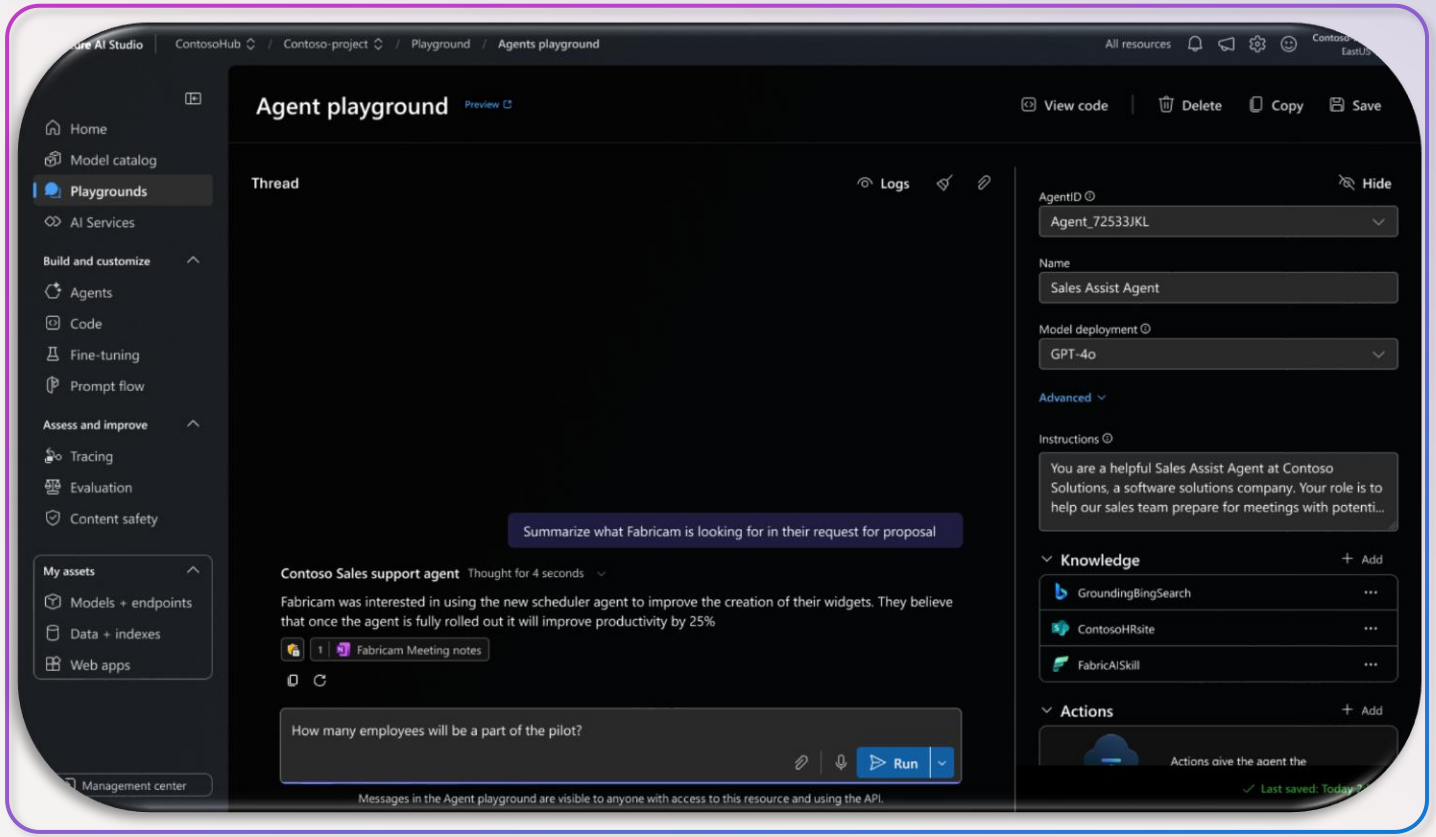
Data Connections



Grounding with private data in Microsoft SharePoint (coming soon)

Data Connections

- Grounding with SharePoint makes your SharePoint content more accessible to your end users.
- Enterprise-grade security features, such as On-behalf-of (OBO) authentication for SharePoint, ensure secure and controlled access for end users.

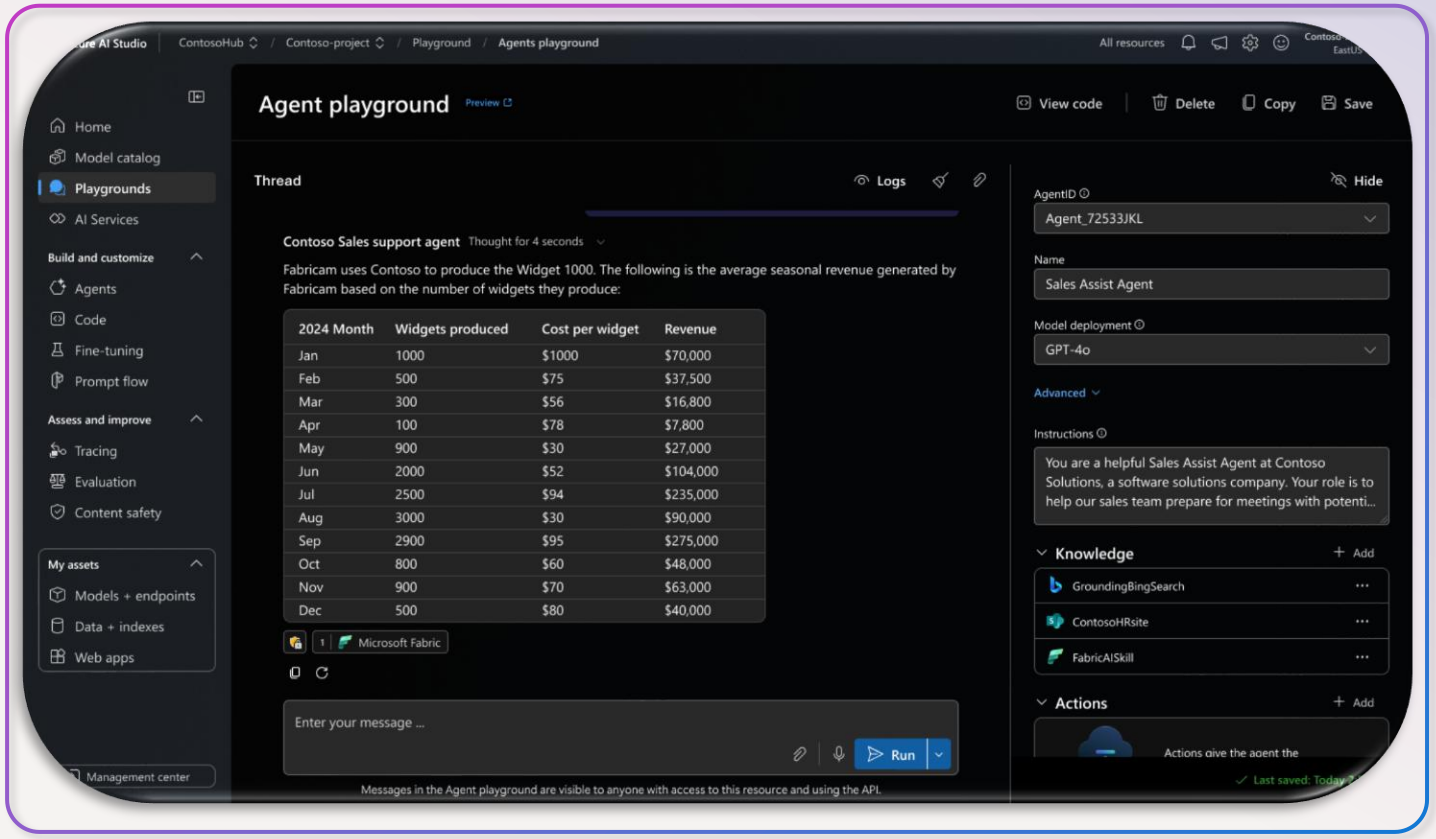


Chat with your structured data in Microsoft Fabric

(coming soon)

- Integrate your agents with Fabric AI Skill to unlock powerful data analysis capabilities.
- Fabric AI Skills can transform enterprise data into conversational Q&A systems, allowing users to interact with data through chat and uncover data-driven and actionable insights effortlessly.
- On-behalf-of (OBO) authentication simplifies access to enterprise data in Fabric while maintaining robust security, ensuring proper access control and enterprise-grade protection.

Data Connections

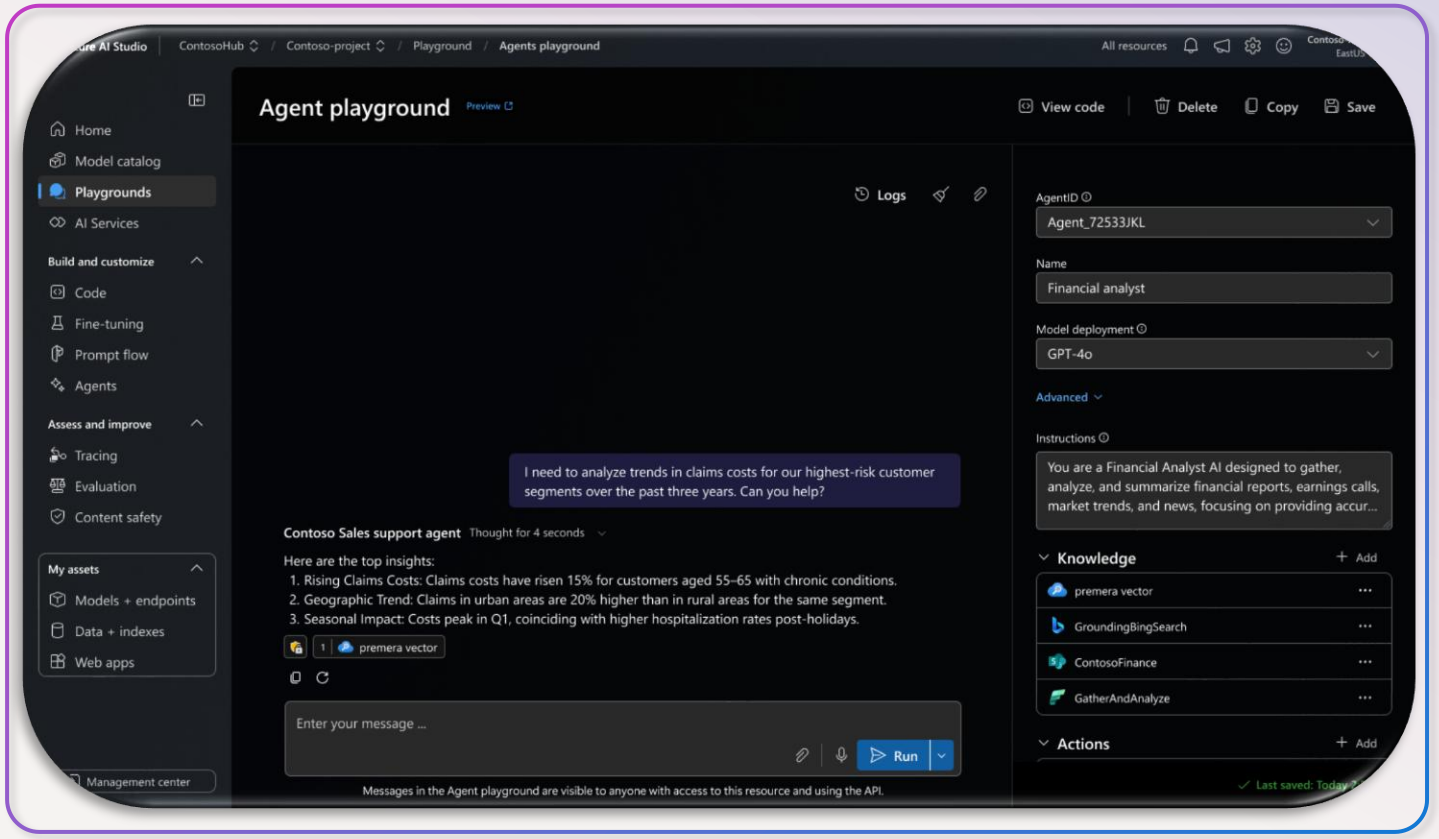


Playground experience in Azure AI Foundry portal is coming soon in January

Data Connections

Incorporate private data in Azure AI Search, Blob, and your local files

- Bring your existing Azure AI Search index or create a new one using the improved File Search tool.
- This tool leverages a built-in ingestion pipeline to process files from your local system or Azure Blob Storage.
- Maintain complete control over your data as your files remain in your own storage, and your Azure AI Search resource ingests them



Flexible Model Selection

Flexible Model Selection

Azure OpenAI Service models

Models as a Service (Serverless API)

Azure AI Services models

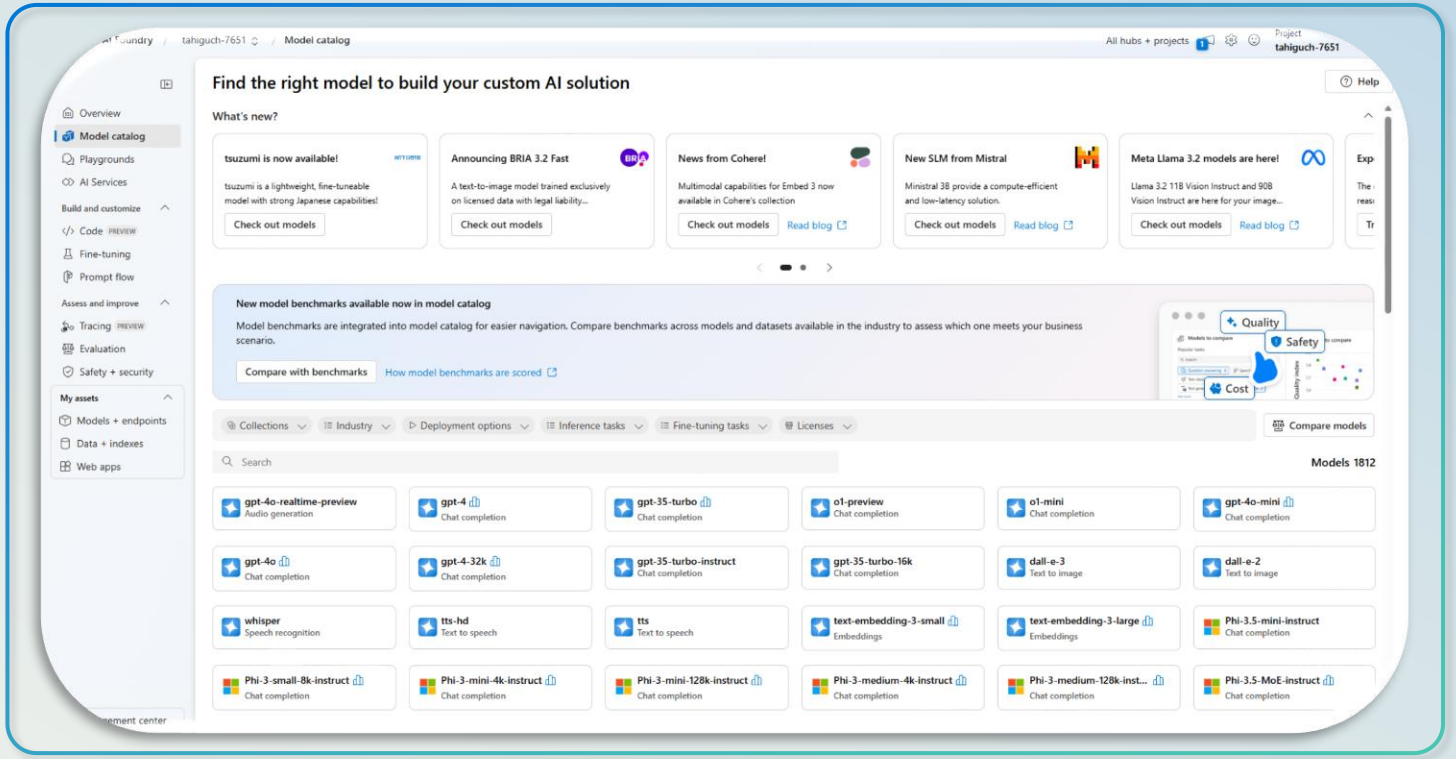
Multi-modal support

Fine-tuned models

Leverage Models as a Service (MaaS)

- Support a variety of models from Azure AI Foundry model catalog to power your agent's reasoning. The supported models include:
  - GPT-4o
  - GPT-4o mini
  - Llama 3.1
  - Mistral Large
  - Cohere Command R+
- Leverage Model Inference API to easily swap models and compare performance to find the best model for your specific needs.

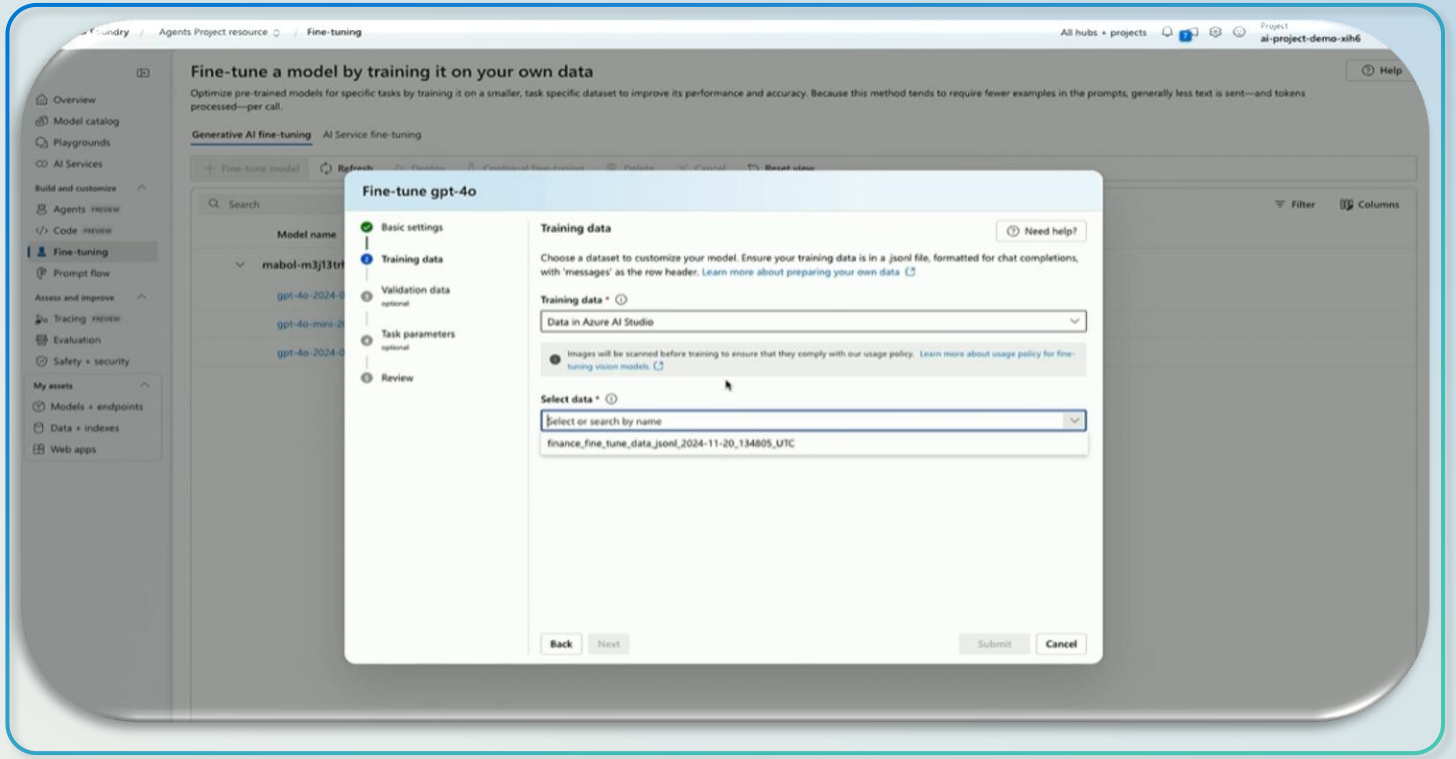
Flexible Model Selection



Use fine-tuned models for your agents

- Customize AI models to address your unique business needs with fine-tuning, enabling you to build task-specific agents while optimizing token costs.
- Through collaborations with partners like Scale AI, you can efficiently label training data and create fine-tuned models that integrate seamlessly with Azure AI Agent Service.
- This process enhances the performance of AI agents, streamlining development and reducing time to production.

Flexible Model Selection

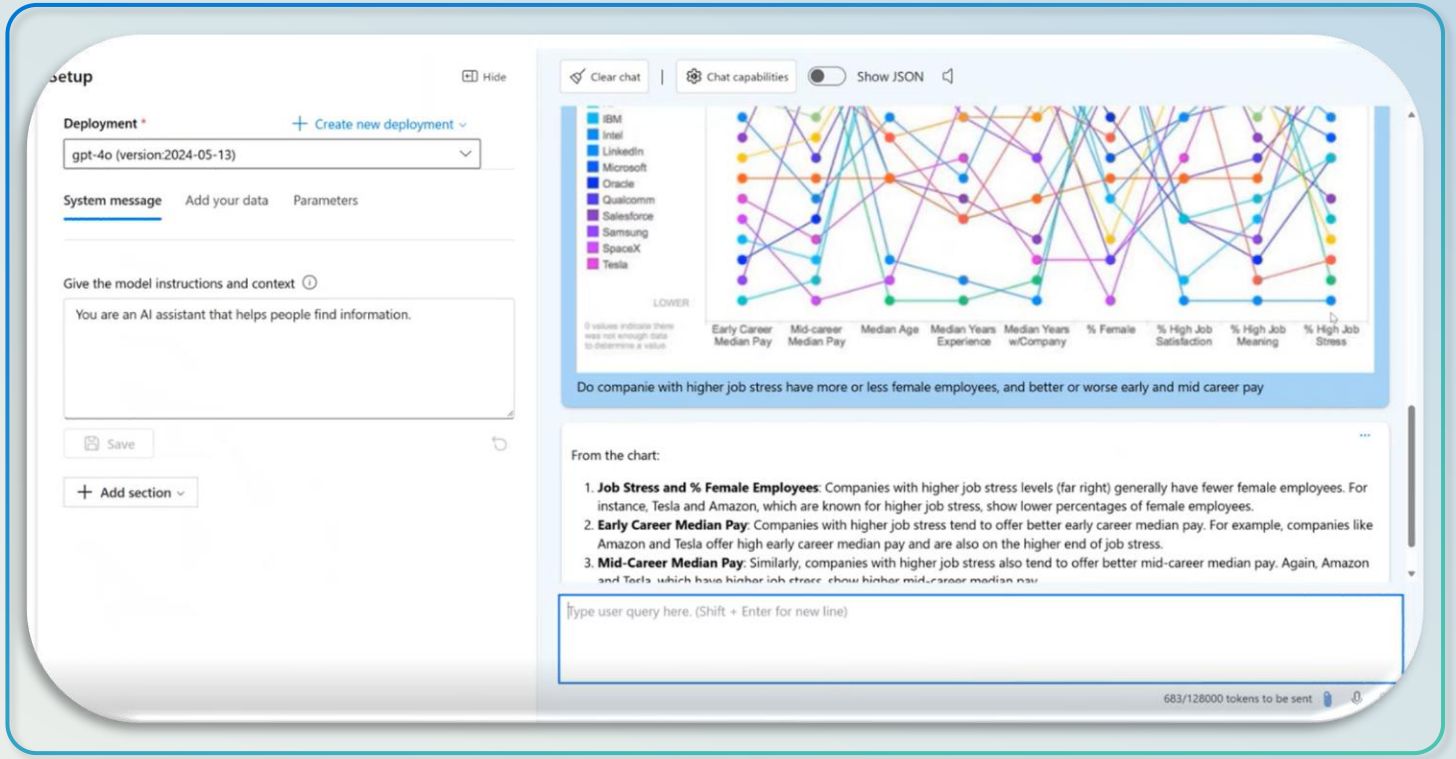


Azure AI Agent Service supports fine-tuned gpt-35-turbo (0125) as of Jan 2025

Multi-modal support across text, image, and audio

Flexible Model Selection

- Unlock new scenarios with multi-modal support, enabling AI agents to process and respond to diverse data formats beyond text, expanding the potential use cases.
- Support for GPT-4o's image and audio modalities so that you can analyze and combine data from various formats to deliver comprehensive insights, make decisions, and provide relevant outputs tailored to specific user needs



Enterprise Readiness

Enterprise Readiness

Bring your own storage

Keyless setup and authentication

Private Virtual Network support

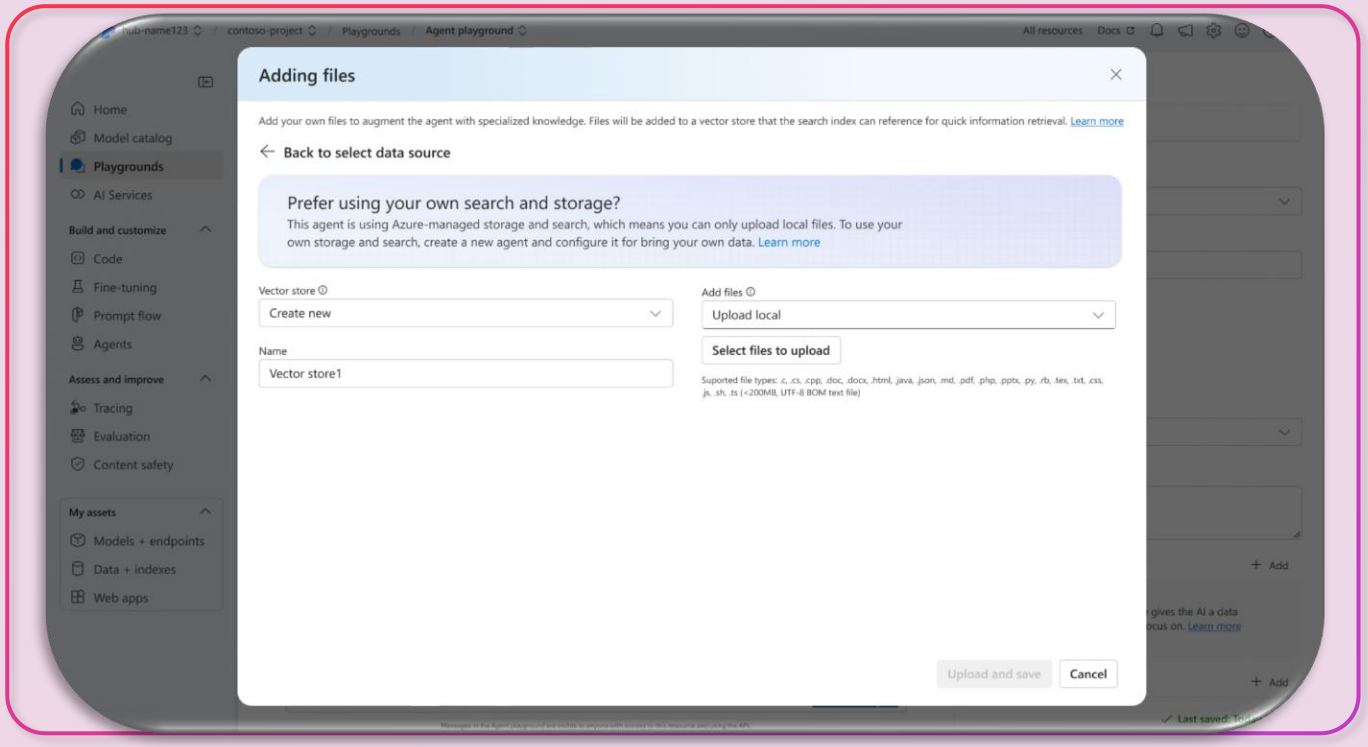
Tracing/ monitoring

Content filters

Bring your own storage and search

Enterprise Readiness

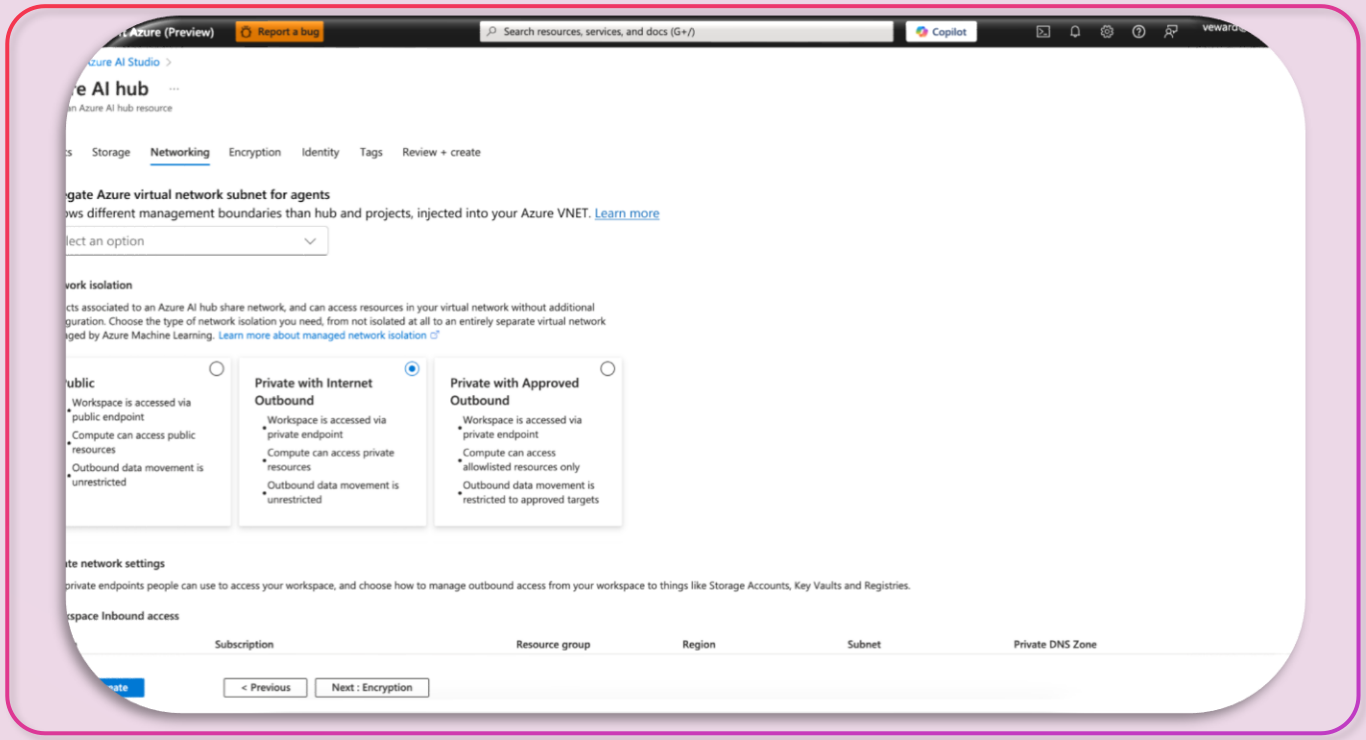
- Bring your own storage account (coming soon) and Azure AI Search resource for custom data handling.
- All files you upload will be stored in your storage account, ensuring you maintain complete control over your data.
- All indexes get created using your connected Azure AI Search resource.



Bring your own virtual network

Enterprise Readiness

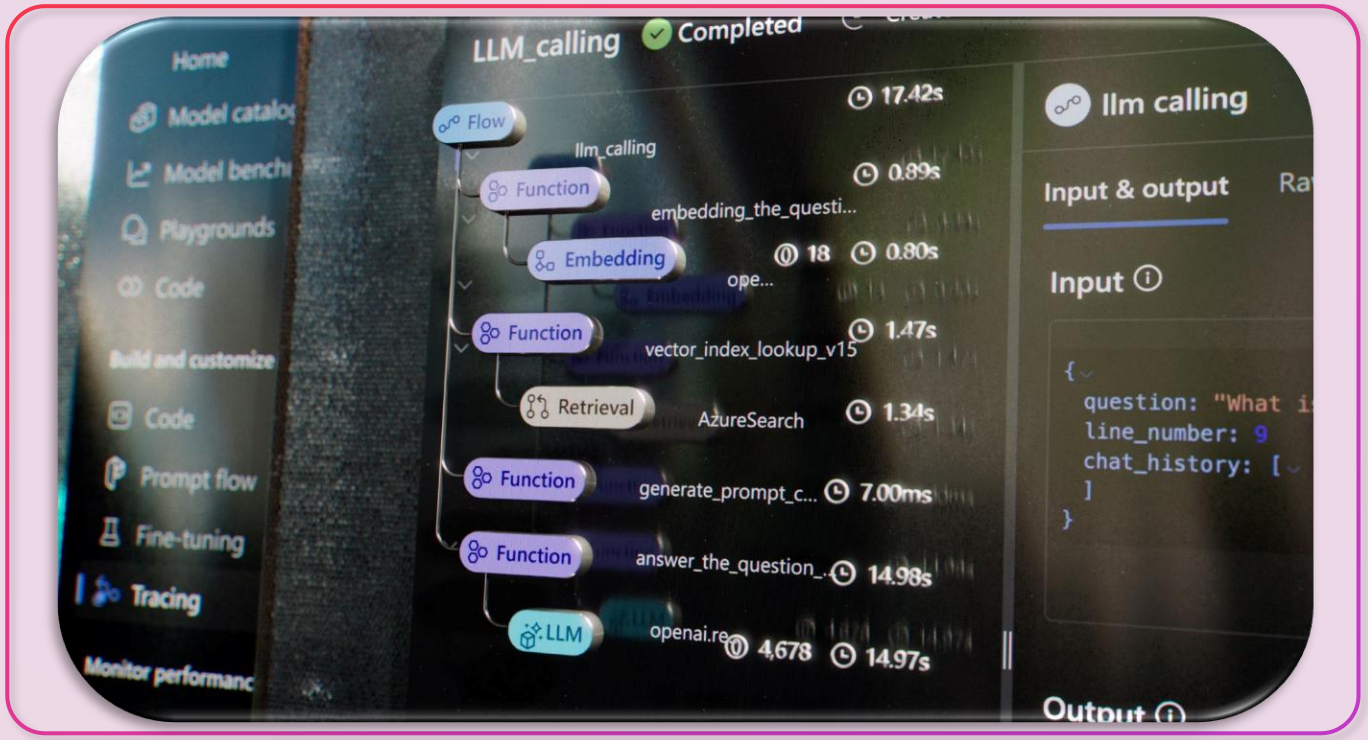
- No public egress foundational infrastructure ensures the right authentication and security for your agents and tools, without you having to do trusted service bypass
- Container injection allows the platform network to host APIs and inject a subnet into your network, enabling local communication of your Azure resources within the same virtual network (VNet).
- If your resources are marked as private and non-discoverable from the Internet, the platform network can still access them, provided the necessary credentials and authorization are in place



Enterprise Readiness

Monitor agent performance with OpenTelemetry-based tracing

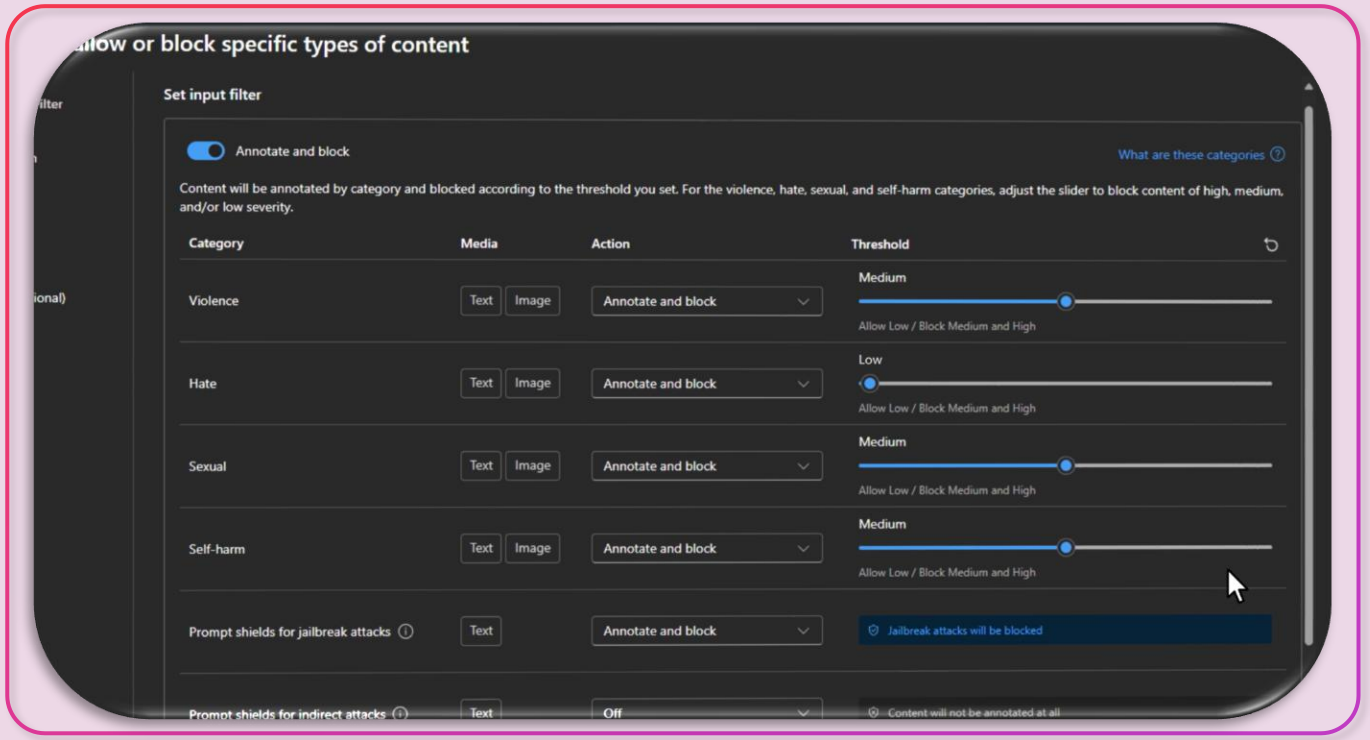
- Gain insights into your AI agent's performance and reliability
- OpenTelemetry-compatible metrics for offline and online evaluation of agent outputs through the Azure AI Foundry SDK
- Add local variables and intermediate results to trace decorator for detailed tracing capabilities for user defined functions
- Options to prevent sensitive or large data logging as per OpenTelemetry standards



Build responsibly with content filters and XPIA mitigation

Enterprise Readiness

- Supports prebuilt and custom content filters that detect harmful content at varying severity levels.
- Prompt shields protect agents against cross-prompt injection attacks from malicious actors.
- As with Azure OpenAI Service, prompts and completions processed by the Azure AI Agent Service are not used to train, retrain, or improve Microsoft or 3rd party products or services without your permission. Customers can delete their stored data when they see fit.



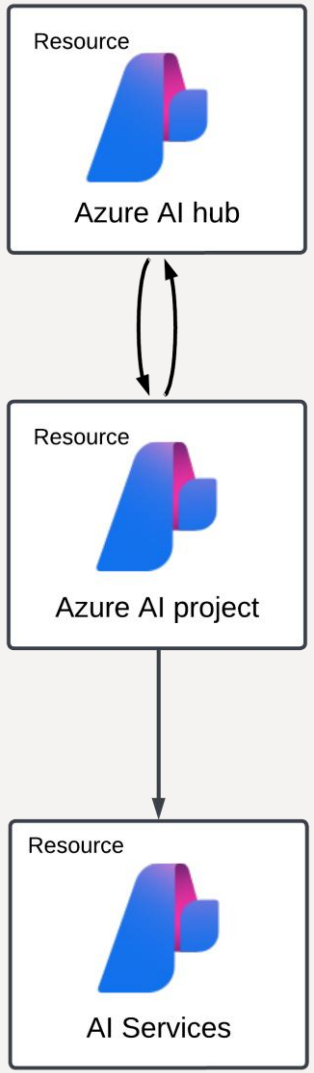
Basic agent setup

- **Overview**: All agents created in this project use multi-tenant search and storage resources fully managed by Microsoft. You don't have visibility or control over these underlying Azure resources.

- **Required customer resources:**
  - AI hub
  - AI project
  - AI Services/AOAI

- **How to use**: Deploy the basic setup template

- **Tool Implications:**
  - File search and code interpreter
    - Uploaded files get stored in Microsoft managed storage
    - Vector stores get created using a Microsoft managed search resource
  - Azure AI Search tool is not supported
  - Azure Blob Storage with file search is not supported



1

Single-agent

Deploy agents with **Azure AI Agent Service**

Managed agent micro-services

2

Multi-agent

Orchestrate them together with **AutoGen** and **Semantic Kernel**

State-of-the-art research SDK

Production-ready and stable SDK

Ideation → **Production**

Overview
Model catalog
Playgrounds
AI Services
Build and customize
Agents PREVIEW
Code PREVIEW
Fine-tuning
Prompt flow
Assess and improve
Tracing PREVIEW
Evaluation
Safety + security
My assets
Models + endpoints
Data + indexes
Web apps
Management center

Agents playground

Connected resource: ai-hub-demo-xih6-connection-AI... | New agent | Delete | View code

Clear chat | Logs | Thread files | JSON response | 0 tokens

New thread started

thread_wEv1lSSIoqArQ5rxUjcREmH2

Type user query here. (Shift + Enter for new line)

Messages in the Agents playground are visible to anyone with access to this resource and using the API.

Setup

Hide

Agent id

asst_L2GCrxIea7gMBDClUYJsYePn

Agent name

LeadGenerationAgent

Azure OpenAI resource connection

ai-hub-demo-xih6-connection-AIServices_aoai

Azure AI Search resource connection

ai-hub-demo-xih6-connection-AISearch

Deployment *

Create new deployment

gpt-4o (version:2024-05-13)

Instructions

You will help generate a personalized email message for a customer who
is interested in automating a business process with AI using Microsoft products.

You will then use this information to perform the following steps in order:

1. **Perform research** – use the Bing search tool to find relevant information
about the industry and business process and how AI can be used to automate it.
It's currently November 21st, 2024, use announcements that just happened at Microsoft Ignite 2024.
Create 3 parallel searches for features from Copilot Studio, Azure AI Foundry, and Microsoft 365 Copilot.

- You **must** ensure the email does not include any references to competitors to Microsoft

Microsoft Azure

Demo for illustrative purposes only.
Actual results may vary based on specific use cases and configurations.

Finance

AI Agent workflows enhance efficiency, accuracy, and customer satisfaction

Fraud Detection

- Transaction data analysis
- Detect unusual patterns
- Prevent fraud
- Customer security

Risk Management

- Analyze market data
- Predict trends
- Insights and recommendations

Customer Service

- Virtual assistants
- Process transactions
- Provide financial advice

Algorithmic Trading

- Analyze market conditions
- Execute trades
- Optimize trading strategies
- Reduce human error

Regulatory Compliance

- Monitor transactions and communications
- Flag suspicious activities
- Automatically report to regulatory authorities