案例研究

情境檢視與分析

案例研究概覽

公司背景:

MicroRetail 微零售有限公司 總部位於台北市，擁有超過 200 間實體門市，是全台具代表性的零售通路之一。儘管名稱中有「micro」這個字，他們在零售市場的影響力卻不容小覷。近年來，MicroRetail 積極推動電商發展與 OMO（虛實整合）策略，希望透過創新技術優化顧客互動體驗與營運效率。然而，目前仍有許多核心系統依賴地端資料中心運作。

案例研究概覽

現狀與公司未來策略:

資訊長 Daniel Liang 積極推動生成式 AI 的導入，他表示：「我們最怕的不是沒人來逛，而是人太多時 API 被點爆。現在真的需要雲端來分擔流量與運算壓力。」Daniel 相信，AI 的應用不僅可以提升服務韌性，更能帶動整體營運成長。

架構師 Tina 與強大的技術夥伴們，已著手啟動第一階段導入計畫，包括：

- 在官方 App 與網站導入 AI 驅動的**內容審核機制**，過濾不當評論與意見回饋

- 建立一個具備**個人化推薦能力**的 AI 聊天機器人，能與顧客互動對話，也能適時推薦商品

這項計畫的目標，是在財政年度第 4 季結束前，**將客服平均回應時間縮短 15%**，進一步提升顧客滿意度與服務效率。

面對母親節、年中慶等高峰檔期的巨大流量挑戰，MicroRetail 期望透過 AI 自動化客服與即時互動分析，讓 AI 接手處理大量重複性問題，使客服人員能專注於高價值的溝通與協助，全面升級服務品質。

最終，MicroRetail 的願景是整合來自實體門市、電商平台與行動 App 的顧客數據，打造真正個人化、即時且智慧的顧客體驗。透過 AI 與雲端的強大能量，這個「微小」品牌，正以智慧與數據在市場中閃耀發光。

客戶目標與需求

一、效能與速度（Performance and Speed）

- 回應速度需低於 2 秒，提升即時互動體驗。
- 系統須支援高效且即時的顧客互動，不影響網站流暢度。

二、個人化體驗（Personalization）

- 根據顧客偏好與購買紀錄進行推薦，點擊率提升目標為 20%。
- 對話式 AI 機器人須具備個人化推薦能力，強化顧客體驗。

三、可靠性與可擴展性（Reliability and Scalability）

- 系統須具備 99.99% 的可用性與低於 1 秒的延遲。
- 能夠支援大量互動的可擴展性架構。
- 基礎架構需具備容錯與災難復原能力。

四、內容審核（Moderation）

- 能自動審核評論與顧客回饋，避免出現不當語言、垃圾訊息或不適當內容。

客戶目標與需求

五、安全性與法規遵循（Security and Compliance）

- 整合私有端點、RBAC、SSO 與資料標註，達成零資料外洩並通過年度資安稽核。
- 符合 GDPR 等隱私法規，並進行季節性稽核。
- 遵循資安最佳實踐確保資料安全。

六、治理與成本控管（Governance and Cost Management）

- 建立治理模型管理 AI 解決方案。
- 成本控管目標為預算偏差控制在 5% 以內。

七、技術與基礎建設（Technology and Infrastructure）

- 支援商用與開源大型語言模型的調用與微調，並能靈活部署以達成商業目標。

八、訓練與支援（Training and Support）

- 團隊對 Azure 不熟悉，需要 Microsoft 合作夥伴協助設計架構、網路與資安配置並提供訓練。

客戶問題與疑慮

- 如何確保 AI 審核能有效過濾不當內容（如髒話、垃圾訊息），同時不影響正常評論？
- 這套 AI 解決方案將如何處理客戶資料的安全性與隱私需求？
- 新系統是否會影響我們現有系統的效能，特別是在高流量期間？
- 我們擔心所謂的「幻覺現象」，也就是 AI 聊天機器人可能會捏造產品資訊。我們可以怎麼預防這種情況？
- 要如何衡量 AI 回覆的品質與準確度，確保符合企業標準？
- 顧客回饋將如何被整合，用來持續提升 AI 的準確性？
- 網路與基礎建設需要怎樣更新，才能支援 AI 整合？
- 能否更主動處理資安問題？有沒有更好的威脅偵測與安全監控自動化方式？

Lab outline

Level - Easy

Exercise one

Set up Azure AI Foundry

Exercise two

Perform prompt engineering

Level - Medium

Exercise Three

Set up Azure prompt flow

Exercise four

Deploy a chatbot to a web app

Exercise five

Content moderation

Exercise one

Set up Azure AI Foundry

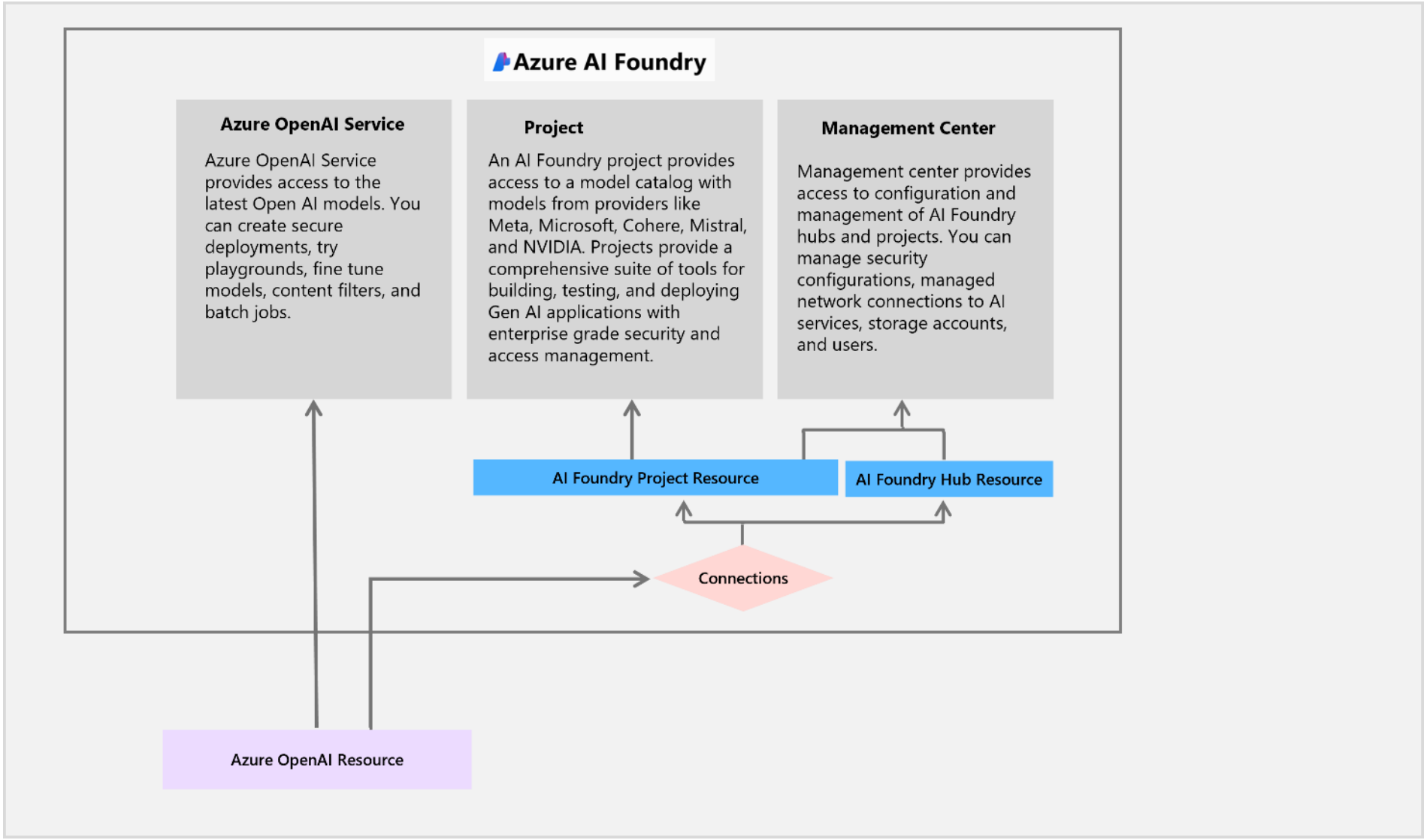
Introduction: Set up Azure AI Foundry

MicroRetail aims to enhance its user experience with a chatbot, ensuring a safer and more engaging environment for customers. Building a chat bot begins with enabling an AI Foundry, creating a Hub, and managing permissions.

After completing this exercise, you'll be able to:

- Set up an Azure AI Foundry hub to manage AI projects and services.
- Deploy AI models essential for chat-based applications.
- Create a search service to enhance chatbot capabilities.

Exercise one concepts



Task 01: Set up a hub and project in Azure AI Foundry

Introduction:

To support **MicroRetail**'s initiative for AI-driven customer engagement, the first step is to establish a structured environment where AI projects can be built and managed. By creating a hub in Azure AI Foundry, MicroRetail can centralize AI models, configurations, and data integrations, ensuring scalability and efficiency for future AI implementations.

Description:

In this task, you'll create a hub within Azure AI Foundry to host your projects and resources. The hub serves as the foundation for deploying AI solutions and integrating AI services that MicroRetail will use to enhance customer engagement. After setting up the hub, you will also create a project within it to organize and manage AI-related workflows effectively.

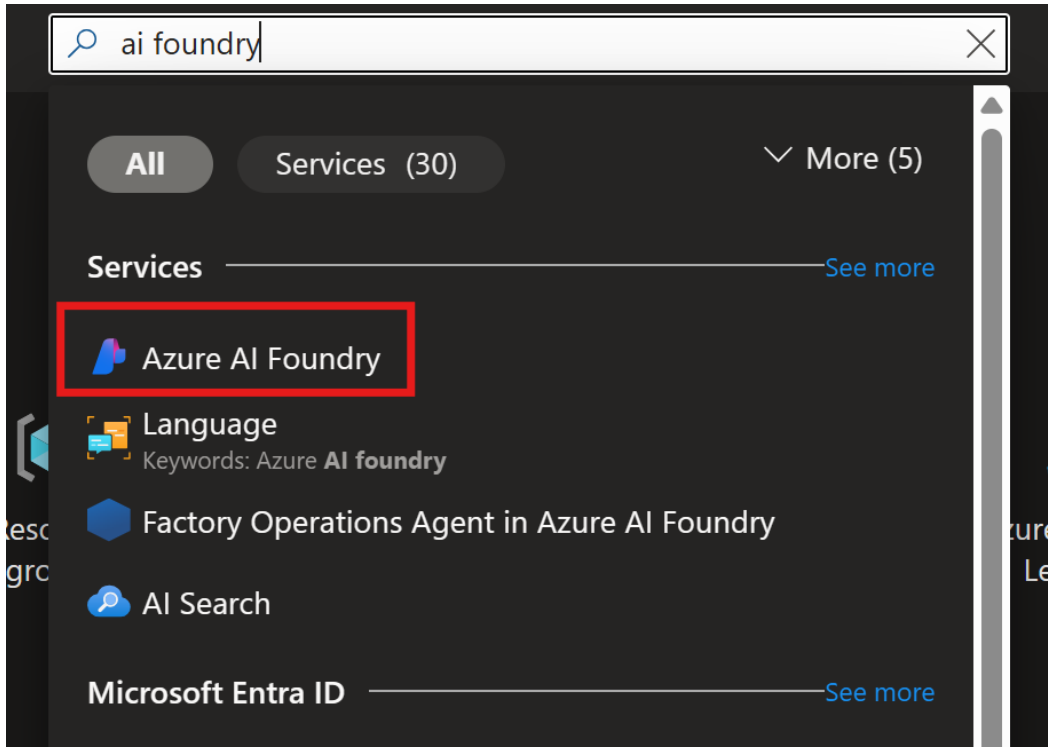
Success Criteria:

- The Azure AI Foundry hub has been created successfully.
- A new project has been set up within the hub.

1-1. Set up a hub and project in Azure AI Foundry

Step 01: Create hub in Azure AI Foundry

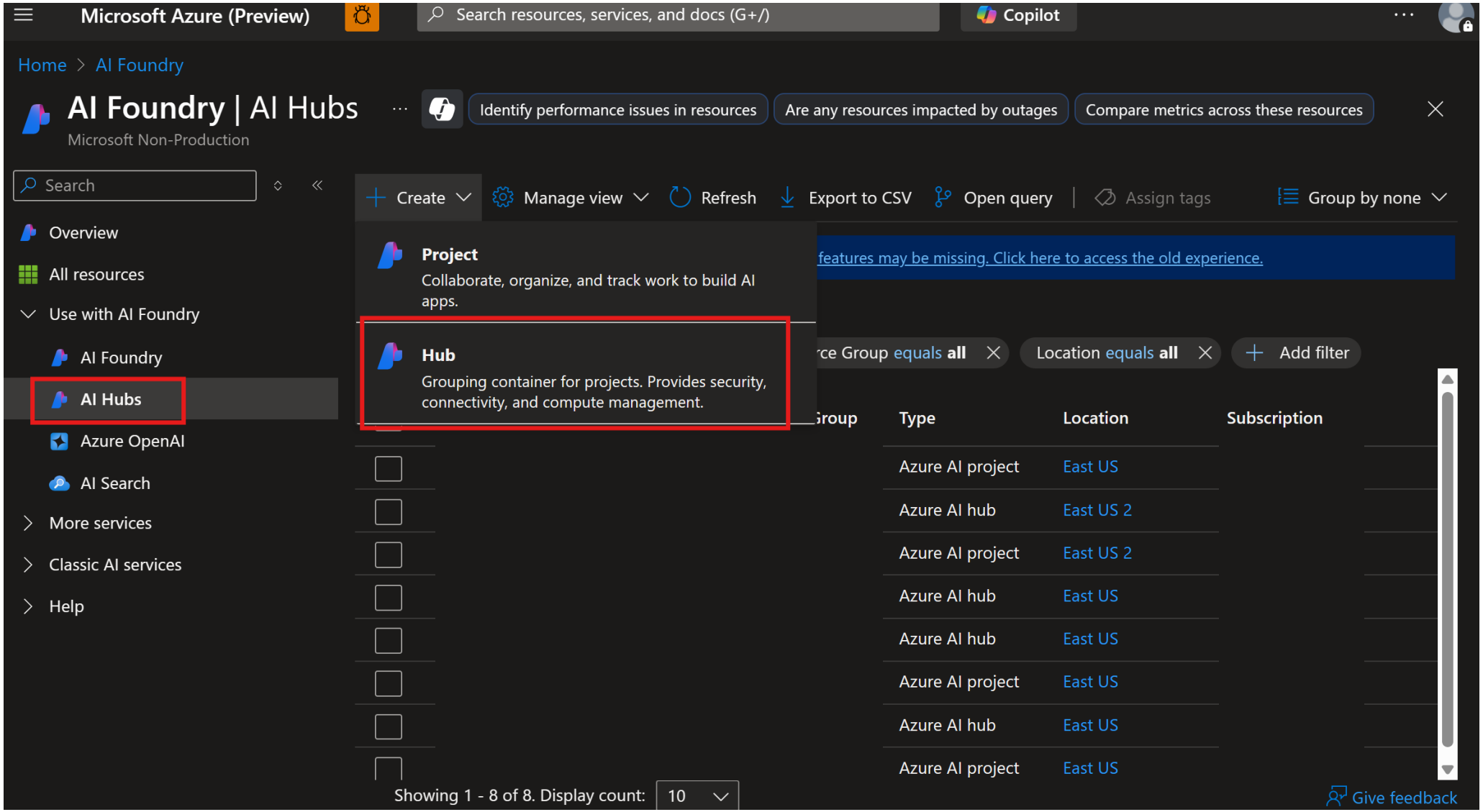
1.Open the Microsoft Edge browser, go to the Azure Portal site, and sign in with your credentials.
2.Once signed in to the portal, select the search bar at the top. Search for **Azure AI** and then select **Azure AI Foundry**.



1-1. Set up a hub and project in Azure AI Foundry

Step 01: Create hub in Azure AI Foundry

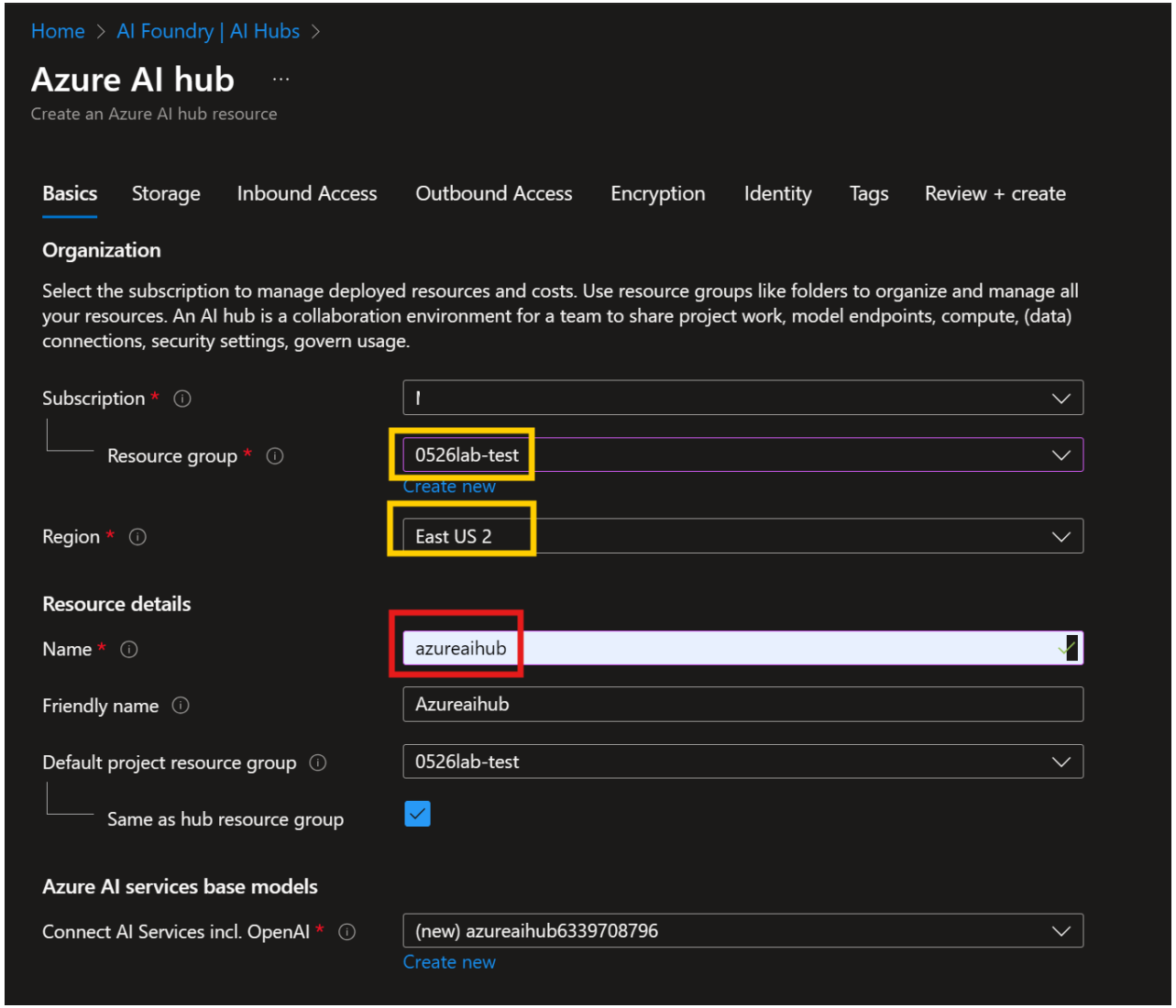
3. From **Azure AI Foundry**, select **+ Create** and then select **Hub**.



1-1. Set up a hub and project in Azure AI Foundry

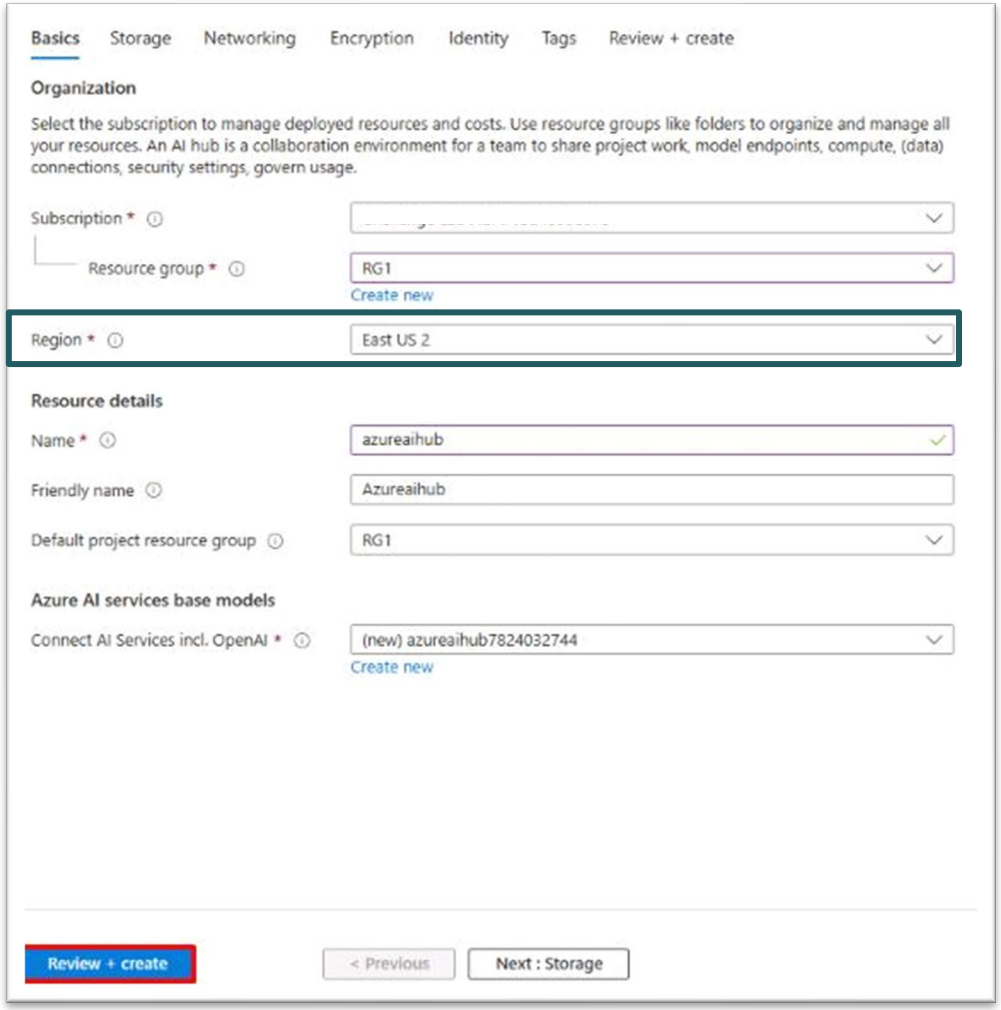
Step 01: Create hub in Azure AI Foundry

4. On the **Azure AI hub** page, select your **resource group and region** (these may differ from the screenshot).
Name the hub **azureaihub** and leave the default setting to create a new AI service model.



5. Select **Review + create**, then select **Create**.
6. Once the deployment is complete, select **Go to resource**.

Choice of region



Important

The availability of models and other resources will depend on the region chosen. Ensure you are using a region that supports Azure AI.

Needed roles for the hub and project

The hub is the overarching development environment.

- **Role needed**: Owner or Contributor
- **Where**: The associated resource group or an existing hub

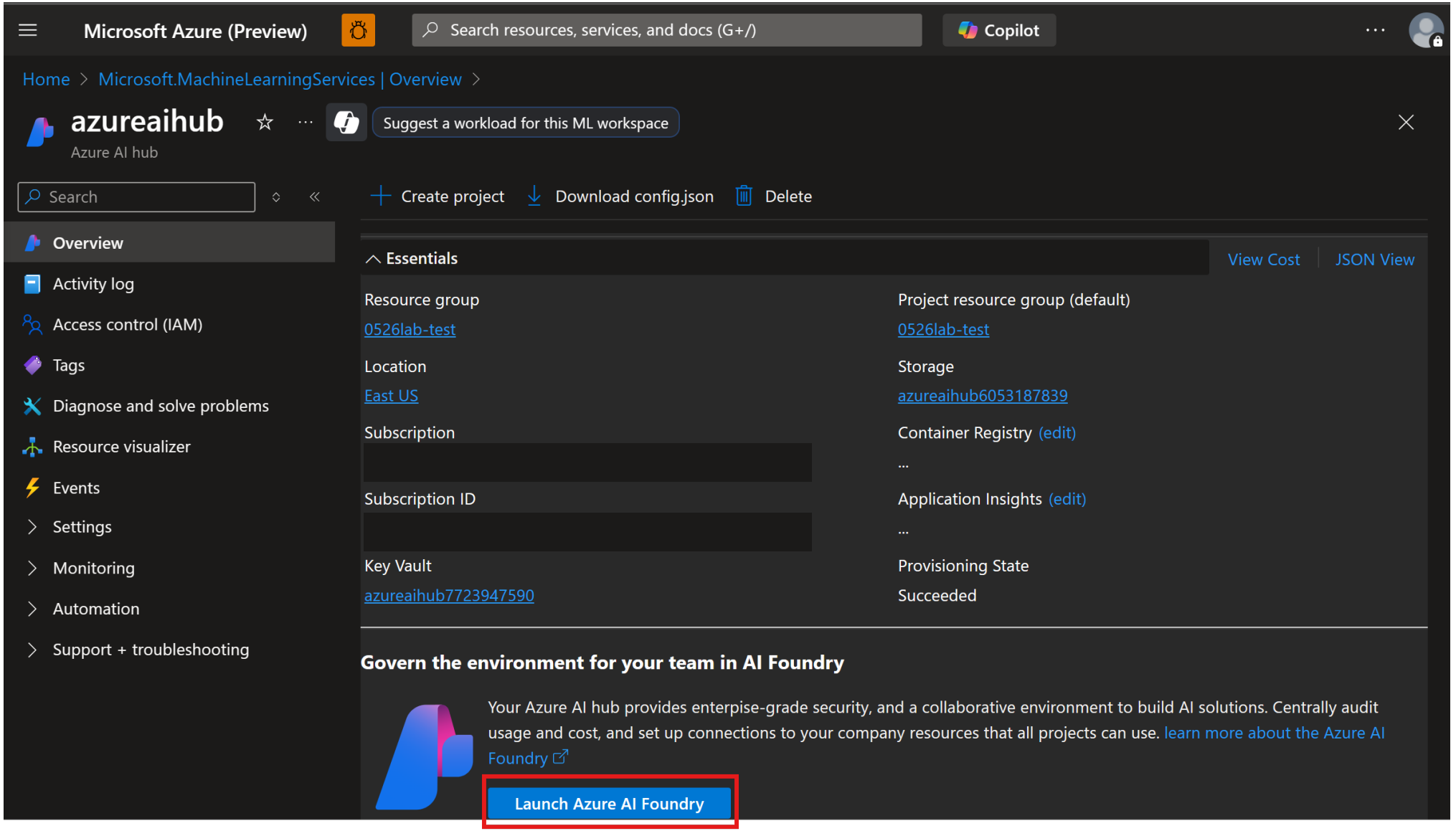
Projects exist within a hub and can have different permissions and resources.

Role needed: Either the Azure AI Developer or Azure AI Inference Deployment Operator

1-1. Set up a hub and project in Azure AI Foundry

Step 02: Create project in Azure AI Foundry

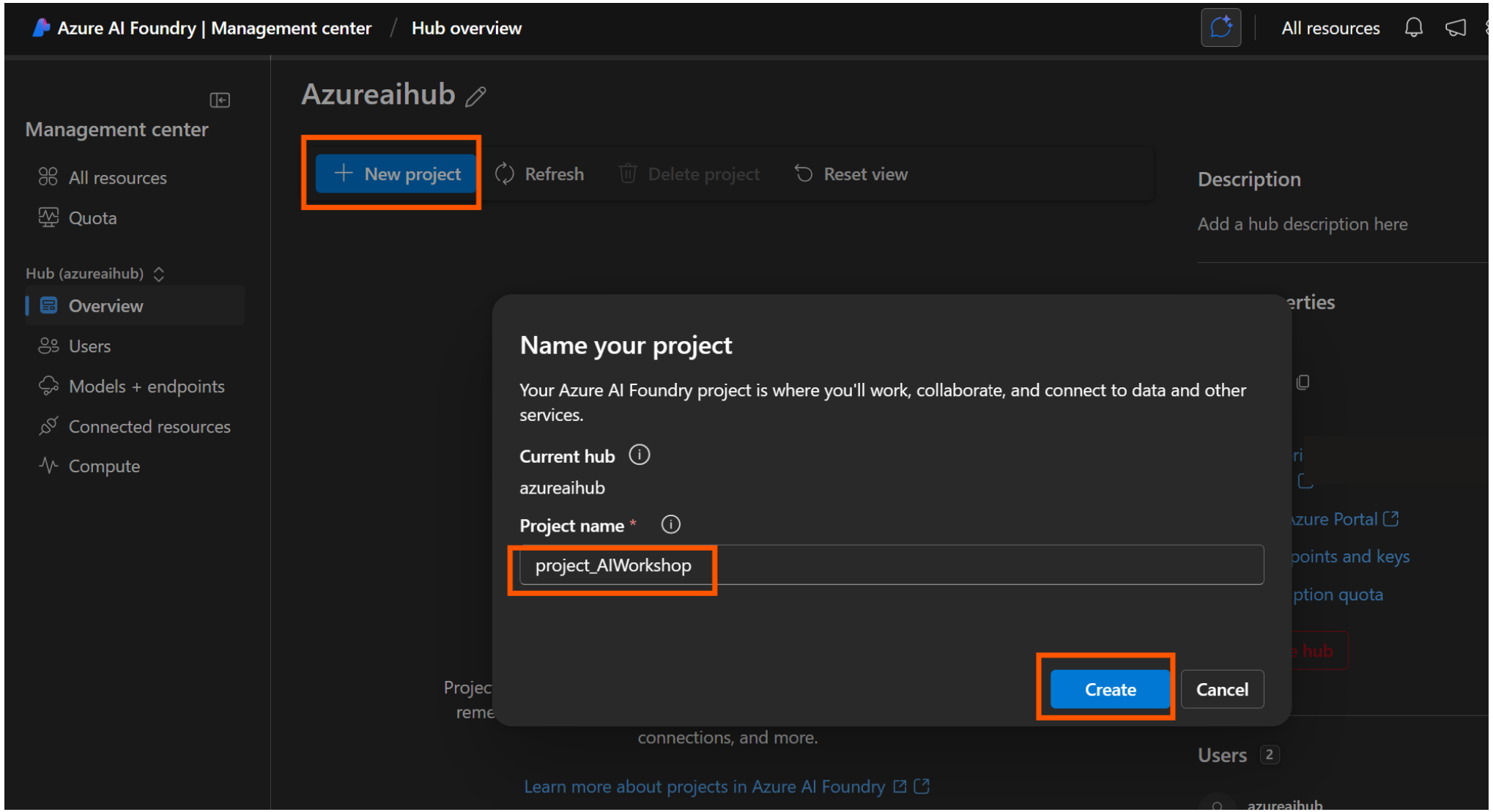
1. From the **azureaihub** page, select **Launch Azure AI Foundry**.



1-1. Set up a hub and project in Azure AI Foundry

Step 02: Create project in Azure AI Foundry

2. Azure AI Foundry will open in a new tab. Select **+ New project** to create a new project. Enter **project_AIWorkshop** and select **Create**.



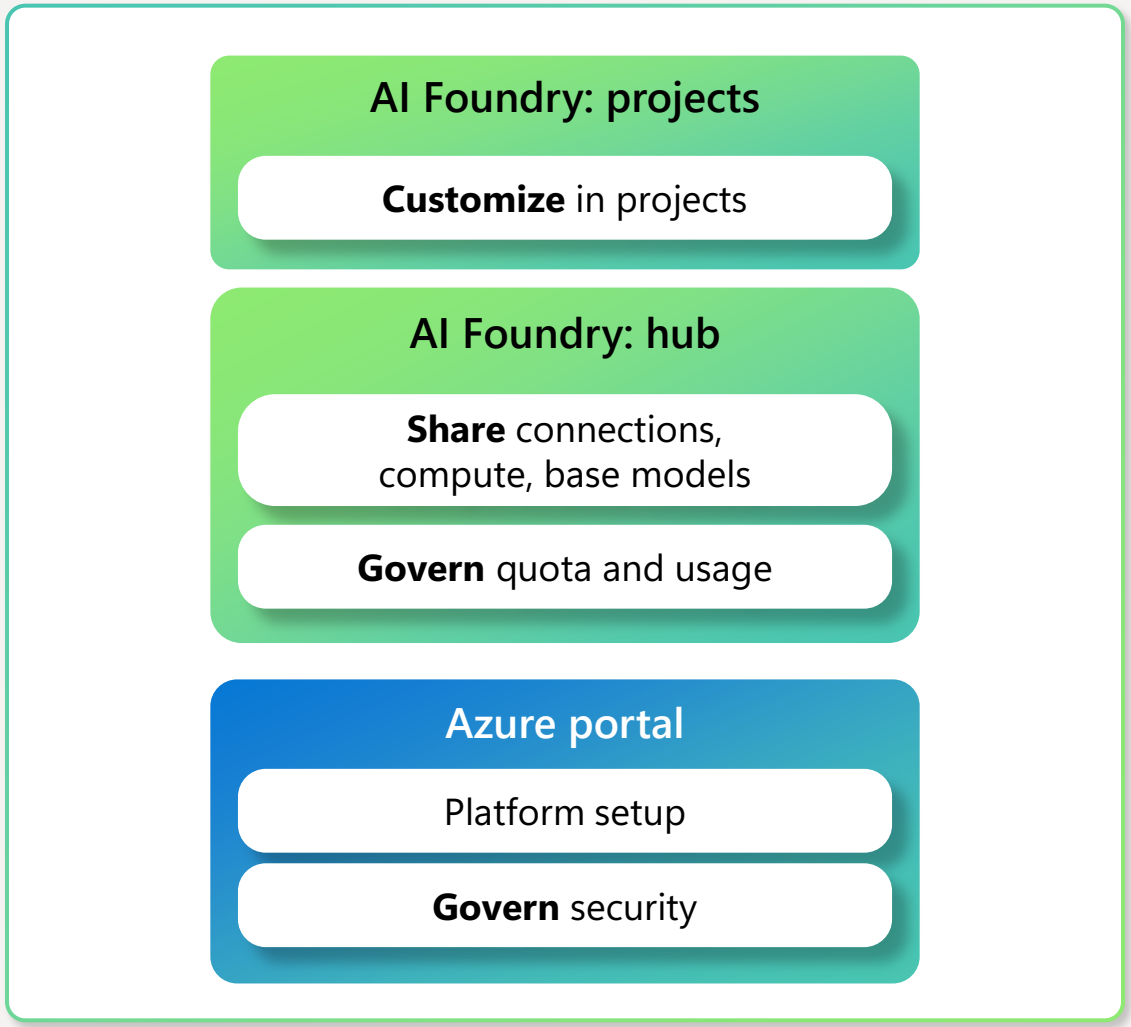
Management layers in Azure AI Foundry

AI Foundry targets three different management needs:

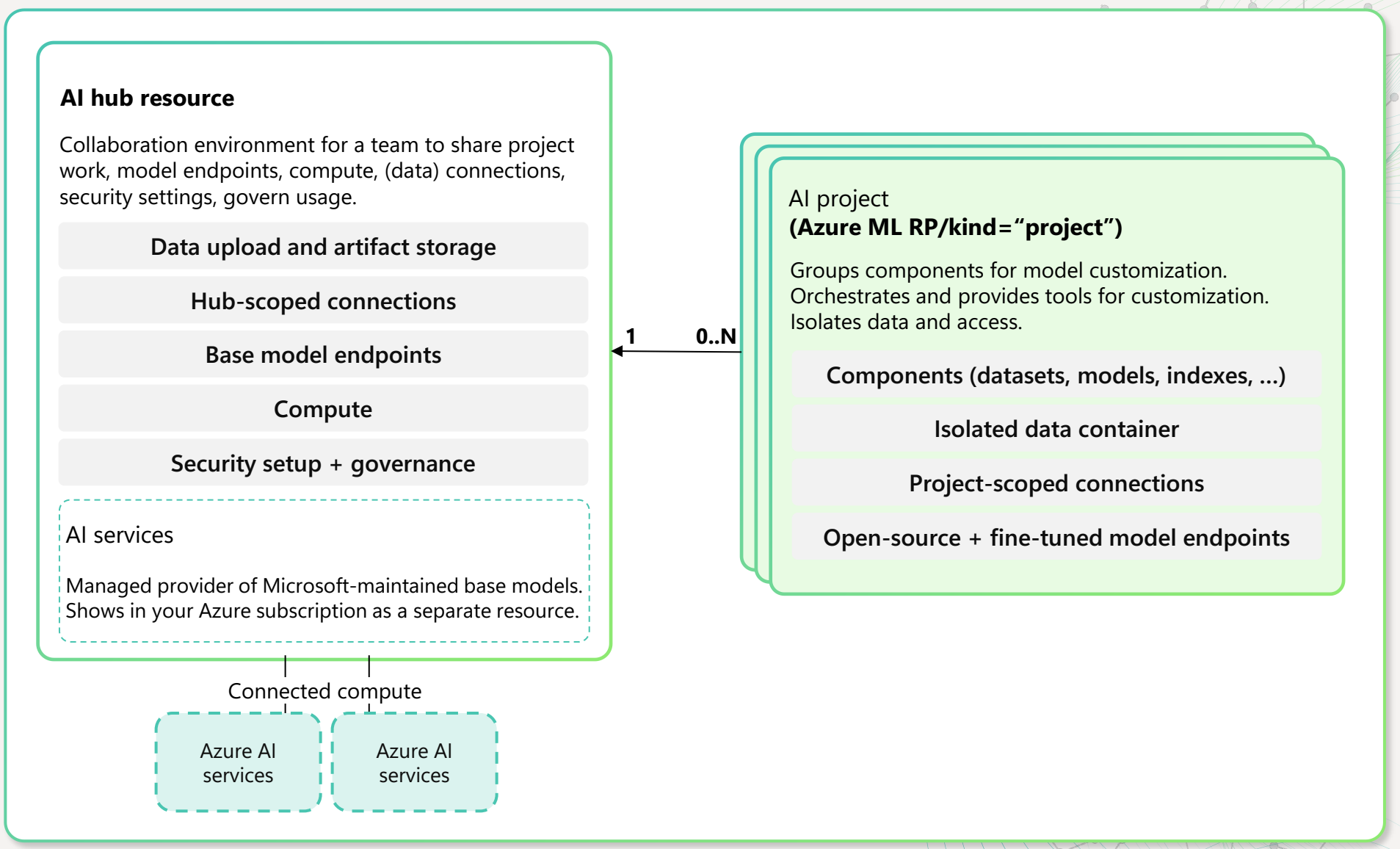
Provide **AI developers and business stakeholders** with a SaaS-like self-serve experience, to allow for rapid AI experimentation

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



Azure AI Foundry Hubs

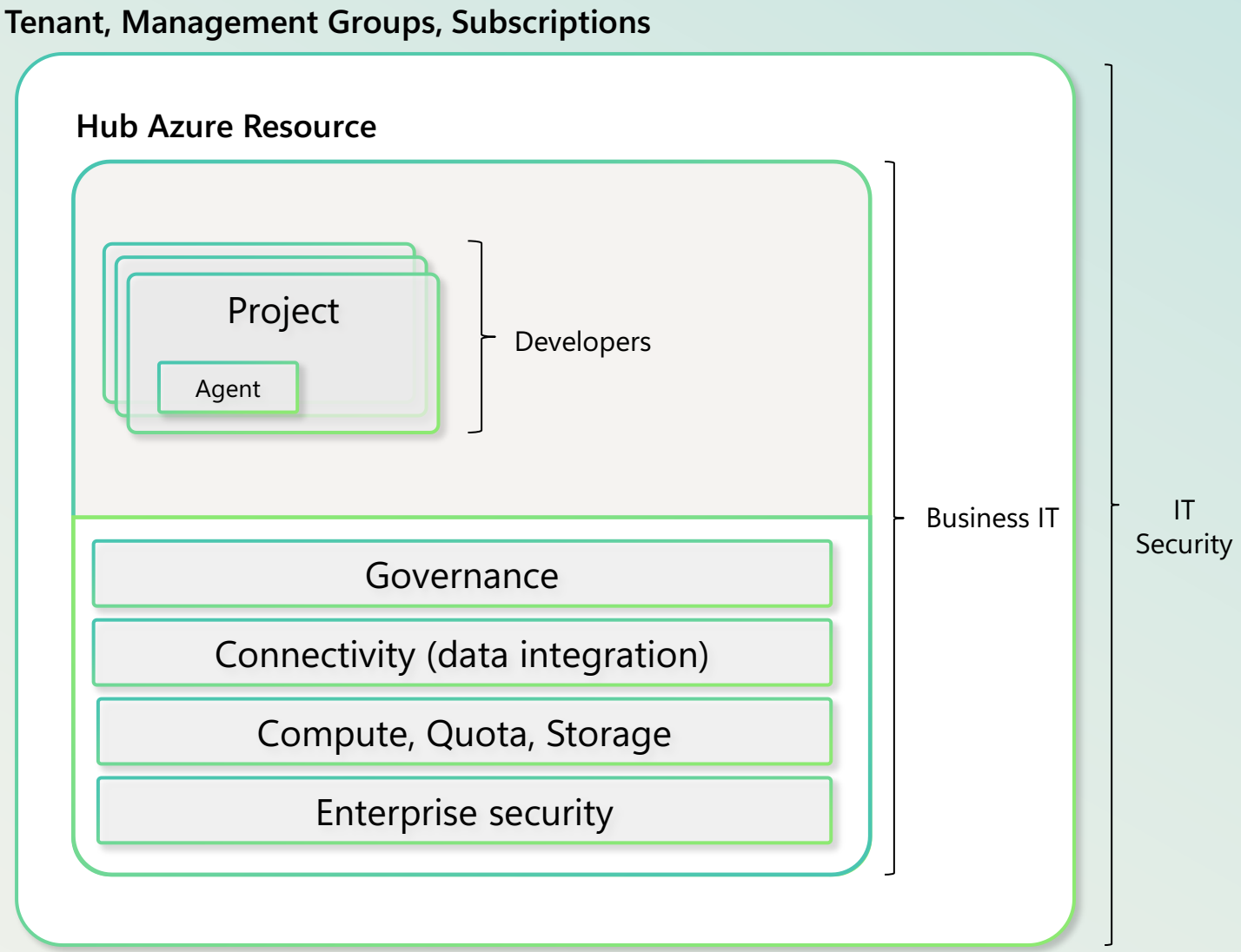


Enterprise setup in AI Foundry

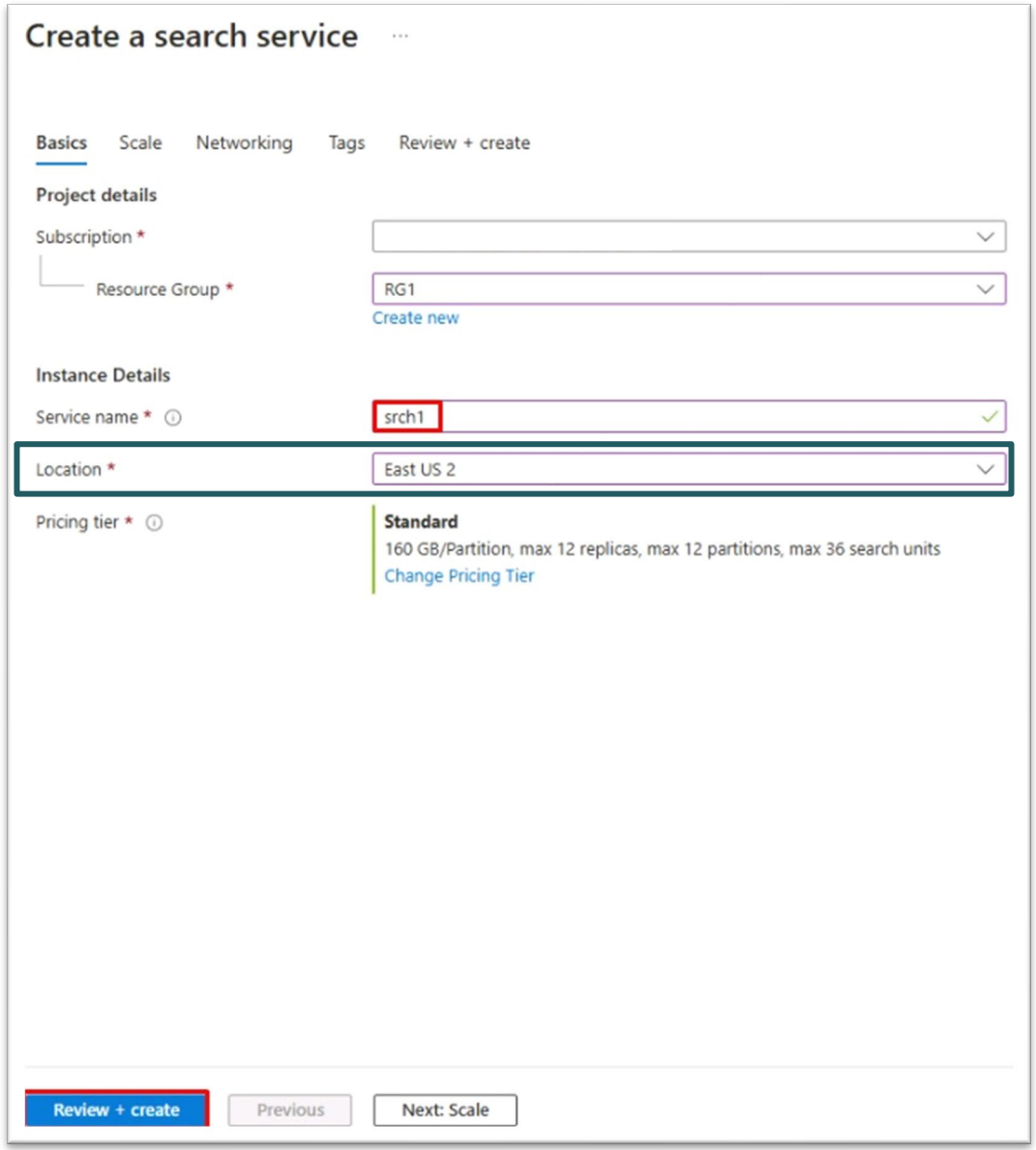
Seeks to balance a self-serve authoring experience for developers in AI Foundry, with granular platform controls for IT using Azure Portal/IaC templates

- **Hub:** configures Azure infrastructure, enterprise security, storage, connectivity for data integration. Groups projects for a team

- **Projects:** Provides a place for developers to collaborate and organize their work for customizing AI models. Container for data upload and access control



Location of search service



Important

Ensure the **Location** is set to the same location where the hub was created earlier.

Task 02: Create AI model deployments

Introduction:

To enable AI-driven customer service and content moderation, MicroRetail requires robust AI models that can process language effectively. By deploying models like gpt-4o-mini for conversation handling and text-embedding-ada-002 for data representation, MicroRetail can power its AI chatbot and optimize search results based on user queries.

Description:

In this task, you'll deploy the **gpt-4o-mini** and **text-embedding-ada-002** models to your project. These models are necessary for creating an AI-assisted chat application.

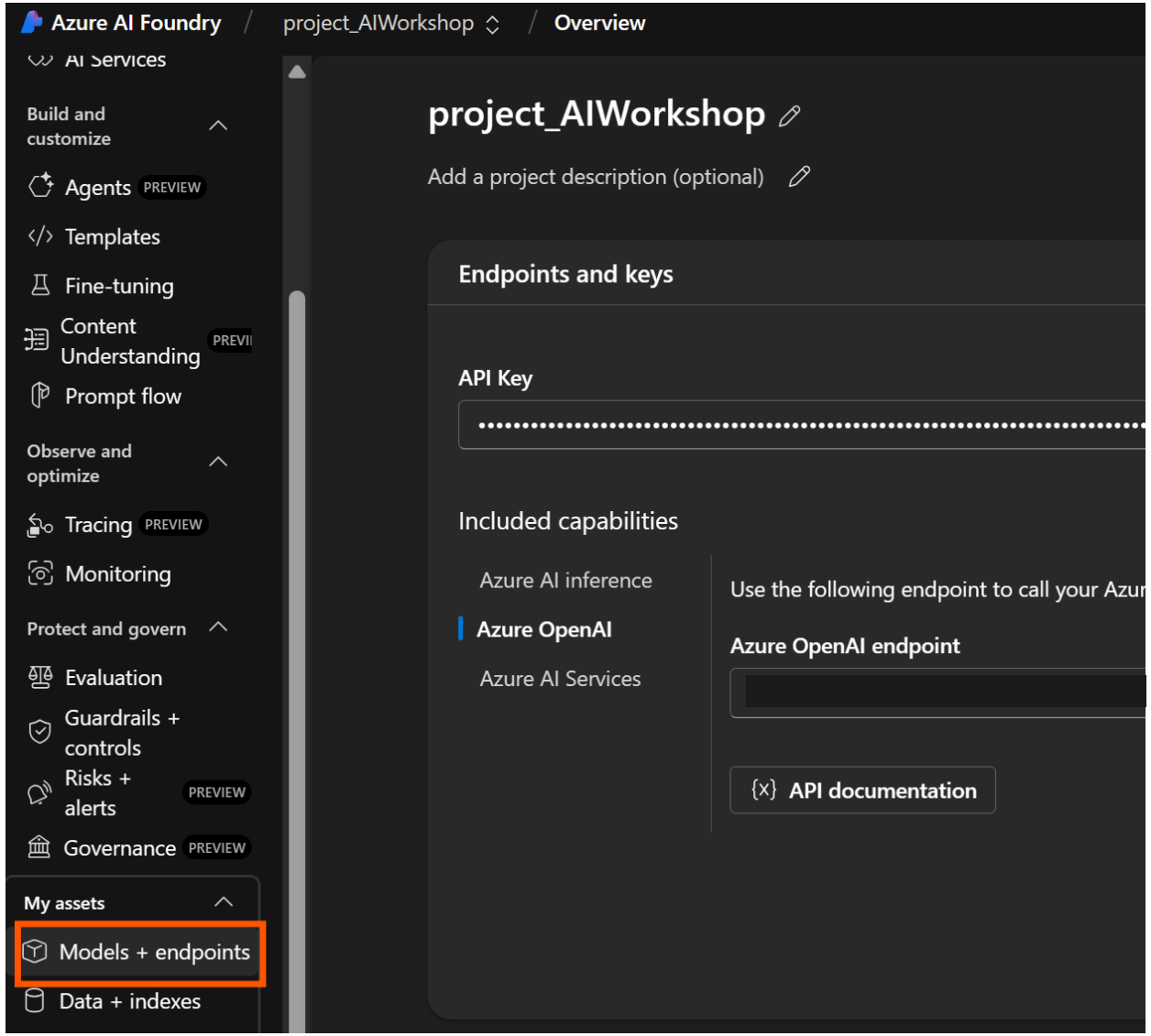
Success Criteria:

- The **gpt-4o-mini** model has been deployed.
- The **text-embedding-ada-002** model has been deployed.

1-2. Create AI model deployments

Step 01: Deploy gpt-4o-mini and text-embedding-ada-002 models

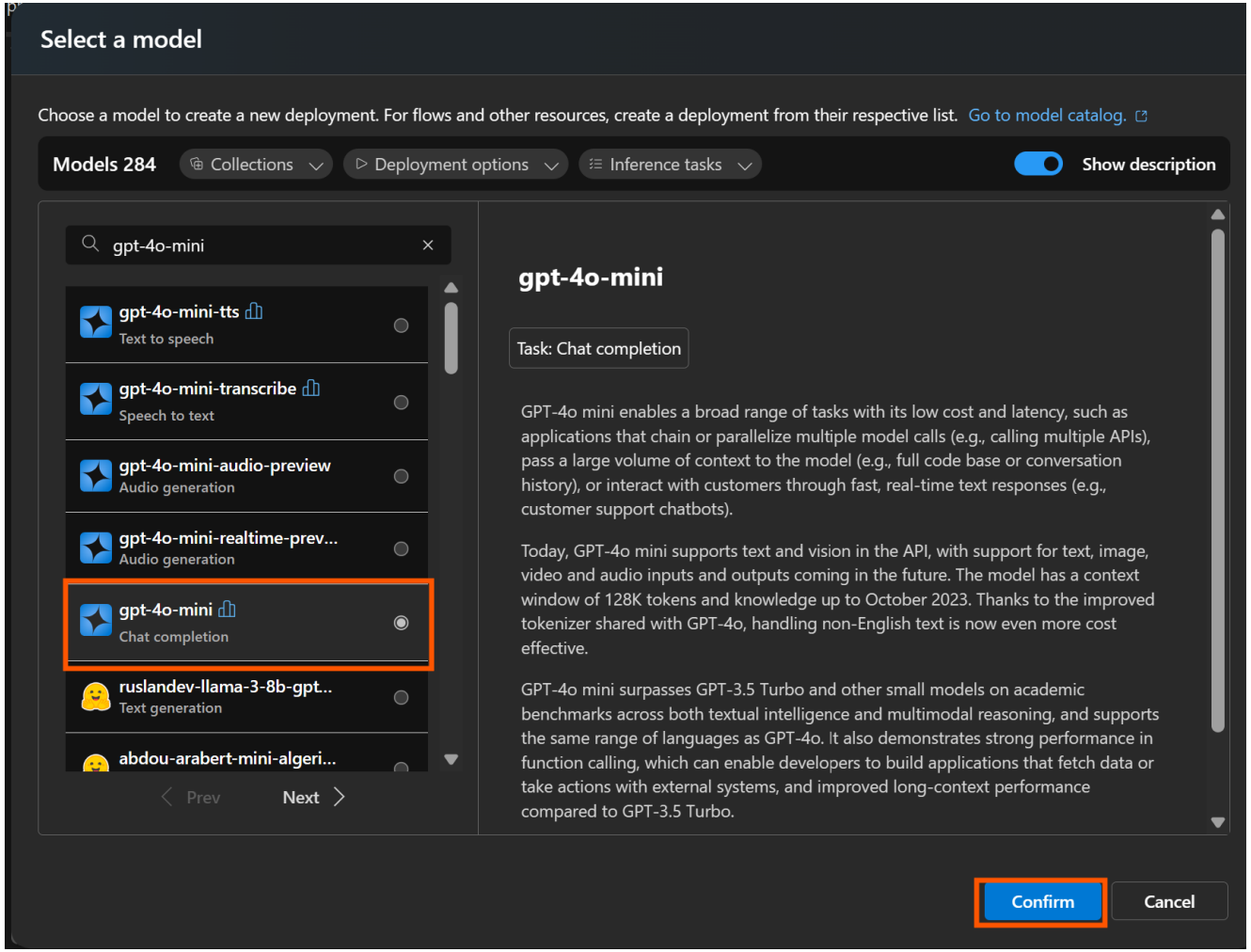
1. From the **Overview** tab on the **project_AIWorkshop** page, select **Models + endpoints** from the left menu.



1-2. Create AI model deployments

Step 01: Deploy gpt-4o-mini and text-embedding-ada-002 models

2. From the **Manage deployments of your models and services** page, select **+ Deploy model** and then select **Deploy base model**.
3. From the **Select a model** page, select **gpt-4o-mini**, then select **Confirm**.



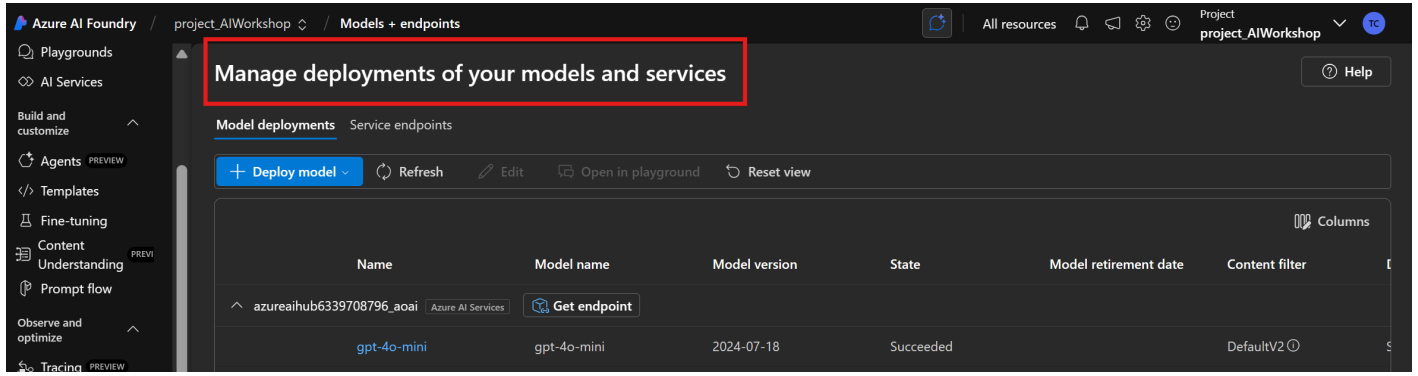
1-2. Create AI model deployments

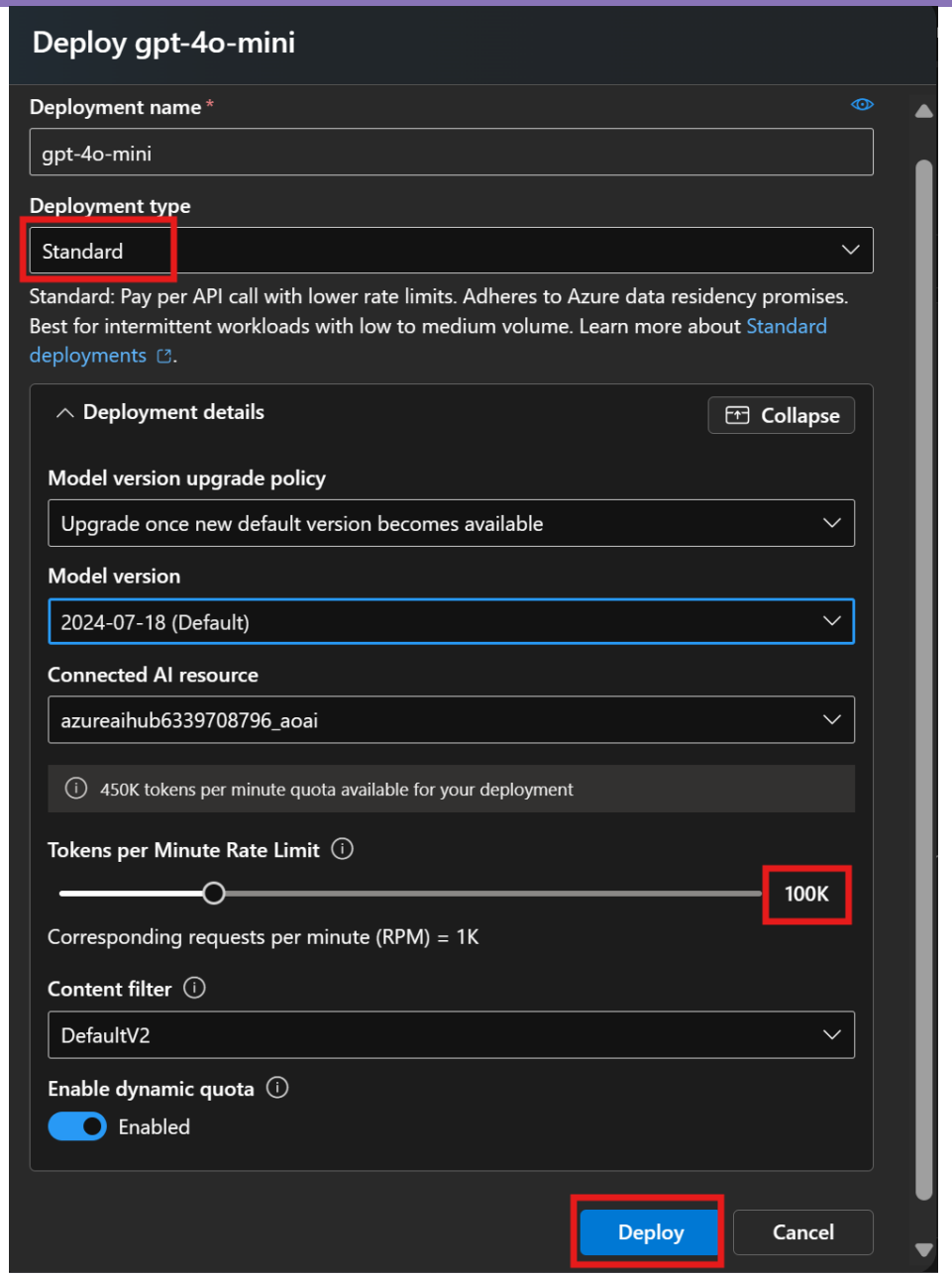
Task 01: Deploy gpt-4o-mini and text-embedding-ada-002 models

4. On the **Deploy model gpt-4o-mini** page, set the **Deployment type** to **Standard**.

5. Move the **Tokens per Minute Rate Limit** slider to around **100K**. Leave all other settings as default and select **Deploy**.

6. Once deployed, select the back button to return to the **Manage deployments of your models and services** page and deploy another base model.

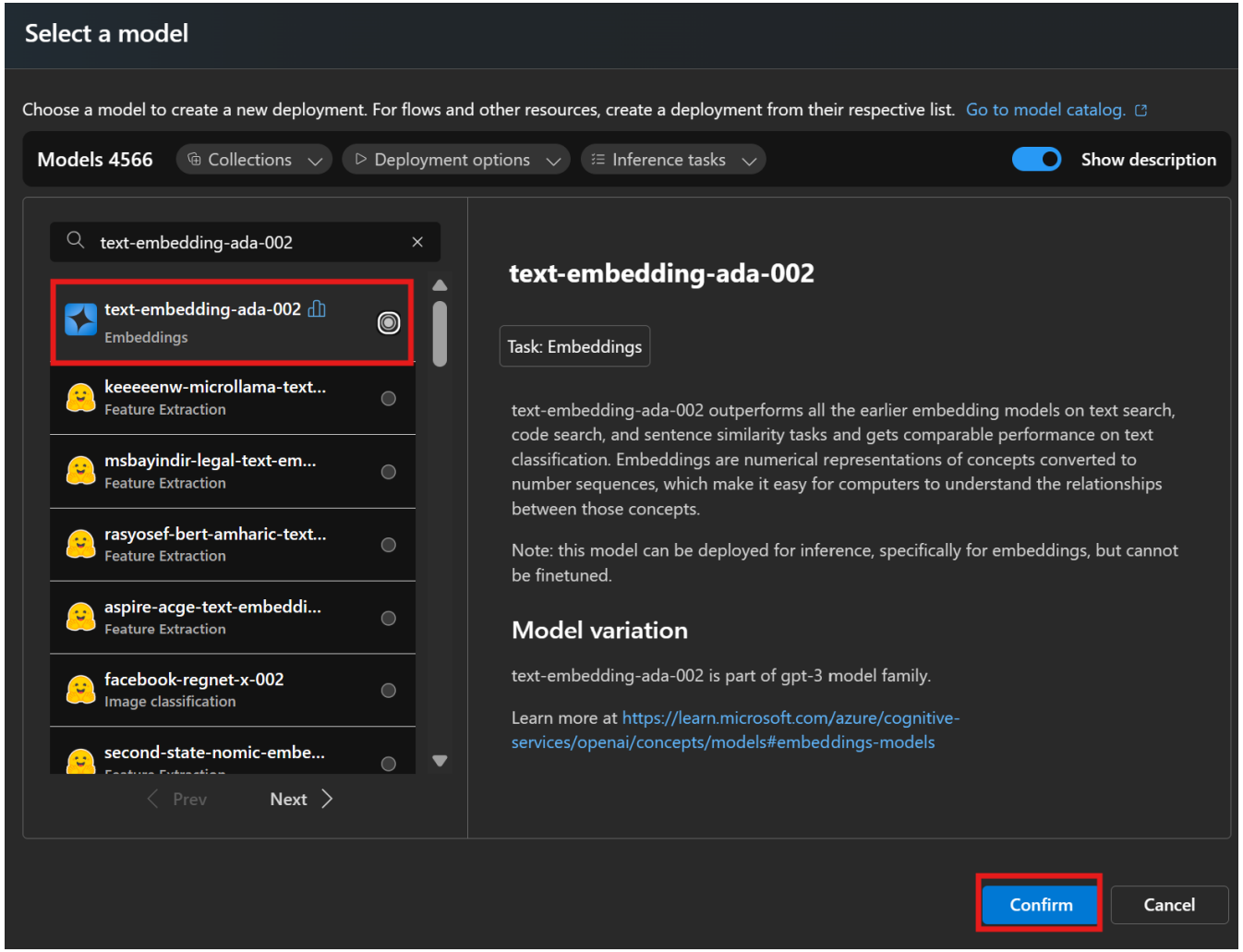




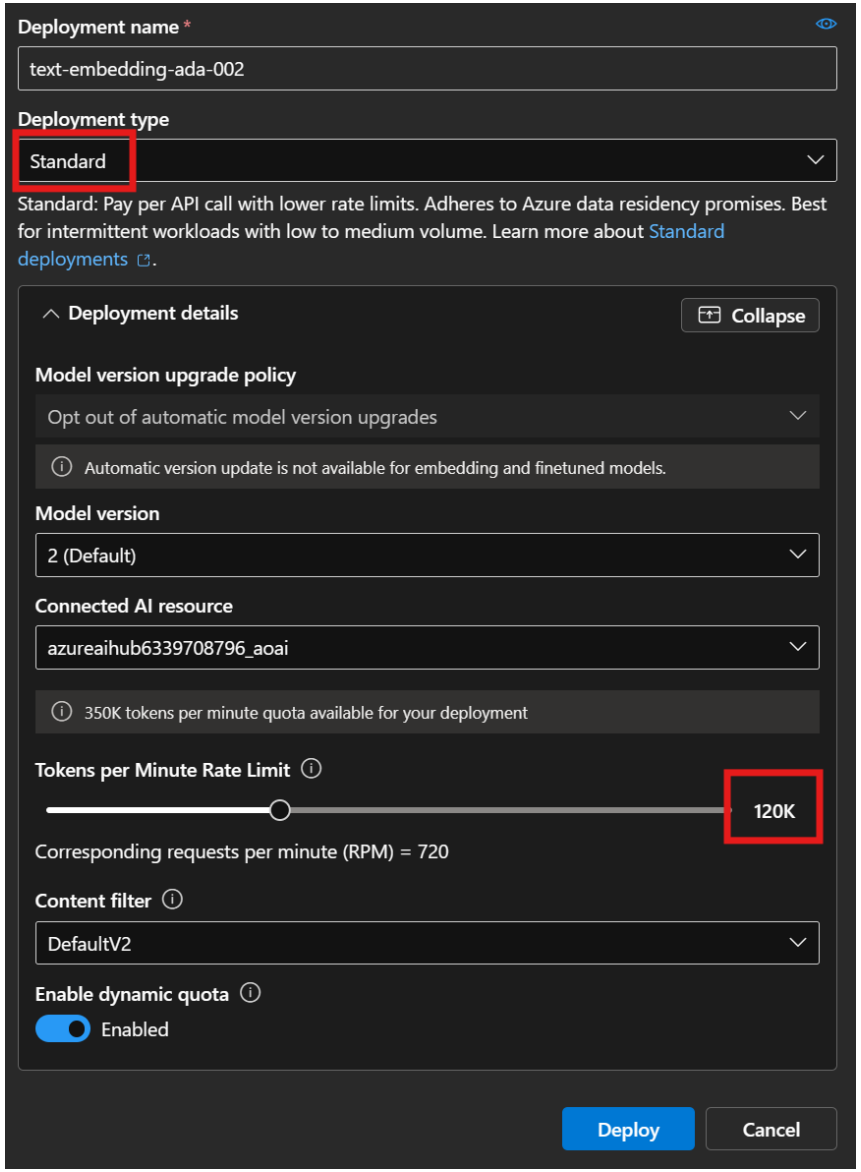
1-2. Create AI model deployments

Step 01: Deploy gpt-4o-mini and text-embedding-ada-002 models

7. For this one, select **text-embedding-ada-002**, then select **Confirm**.



8. Ensure the **Deployment type** is set to **Standard**, then select **Deploy**.



Task 03: Create a search service and add data connections

Introduction:

MicroRetail's chatbot must provide accurate and context-aware responses. To achieve this, an AI-powered search service is required. By implementing Azure AI Search, MicroRetail can index and retrieve relevant information from a structured data repository, ensuring customers receive the most relevant responses based on their queries.

Description:

In this task, you'll create an Azure AI Search service resource and connect it to your Azure AI project. The AI Search service enhances chatbot functionality by allowing it to pull relevant information from customer reviews, product catalogs, and historical interactions.

Success Criteria:

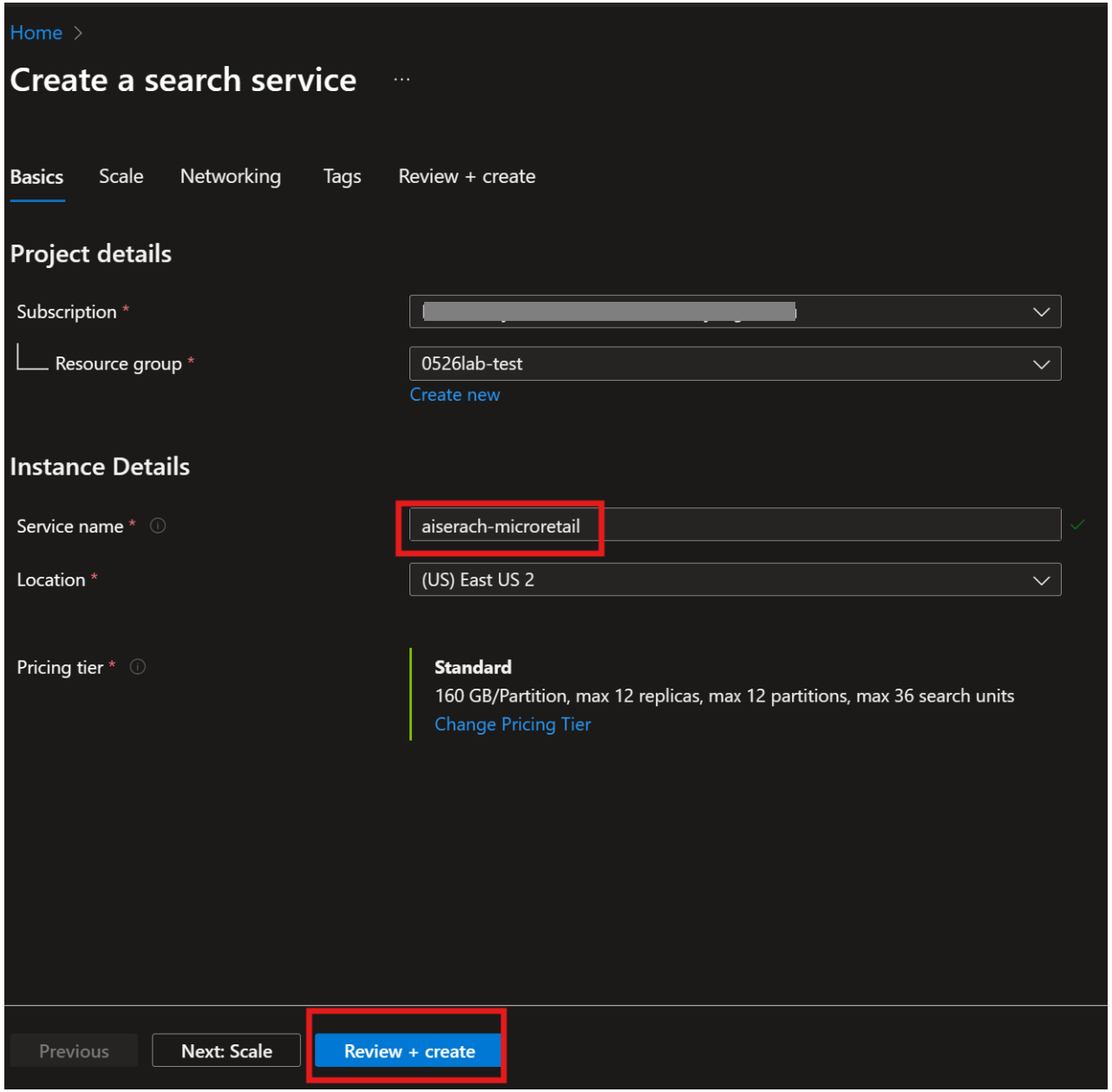
- The Azure AI Search service has been created.
- The Azure AI Search service has been connected to the Azure AI project.

1-3. Create a search service and add data connections

Step 01: Create the azureaiss search service

1. To create a search service, return to the Azure portal tab.
2. In the search bar at the top, search for **search** , then select **AI Search**.
3. From the **AI Search** page, select **Create search service**.
4. On the **Create a search service** page, select your resource group and location (these may differ from the screenshot).
   Name the search service **aiserach-microretail**, select **Review + create**, then select **Create**.

** Ensure the **Location** is set to the same location where the hub was created earlier.

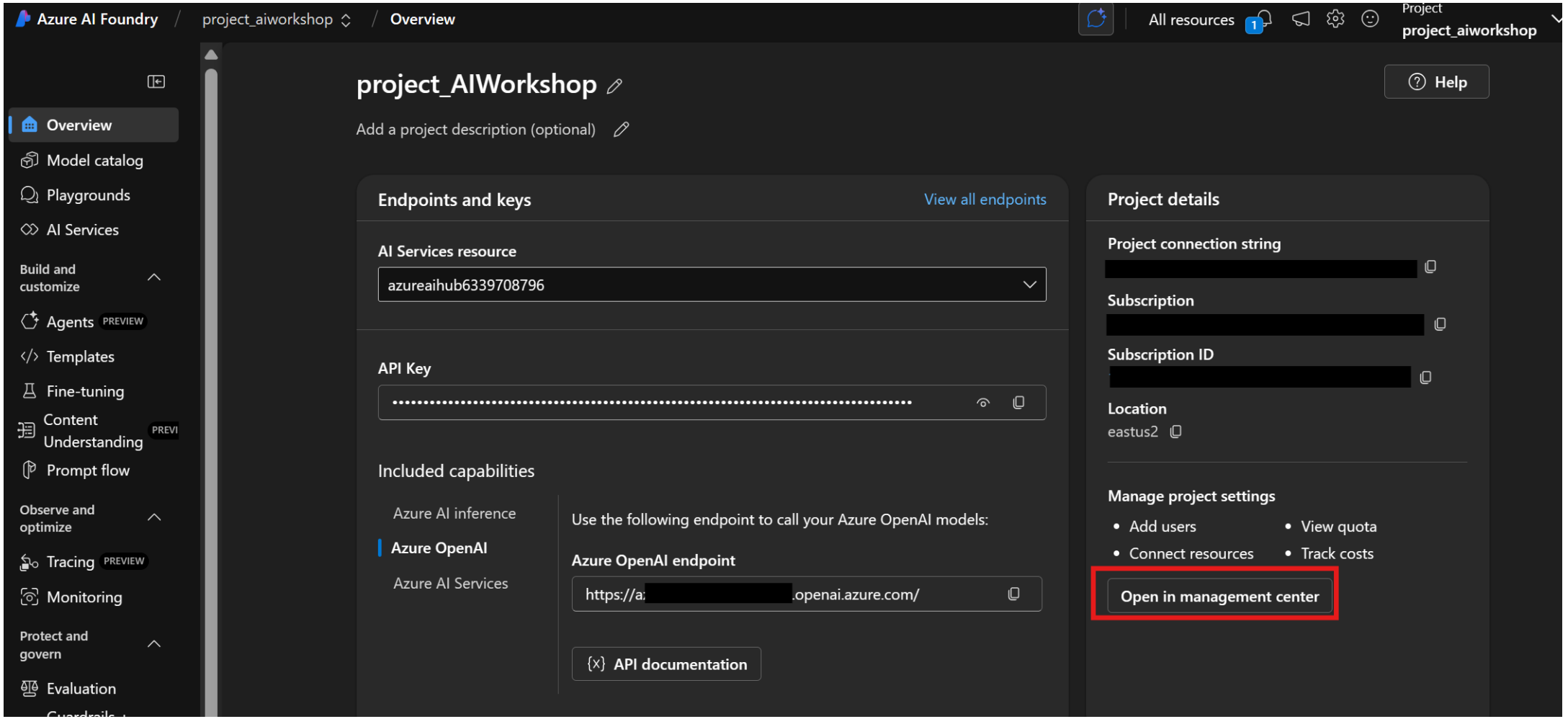


1-3. Create a search service and add data connections

(Optional)

Step 01: Create the azureaiss search service

5. Return to the tab with Azure AI Overview . On the right of the Overview page, select **Open in management center**.

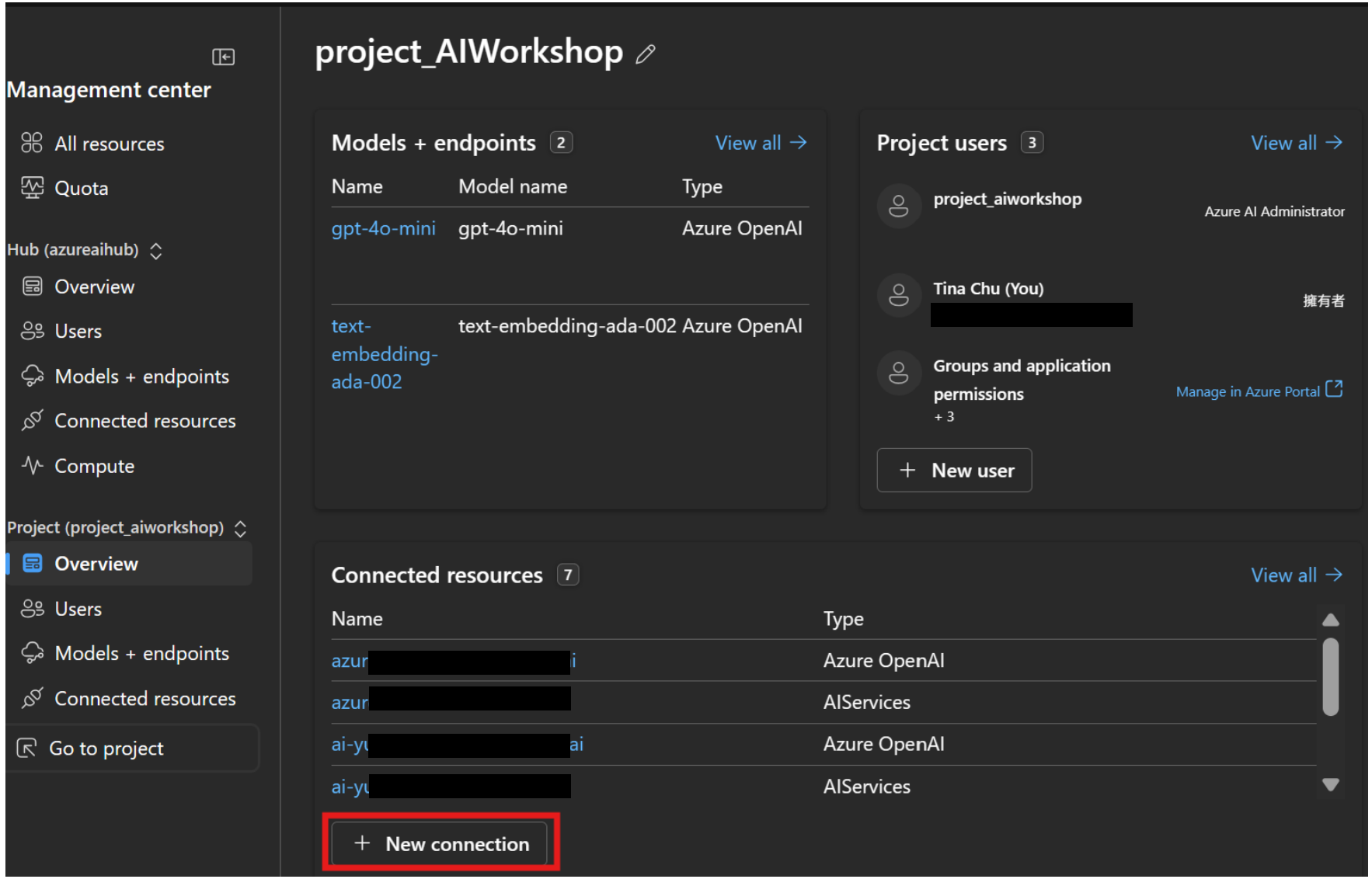


1-3. Create a search service and add data connections

(Optional)

Step 01: Create the azureaiss search service

6. From the **Management center** for **project_AIWorkshop**, under **Connected resources**, select **+ New connection**.

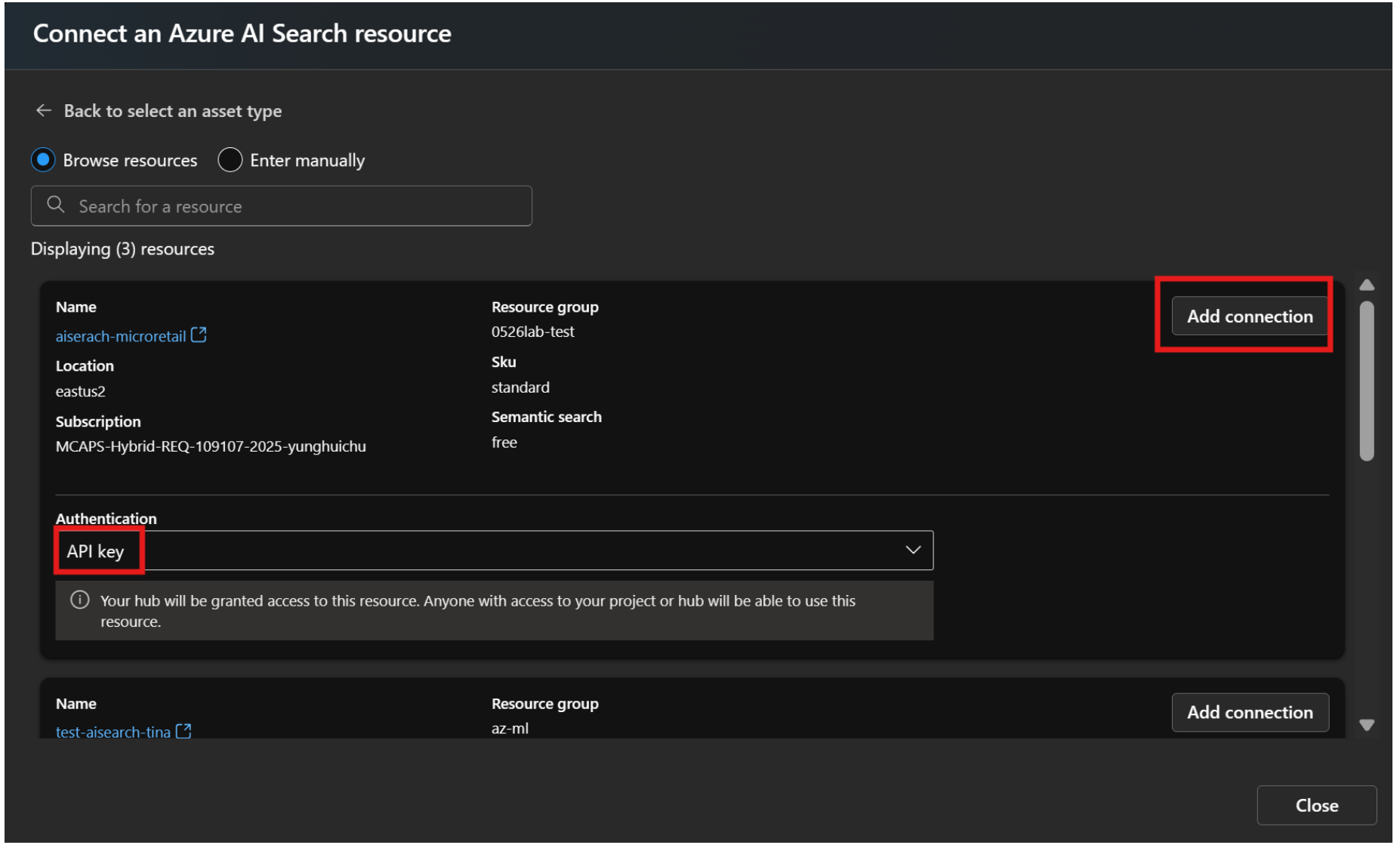


1-3. Create a search service and add data connections

(Optional)

Step 01: Create the azureaiss search service

7. On the **Add a connection to external assets** page, select **Azure AI Search**, then select **Add connection** next to the **aiserach-microretail** service.



Exercise two

Perform prompt engineering

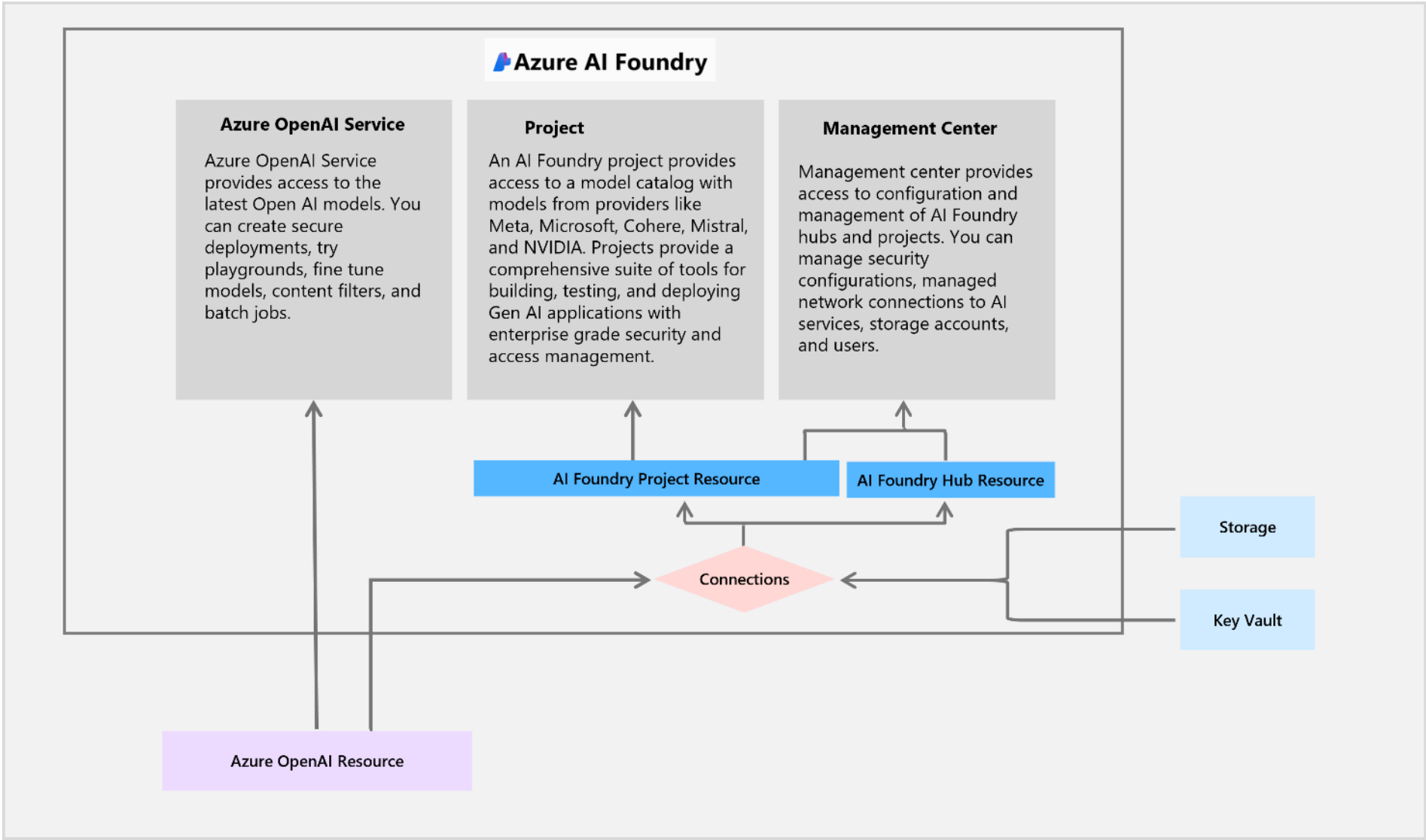
Introduction: Perform prompt engineering

In addition to providing a safe and engaging environment MicroRetail aims to provide a personalized online shopping experience for its customers. To achieve this, MicroRetail will introduce its own data and vector stores in the chat playground.

After completing this exercise, you'll be able to, within Azure AI Foundry:

- Manage compute resources for AI workflows.
- Add data to AI projects and create a search index.
- Use the Azure AI Playground to interact with indexed data.

Exercise two architecture



Task 01: Manage compute at the hub level

Introduction:

To ensure smooth execution of AI workflows, **MicroRetail** must provision appropriate compute resources. Managing compute at the hub level allows for efficient processing of AI models and prompt flows, supporting the company's AI-driven chatbot and customer interaction initiatives.

Description:

In this task, you'll create a compute instance to use in running prompt flows. This ensures the necessary processing power is available to execute AI workloads effectively.

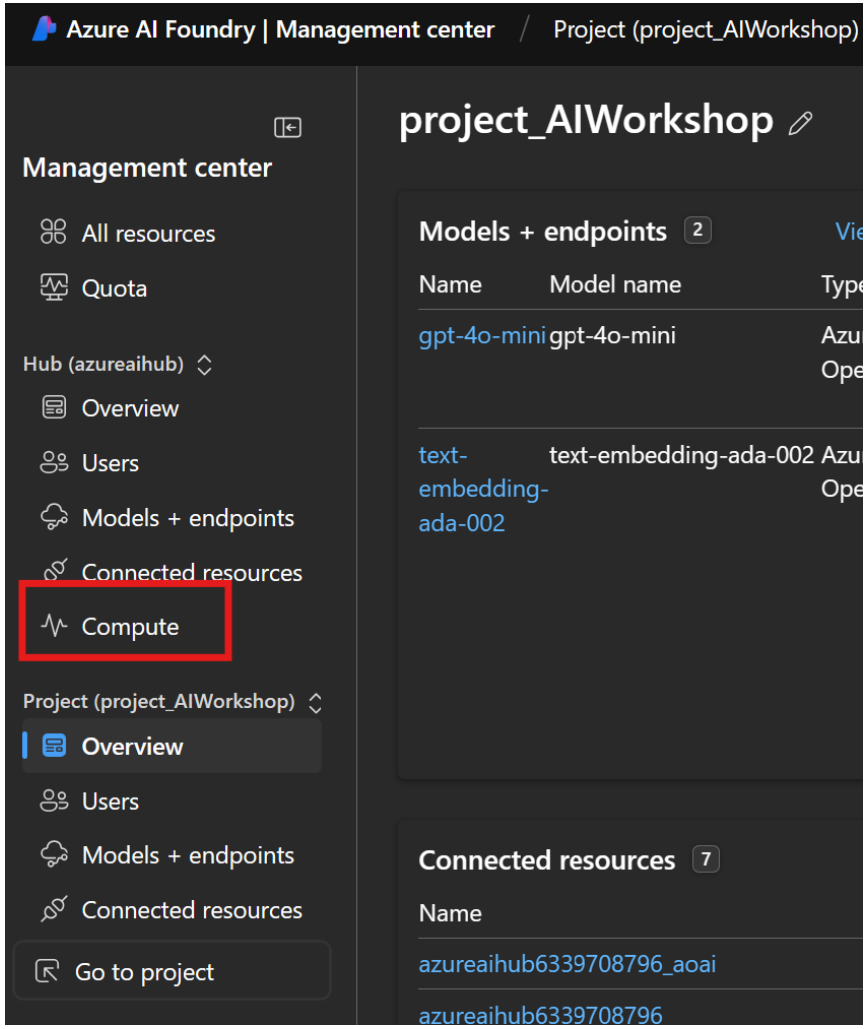
Success Criteria:

- The compute instance has been created and is running.

2-1. Manage compute at the hub level

Step 01: Create the compute instance resource in hub

1. From the **Azure AI Foundry Management center**, select **Compute** under the **Hub** section from the left menu.



IMPORTANT

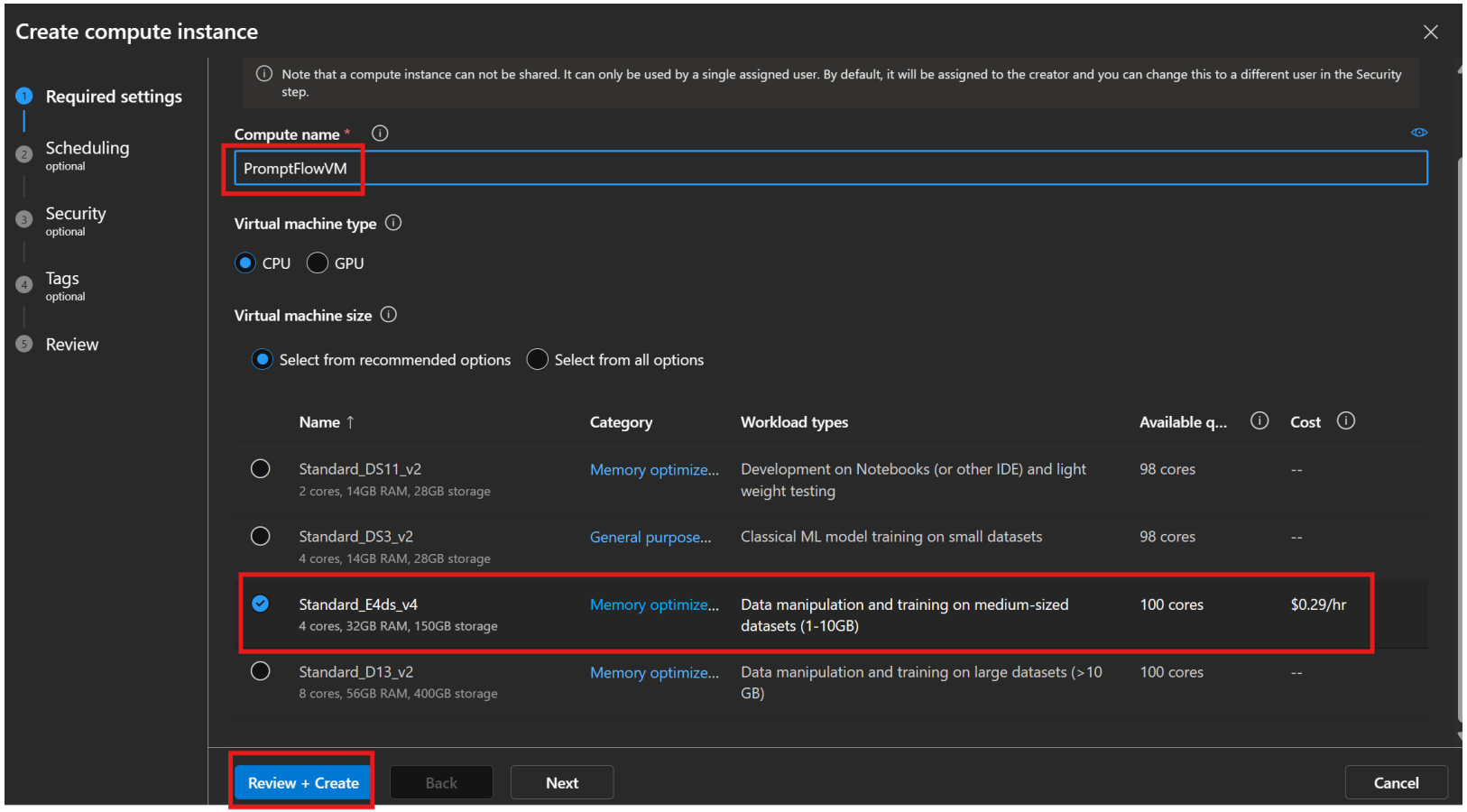
A compute instance is a virtual machine that is used to run prompt flows within projects. It's created at the hub level and accessible by any projects with proper permissions.

2. From the **Manage compute resources in this hub** page, select **+ New**.

2-1. Manage compute at the hub level

Step 01: Create the compute instance resource in hub

3. Name the compute instance **PromptFlowVM**, leave the other settings as default, then select **Review + Create**.



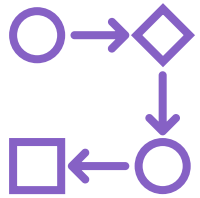
4. Once the compute instance is created, it should show a status of **Running**.

IMPORTANT

The compute instance can be used for various different tasks such as creating an index, building an open source model, or executing complex prompt flows.

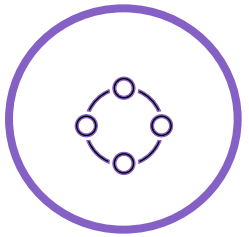
Compute instance

What it is:



A virtual machine used to run prompt flows within projects

Accessible by any projects with proper permissions



Created at the hub level

You will use the compute instance later to run prompt flows.

Task 02: Add data and create a search index

Introduction:

To enable personalized AI interactions, **MicroRetail** needs to integrate structured data that enhances chatbot responses. By indexing this data, the AI model can retrieve and process relevant customer and product information efficiently.

Description:

In this task, you'll add data to your project and index it for retrieval. This allows the chatbot to access relevant product details and improve customer interaction quality.

Success Criteria:

- **products.xlsx** has been created.
- **products.xlsx** has been indexed successfully.

2-1. Add data and create a search index

Step 01: Download products.xlsx file for use in Playground

1. Download the **products.xlsx** file HERE.

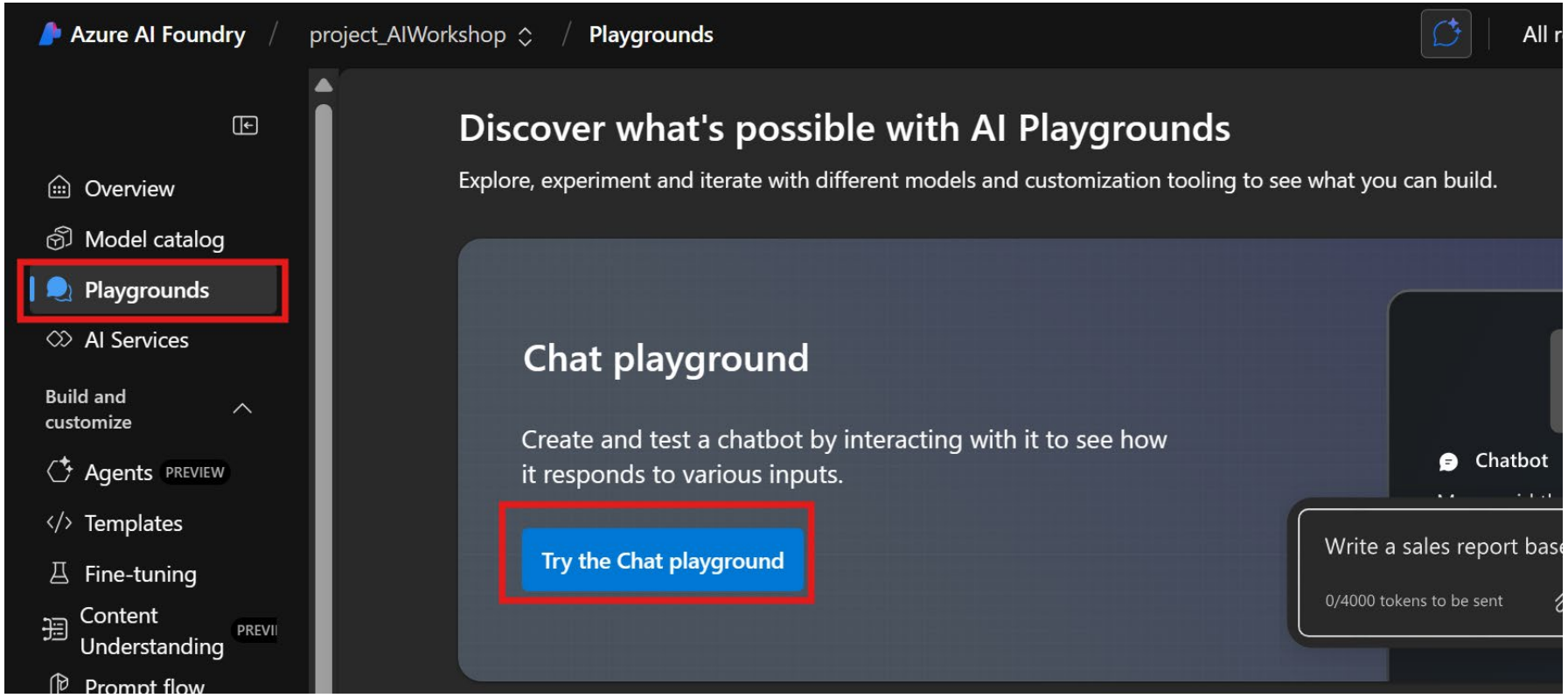
id	name	price	category	brand	description
1	TrailMaster X4 Tent	250	Tents	OutdoorLiving	Unveiling the TrailMaster X4 Tent from OutdoorLiving, your
2	Adventurer Pro Backpack	90	Backpacks	HikeMate	Venture into the wilderness with the HikeMate's Adventurer
3	Summit Breeze Jacket	120	Hiking Clothing	MountainStyle	Discover the joy of hiking with MountainStyle's Summit Bree
4	TrekReady Hiking Boots	140	Hiking Footwear	TrekReady	Introducing the TrekReady Hiking Boots - stepping up your
5	BaseCamp Folding Table	60	Camping Tables	CampBuddy	CampBuddy's BaseCamp Folding Table is an adventurer's
6	EcoFire Camping Stove	80	Camping Stoves	EcoFire	Introducing EcoFire's Camping Stove, your ultimate compa
7	CozyNights Sleeping Bag	100	Sleeping Bags	CozyNights	Embrace the great outdoors in any season with the lightwei
8	Alpine Explorer Tent	350	Tents	AlpineGear	Welcome to the joy of camping with the Alpine Explorer Te
9	SummitClimber Backpack	120	Backpacks	HikeMate	Adventure waits for no one! Introducing the HikeMate Sumr
10	TrailBlaze Hiking Pants	75	Hiking Clothing	MountainStyle	Meet the TrailBlaze Hiking Pants from MountainStyle, the s
11	TrailWalker Hiking Shoes	110	Hiking Footwear	TrekReady	Meet the TrekReady TrailWalker Hiking Shoes, the ideal co
12	TrekMaster Camping Chair	50	Camping Tables	CampBuddy	Gravitate towards comfort with the TrekMaster Camping Cl
13	PowerBurner Camping Stove	100	Camping Stoves	PowerBurner	Unleash your inner explorer with the PowerBurner Dual Bur
14	MountainDream Sleeping Bag	130	Sleeping Bags	MountainDream	Meet the MountainDream Sleeping Bag: your new must-hav
15	SkyView 2-Person Tent	200	Tents	OutdoorLiving	Introducing the OutdoorLiving SkyView 2-Person Tent, a pe
16	TrailLite Daypack	60	Backpacks	HikeMate	Step up your hiking game with HikeMate's TrailLite Daypac
17	RainGuard Hiking Jacket	110	Hiking Clothing	MountainStyle	Introducing the MountainStyle RainGuard Hiking Jacket - th
18	TrekStar Hiking Sandals	70	Hiking Footwear	TrekReady	Meet the TrekStar Hiking Sandals from TrekReady - the ult
19	Adventure Dining Table	90	Camping Tables	CampBuddy	Discover the joy of outdoor adventures with the CampBudd
20	CompactCook Camping Stove	60	Camping Stoves	CompactCook	Step into the great outdoors with the CompactCook Campi

2-1. Add data and create a search index

Step 02: Create vector index for gpt-4o-mini model

1. Return to your browser tab with Azure AI Foundry. From the **Management center** for **project_AIWorkshop**, select **Go to project** at the bottom of the left menu.

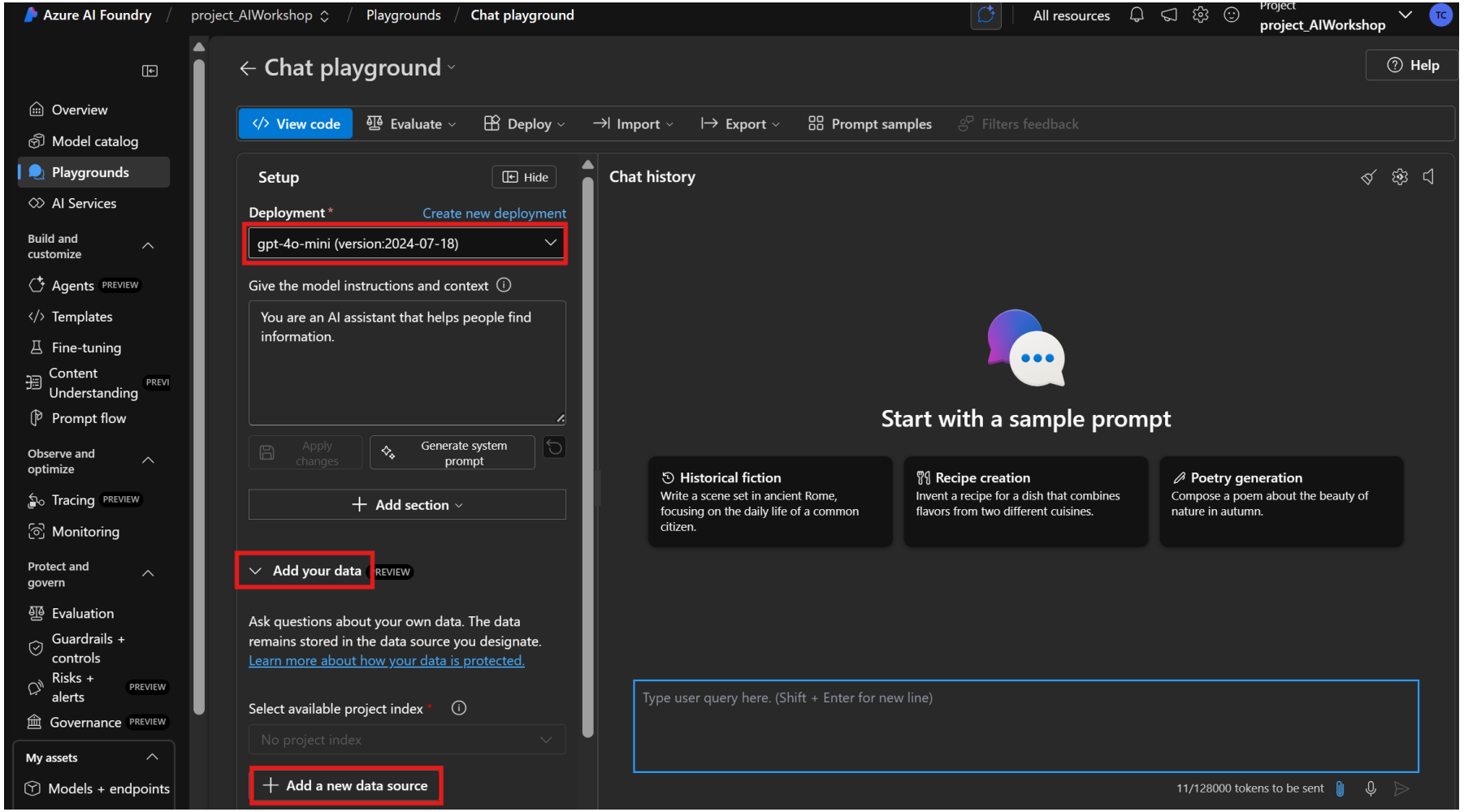
2. From the left menu of the **project_AIWorkshop** page, select **Playgrounds**.
3. At the top of the page, select **Try the Chat playground**.



2-1. Add data and create a search index

Step 02: Create vector index for gpt-4o-mini model

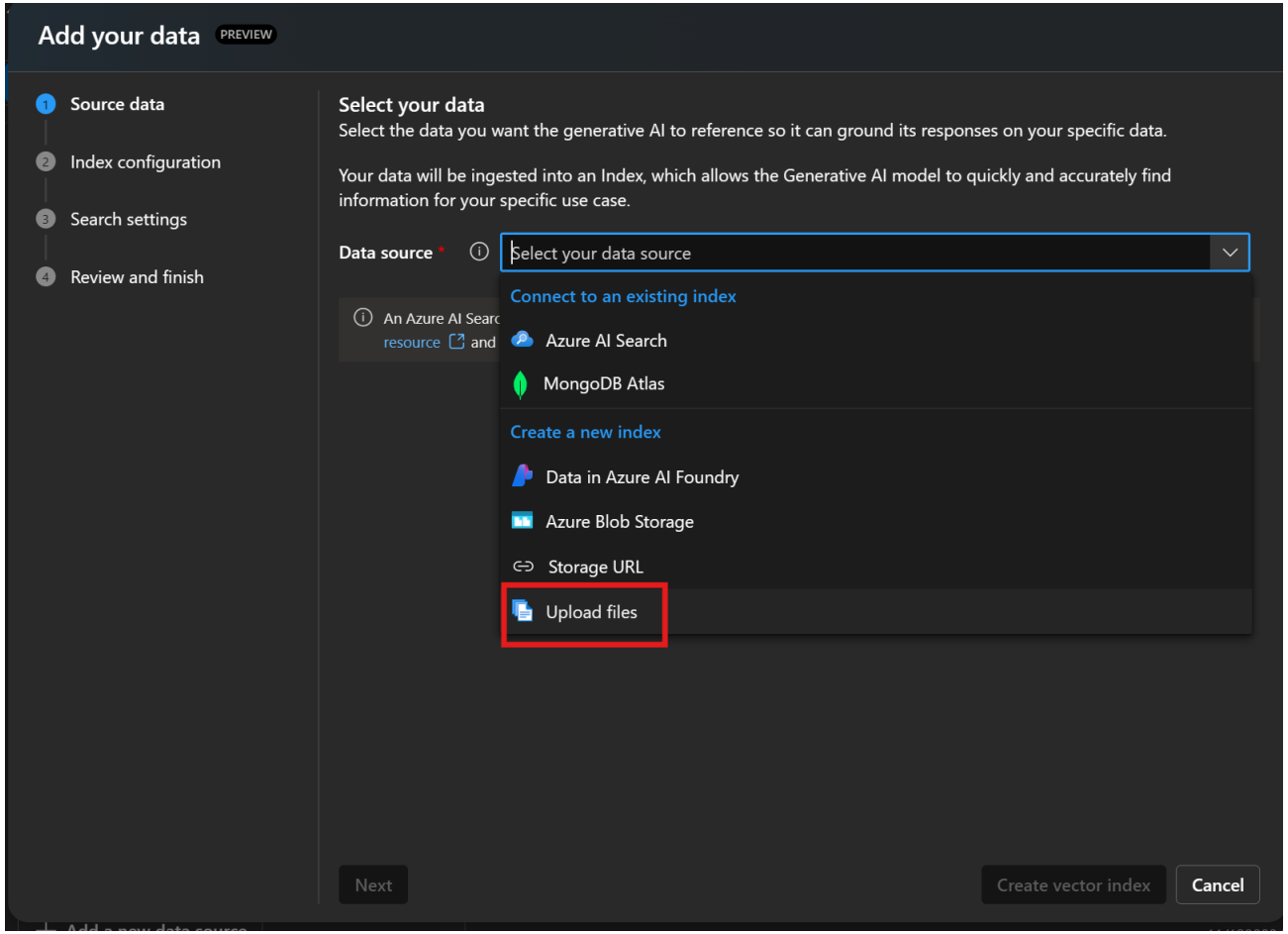
4. From the **Chat playground** page, ensure the **Deployment** is set to the **gpt-4o-mini** model, then select **+ Add a new data source** under the **Add your data** dropdown menu.



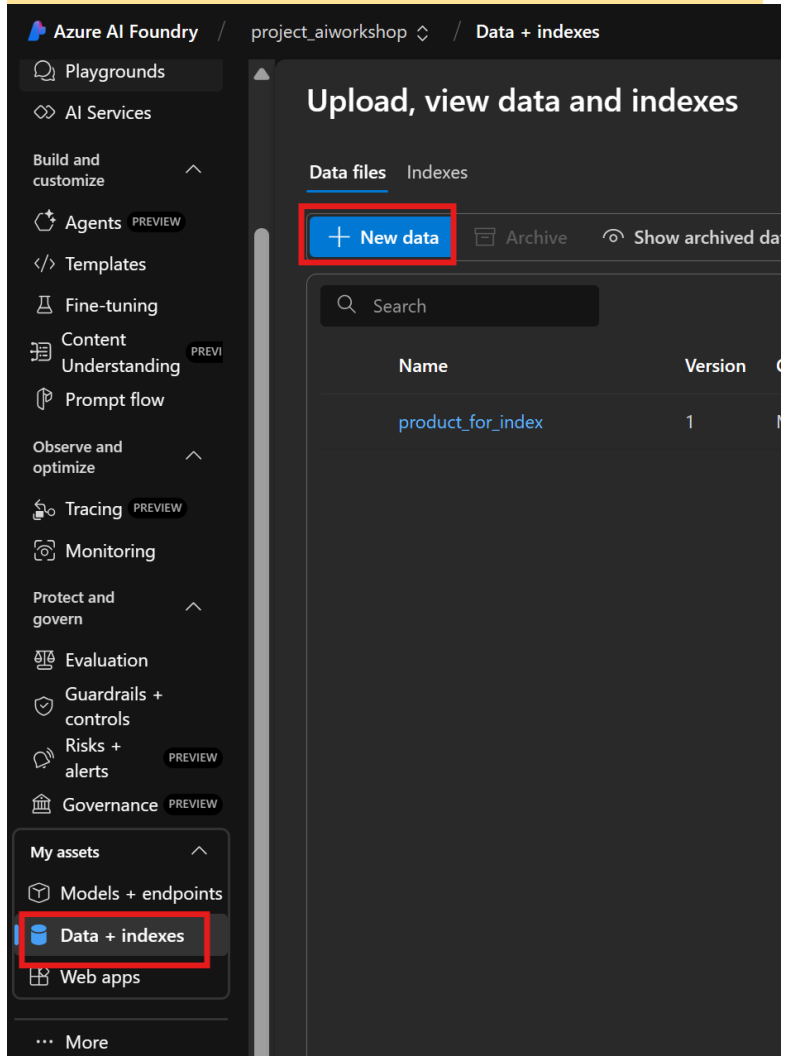
2-1. Add data and create a search index

Step 02: Create vector index for gpt-4o-mini model

5. From the **Source data** tab on the **Add your data** page, select the **Data source** dropdown menu, then select **Upload files**.
6. Select the **Upload** dropdown menu, then select **Upload files**.
7. Select the **products.xlsx** file,then select **Open**. Select **Next** once it has been uploaded.



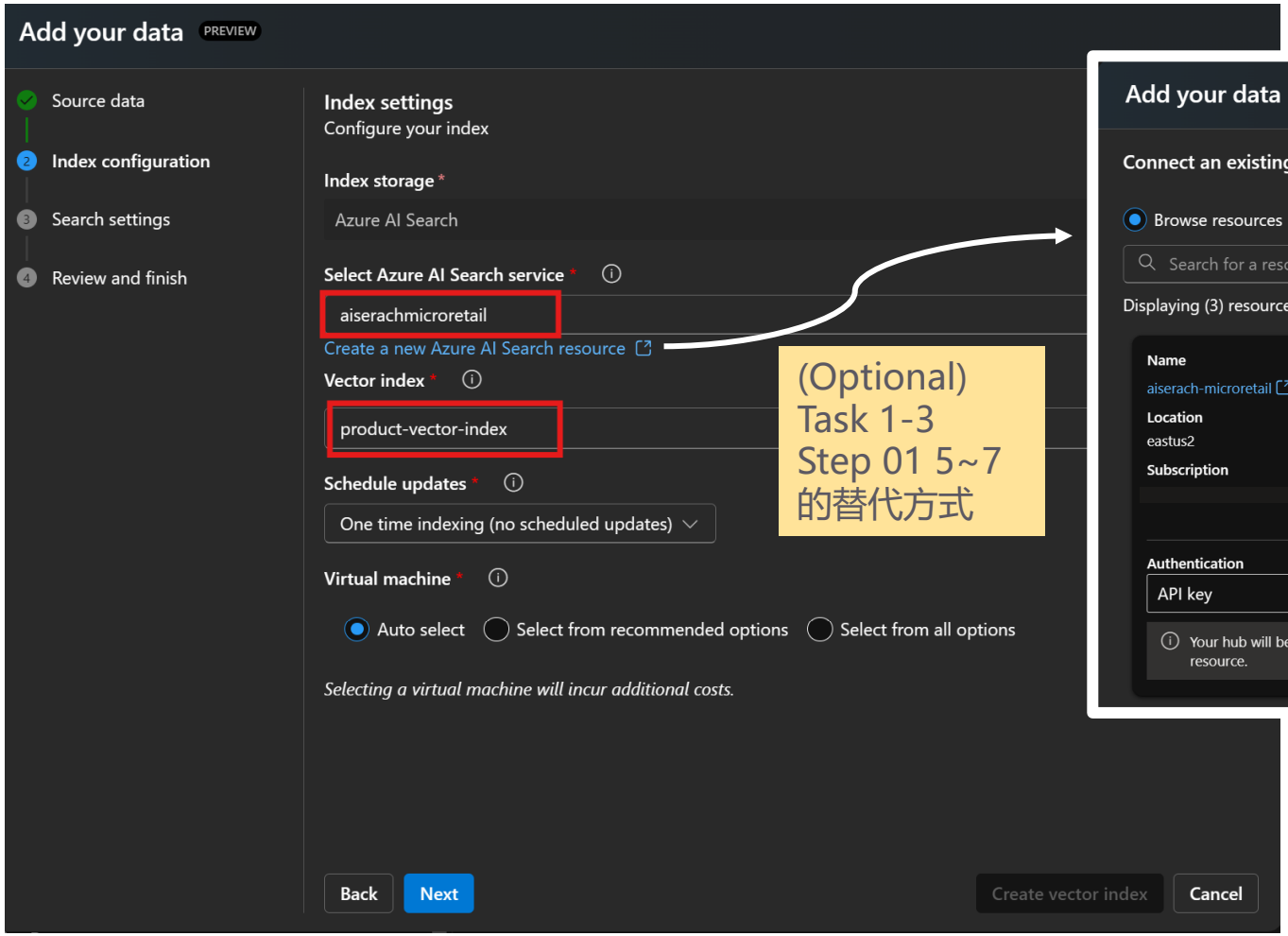
Method 2: 資料於AI Foundry管理



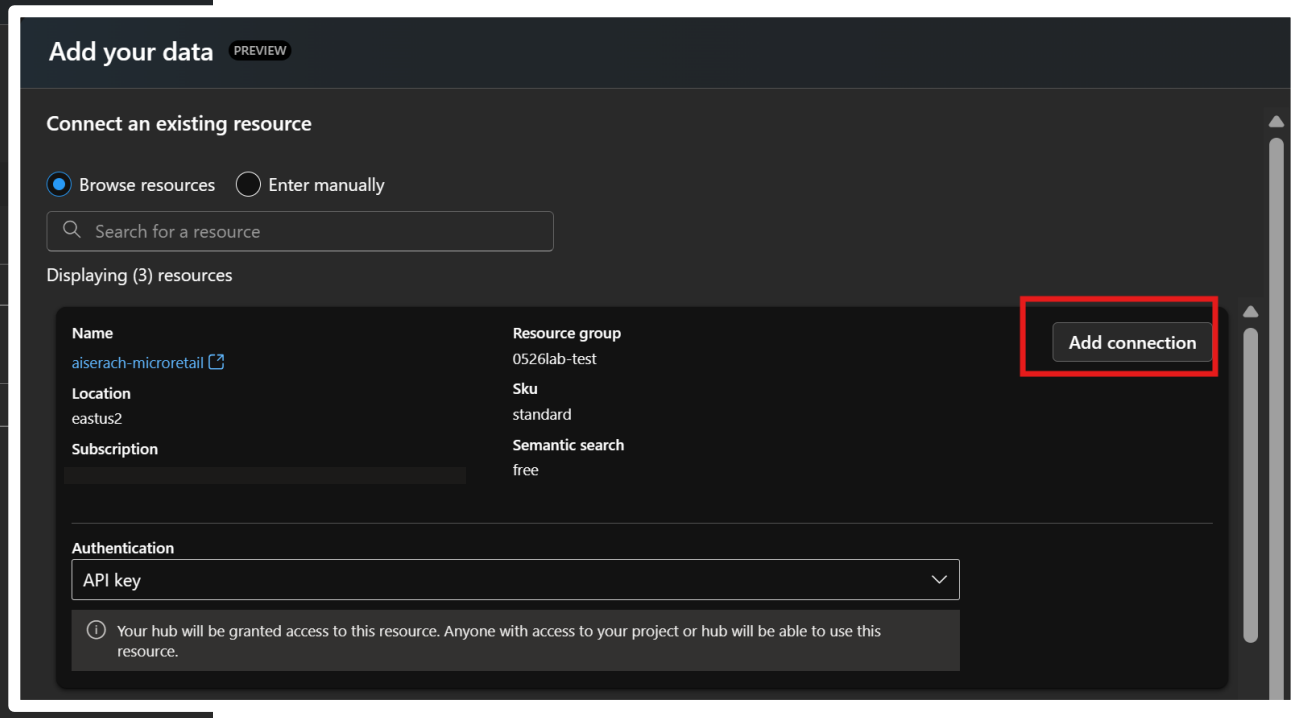
2-1. Add data and create a search index

Step 02: Create vector index for gpt-4o-mini model

8. From the **Index configuration** tab, select the **Select Azure AI Search service** dropdown menu ,then select **aiserach-microretail**.
9. Enter **product-vector-index** for the **Vector index** name, then select **Next**.



(Optional)
Task 1-3
Step 01 5~7
的替代方式

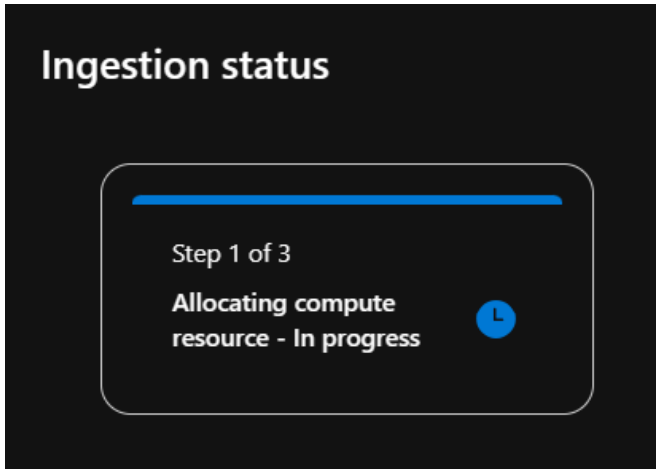


2-1. Add data and create a search index

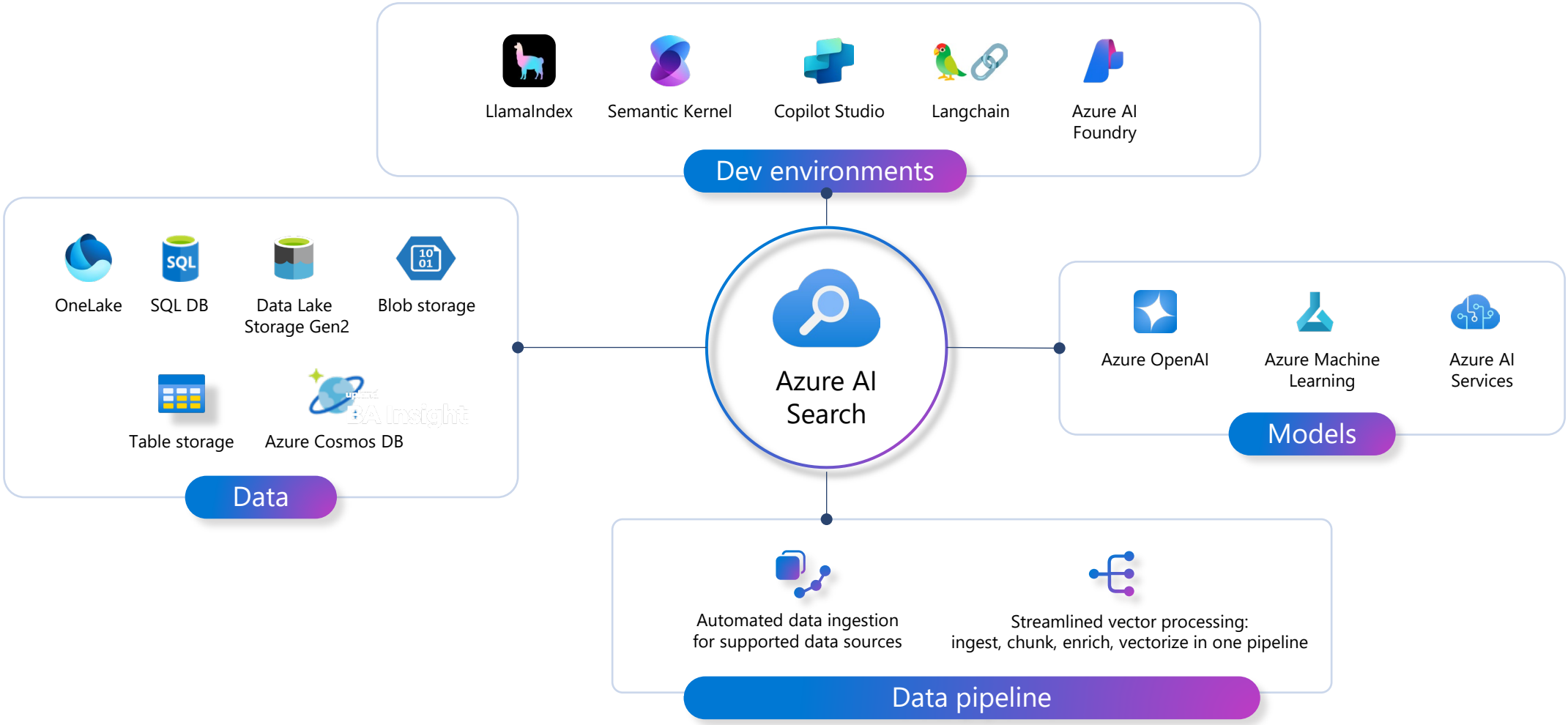
Step 02: Create vector index for gpt-4o-mini model

10. From the **Search settings** tab, leave the default settings, then select **Next**.
11. From the **Review and finish** tab, select **Create vector index**.

> Note:
> Wait for the index to be created. This should take about 3-5 minutes.
> The Ingestion status will show you the progress.



Seamless integrations for your GenAI deployments



Task 03: Use Playground to chat with your data

Introduction:

With indexed data in place, **MicroRetail** can leverage the Azure AI Playground to test and refine chatbot interactions. The Playground enables real-time testing, ensuring the chatbot returns relevant responses based on customer inquiries.

Description:

In this task, you'll use the Azure AI Playground to chat with your indexed data. This will allow you to validate response accuracy and optimize chatbot interactions.

Success Criteria:

- The chat returns output based on the indexed data when queried.

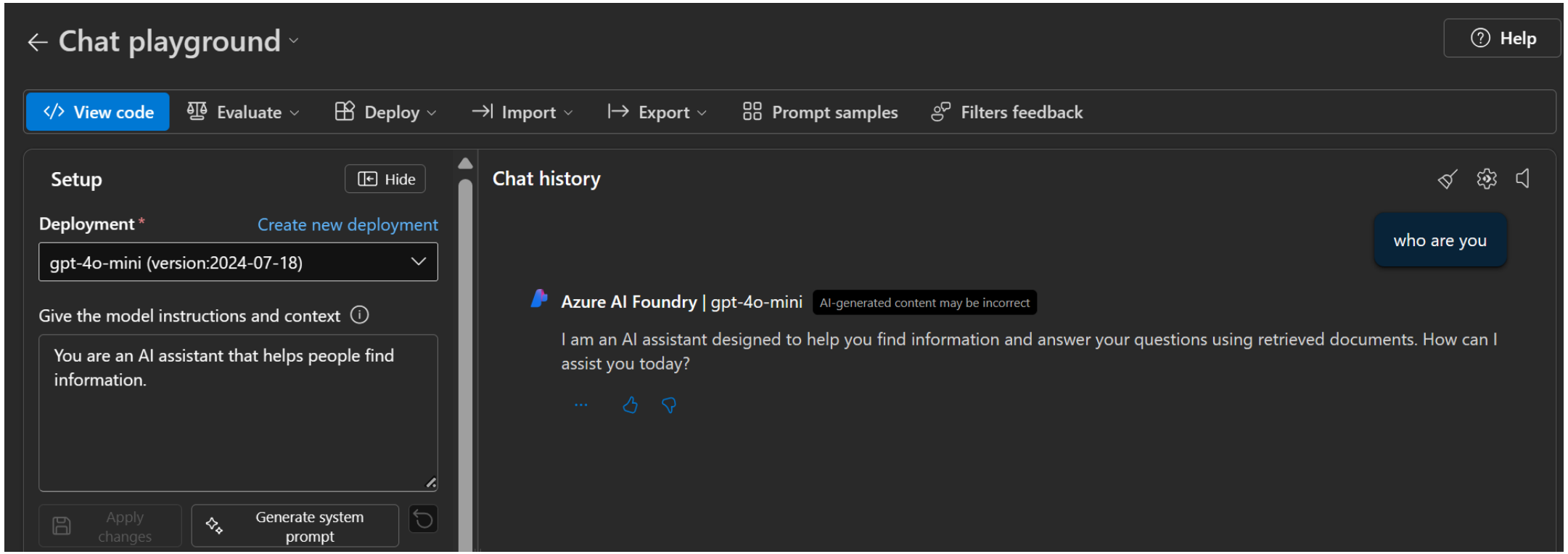
2-1. Use Playground to chat with your data

Step 01: Use the system message in the Playground to test responses

1. Once the data has been indexed, select the chat box on the right of the playground and enter the following query:

```
Who are you?
```

2. Select **Enter** and take note of the response. This is referencing the information provided in the system message, shown to the left.

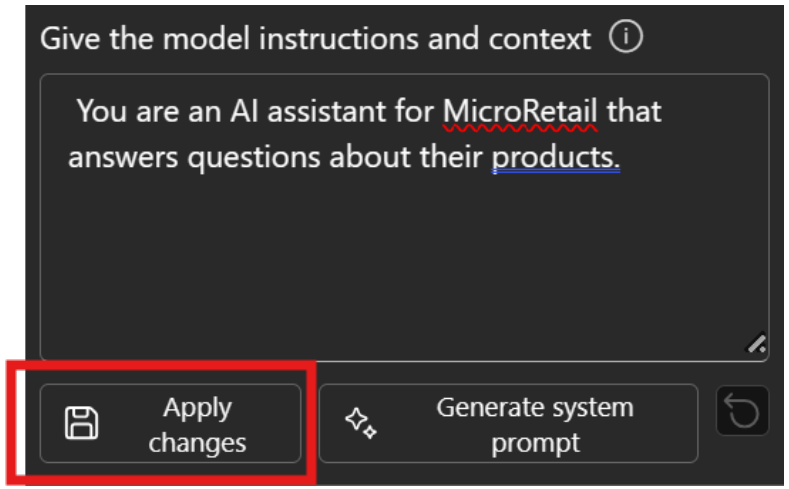


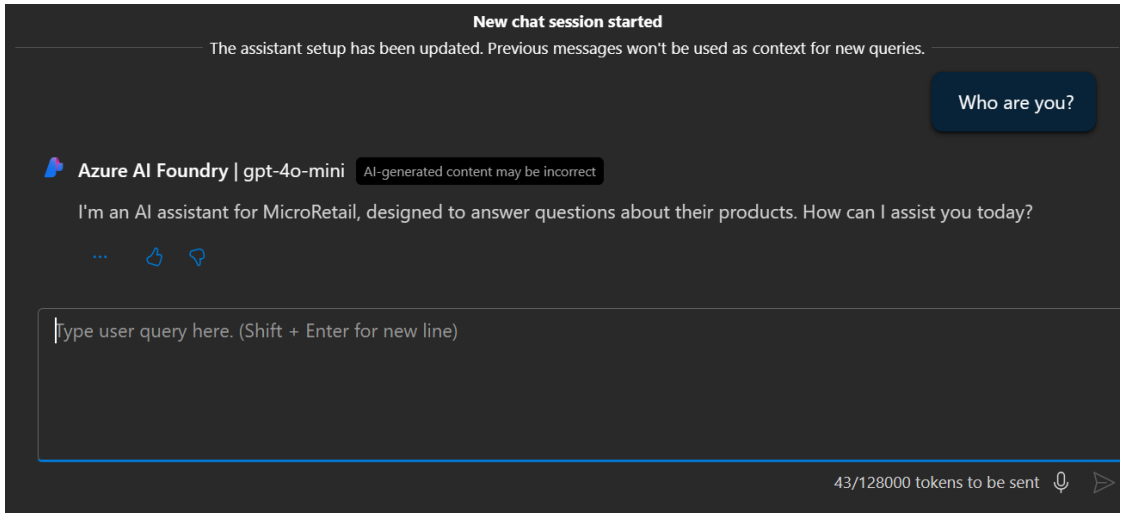
2-1. Use Playground to chat with your data

Step 01: Use the system message in the Playground to test responses

3. Change the system message to the following, then select **Apply changes**:

```
You are an AI assistant for MicroRetail that answers questions about their products.
```





4. Run the **Who are you?** query again and note the updated response.

The system message can provide more than just context for the purpose of the tool, it can also be used to influence how the model behaves and responds to queries.

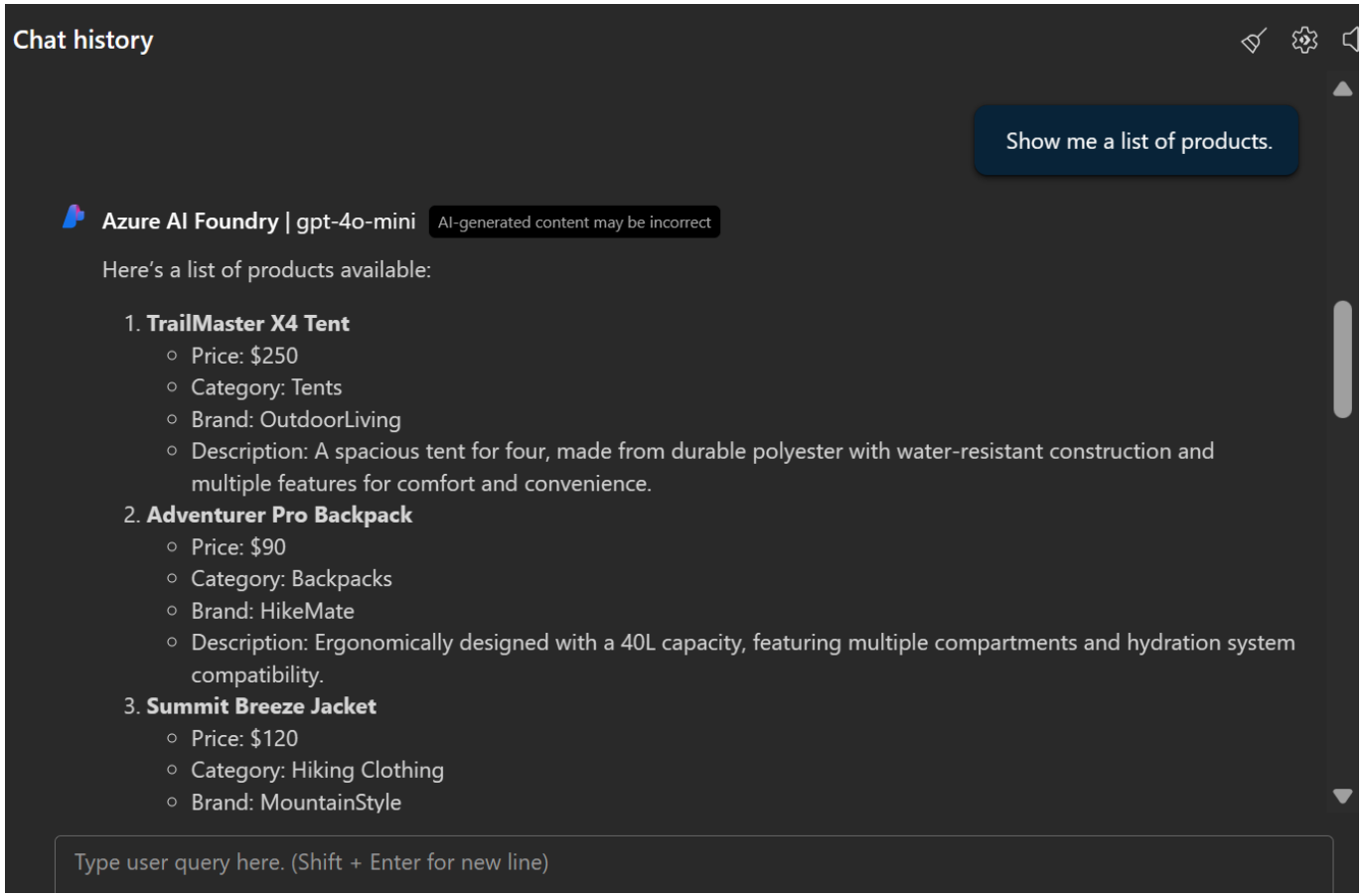
2-1. Use Playground to chat with your data

Step 01: Use the system message in the Playground to test responses

5. Next, run the following query:

```
Show me a list of products.
```

The response should return a list of products pulled from the **products.xlsx** file.



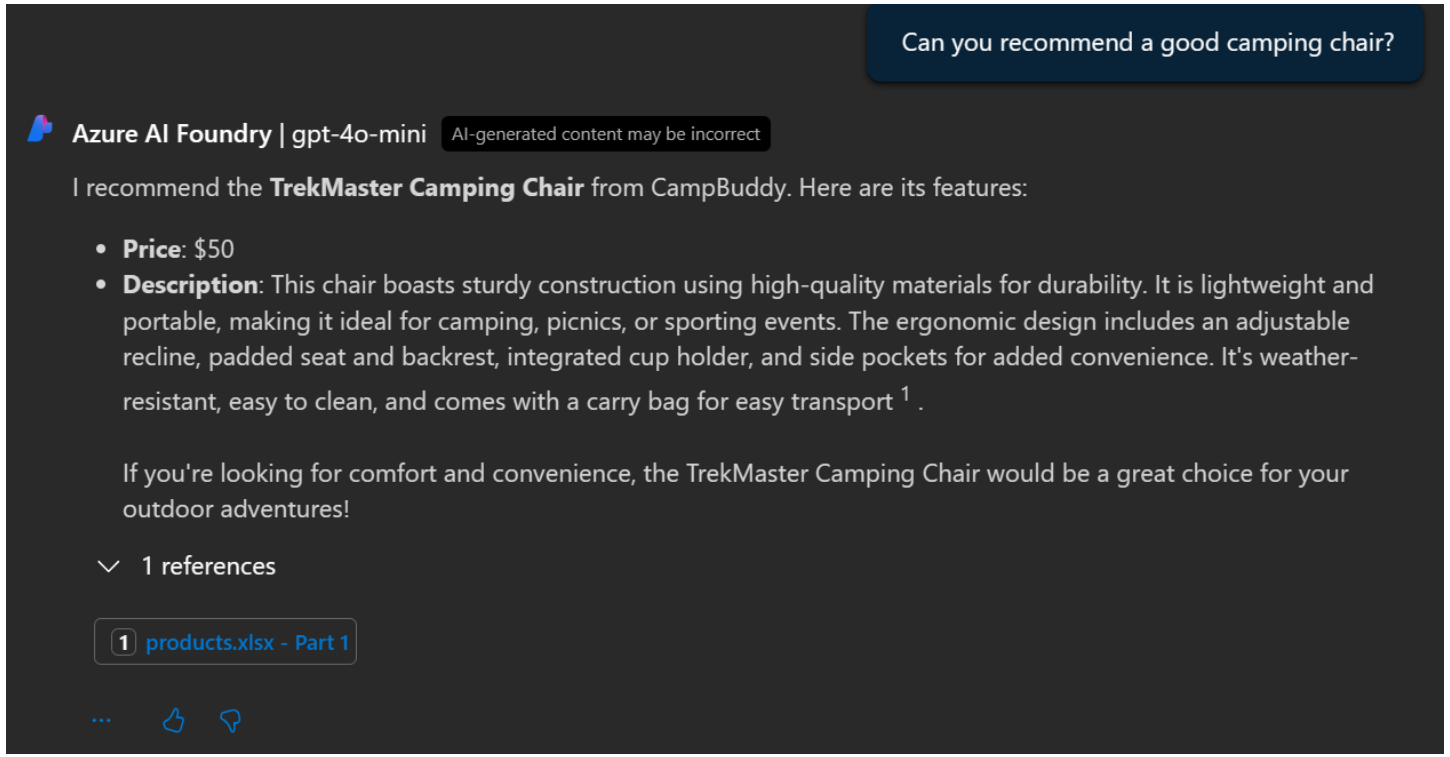
2-1. Use Playground to chat with your data

Step 01: Use the system message in the Playground to test responses

6. Next, run the following query to narrow down the products to what we're looking for:

```
Can you recommend a good camping chair?
```

The response should reference the TrekMaster Camping Chair and provide information about it.



The chat history keeps track of the entire conversation with the chatbot. In the previous question, we didn't specify what item we wanted the price for, but the chat was able to use context from chat history to properly answer the query.

System message and chat history

System message

- provides more than just context for the purpose of the tool, it can also be used to influence how the model behaves and responds to queries.

Chat history

- Keeps track of the entire conversation with the chatbot.

The chat is able to use context from chat history to properly answer the query.

Exercise three

Set up Azure Prompt Flow

Introduction: Set up Azure Prompt Flow

MicroRetail aims to leverage AI-driven solutions to reduce customer service response times and proactively address common issues through real-time analysis. By adhering to established best practices in Prompt Flow and content safety, MicroRetail is well-positioned to achieve these objectives.

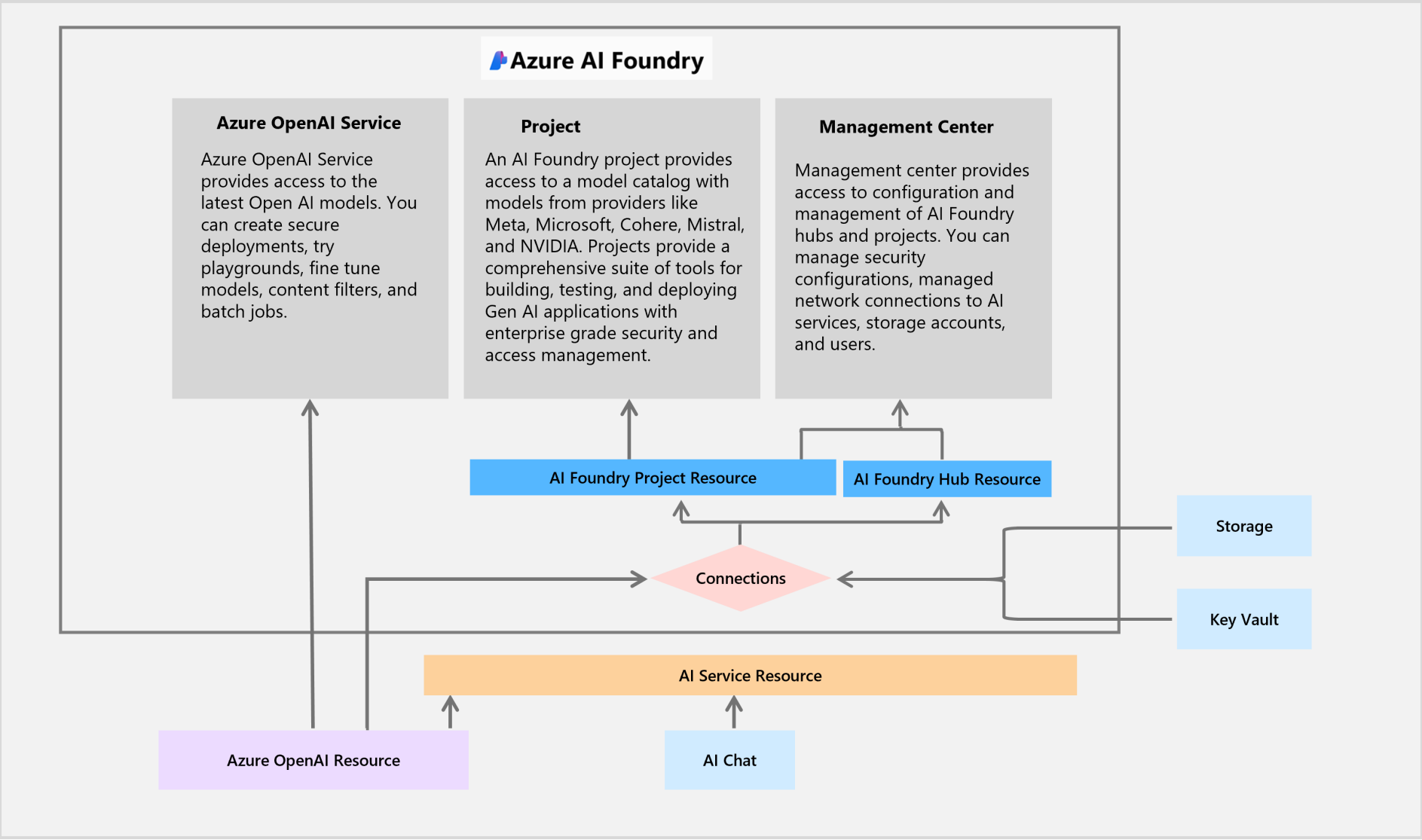
After completing this exercise, you'll be able to:

Create and manage AI-driven chat flows.

Test and refine chat interactions for accuracy and efficiency.

Evaluate chatbot responses based on predefined criteria.

Exercise three architecture



Task 01: Create a chat flow with chat history

Introduction:

To enhance customer support and engagement, **MicroRetail** is implementing chat history tracking within its AI chatbot. Maintaining chat history allows for more context-aware interactions, improving user experience and reducing repetitive queries.

Description:

In this task, you'll create a chat flow using your existing playground configuration as a template. This structured approach ensures consistency in chatbot interactions and enables tracking of previous conversations for better response accuracy.

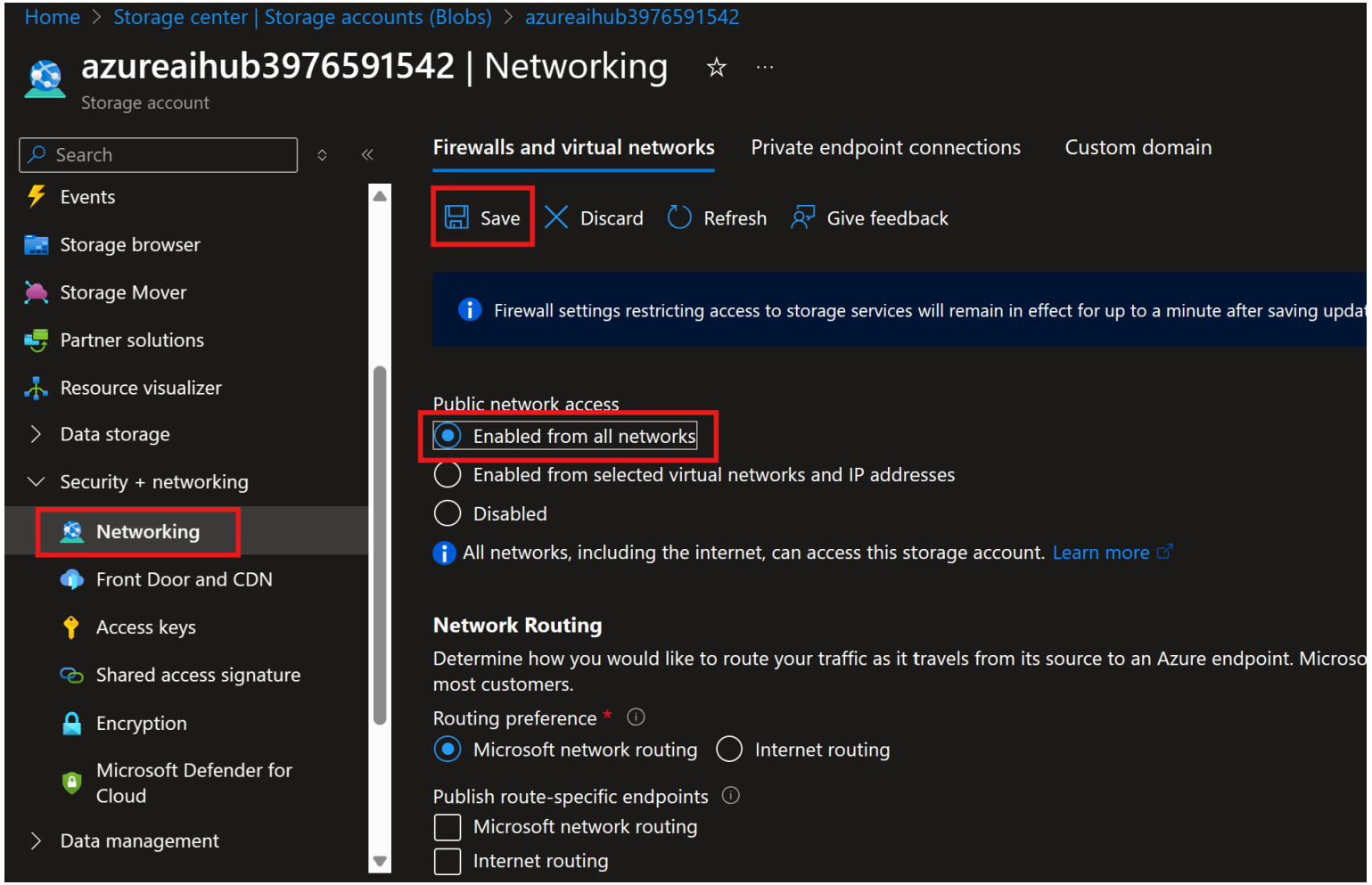
Success Criteria:

- The chat flow has been created successfully.

3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

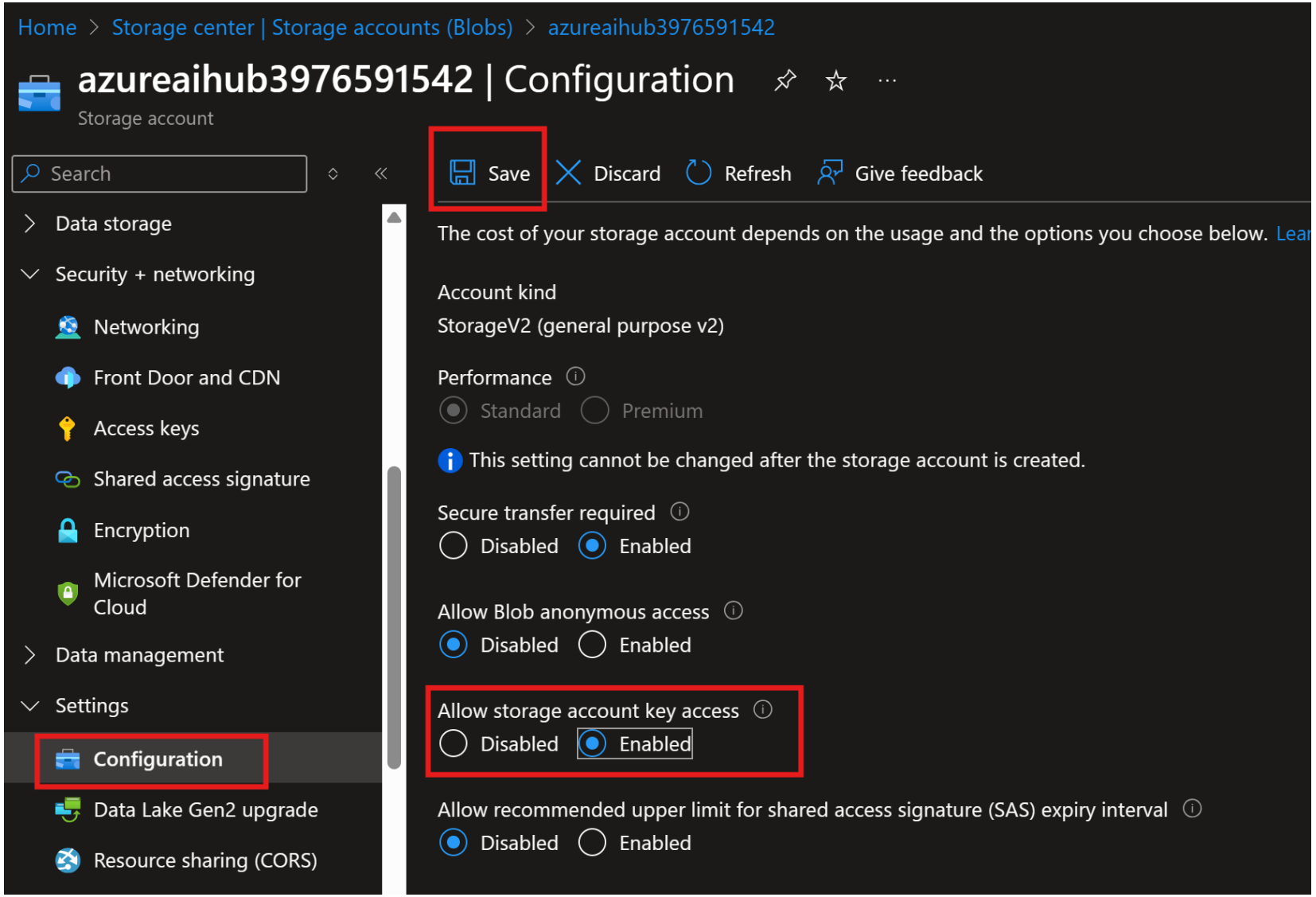
1. Return to the tab with Azure portal.
2. In the search bar at the top, enter **storage** and then select **Storage accounts**.
3. Select the storage account, then select **Security + networking** > **Networking** from the left menu.
4. For **Public network access**, select **Enabled from all networks**, then select **Save**.



3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

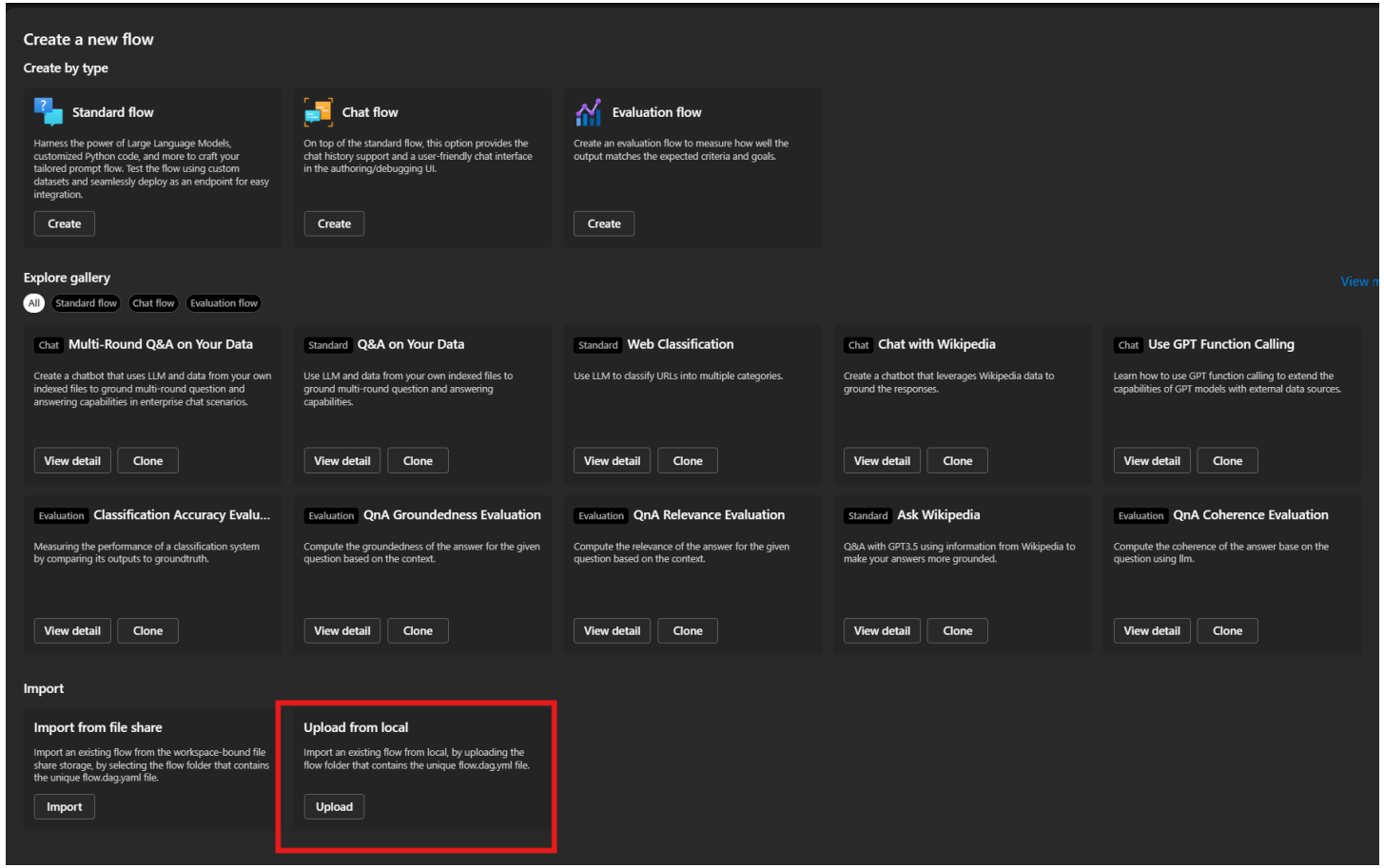
5. Select **Settings** > **Configuration** from the left menu.
6. For **allow storage account key access**, select **Enabled**, then select **Save**.

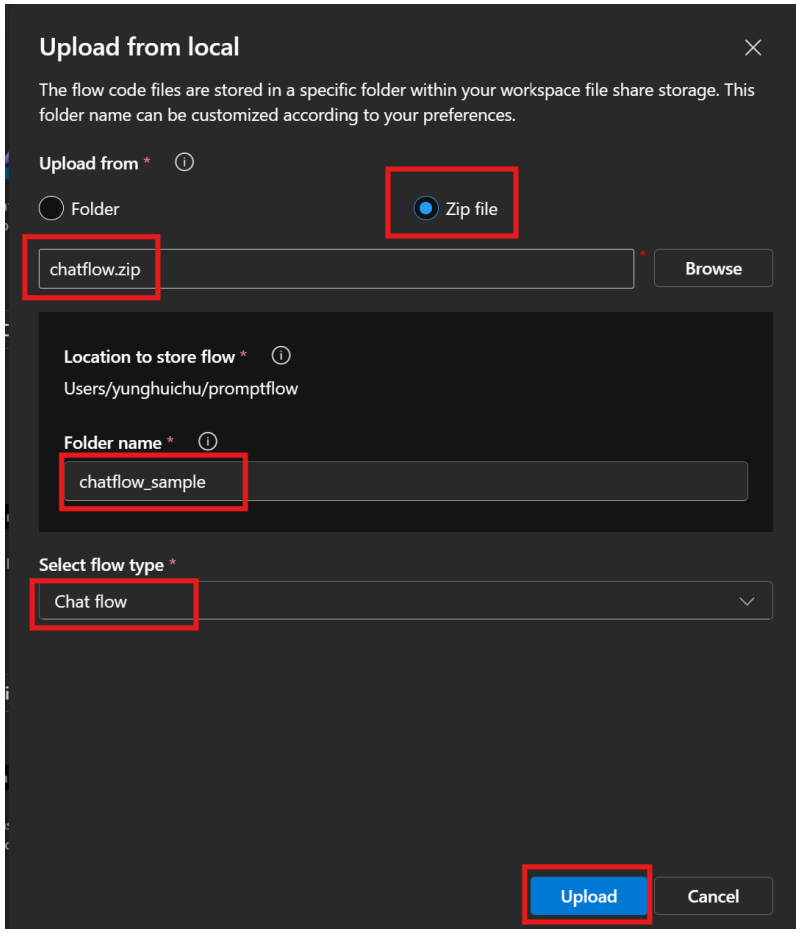


3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

7. Return to the tab with **Chat playground** and select **Prompt flow** from the left menu.
8. Select **+ Create**, then select **Upload** under **Upload from local**.
9. Download the **chatflow.zip** file HERE.
10. On the **Upload from local** pane, select **Zip file**, then select **Browse** and upload the **chatflow.zip** file.
11. Name the folder **chatflow_sample**, set the **Select flow type** field to **Chat flow**, and select **Upload**.

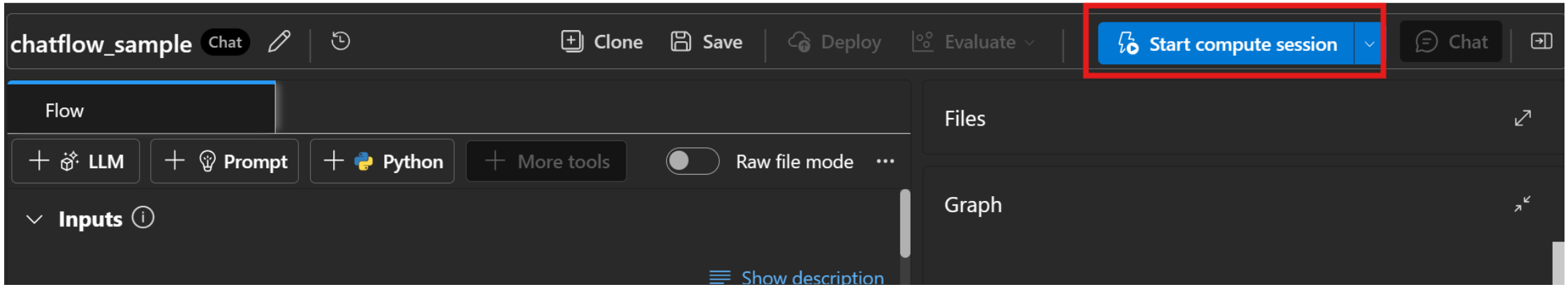




3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

12. Once the chat flow is created, select **Start compute session** in the top right.

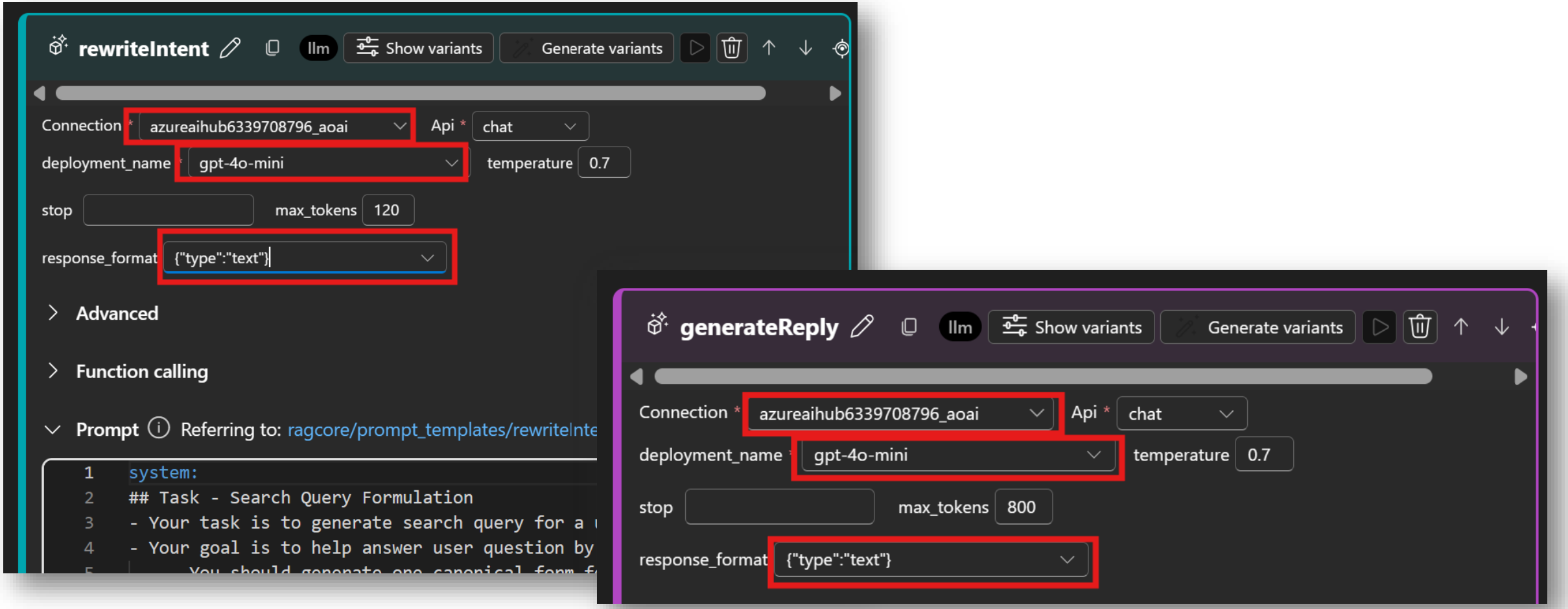


The prompt flow page contains two main panes. On the left is the flow pane, which allows you to add and configure new flows with LLMs, prompts, and various Python tools. The right pane contains a graph, allowing you to easily visualize the flow of different nodes. The graph will update dynamically as updates are made in the flow pane.

3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

13. In the left pane, find the **rewriteIntent** node and connect it to your hub. Ensure the deployment is set to **gpt-4o-mini** and the response it set to **text**.
14. Next, find the **generateReply** node and configure the same settings.

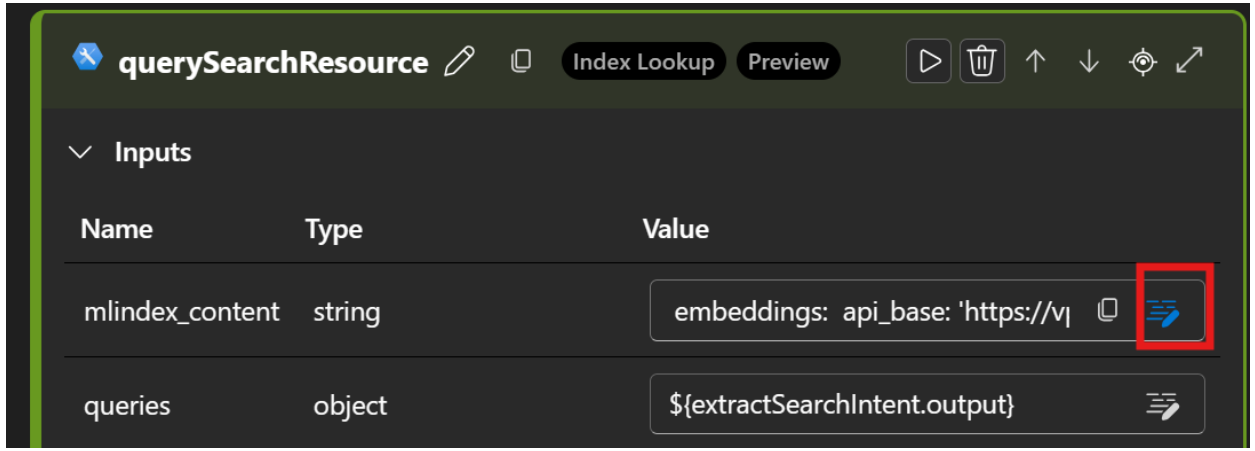


**Ensure the compute instance has started before continuing.

3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

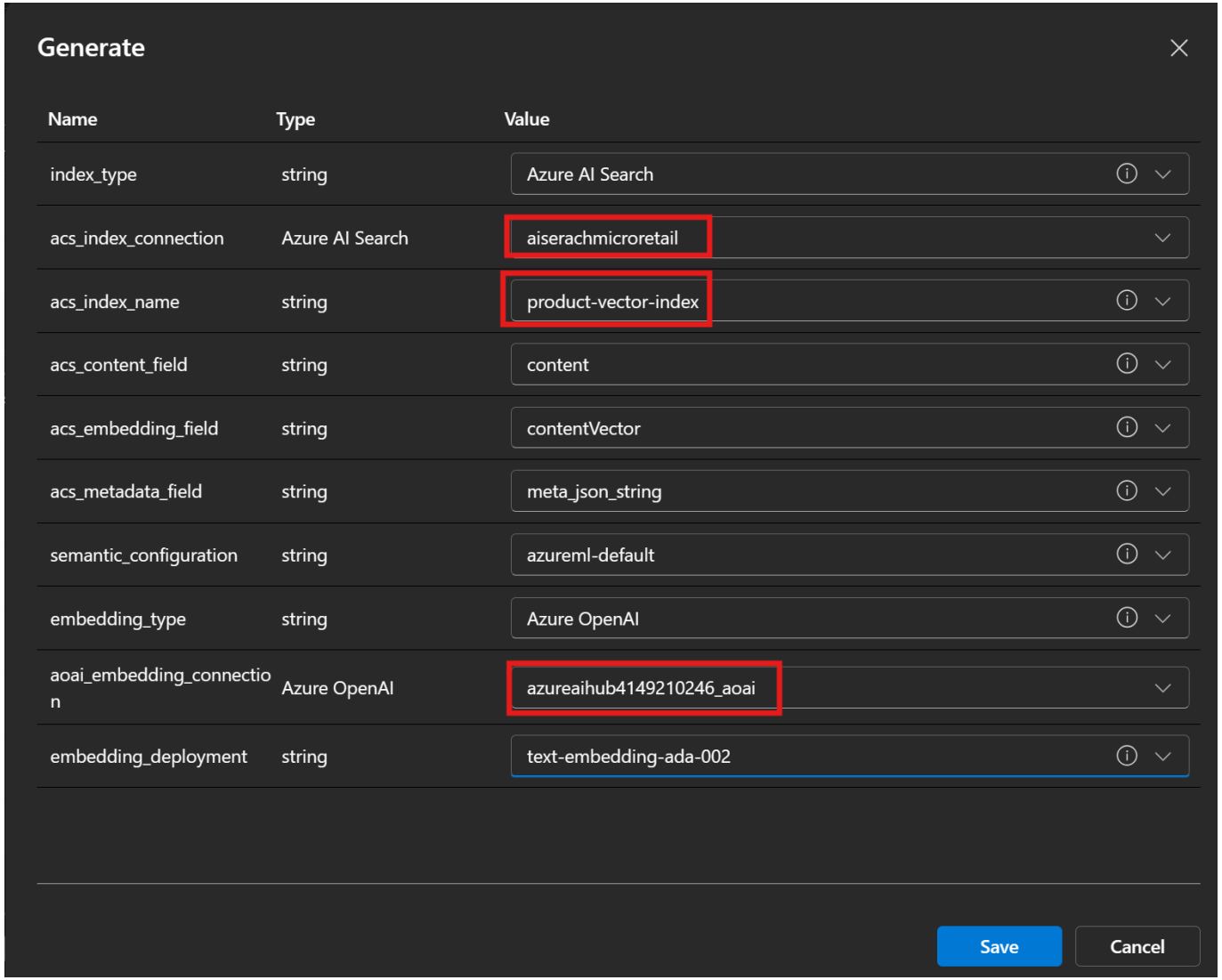
15. Next, find the **querySearchResource** node and select the edit icon next to **mlindex_content**.



3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

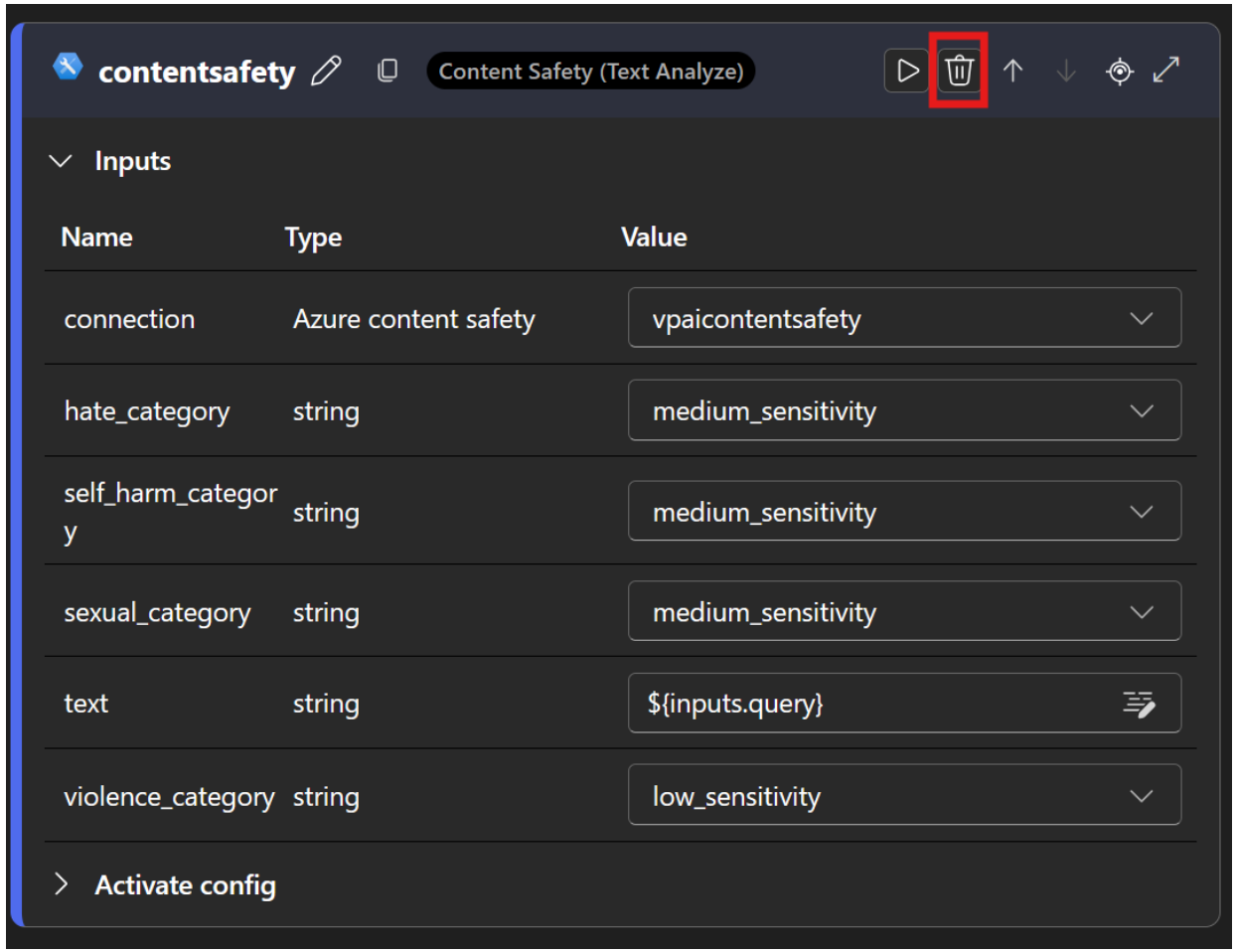
16. Set the **acs_index_connection** and set it to the **aiserach-microretail** index. Ensure the remaining fields match the screenshot below and select **Save**.



3-1. Create a chat flow with chat history

Step 01: Create prompt flow using Playground as a template

17. Find the **contentsafety** node at the bottom and delete it for now. We'll add this in a later step.



Prompt Flow Types: Standard vs. Chat

Feature	Standard Flow	Chat Flow
Purpose	General-purpose LLM applications	Designed for chat-based applications
Input/Output	Single-turn input/output	Multi-turn with chat_history, chat_input, and chat_output
UI Experience	Traditional form-based input	Chat-like interface for development and testing
Use Case Examples	Document summarization, classification, Q&A	Virtual assistants, customer support bots, multi-turn Q&A
Debugging	Standard logs and outputs	Simulates real-time chat interactions for debugging

Prompt flow in Azure AI Foundry portal - Azure AI Foundry | Microsoft Learn

Notes on Prompt Flow in Chat playground

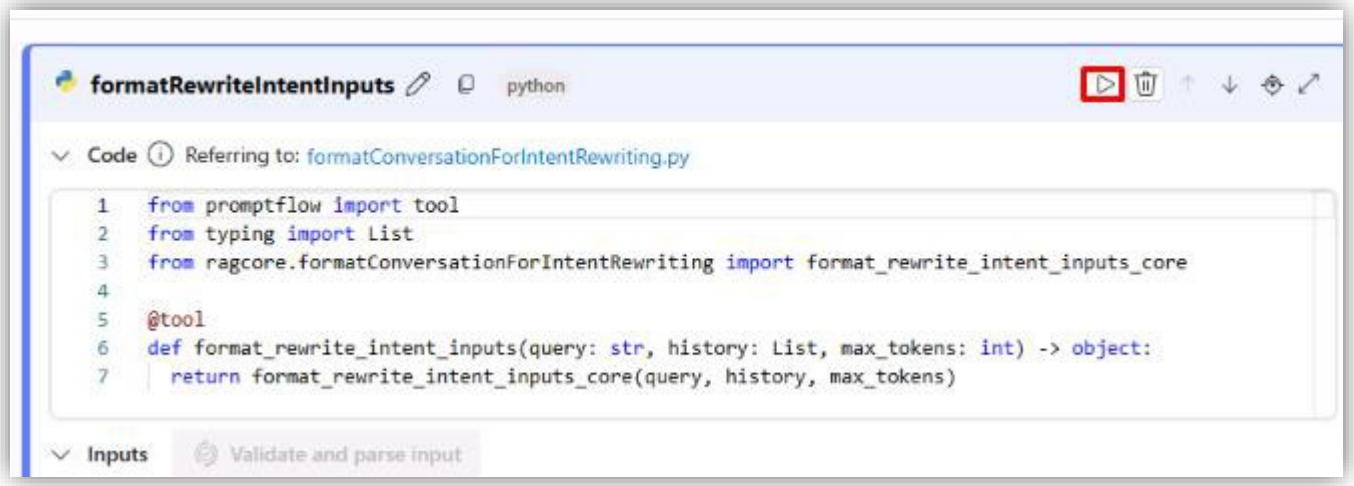
Efficiency

- Deploying a chat flow from the playground using its configuration as a template is quicker and requires less configuration than creating the prompt flow manually.

Two panes

- Flow pane (left) - add and configure new flows with LLMs, prompts, and various Python tools.
- Graph (right) - easily visualize the flow of different nodes.

Location of search service



Test a single flow at a time and see the status of the test in the graph

Testing Prompt Flow

Outputs

- Use the Outputs dropdown menu to test the connection of your flow

As multiple items in a flow link together, ensuring the correct input and output values for each is important.

- Check the outputs of the different flows to get a better idea of how information is being linked and passed between them.

Task 02: Test the chat flow

Introduction:

To ensure optimal performance, **MicroRetail** needs to test its chat flow before deployment. Testing allows the identification of errors, inconsistencies, and areas for improvement, ensuring a seamless user experience..

Description:

In this task, you'll test various elements of the chat flow to ensure it's ready for deployment. This includes verifying input/output accuracy, response latency, and conversation coherence.

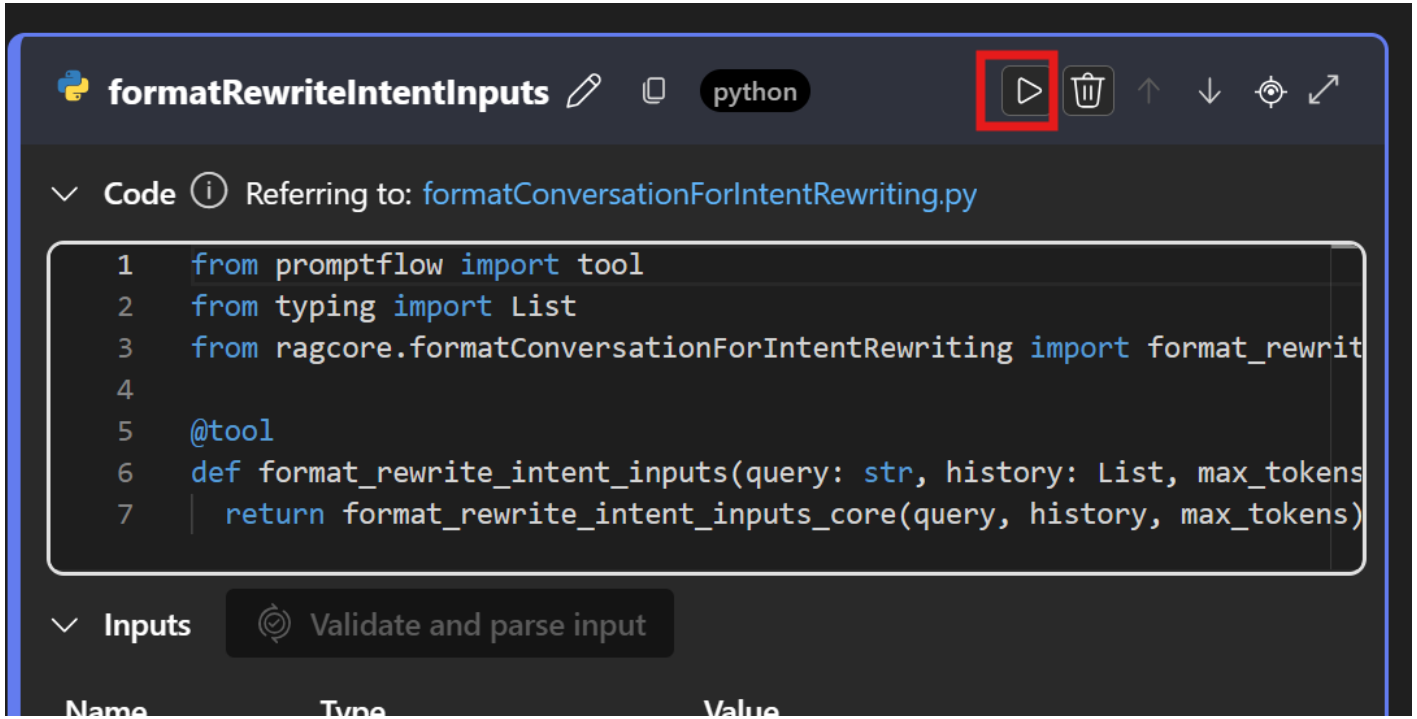
Success Criteria:

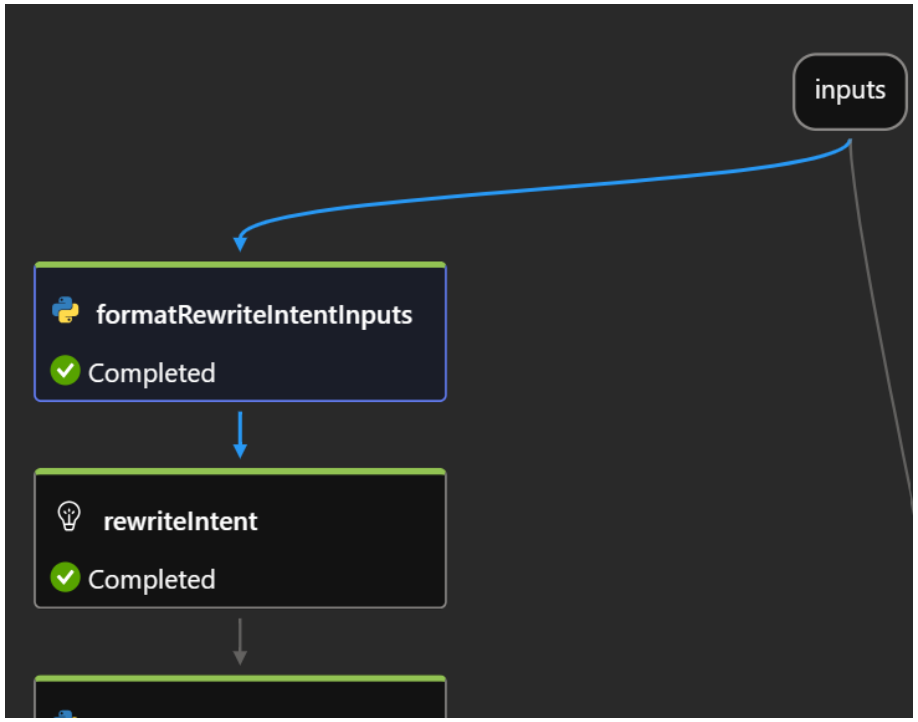
- The chat flow is working properly.

3-2. Test the chat flow

Step 01: Test the individual items using the flow pane

1. Once the compute session is running, scroll to the first item in the flow pane on the left, **formatRewriteIntentInputs** and select the play button in the upper right.



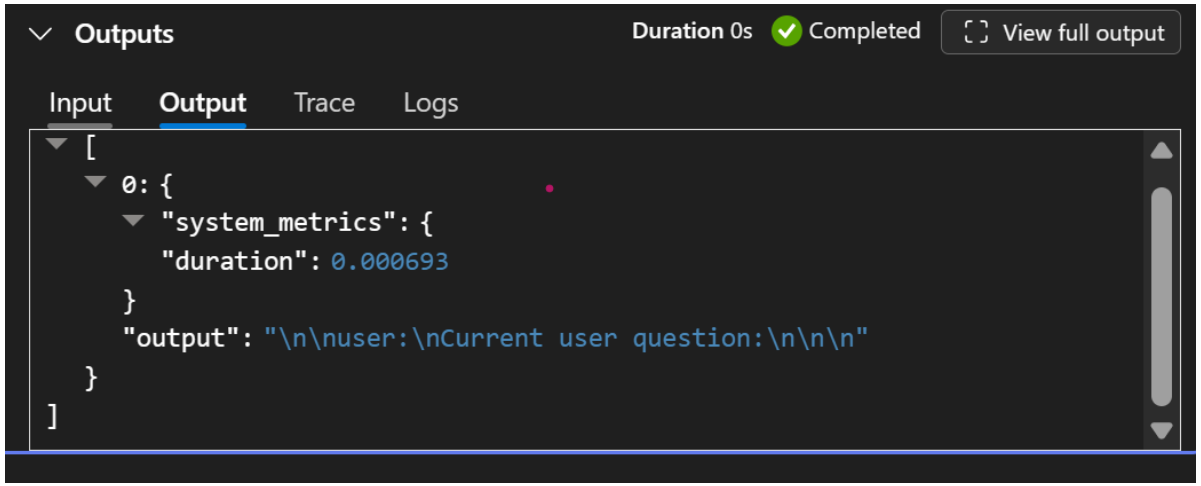


This will allow you to test a single flow at a time. This is useful when building a new flow and connecting the individual elements. The graph to the right will also show the status of the test.

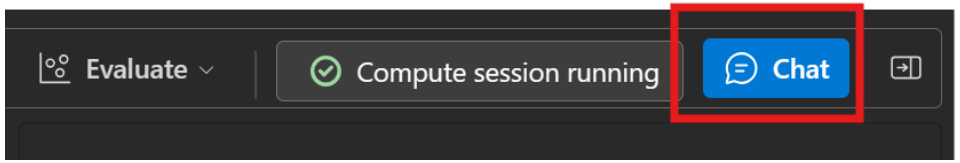
3-2. Test the chat flow

Step 01: Test the individual items using the flow pane

2. Once the test has completed, select the **Outputs** dropdown at the bottom of the **formatRewriteIntentInputs** flow.



3. To test the entire flow, select the **Chat** button in the upper right. This will bring up a chat window window over the flow visualization graph



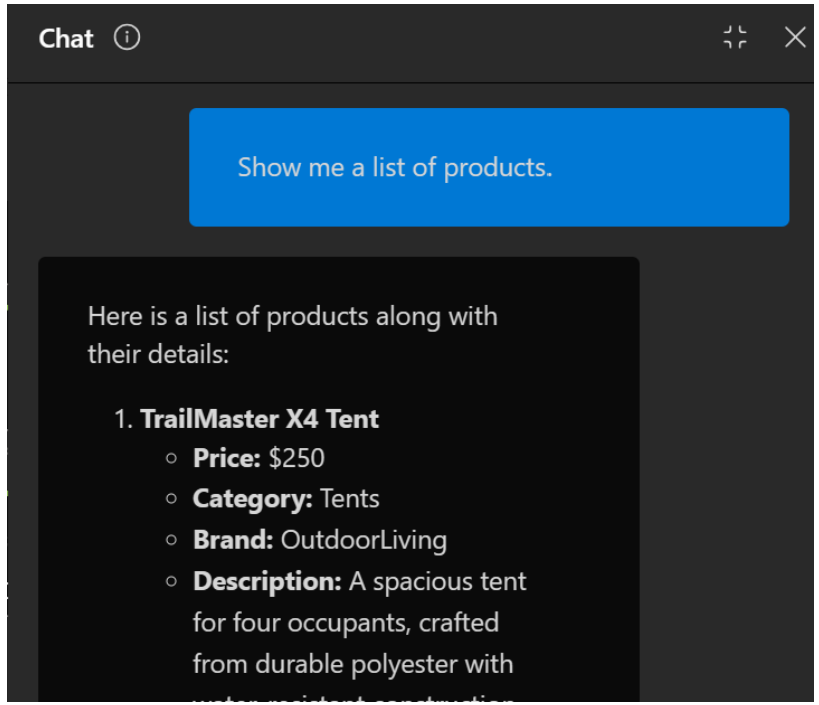
3-2. Test the chat flow

Step 01: Test the individual items using the flow pane

4. In the chat, run the following query to test the overall functionality of the flow:

```
Show me a list of products.
```

The response should show a list of products, like in the playground.



Each of the items in the flow pane will now have output data relevant to an actual test. You can check the outputs of the different flows to get a better idea of how information is being linked and passed between them.

5. Close the chat window with the **X** in the upper right to return to the flow visualization graph. Note that all flow items show as **Completed** as a result of running the entire flow.

Task 03: Evaluate the chat flow

Introduction:

Evaluating chatbot interactions ensures that responses align with **MicroRetail's** customer service goals. Using predefined criteria, Adatum can measure and refine chatbot accuracy, relevance, and coherence.

Description:

In this task, you'll assess chatbot performance using key evaluation metrics. By setting up automatic evaluation for Groundedness, Relevance, Coherence, Fluency, and Similarity, **MicroRetail** can ensure that the chatbot maintains high-quality interactions and meets user expectations.

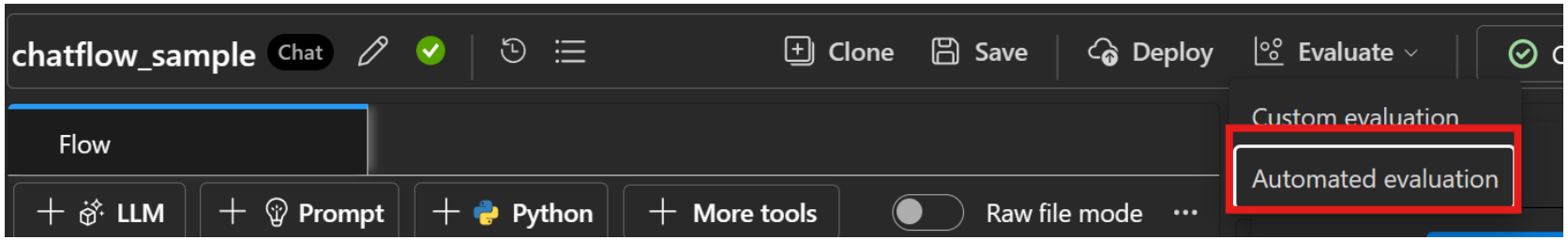
Success Criteria:

- The flow evaluation is completed successfully.

3-3. Evaluate the chat flow

Step 01: Set up the automatic evaluation for Groundedness, Relevance, Coherence, Fluency, and Similarity.

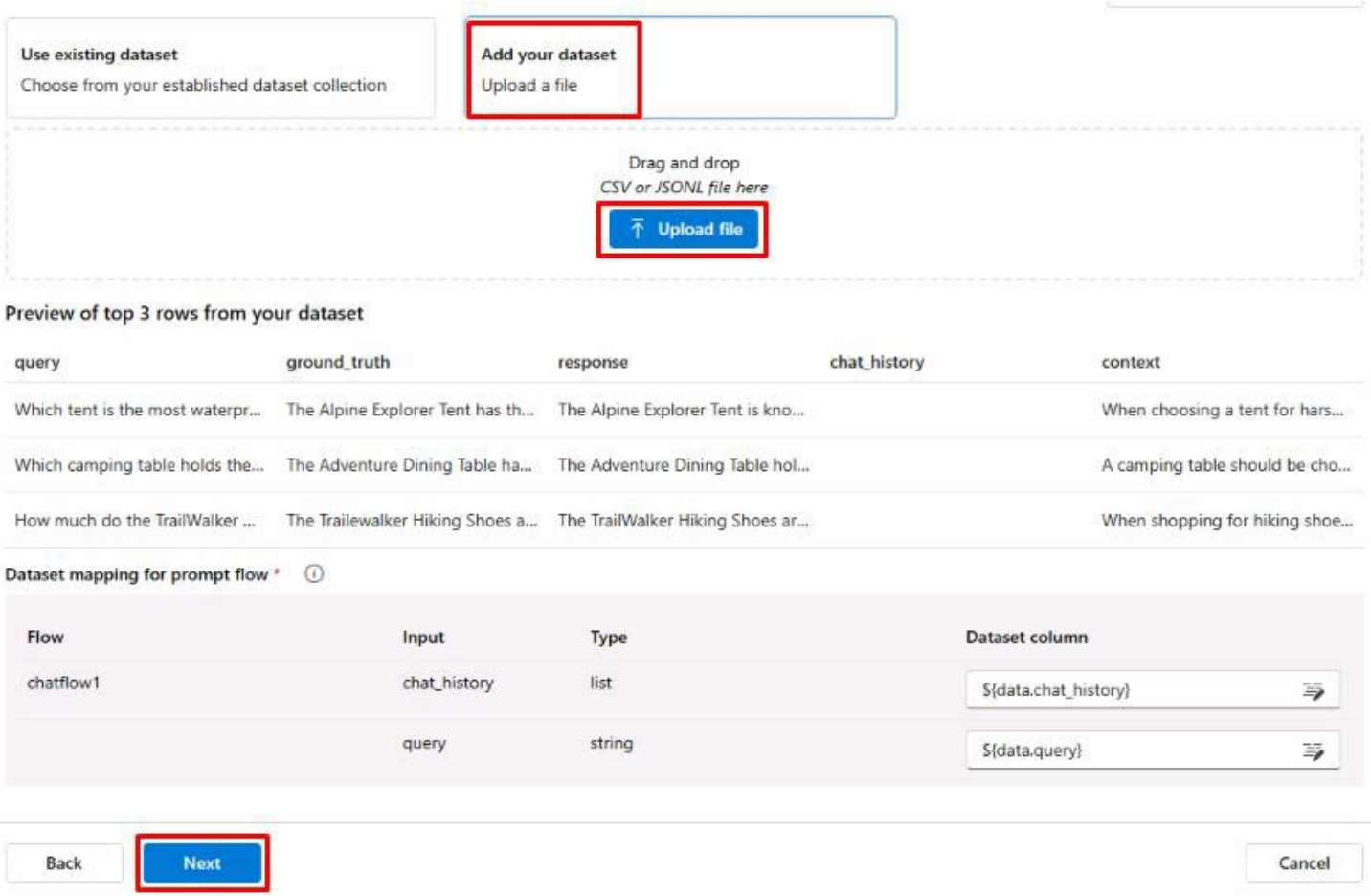
1. Download the **eval.jsonl** file HERE.
2. On the **chatflow_sample** page, select **Evaluate** and then select **Automated evaluation**.



3-3. Evaluate the chat flow

Step 01: Set up the automatic evaluation for Groundedness, Relevance, Coherence, Fluency, and Similarity.

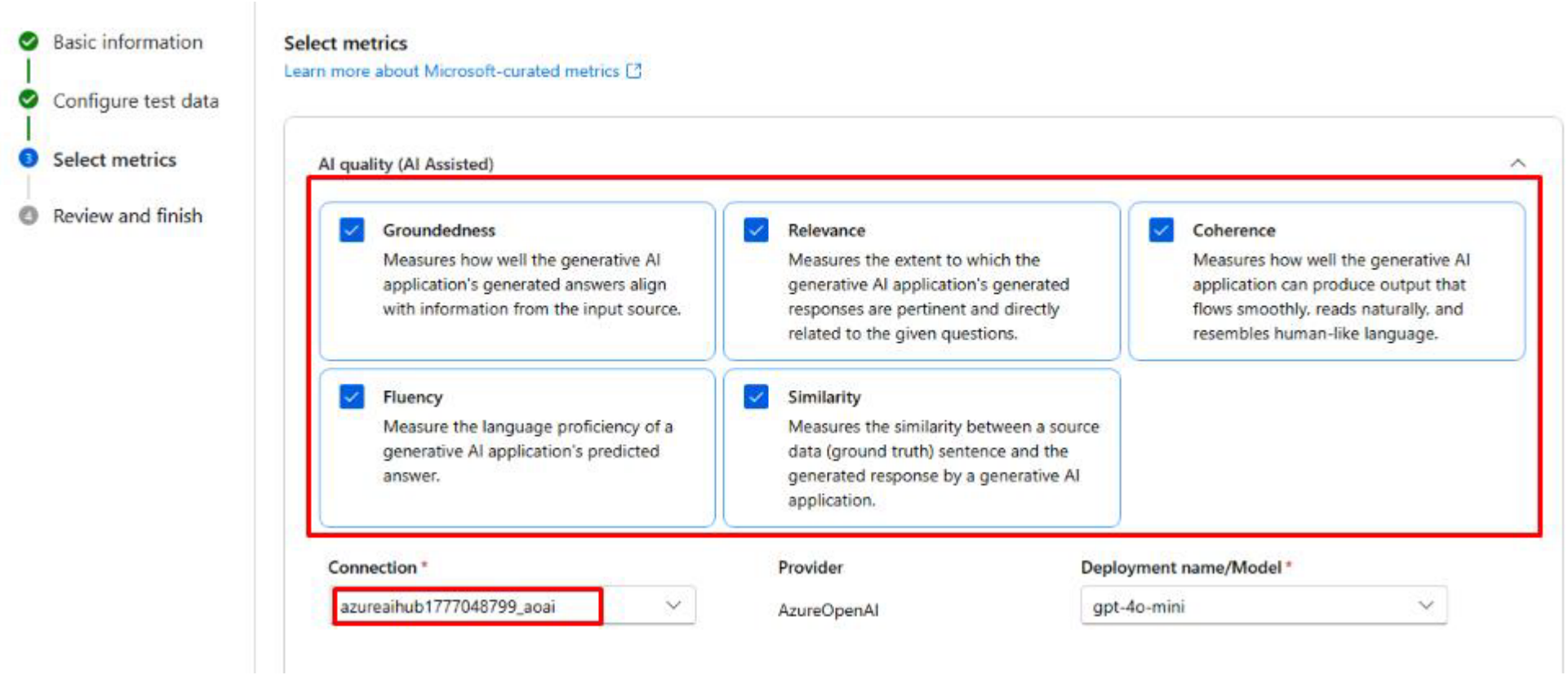
3. On the **Basic information** tab, set the **Evaluation name** to **eval_aiworkshop** and select **Next**.
4. On the **Configure test data** tab, select **+ Add your dataset**.
5. Select **Upload file** and select the **eval.jsonl** file.
6. Once the data loads, ensure the **chat_history** value is set to **${data.chat_history}** and the **query** value is set to **${data.query}**, then select **Next**.



3-3. Evaluate the chat flow

Step 01: Set up the automatic evaluation for Groundedness, Relevance, Coherence, Fluency, and Similarity.

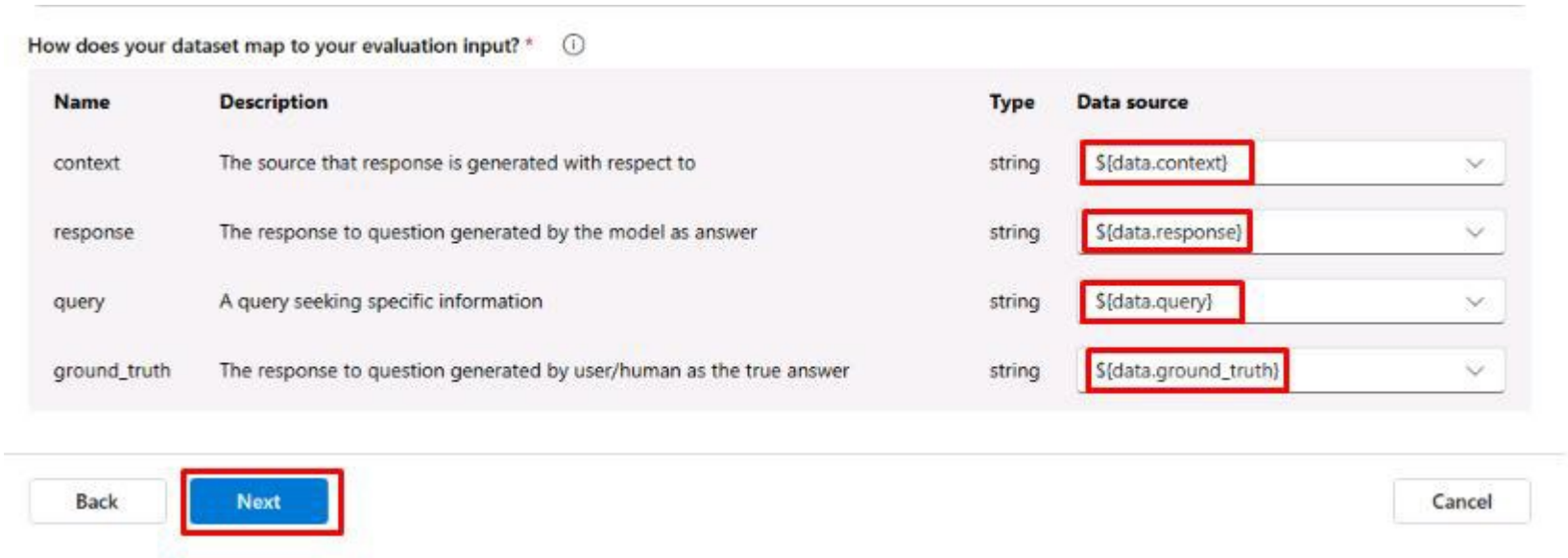
7. On the **Select metrics** tab, select the checkboxes for **Groundedness**, **Relevance**, **Coherence**, **Fluency**, and **Similarity**.
8. Select your connection from the **Connection** dropdown menu. The **gpt-4o-mini** model should be automatically selected.



3-3. Evaluate the chat flow

Step 01: Set up the automatic evaluation for Groundedness, Relevance, Coherence, Fluency, and Similarity.

9. Scroll to the bottom and ensure the data mapping is correct, then select **Next**:



3-3. Evaluate the chat flow

Step 01: Set up the automatic evaluation for Groundedness, Relevance, Coherence, Fluency, and Similarity.

10. Select **Submit** and wait for the evaluation to finish.

Once the evaluation is finished, you'll see scores for the metrics that were selected. These scores are based on the chat flow's response to the test data. You can scroll down to see more detailed information on the metric scores and the reasoning behind those scores.

Models and metrics

Models

- Advanced models are normally used for evaluation.
- However for this evaluation, we'll use the compact gpt-4o-mini model that our chat flow uses

Metrics

- Scores based on chat flow's response to test data.
- More detail is available on the metric scores and the reasoning behind them.

Task 04: Introduction to Azure AI Agent Service

Introduction:

Azure is introducing the Azure AI Agent Service to help developers create secure, stateful AI agents that automate complex business processes. This service integrates models, tools, and data sources, making it easier for businesses to deploy autonomous agents for a range of tasks like scheduling, market research, customer service, and code management.

The service leverages a wide array of tools, including Azure Logic Apps, Azure Functions, OpenAPI 3.0, and more. It also provides secure data grounding via Bing, SharePoint, and Azure Blob, ensuring agents have accurate, up-to-date information. With multi-modal support and a flexible model selection, including OpenAI and other top-tier providers, developers can optimize their agents for specific tasks.

Azure AI Agent Service is built for enterprise use, offering features like secure data handling, compliance, and monitoring. It's designed for scalability and reliability, with built-in performance tracking and content filters to ensure safe, high-quality outputs. Additionally, the service supports multi-agent orchestration for more complex workflows.

This platform is designed to simplify agent creation, allowing businesses to rapidly develop, deploy, and manage AI agents to improve efficiency and productivity.

Azure AI Foundry – Agent Service

Built-in enterprise readiness

BYO-file storage

BYO-search index

OBO Authorization Support

Enhanced Observability

Extensive Ecosystem of Tools

Knowledge

- Microsoft Fabric
- SharePoint
- Bing Search
- Azure AI Search
- Your own licensed data
- Files (local or Azure Blob)
- File Search
- Code Interpreter

Actions

- Azure Logic Apps
- OpenAPI 3.0 Specified Tools
- Azure Functions

Model Catalog

Azure OpenAI Service
(GPT-4o, GPT-4o mini)

Models-as-a-Service

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

Task 04: Introduction to Azure AI Agent Service

Description:

In this task, you'll learn how to set up and deploy the Azure AI Agent Service. By following the steps, you will understand how to integrate various tools and data sources to create secure, stateful AI agents that can automate complex business processes.

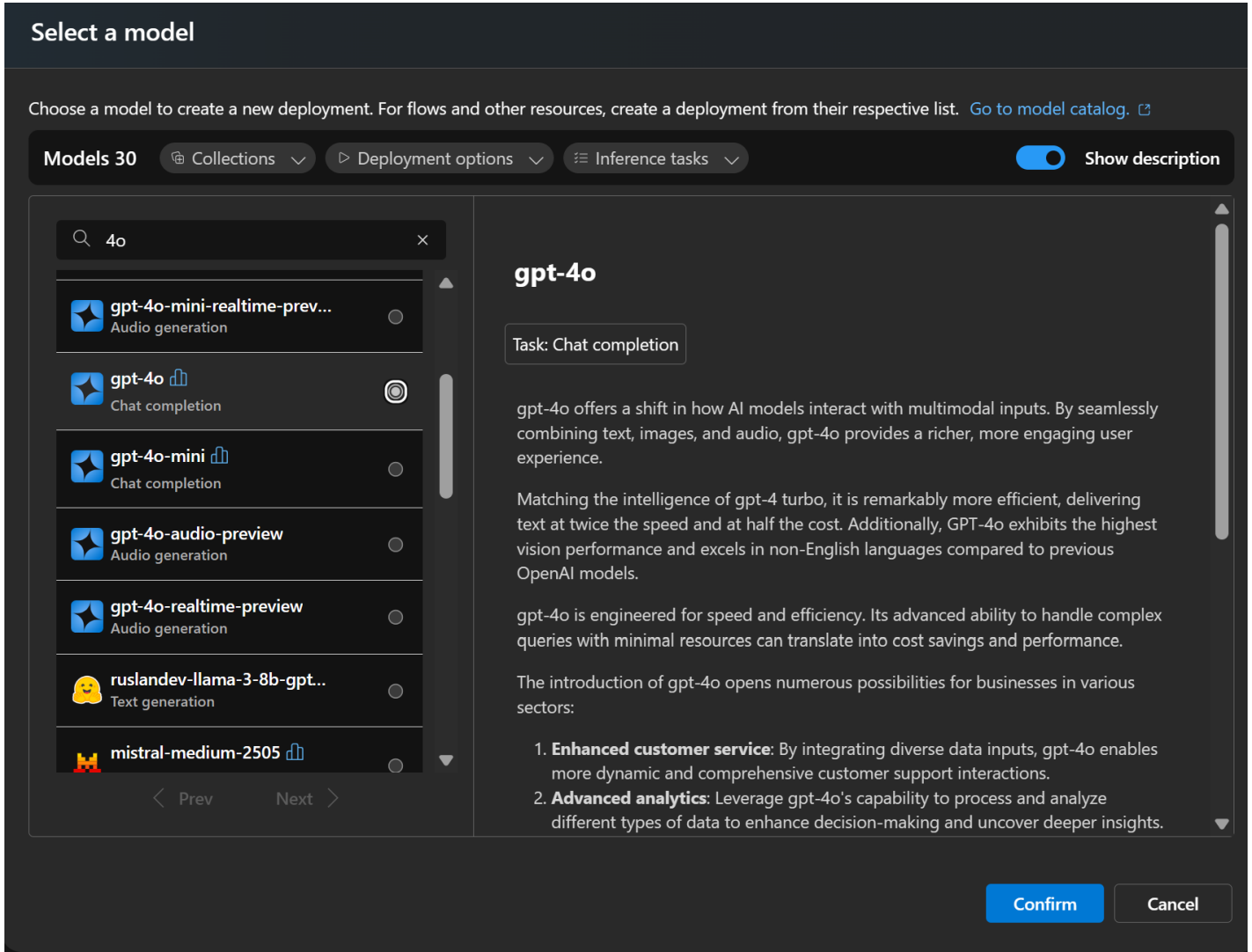
Success Criteria:

- The Azure AI Agent Service is set up and deployed successfully.
- The AI agents created are able to automate tasks like scheduling, market research, customer service, and code management.

3-4. Introduction to Azure AI Agent Service

Step 01: Set up Azure AI Agent Service

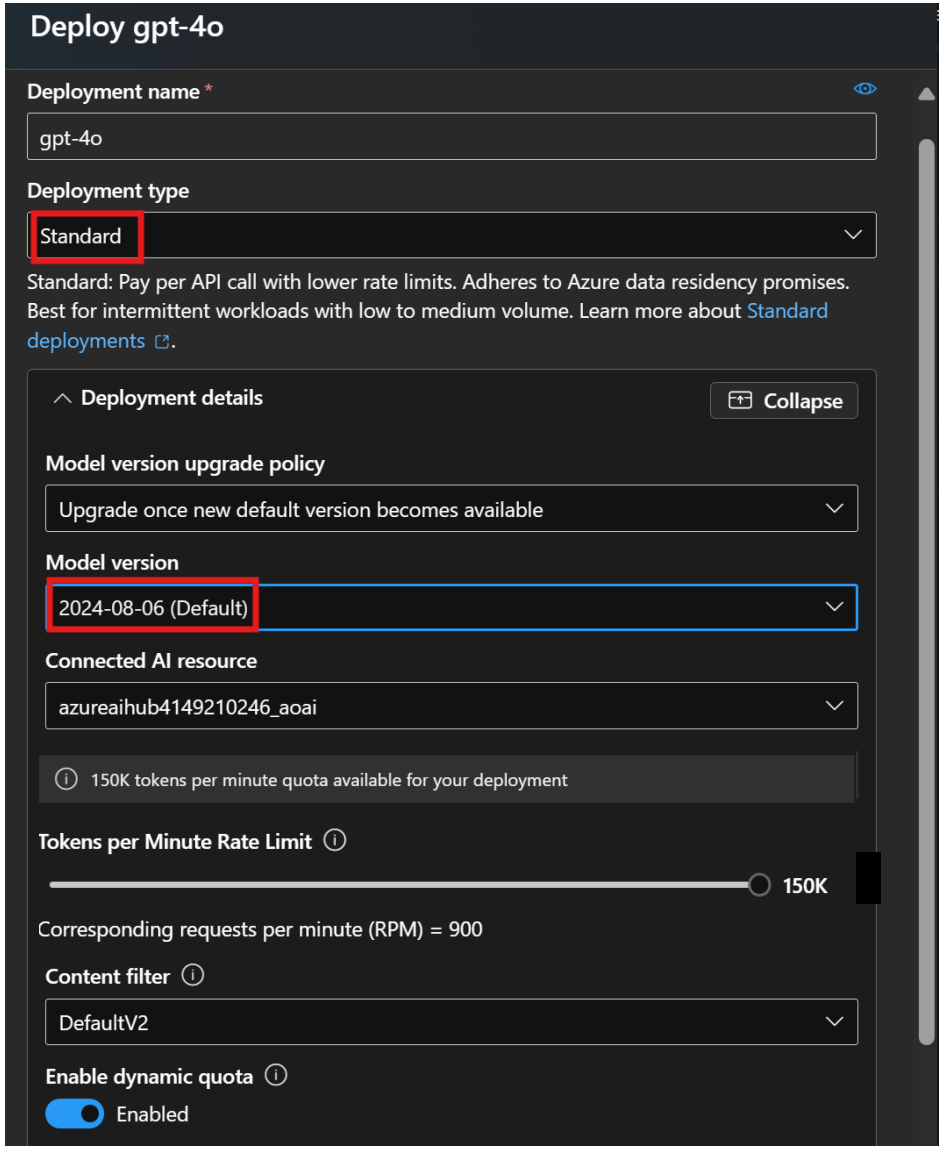
1. From the **project_AIWorkshop** page, select **Models + endpoints** from the left menu.
2. Select **+ Deploy model**, then select **Deploy base model**.
3. Select the **gpt-4o** model, then select **Confirm**.



3-4. Introduction to Azure AI Agent Service

Step 01: Set up Azure AI Agent Service

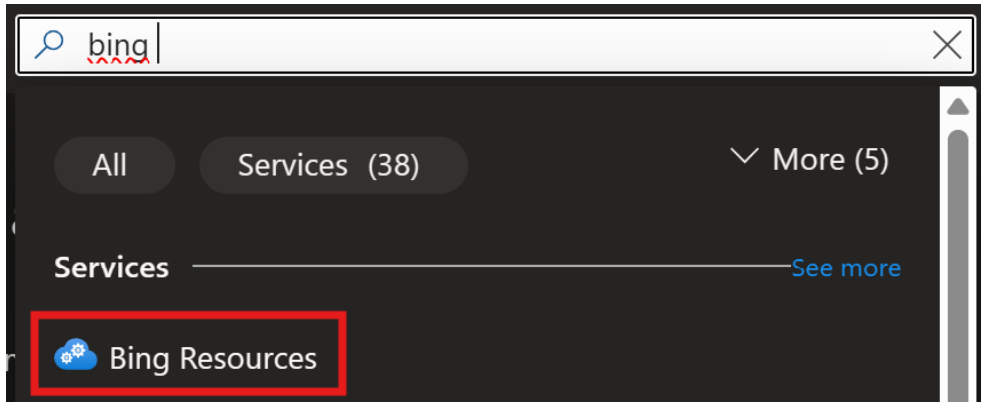
4. Set the **Deployment type** to **Standard**, set the **Tokens per Minute Rate Limit** to **150K**, then select **Deploy**.



3-4. Introduction to Azure AI Agent Service

Step 01: Set up Azure AI Agent Service

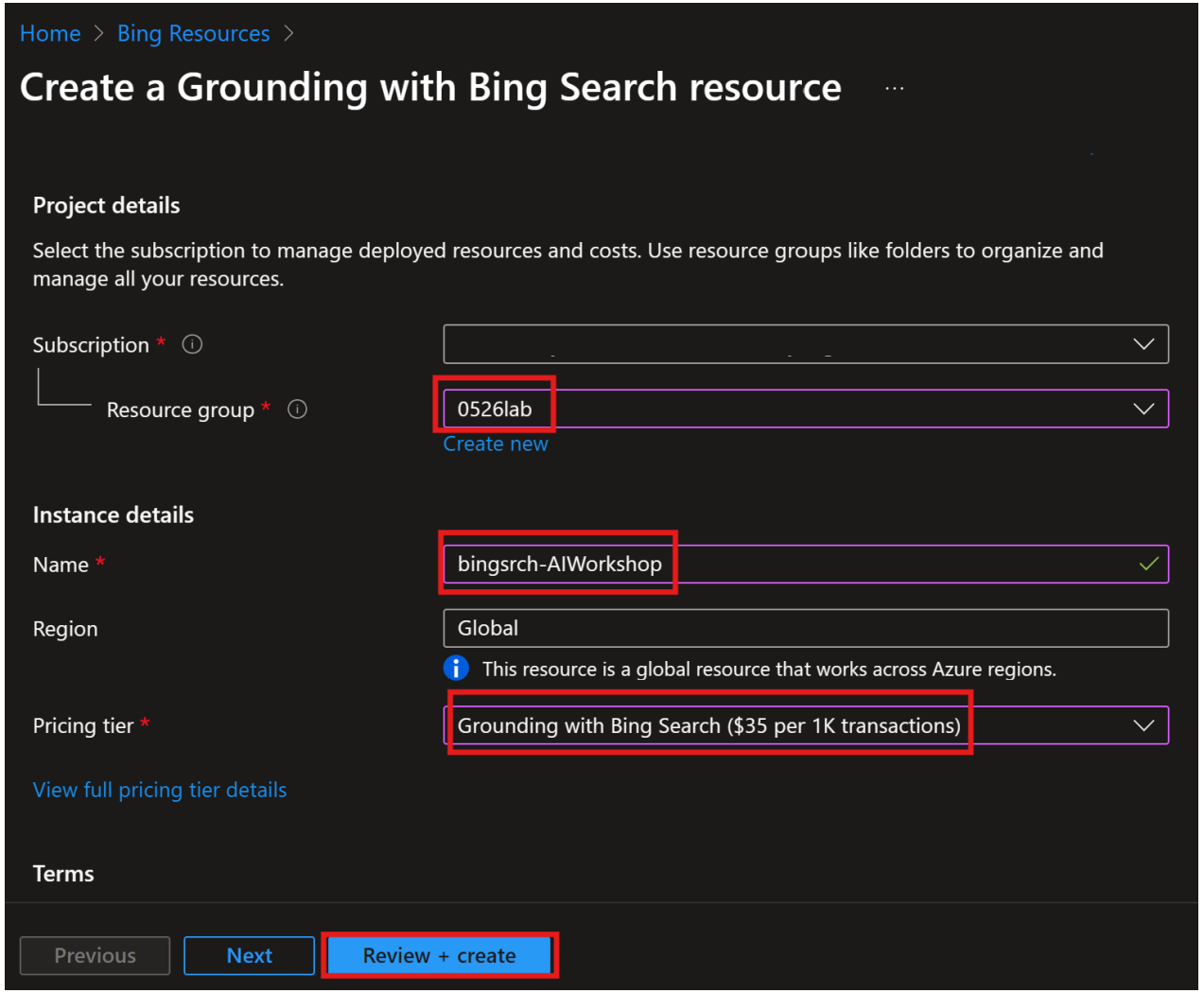
5. Switch to the tab with Azure portal.
6. In the search bar at the top, search for **bing** and then select **Bing Resources**.



3-4. Introduction to Azure AI Agent Service

Step 01: Set up Azure AI Agent Service

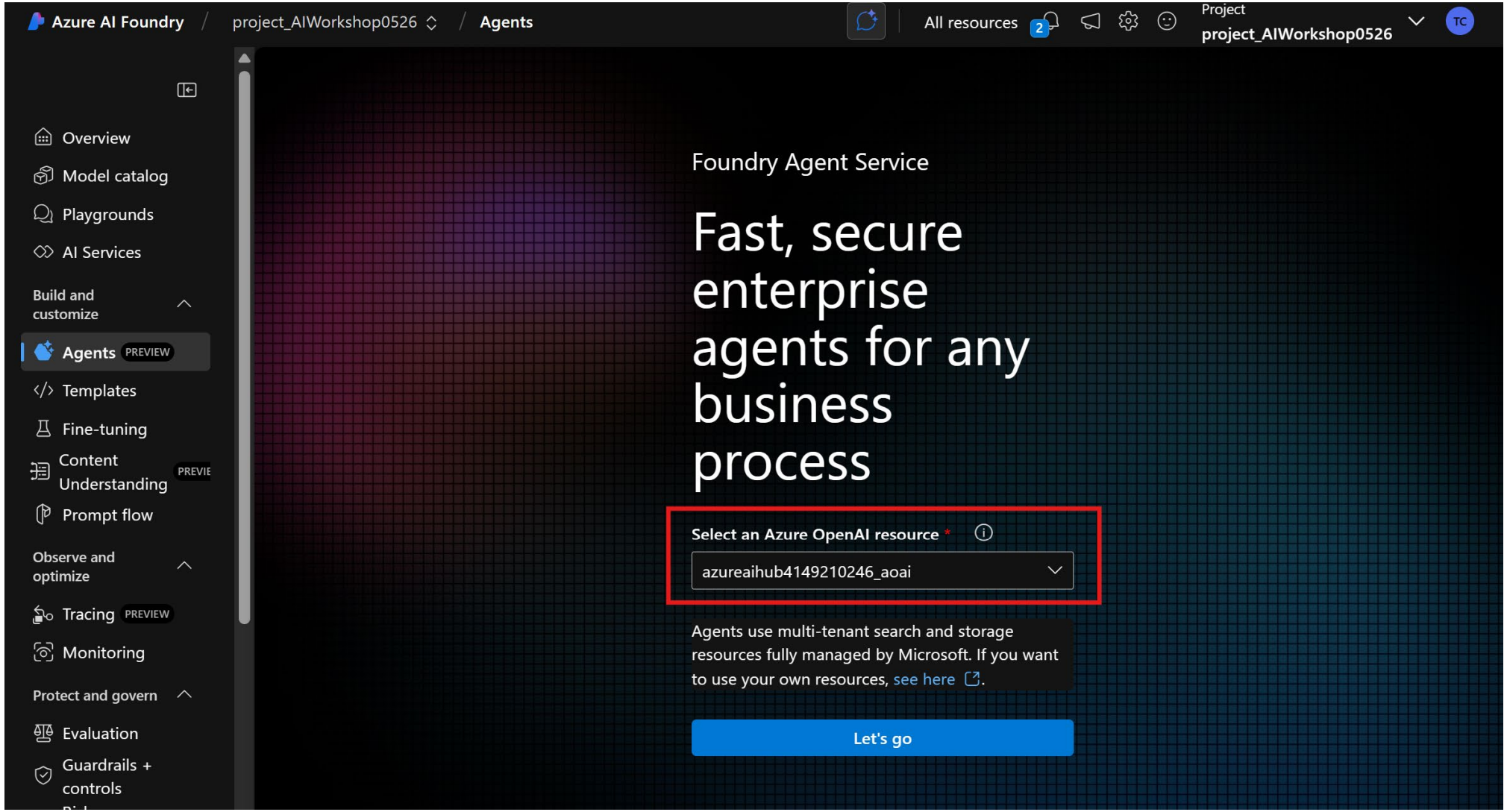
7. From the **Bing Resources** page, select **+ Add**, then select **+ Grounding with Bing Search**.
8. On the **Create a Grounding with Bing Search resource** page, select your resource group and pricing tier. Give it a name of **bingsrch-AIWorkshop** and select **Review + Create**, then select **Create**.



3-4. Introduction to Azure AI Agent Service

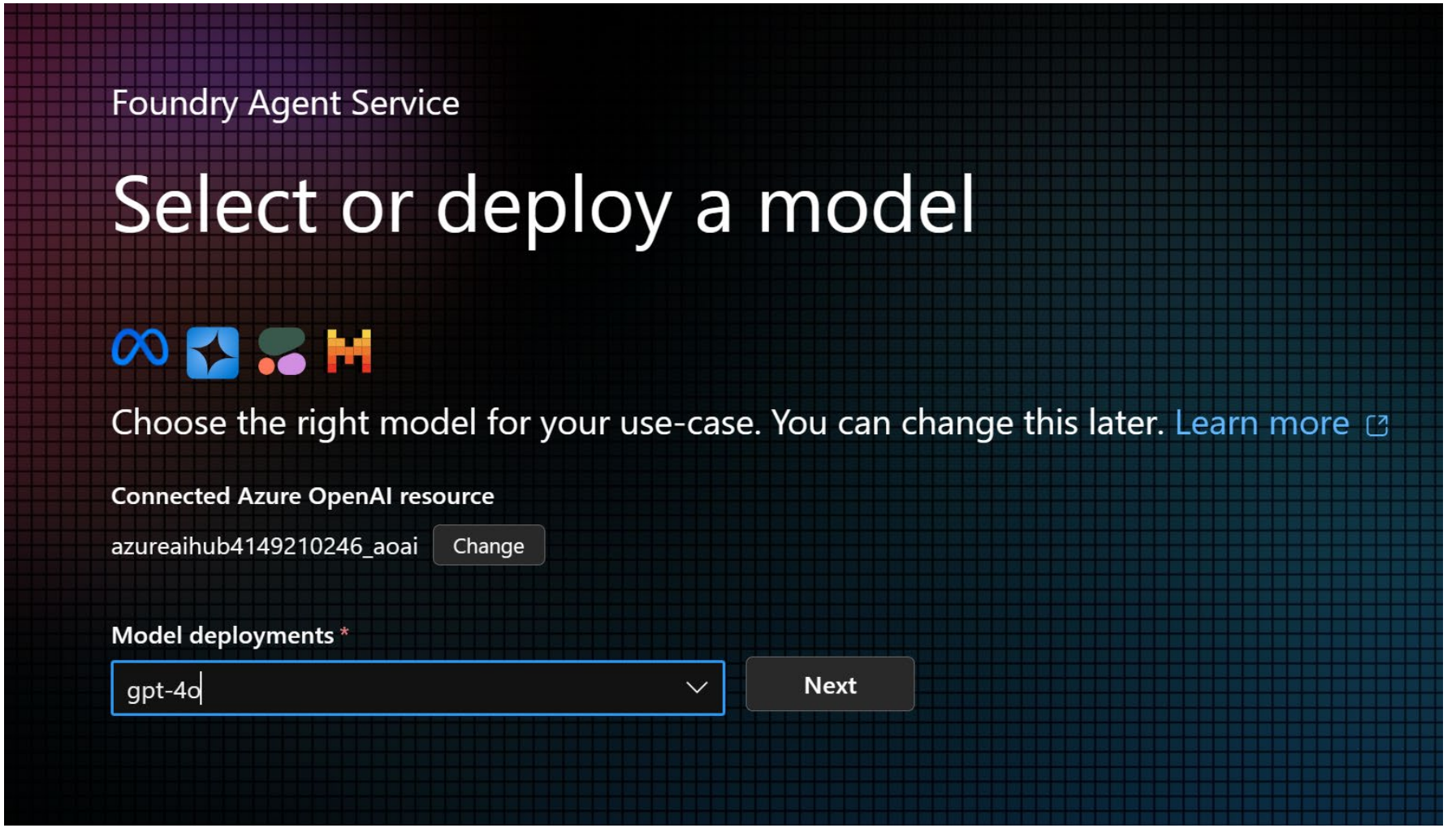
Step 01: Set up Azure AI Agent Service

9. Return to the tab with the Azure AI model deployment and select **Agents** from the left menu.
10. Under **Select an Azure OpenAI Service resource**, select your hub and select **Let's go**.



Step 01: Set up Azure AI Agent Service

11. Under **Select or deploy a model**, select **gpt-4o** then select **Next**.



Step 01: Set up Azure AI Agent Service

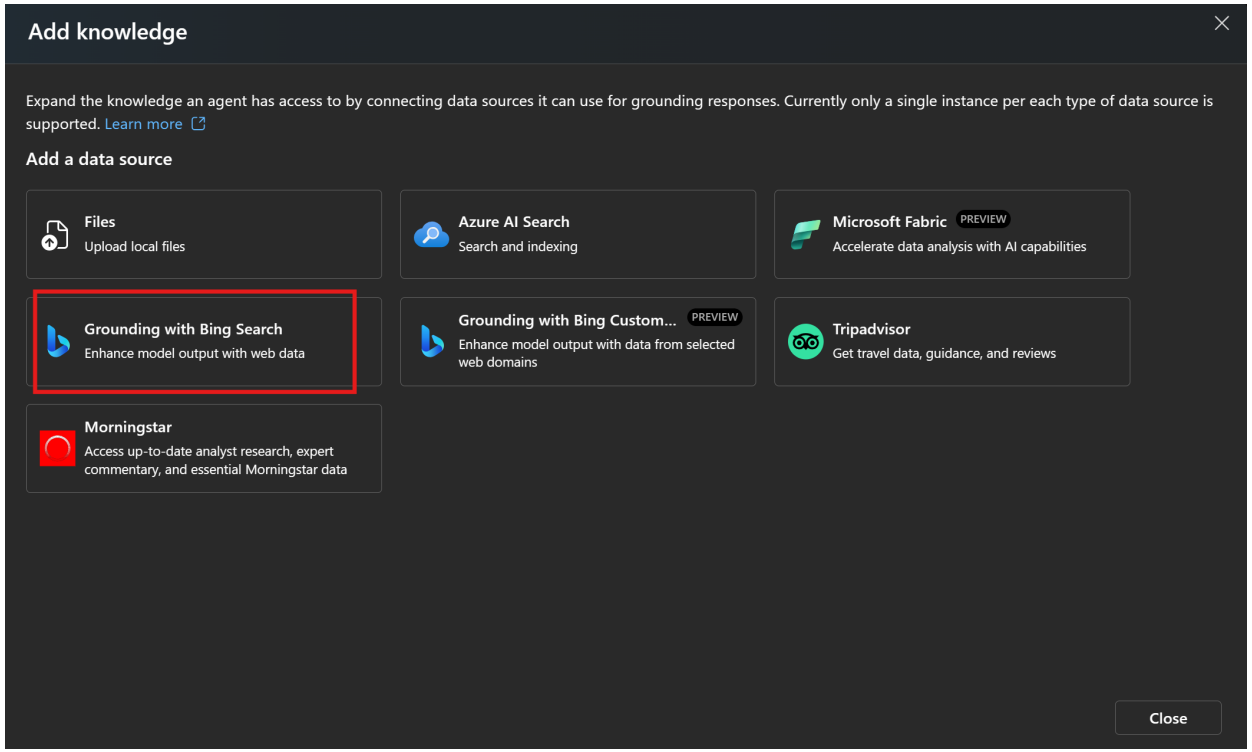
12. Select the agent to open the **Setup** pane.
13. In the **Instructions** field of the **Setup** pane, enter the following:

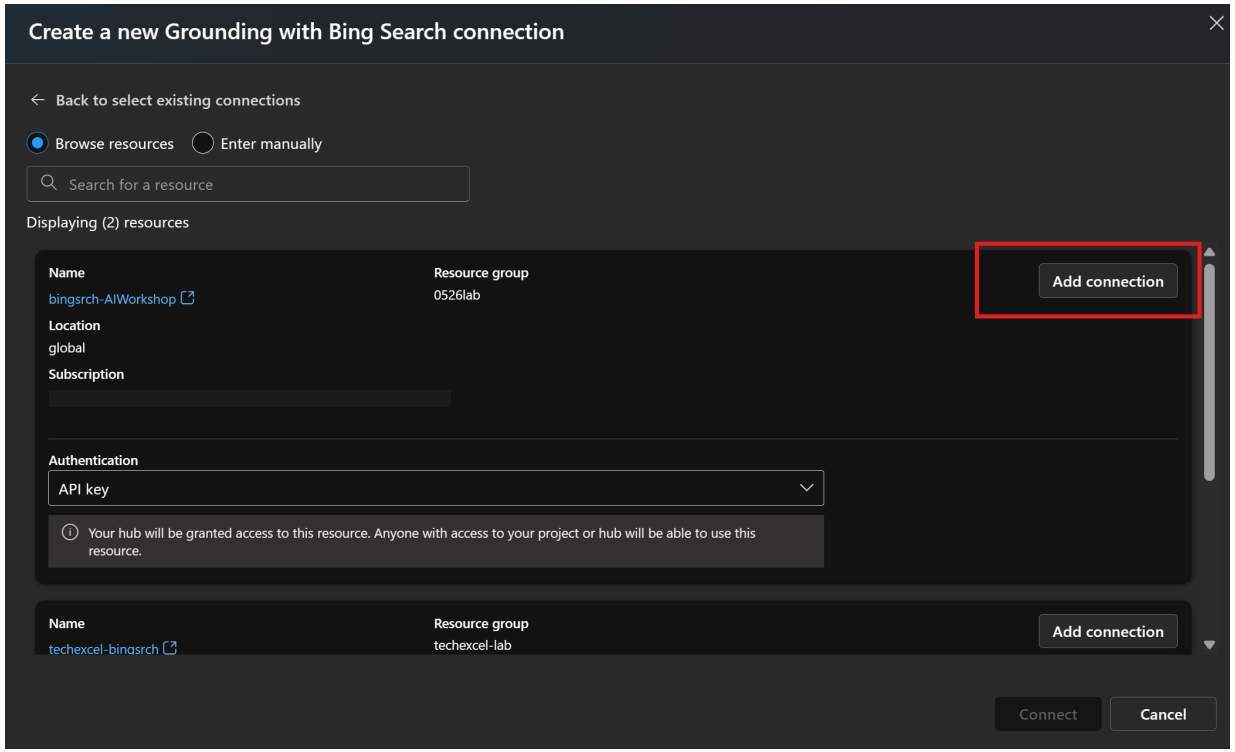
```
Understand User Query:
Analyze the user's query to identify if it requires real-time information (e.g.,
weather, date, news).

Use Bing Search Tool for Real-Time Data:
If the query involves up-to-date information, use the Bing Search tool to retrieve
relevant data.

Craft a Clear, Concise Response:
Extract the relevant information (e.g., temperature, news) and provide the answer in
a simple and direct way.

Ask for Clarification if Needed:
If the query is vague or missing details (e.g., location for weather), ask the user
for more information.
```

3-4. Introduction to Azure AI Agent Service

Step 01: Set up Azure AI Agent Service

14. Under **Knowledge** in the **Setup** pane, select **+ Add**, then select **Grounding with Bing Search**.
15. Select the **+ Create connection** connection, then select **Add connection** next to the **bingsrch** resource.



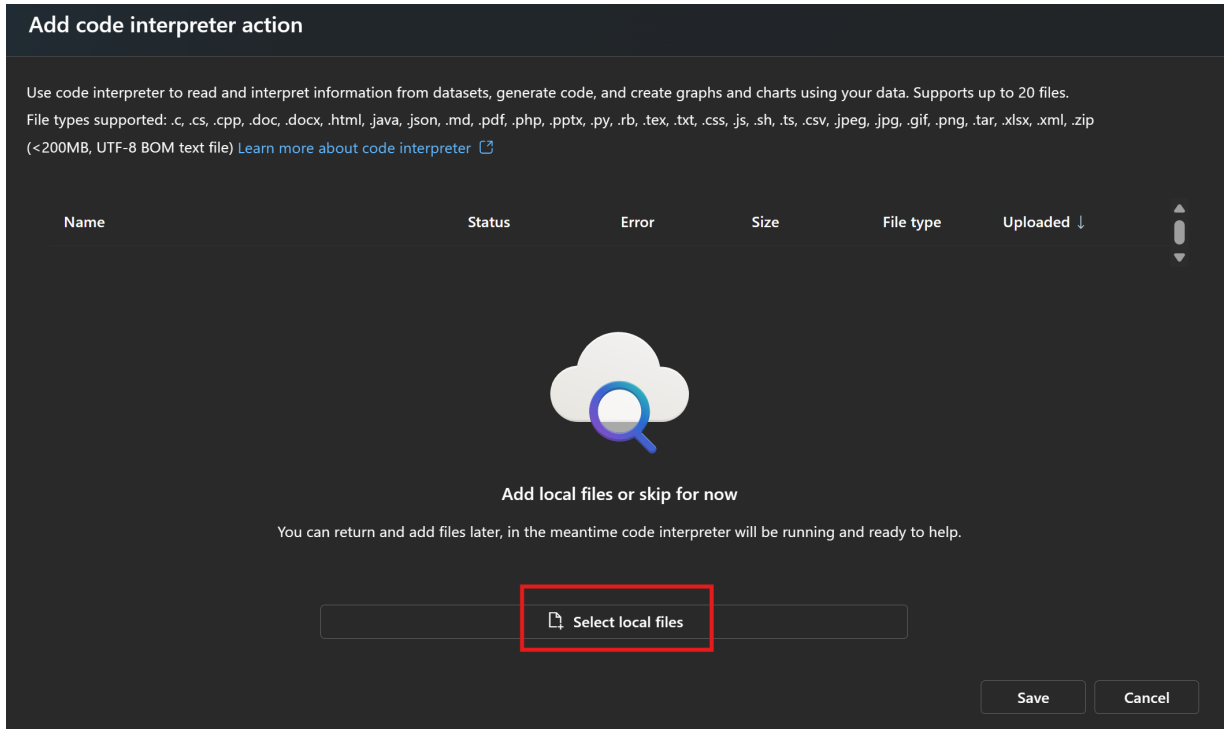


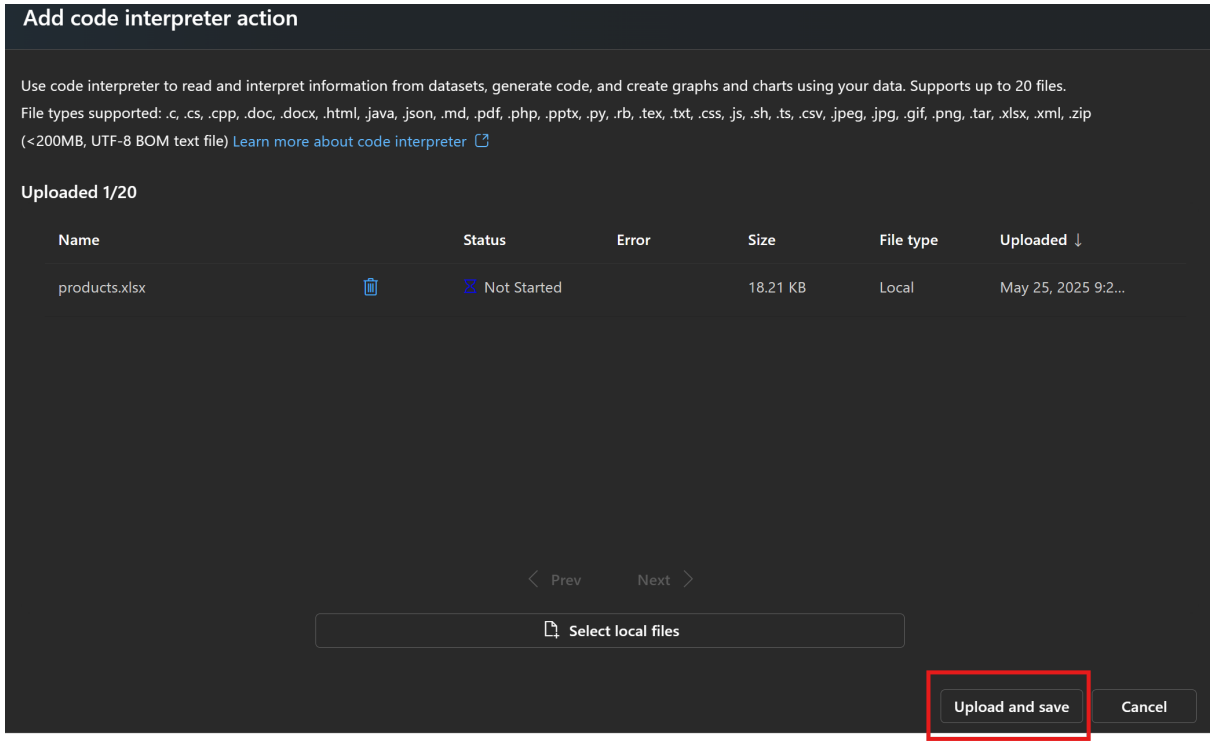
The **Knowledge** section allows you to specify the source of information for the agent. In this case, we're using the **Grounding with Bing Search** service to retrieve up-to-date information from external sources, beyond our local dataset. You can also set the source to a pre-existing search index or local data. If you want the agent to be able to retrieve both local and web-based results, you can add a connection for each.

3-4. Introduction to Azure AI Agent Service

Step 01: Set up Azure AI Agent Service

16. Under **Actions** in the **Setup** pane, select **+ Add**, then select **Code interpreter**.
17. On the **Add code interpreter action** page, select **Select local files** and then select the **products.xlsx** file created earlier.
18. Select **Upload and Save**.



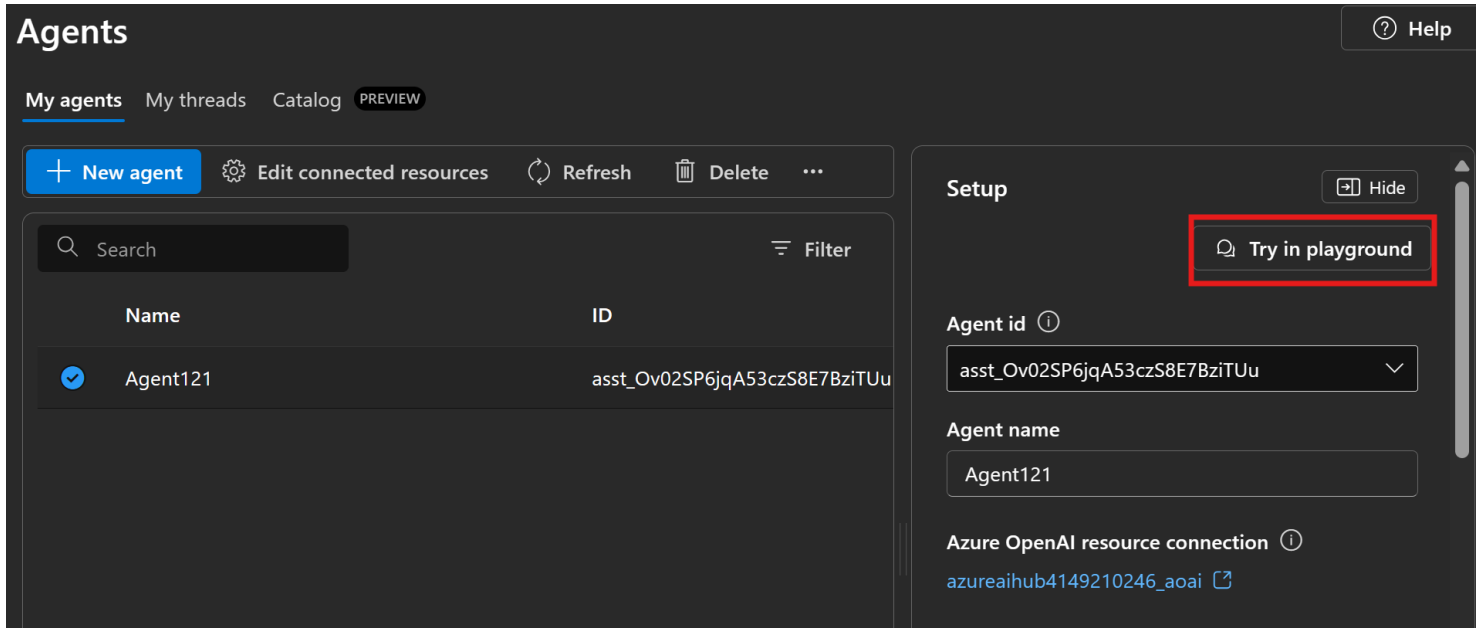


The **Actions** section allows you to specify additional tasks for the agent beyond simple data retrieval. The **Code interpreter** tool can be used for tasks like performing calculations or creating visualizations from your data.

3-4. Introduction to Azure AI Agent Service

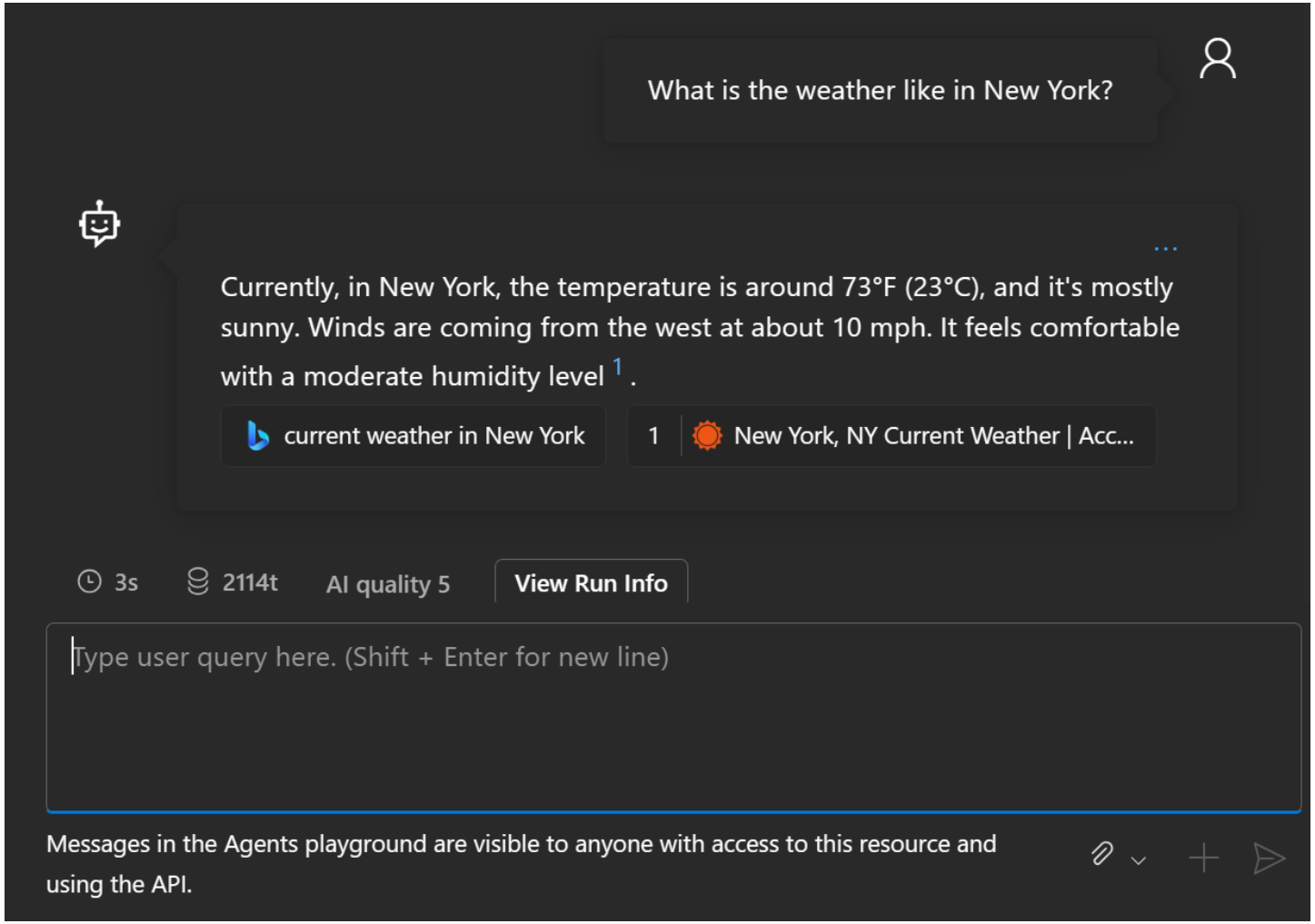
Step 01: Set up Azure AI Agent Service

19. From the upper right of the **Setup** pane, select **Try in playground**.



Step 01: Set up Azure AI Agent Service

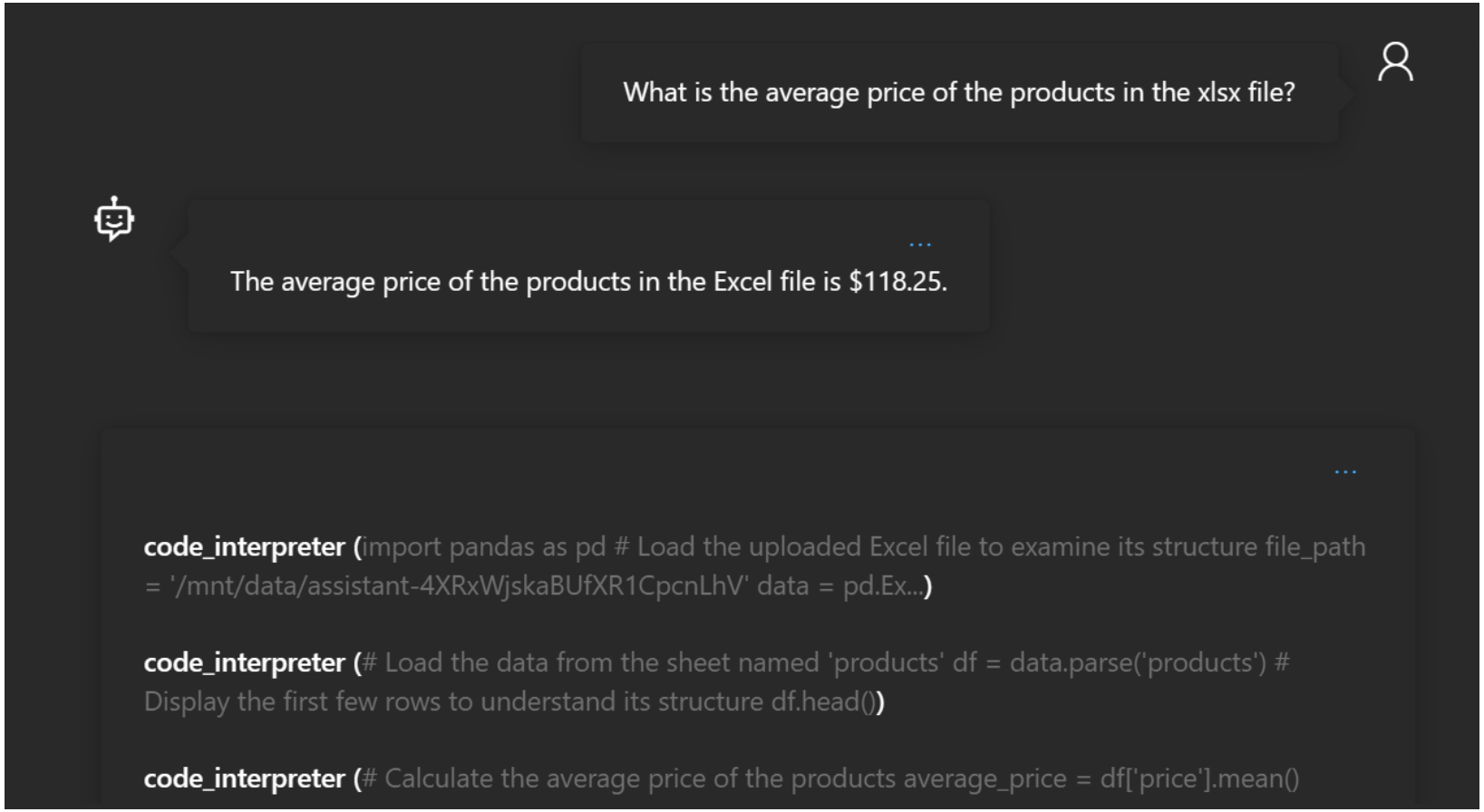
20. In the **Agents playground** chat, enter **What is the weather like in New York?**



The gpt-4o model doesn't have direct access to the current date. However, by using the Bing Search tool, the agent can retrieve up-to-date information for time-sensitive queries.

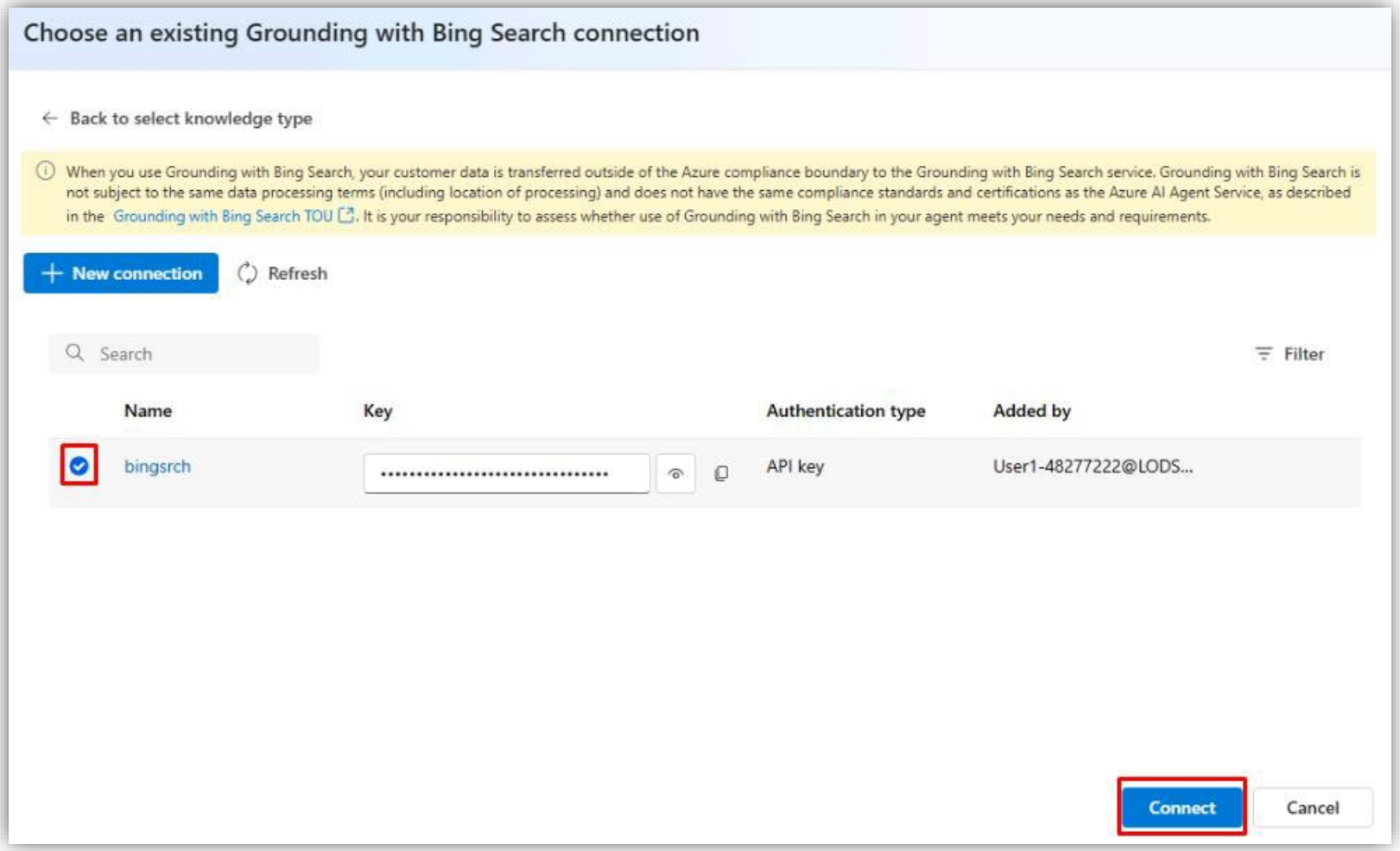
Step 01: Set up Azure AI Agent Service

21. In the chat, enter **What is the average price of the products in the xlsx file?**



The code interpreter tool allows for more complex queries about your data. In this case, we used it to retrieve the average price of the products in our data set.

Knowledge section



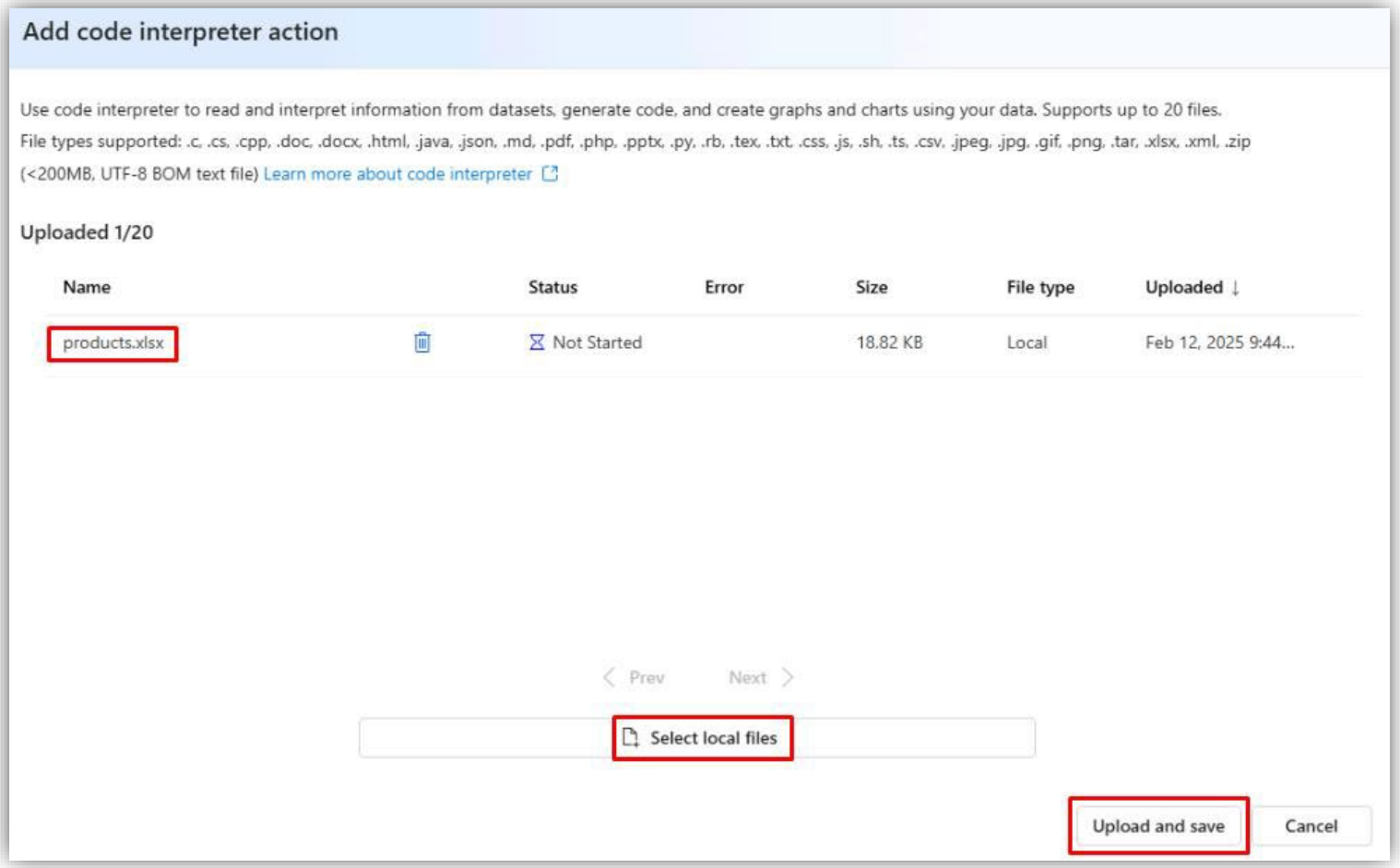
Allows you to specify the source of information for the agent.

You can set the connection to:

- Grounding with Bing Search service to retrieve up-to-date information from external sources
- A pre-existing search index or local data.
- Separate connections for local and web-based results to allow agent to retrieve from both.

Actions section

Allows you to specify additional tasks for the agent beyond simple data retrieval



Code interpreter tool

- Perform calculations
- Create visualizations from your data
- Etc.

Exercise four

Deploy a ChatBot

Introduction: Deploy a Chatbot

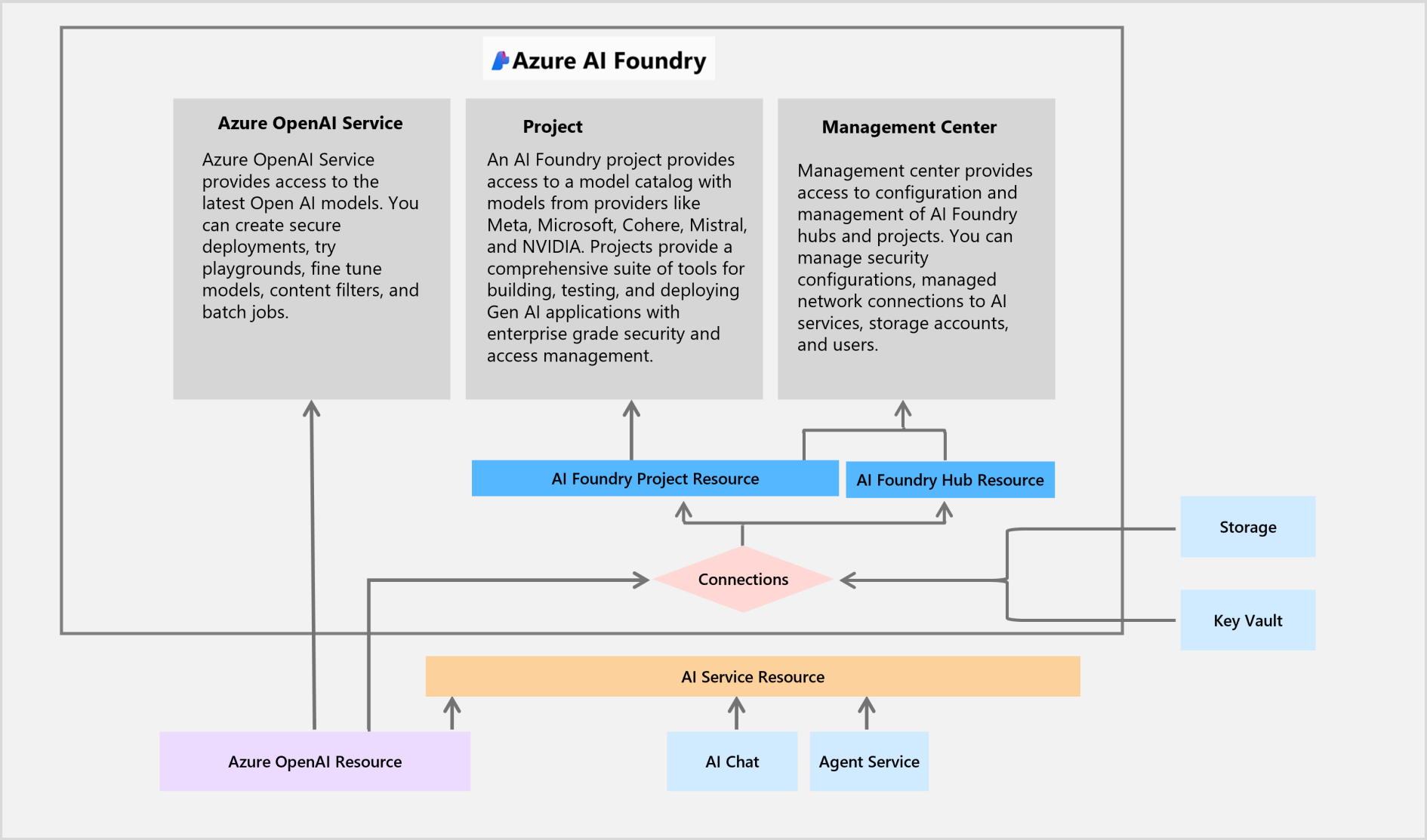
MicroRetail is ready to deploy the new AI chatbot, offering a cohesive and personalized customer service experience.

After completing this exercise, you'll be able to:

Deploy an AI chatbot to a web application.

Validate and test chatbot interactions in a live environment.

Exercise four architecture



Task 01: Deploy the chatbot

Introduction:

Deploying the chatbot enables **MicroRetail** to provide seamless AI-powered customer support on its website and mobile platforms. Ensuring a successful deployment is crucial for enhancing user engagement and optimizing response times.

Description:

In this task, you'll deploy the chatbot using the configured chat flow. This deployment will allow real-time customer interactions, enhancing the overall service experience.

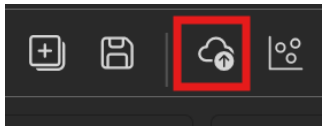
Success Criteria:

- The chatbot has been successfully deployed.

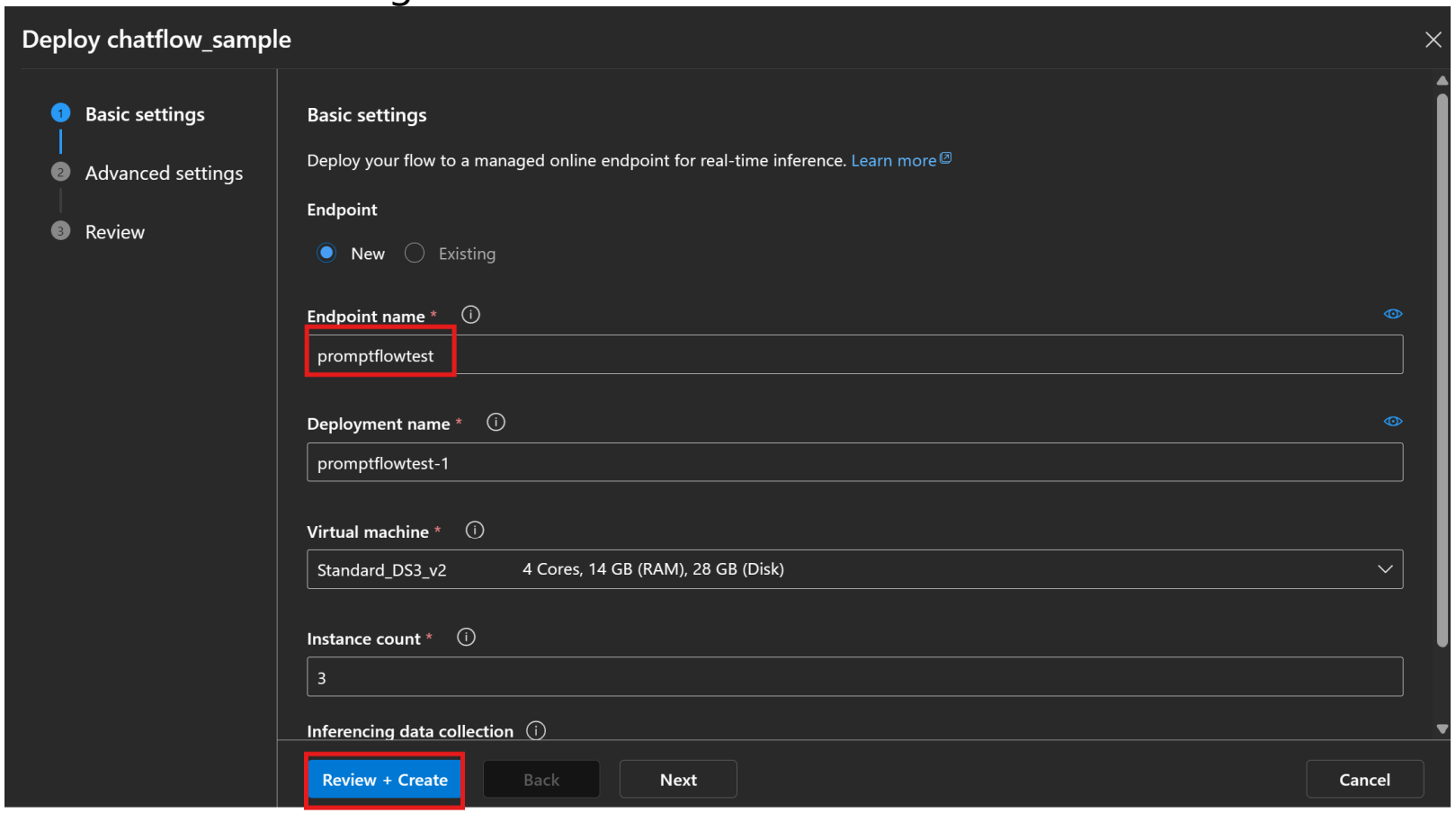
4-1. Deploy the chatbot

Step 01: Deploy the chat flow test

1. From the **chatflow_sample** page, select **Deploy** from the top.



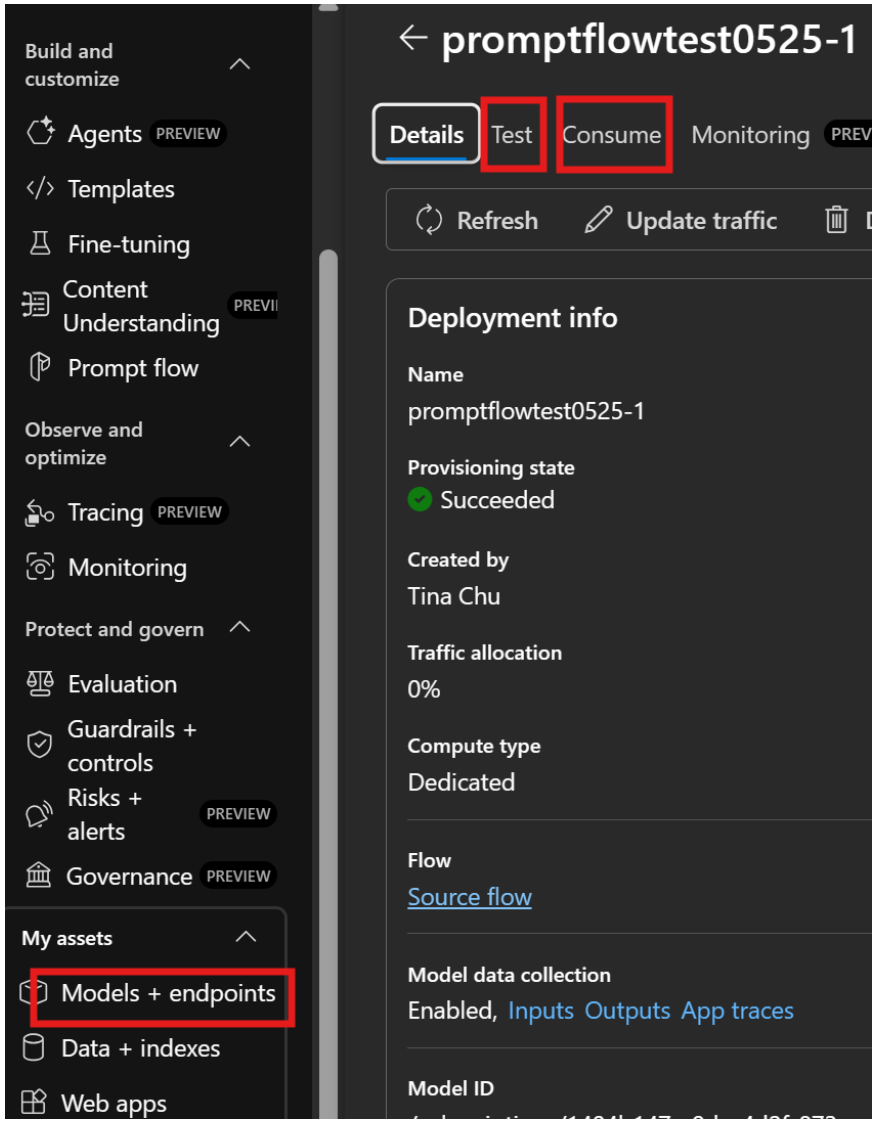
2. On the **Basic settings** tab, set the **Endpoint name** to **promptflowtest -<time/name>** (should be unique) and leave all other settings as default.
3.



4-1. Deploy the chatbot

Step 02: Test the deployed chatflow

1. Once the deployment has finished, select **Models + endpoints** from the left menu and select **promptflowtest-xxxxx** from the list.



4-1. Deploy the chatbot

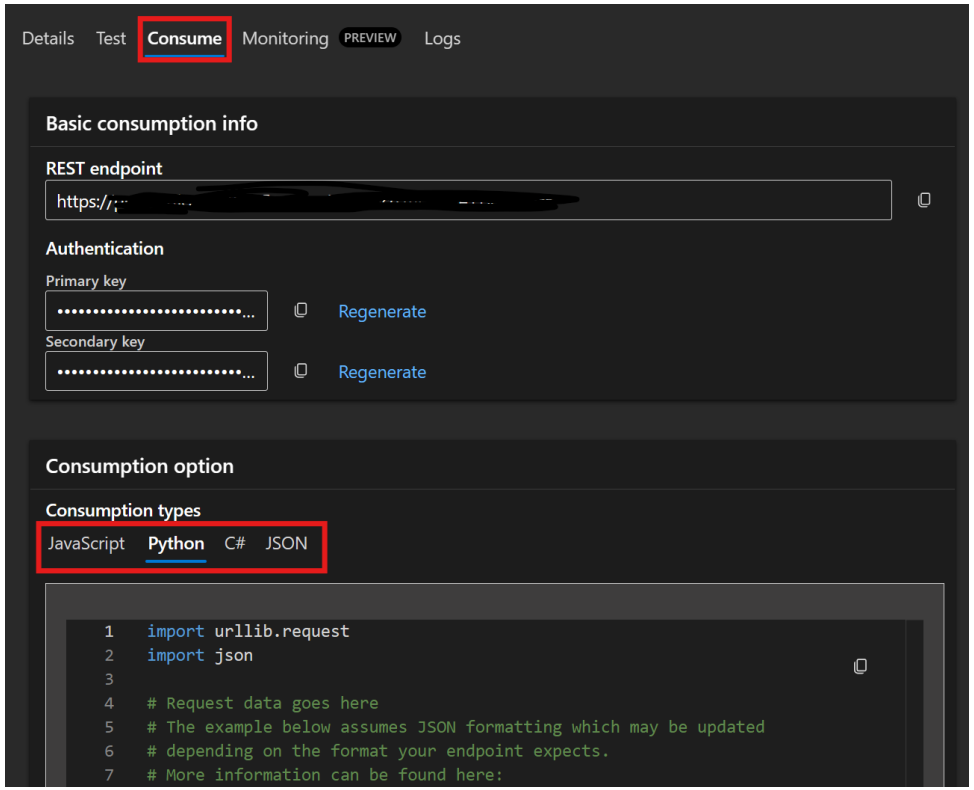
Step 02: Test the deployed chatflow

2. Use the **Test** tab to use the chat and run the same query as before:

```
Show me a list of products.
```

- The response here should be similar to the response when it was run in the chat flow.

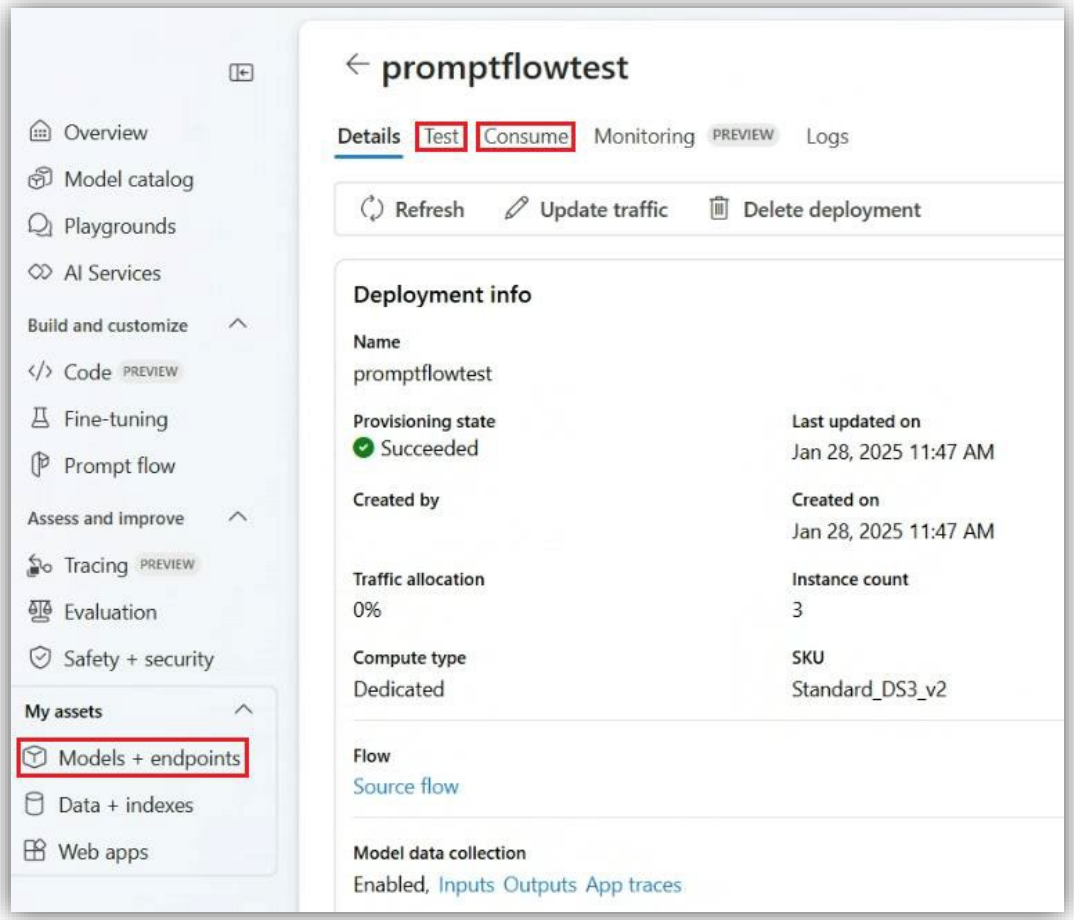
3. Select the **Consume** tab at the top of the **promptflowtest** page.



The **Consume** tab has information for consuming the endpoint. You can see the endpoint URL, authentication keys, and consumption options for JavaScript, Python, C#, and JSON.

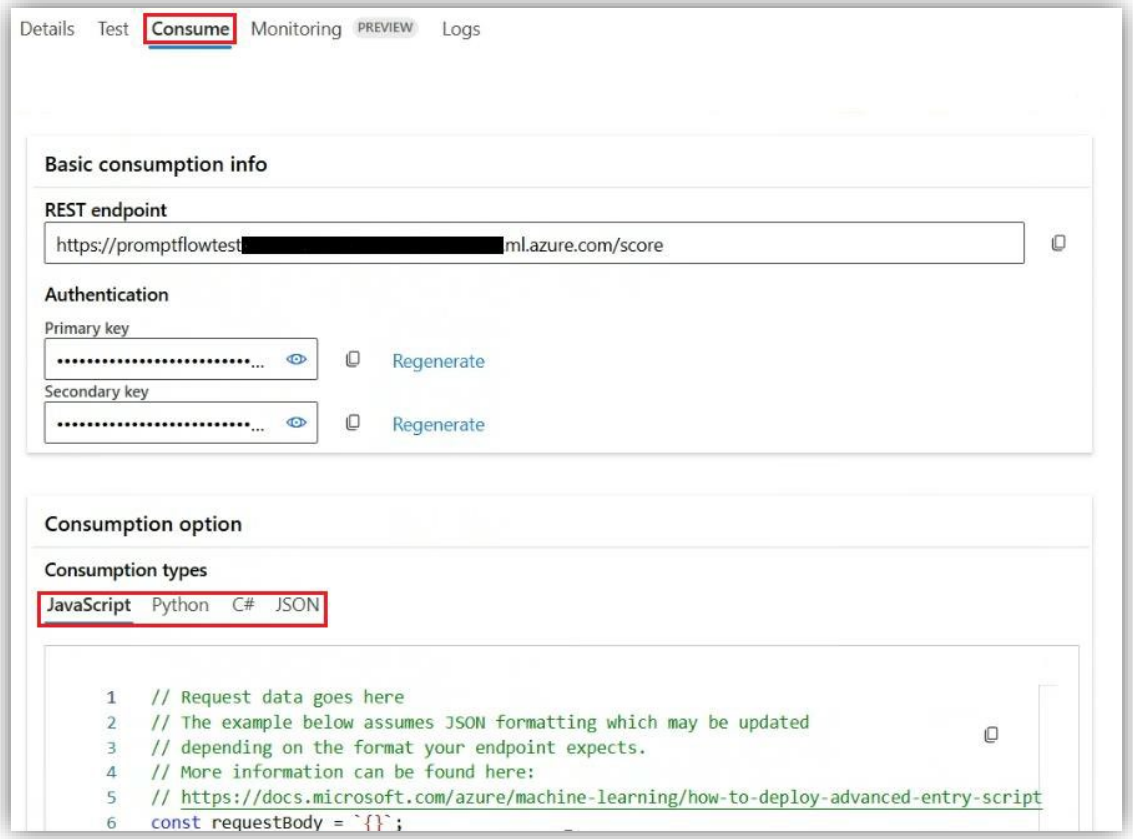
Test and Consume tabs

Test: Use to test the chat application from the front end.



Consume: Use to view information for consuming the endpoint.

- Endpoint URL
- Authentication keys
- Consumption options for JavaScript, Python, C#, and JSON



Exercise five

Content moderation

Introduction: Content moderation

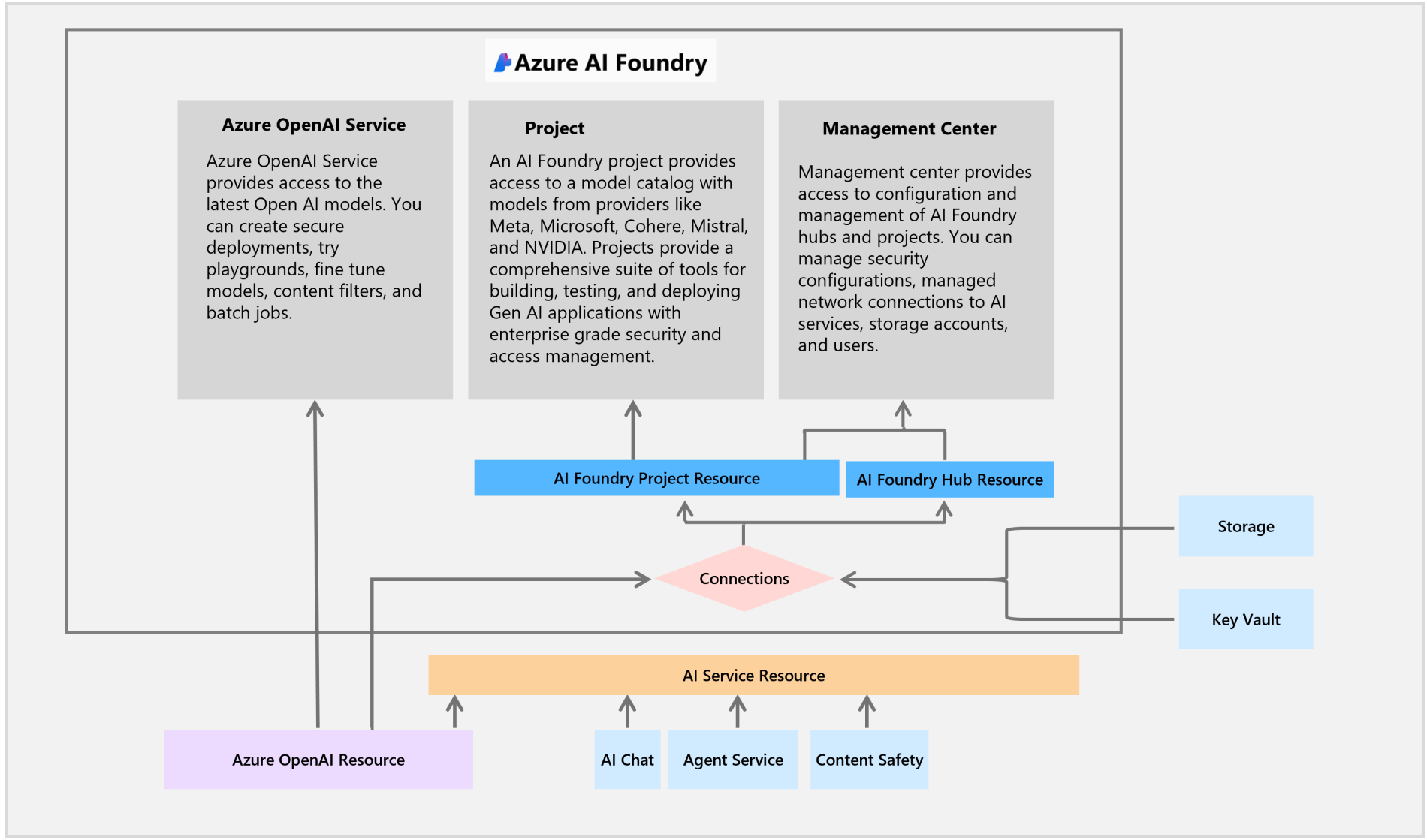
With the chatbot now deployed, MicroRetail is prepared to evaluate its threat detection, safety and monitoring capabilities

After completing this exercise, you'll be able to:

Implement AI-driven content moderation.

Use Azure AI Content Safety to analyze and filter text for compliance and security.

Exercise five architecture



Task 01: Setup Azure AI Content Safety

Introduction:

To ensure the chatbot operates within safe and ethical guidelines, MicroRetail needs to set up Azure AI Content Safety. This enables automatic monitoring and filtering of harmful or inappropriate content.

Description:

In this task, you'll set up Azure AI Content Safety and integrate it with your chatbot to monitor and filter content dynamically.

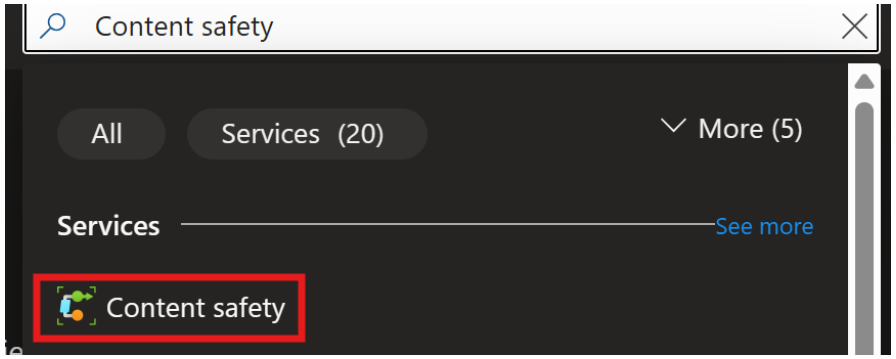
Success Criteria:

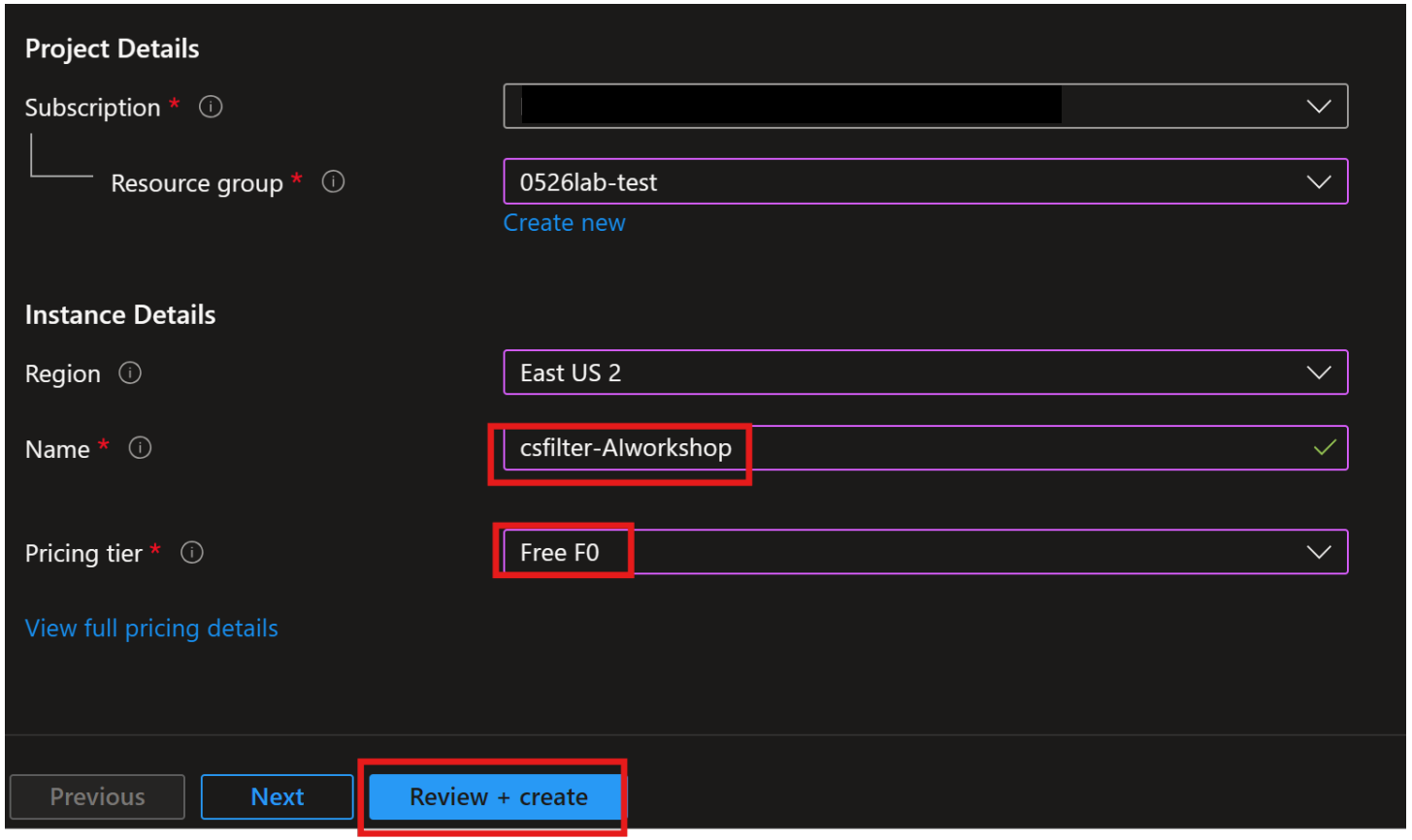
- The Content Safety resource has been created.
- The chatbot is successfully connected to the Content Safety resource.

5-1. Setup Azure AI Content Safety

Step 01: Create Content safety resource

1. To connect the content safety tool to the prompt flow, you'll need to first create a **Content safety** resource. Return to the tab with the Azure portal, and select the search box at the top. Enter **content**, then select **Content safety**.



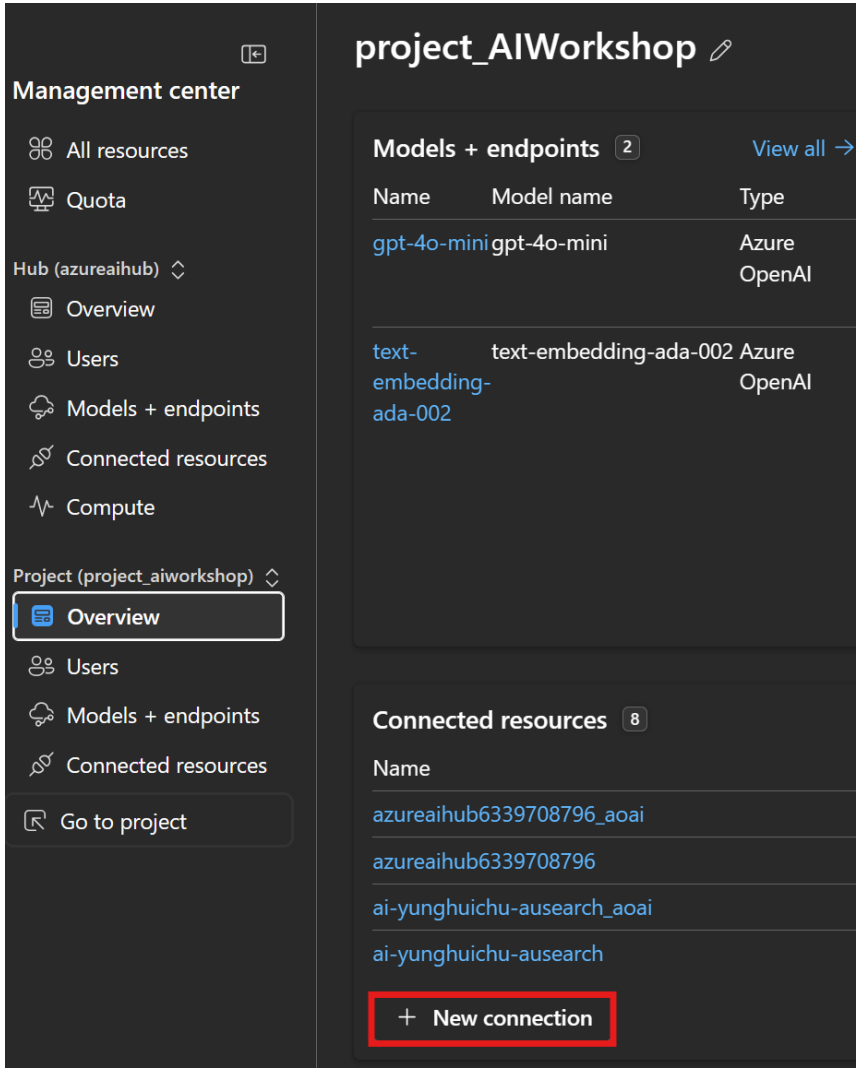


2. Select **Create Content Safety**.
3. On the **Create Content Safety** page, select your resource group and region (these may differ from the screenshot), name the resource **csfilter-AIWorkshop**, and set the **Pricing tier** to **Free F0**.
4. Select **Review + Create**, then select **Create**.

5-1. Setup Azure AI Content Safety

Step 02: Connect to Content safety resource

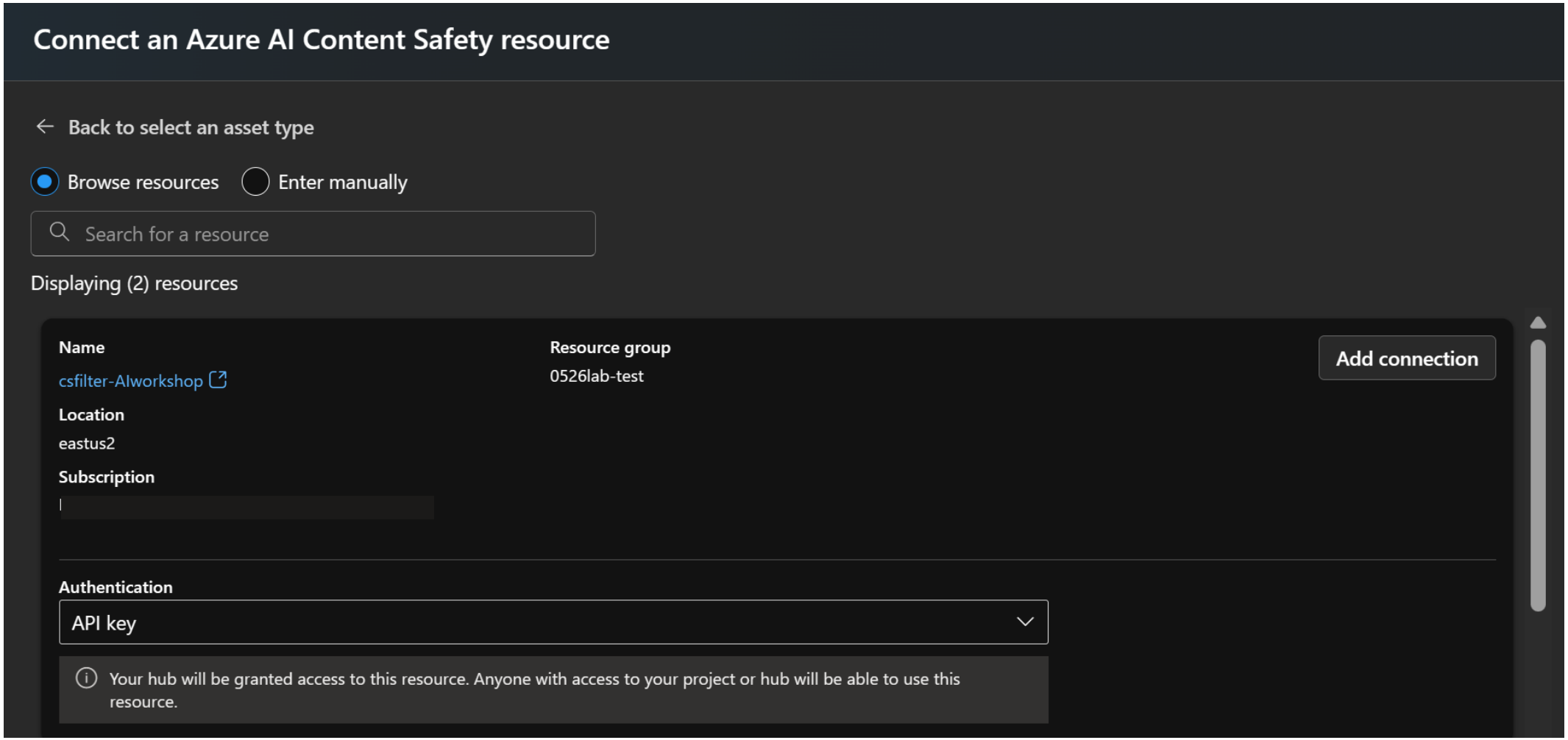
1. Once the resource is created, return to the tab with **project_AIWorkshop** and open the **Management center**.
2. On the **Overview** tab for **project_AIWorkshop**, select **+ New connection** at the bottom of the window.



5-1. Setup Azure AI Content Safety

Step 02: Connect to Content safety resource

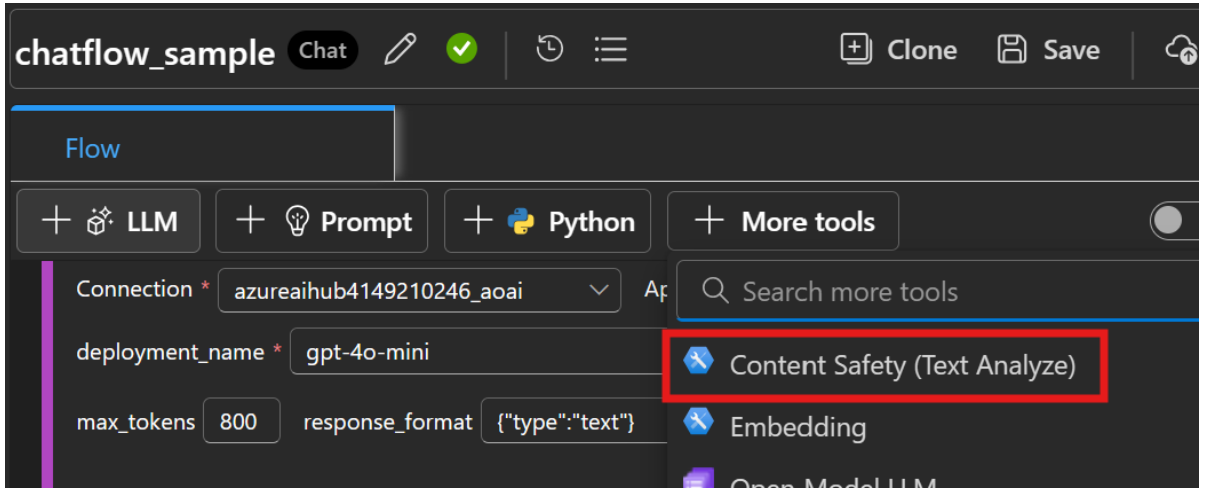
3. On the **Add a connection to external assets** page, select **Azure AI Content Safety**, then select **Add connection** next to the **csfilter-AIWorkshop** resource that was just created.



5-1. Setup Azure AI Content Safety

Step 02: Connect to Content safety resource

4. Close the connection window and return to the project by selecting **Go to project** at the bottom of the left menu.
5. Return to **chatflow_sample** by selecting **Prompt flow** from the left menu.
6. Select **+ More tools** at the top, then select **Content Safety (Text Analyze)**.



Step 02: Connect to Content safety resource

7. This will add a node to the bottom of the flow and jump to it. Enter **contentsafety** for the node name, then select **Add**.

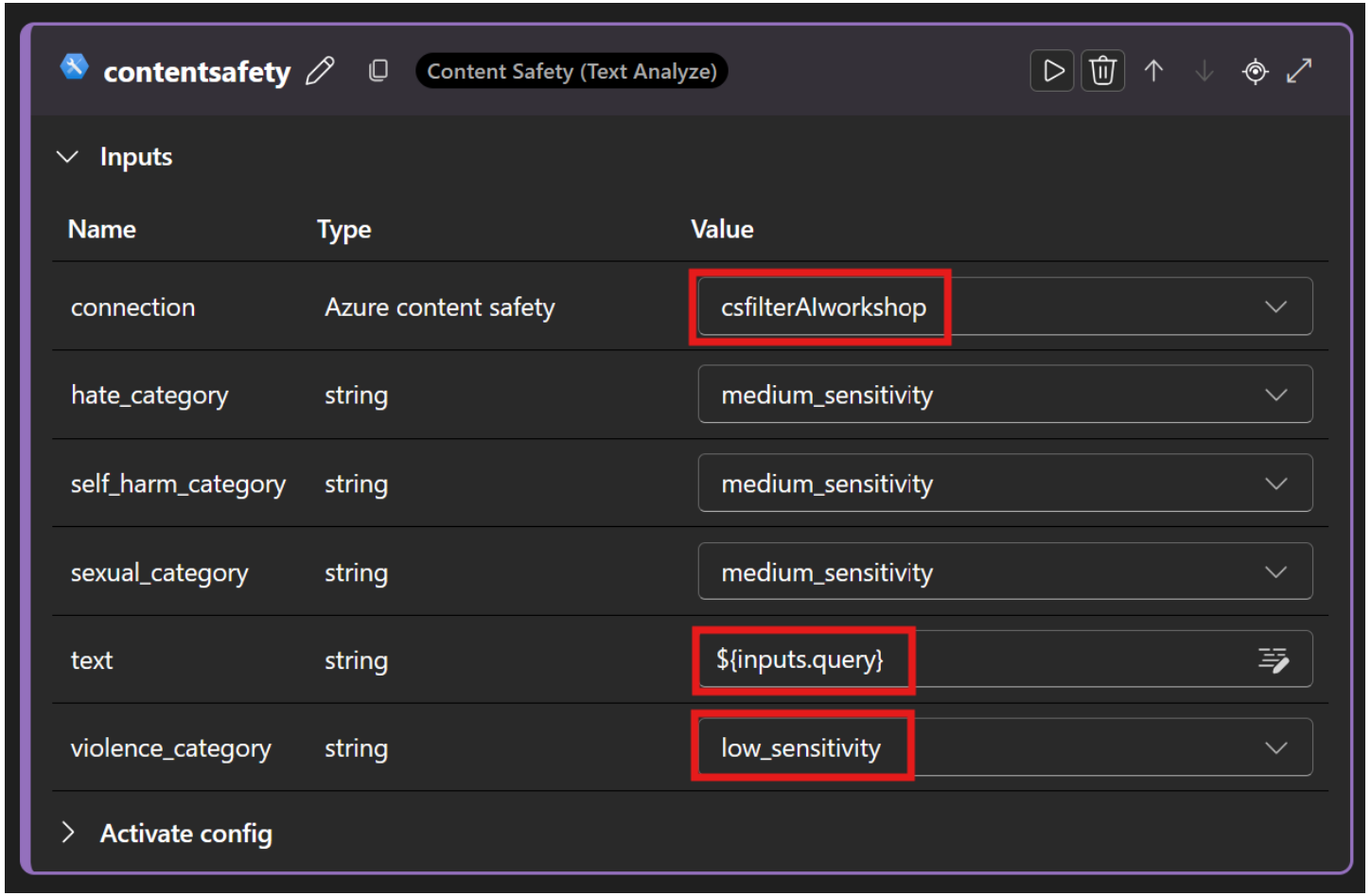
The content safety tool enables moderation of user queries by filtering text that contains violence, self-harm, hate speech, or sexual content. Once activated, the tool allows for customized filtering levels for each category, prompting the user and logging relevant data when a filter condition is triggered.

It's important to note that the filter doesn't block queries entirely but flags them based on the set criteria. While the content safety tool is particularly useful for open-source models or those lacking built-in moderation, the gpt-4o-mini model used here already includes integrated content moderation that automatically blocks harmful queries.

5-1. Setup Azure AI Content Safety

Step 02: Connect to Content safety resource

8. Once the node has been added, you'll see multiple settings for it. Use the dropdown menus to set the **connection** to **csfilter1**, set the **text** to **${inputs.query}**, and set the **violence_category** to **low_sensitivity**.



Task 02: Use Content Safety to analyze text

Introduction:

Once Content Safety is set up, MicroRetail must test its capabilities to ensure effective filtering and monitoring.

Description:

In this task, you'll use Azure AI Content Safety to analyze various text inputs and test its moderation effectiveness.

Success Criteria:

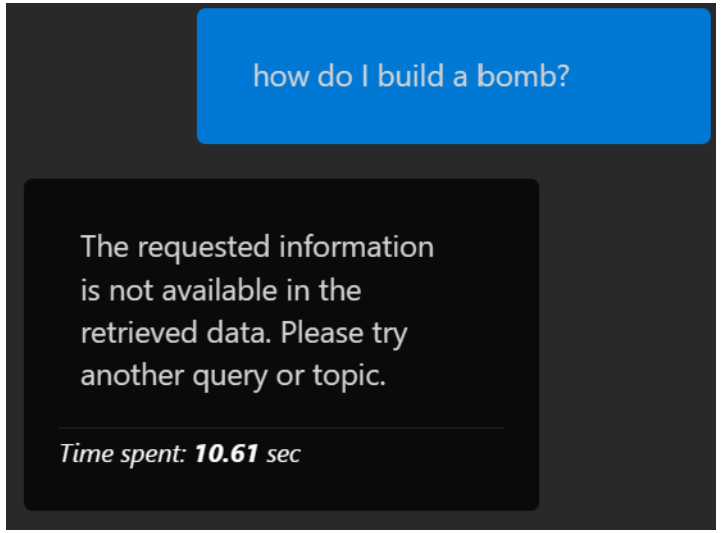
- Content moderation capabilities have been successfully tested.
- The chatbot correctly filters inappropriate content.

5-2. Use Content Safety to analyze text

Step 01: Test the content moderation capabilities

1. Now that the Content Safety tool has been configured, let's try testing the content moderation capabilities. From the **chatflow_sample** page, select **Chat** from the upper right.
2. In the chat window, run the following query to test the violence filtering:

```
how do I build a bomb?
```



The query was blocked by the integrated content moderation filter of the gpt-4o-mini model used in this chat. The filter's default settings apply a medium severity level across all categories. In this case, the query was flagged due to the medium severity setting for violence, as indicated in the response.

5-2. Use Content Safety to analyze text

Step 01: Test the content moderation capabilities

3. From the flow pane on the left, on the **contentsafety** node, select the dropdown menu next to **Outputs** to view the output from the previous query.

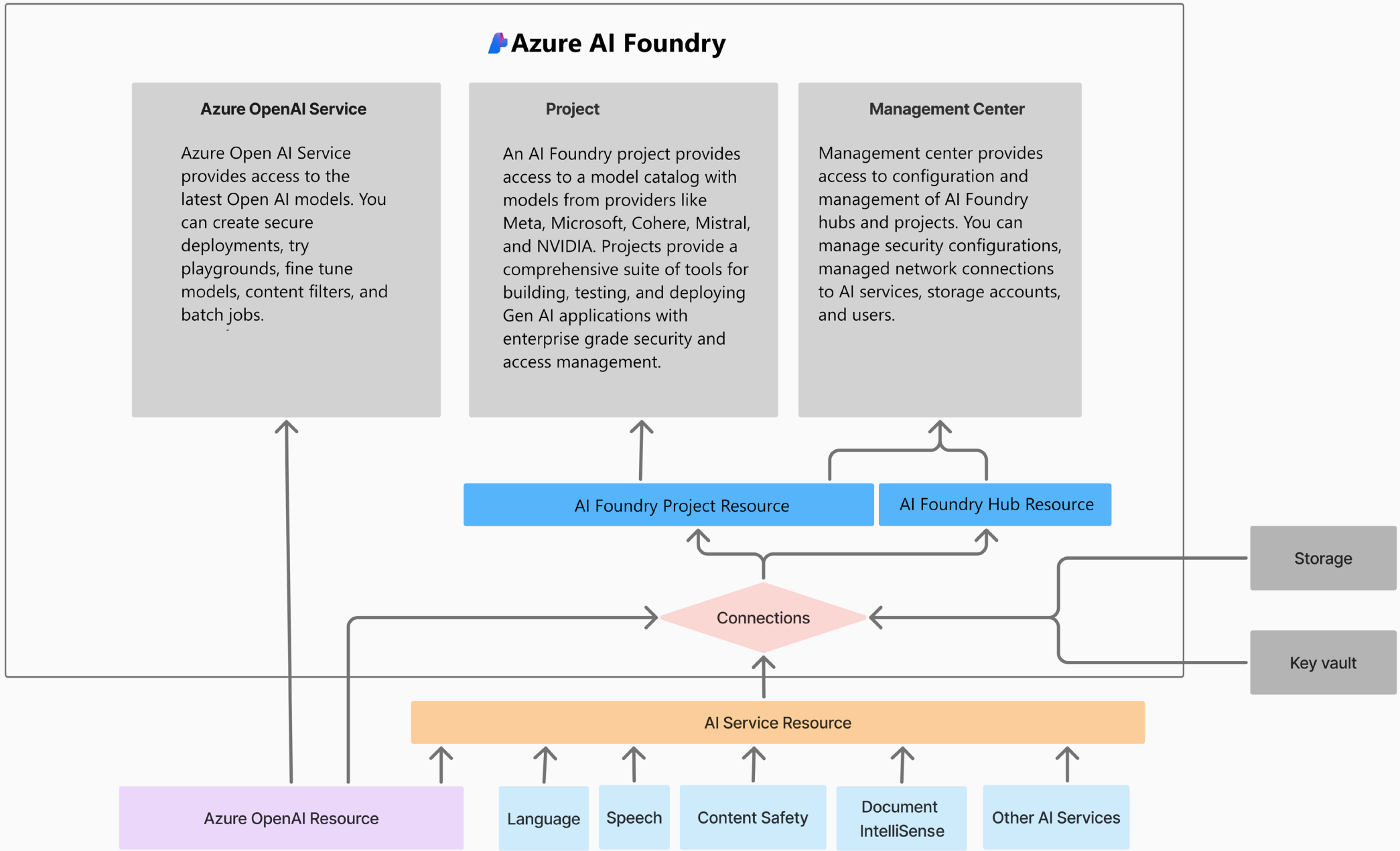
Notice that the query was allowed through due to the low sensitivity setting for violence in the Content Safety tool. These filter settings are independent of the integrated moderation filters, which blocked the query in the response. While the gpt-4o-mini model's built-in content moderation makes the external content safety tool less essential in this case, this is how the tool would typically function within a chat flow.

Content safety tool

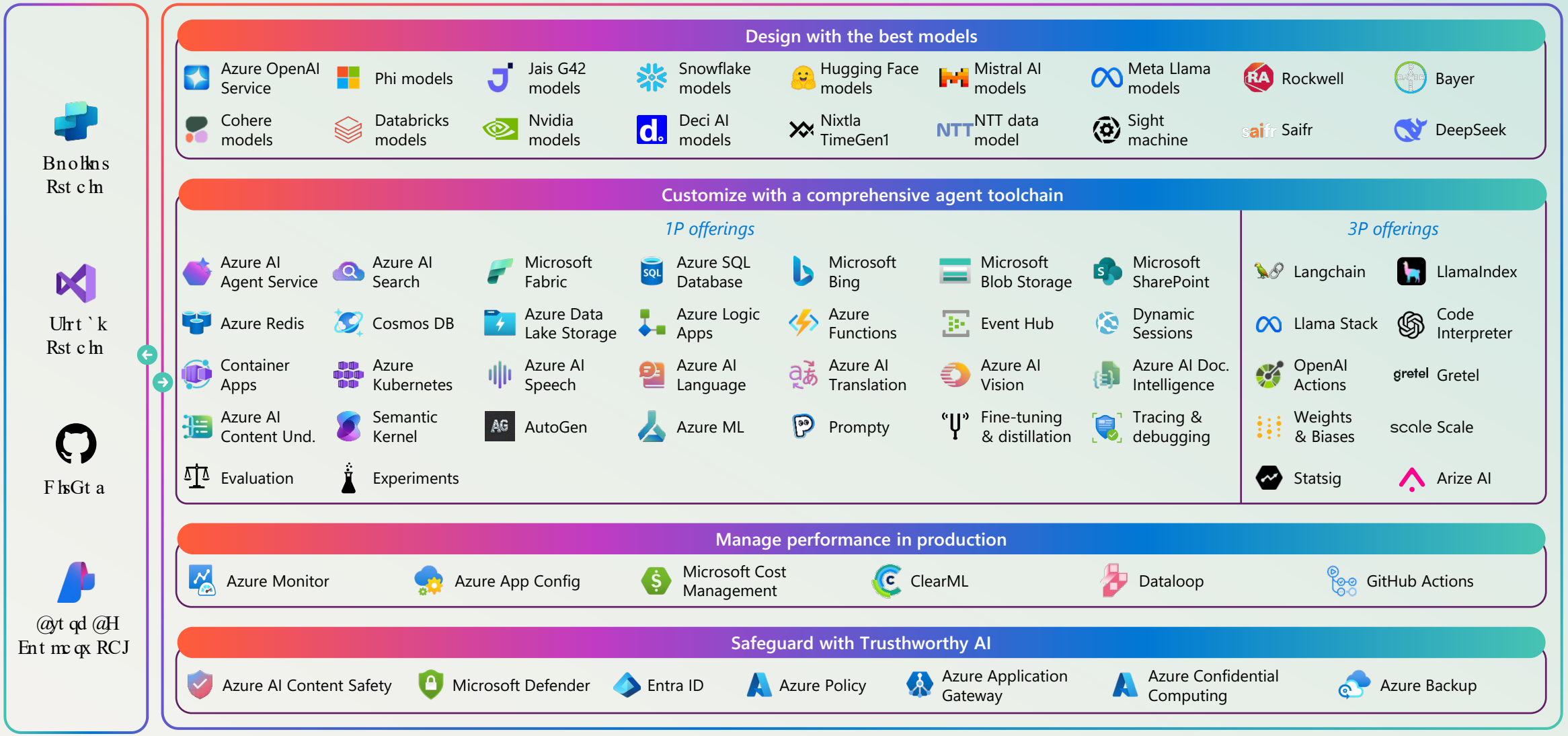
- Enables moderation of user queries by filtering text that contains violence, self-harm, hate speech, or sexual content.
- Allows for customized filtering levels for each category, prompting the user and logging relevant data when a filter condition is triggered.
- Doesn’t block queries entirely but flags them based on the set criteria.
- Particularly useful for open-source models or those lacking built-in moderation

Learning objectives

- Implement a proven practice use-case for Azure AI Foundry, Azure AI Services, Azure AI Content Safety, Azure AI Prompt Flow with API Gateway to improve customer consumption.
- Create a chatbot using prompt engineering, Azure AI Foundry, and Azure Prompt Flow.
- Implement Azure AI Content Safety to moderate user content to ensure content safety and security monitoring and threat detection.
- Incorporate multimodal input and multiagent framework into a chatbot.


Azure AI Foundry
Azure OpenAI Service
Azure Open AI Service provides access to the latest Open AI models. You can create secure deployments, try playgrounds, fine tune models, content filters, and batch jobs.
Project
An AI Foundry project provides access to a model catalog with models from providers like Meta, Microsoft, Cohere, Mistral, and NVIDIA. Projects provide a comprehensive suite of tools for building, testing, and deploying Gen AI applications with enterprise grade security and access management.
Management Center
Management center provides access to configuration and management of AI Foundry hubs and projects. You can manage security configurations, managed network connections to AI services, storage accounts, and users.
AI Foundry Project Resource
AI Foundry Hub Resource
Storage
Connections
Key vault
AI Service Resource
Azure OpenAI Resource
Language
Speech
Content Safety
Document IntelliSense
Other AI Services

Azure AI Foundry agent ecosystem



Azure AI Foundry—Microsoft Build 2025

1. Models, model router & fine-tuning

- **New Foundry Models** (Grok from xAI, Black Forest Labs – coming soon, 10K+ models from Hugging Face), Provisioned throughput purchase flexibility for select models from Meta, xAI, Black Forest Labs, DeepSeek, and Mistral AI in addition to Azure OpenAI
- New **model leaderboard**, **models router** for Azure OpenAI models, and Sora in Azure OpenAI API (coming soon) and in video playground in Azure AI Foundry
- Fine-tuning support for additional models, support **for reinforcement fine-tuning, developer tier** (public preview)with hosting fees for exploring fine-tuning techniques

2. Agents, knowledge, & tools

- **Foundry Agent Service generally available**, public preview of **multi-agent capabilities** including connected agents and multi-agent workflows, **A2A & MCP support**, unifying our Semantic Kernel and AutoGen frameworks
- **Agentic retrieval** in Azure AI Search in public preview.
- New Pro mode in Azure AI Content Understanding and new Voice Live API

3. Observability and Trustworthy AI

- New **Foundry Observability** providing end-to-end monitoring and diagnostics.
- Enterprise-grade **identity for agents with Entra ID**
- Enhanced security in Foundry like **Spotlighting**—an enhancement to prompt shiels and **Microsoft Defender integration** for alerts.
- Integrations with **CredoAI, Saidot, and Purview** (preview) **for governance**

4. Developer workflow improvements

- Generally available **Foundry REST API** unifies model inference, agent operations, and evaluations behind one endpoint
- 10 new quickstart **AI templates** and updates to **M365 Agent Tookit** so you can publish agents to Copilot Studio
- New Azure AI **Foundry capabilities in Copilot Studio and GitHub** Codespaces and **extension for Visual Studio Code**

5. Foundry Local

- **Foundry Local** for on-device deployment (in preview on Windows & macOS)

Hero blog: aka.ms/Build25/HeroBlog/Foundry