

# Tutorial to create your own Custom Gen ai application using Large/Small Language models.

**Dated:** 01/01/2025

## **Audience:**

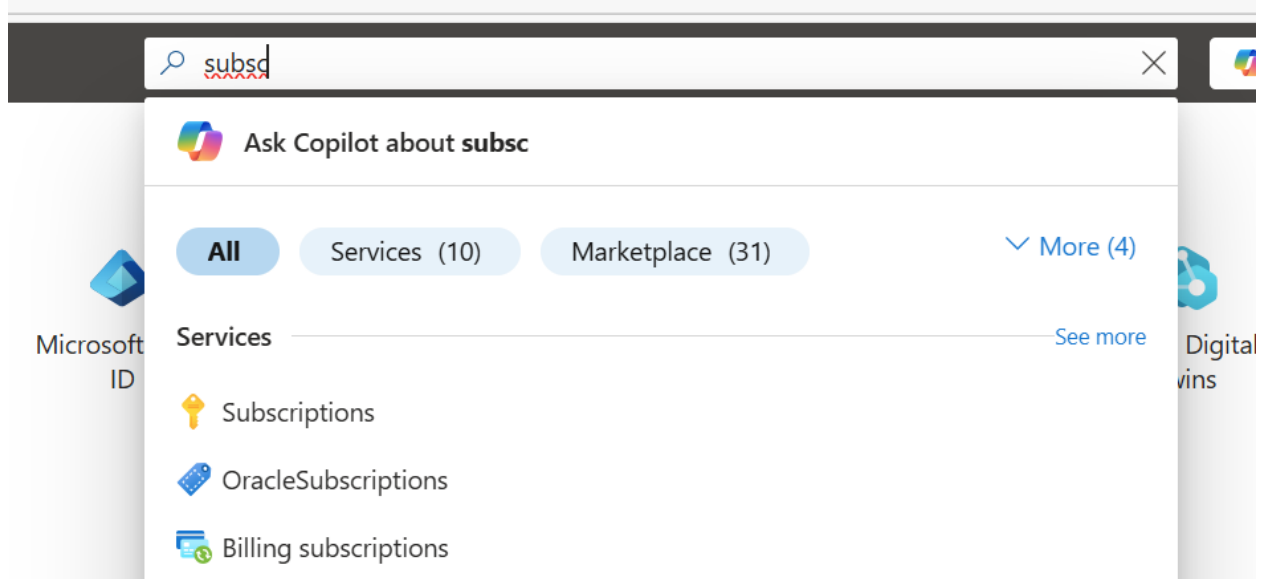
Low Code Developers

AI Engineers

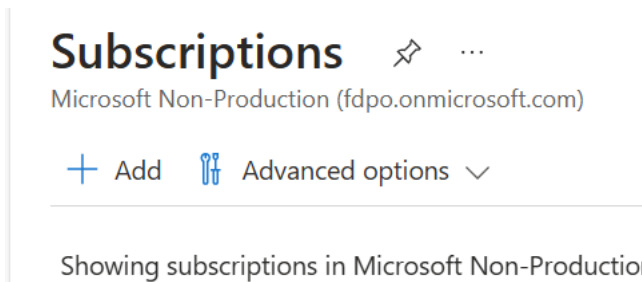
Business Enthusiast

## Pre-requisite:

1. Azure Account:
  - a. Go to [Create Your Azure Free Account Or Pay As You Go | Microsoft Azure](#)
2. Before getting Azure account, we need a email account either work based on outlook or Hotmail or other email providers.
3. Click Pay as you go or sign up for free account
4. Follow the instructions with your information and credit card details.
5. Once account is created, go to <https://portal.azure.com>
6. Login with your email used for signup.
7. On the search bar in the top type in: Subscriptions, if you type partial it will populate as well. Check the image below:



8. Select the subscriptions
9. Now Click Add in the subscriptions page, a sample image is provided below how you will see it.



10. Follow the instructions to create a subscription with Microsoft Azure plan and select your billing profile.
11. For Plan type select Microsoft Azure plan.
12. Choose a billing profile created from the above sign up.
13. Give a name to the subscription and complete the process.
14. Once subscription is ready then choose the subscription in the portal.
15. We are going to create 4 resources, those are:
  - a. Resource Group
  - b. Azure AI Foundry
  - c. Azure Open AI Service
  - d. Azure AI Service

## Steps to Create Resources:

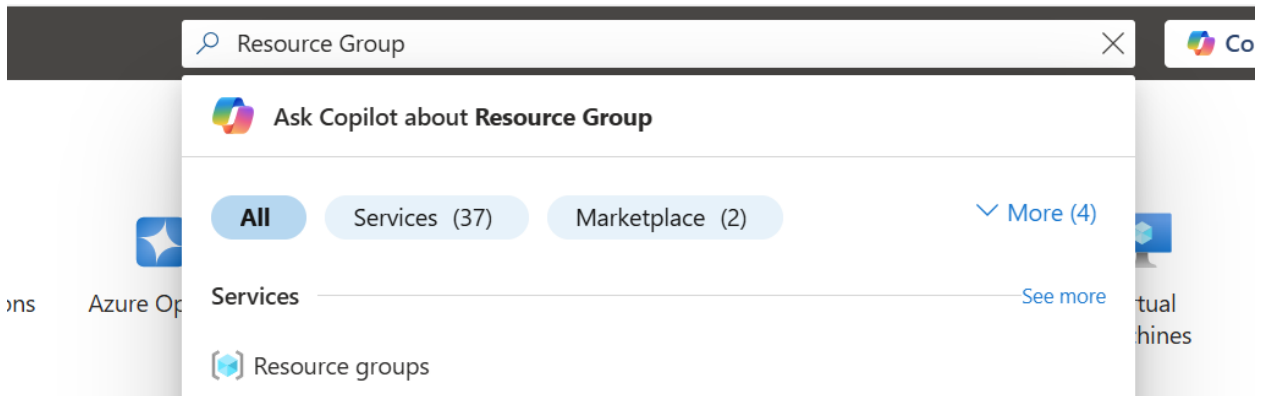
Now to create a resource group. Before creating a resource group pick a region to use like US central or East US etc. Resource group are logical separation provided in portal for us to organize resources or assets based on workload or functionality of workload.

## Creating a Resource Group

- 1) Log into Azure Portal <https://portal.azure.com>, use the above email address which was used to create azure account.
- 2) In the home page if you see Resource group like below then select:



- 3) If there is no Resource group, then go to Search bar in the top center of the page and type Resource Group



- 4) Select Resource group
- 5) Now click Create button like below:

[+ Create](#)

- 6) Fill the details like subscription, which you can choose from list created above.

## Create a resource group ...

Basics   Tags   Review + create

**Resource group** - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. [Learn more](#)

Subscription * ⓘ	<input type="text" value="Subscription"/>
Resource group name * ⓘ	<input type="text" value="WIMEPTraining"/>
Region * ⓘ	<input type="text" value="(US) Central US"/>

- 7) Provide a name, all in small letters without any spaces.
- 8) Select a region to use in the above I have selected Central US
- 9) Then click Review and Create at the bottom of the page

[Previous](#)
[Next](#)
[Review + create](#)

- 10) After validation it will show the page to create so click Create like below

---

Previous

Next

Create

---

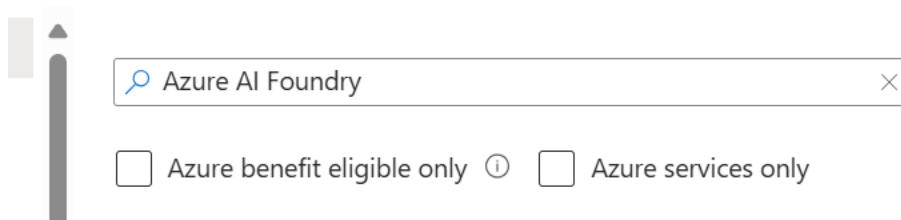
- 11) Wait for the resource to create.
- 12) Now Go to home page and select the Resource group created above from the recent activity section, or Search for resource group that we created in the top search bar. Click into the resource group and should see empty one.

Now it's time to create AI Foundry resources:

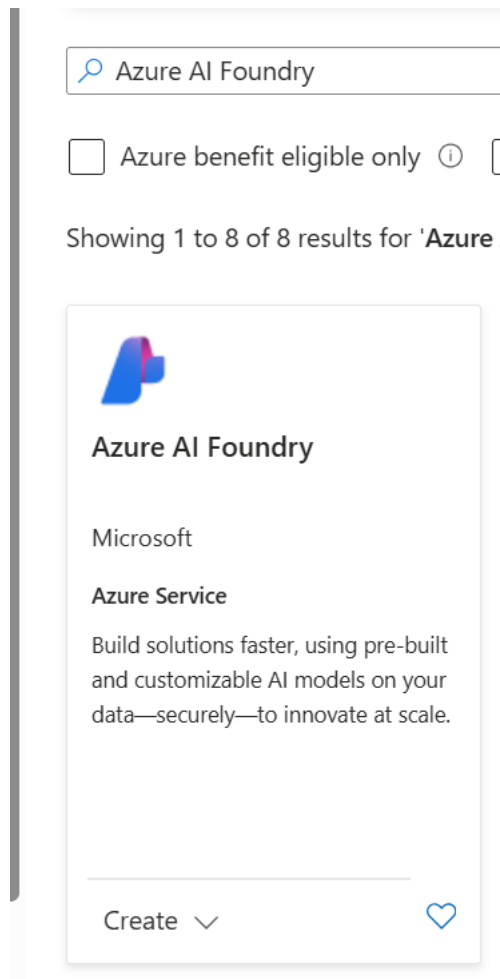
- 1) First we create Azure AI Foundry which is the interaction layer to create the gen ai application.
- 2) Click Create in the resource group selected like below:

 Create

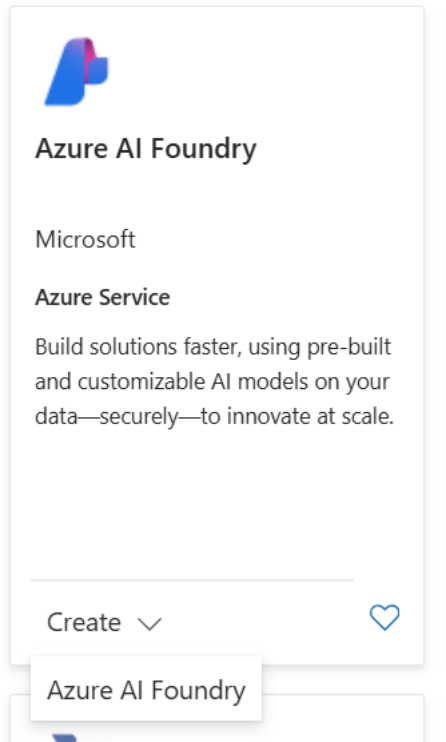
- 3) On the search bar shown below type in Azure AI Foundry and should see a one in the list to choose from



- 4) Or press enter or click search button to find the resource.
- 5) Here is the resource to create:



6) Now click create on the Azure AI foundry resource:



- 7) Click Azure AI Foundry
- 8) In the next screen, choose the subscription, should be autopopulated.
- 9) Then fill the remaining details like Resource group, region to use, name for Azure AI foundry like below and leave the other in tact.

# Azure AI hub

Create an Azure AI hub resource

**Basics** Storage Networking Encryption Identity Tags Review + create

## Organization

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources. An AI hub is a collaboration environment for a team to share project work, model endpoints, compute, (data) connections, security settings, govern usage.

Subscription \* ⓘ

Resource group \* ⓘ  [Create new](#)

Region \* ⓘ

## Resource details

Name \* ⓘ  ✓

Friendly name ⓘ

Default project resource group ⓘ

## Azure AI services base models

Connect AI Services incl. OpenAI \* ⓘ  [Create new](#)

**Review + create**

< Previous

Next : Storage

- 10) Click Review and Create on the bottom left of the page, once all the details are filled. For Resource group you can select the one created. Name is something we need to fill and friendly name will fill the name as name.
- 11) Wait for validation to complete and then click Create like below image:

## Azure AI hub ...

Create an Azure AI hub resource

✓ Validation passed

Create

< Previous

Next >

12) Validation should be passed, if you are using a corporate tenant you might not have permission to create, in that case talk to the administrator who can help you create.

13) Wait for the resource to create.

## ... Deployment is in progress

14) Once the deployment is complete, here is what you will see in the screen:





✓ Your deployment is complete

15) Once the resource is created go back to Resource group. You can click the home link shown below:


[Home](#) >

16) Once in home page, select the resource group in recent list and go inside and should see the AI hub or AI foundry that we created above.



Resources		Recommendations
<input type="text" value="Filter for any field..."/>		Type equals <b>all</b> <input type="button" value="X"/>
<input type="checkbox"/>	Name ↑	
<input type="checkbox"/>	 WIMEPAIHub	
<input type="checkbox"/>	 wimepaihub3564589699	
<input type="checkbox"/>	 wimepaihub4696732790	
<input type="checkbox"/>	 wimepaihub5948912238	


Now it's time to create Azure Open AI resource.

- 1) Click on the Create on the resource group:  Create
- 2) On the Search bar type: Azure Openai, and select azure openai in the list and press enter or click search button and here is what you should see:

 azure openai

☐ Azure benefit eligible only ⓘ

Showing 1 to 20 of 157 results for 'azur




### Azure OpenAI

Microsoft

**Azure Service**

The Azure OpenAI service.

---

Create ▾ 

- 3) Click Create on the resource page
- 4) Below are sample screen shot shown how to fill the page:

# Create Azure OpenAI ...

1 Basics 2 Network 3 Tags 4 Review + submit

Azure OpenAI Service provides access to OpenAI's powerful language models, including all the latest OpenAI models. These models can be easily adapted to your specific tasks, including but not limited to content generation, summarization, image understanding, semantic search, and natural language to code translation. Top use cases include Call Centers, Virtual Assistants, Accessibility, Content Generation, and Code Development. The service also features the Assistants API, Fine Tuning capabilities and many ways to connect your data to the service for conversational experiences. The service can be scaled through Standard (tokens) and Provisioned (PTUs) deployment types.

[Learn more](#)

## Project Details

Subscription \* ⓘ

Resource group \* ⓘ

[Create new](#)

## Instance Details

Region ⓘ

Name \* ⓘ

Pricing tier \* ⓘ

[View full pricing details](#)

## Content review policy

To detect and mitigate harmful use of the Azure OpenAI Service, Microsoft logs the content you send to the Completions and image generations APIs as well as the content it sends back. If content is flagged by the service's filters, it may be reviewed by a Microsoft full-time employee.

[Learn more about how Microsoft processes, uses, and stores your data](#)

[Apply for modified content filters and abuse monitoring](#)

Previous

Next

- 5) Provide a Name for the resource, has to be unique and also select Central US for region, in the above example I choose Japan east since I don't have quota in other regions. But for your training please Select Central US or a region in US.
- 6) Click Next
- 7) Keep Clicking Next on the other few pages until Final validation and then click Create as shown below

Previous

Next

Create

8) Wait for the resource to be created.

...
 Deployment is in progress




- 9) Resource creation usually takes like 2 to 5 minutes.
- 10) Once completed, here is what will be seen in the screen:

✓
 Your deployment is complete






- 11) Now lets create Azure AI Search.
- 12) Go back to Home and select the Resource group:

Recent

Favorite

Name	Type
 WIMEPTraining	Resource group
 WIMEPopenai1	Azure OpenAI
 WIMEPAIHub	Azure Machine Learning workspace

13) Click the WIMEPTraining Resource group and you will see all the resource that we created until now.

<input type="checkbox"/>	 WIMEPAIHub
<input type="checkbox"/>	 wimepaihub3564589699
<input type="checkbox"/>	 wimepaihub4696732790
<input type="checkbox"/>	 wimepaihub5948912238
<input type="checkbox"/>	 WIMEPopenai1

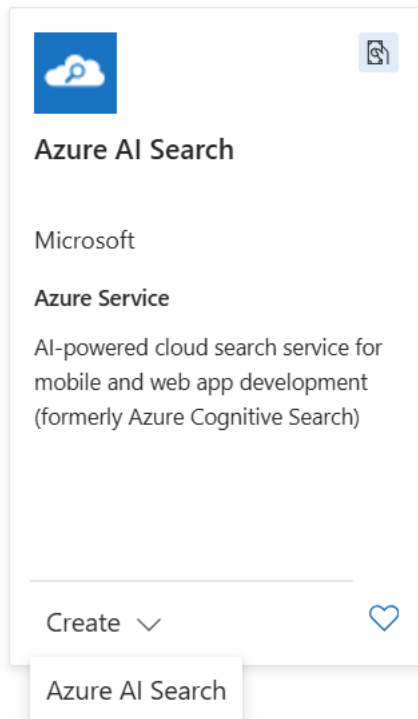
## Create Azure AI Search Resource:

- 1) Now click Create on the top of the resource group page: [+ Create](#)
- 2) In the search bar type Azure AI Search.

A screenshot of the Azure portal search bar. The text 'azure ai search' is entered into the search field. To the left of the text is a magnifying glass icon, and to the right is a close 'X' button.

☐ Azure benefit eligible only ⓘ ☒ Azure services only

- 3) Scroll down until we see Azure AI Search as below image:



- 4) Then click Create and choose Azure AI Services. If you can see the AI Search on under the search bar click Azure Services only and should popup.

Home > WIMEPTraining > Marketplace >

Create a search service ...

Basics

Scale

Networking

Tags

Review + create

Project details

Subscription \*

WIMEPTraining

Create new

Resource Group \*

WIMEPTraining

Instance Details

Service name \* ⓘ

wimepaisearch

Location \*

Central US

Pricing tier \* ⓘ

Standard

160 GB/Partition, max 12 replicas

Change Pricing Tier

Select Pricing Tier

Browse available skus and their features

Sku	Offering	Indexes	Indexers	Vector quota
F	Free	3	3	25 MB ⓘ
B	Basic	15	15	5 GB
S	Standard	50	50	35 GB/Partition
S2 ⓘ	Standard	200	200	150 GB/Partition
S3 ⓘ	Standard	200	200	300 GB/Partition
S3HD ⓘ	High-density	1000	0	300 GB/Partition
L1 ⓘ	Storage Optimized	10	10	150 GB/Partition
L2 ⓘ	Storage Optimized	10	10	300 GB/Partition

⚠ Skus S2, S3, S3HD, L1, L2 are unavailable to select due to high demand.

ℹ Higher storage limits are available for new services in this region at no additional cost.

- From the above image, please give a name for the search services under Name\* section.
- Then in the Pricing tier choose Basic.
- Click the check box for the terms and conditions.
- Then Click Review and Create

# Create a search service ...

Basics   Scale   Networking   Tags   Review + create

## Project details

Subscription \*

Resource Group \*

WIMEPTraining

Create new

## Instance Details

Service name \* ⓘ

Location \*

Pricing tier \* ⓘ

wimepaisearch

Central US

Basic


15 GB/Partition, max 3 replicas, max 3 partitions, max 9 search units

Change Pricing Tier

9)

Wait for validation to complete and then click Create like below

# Create a search service ...

 Validation Success

10) Wait for deployment to be completed

Your deployment is complete

11) Now we have created all the resources needed to start our development journey.

12) Next Section is on how to create the application. For creating a generative ai application we need to get a pdf document. This could be any one we can use to ask or chat with the pdf document.

13) Time to grab the pdf document. Once you have the document follow the instructions below.

14) Now Go back to home and resource group and should see all the resources

<input type="checkbox"/>	Name ↑		Type
<input type="checkbox"/>	WIMEPAIHub	...	Azure AI hub
<input type="checkbox"/>	wimepaihub3564589699	...	Key vault
<input type="checkbox"/>	wimepaihub4696732790	...	Azure AI services
<input type="checkbox"/>	wimepaihub5948912238	...	Storage account
<input type="checkbox"/>	wimepaisearch	...	Search service
<input type="checkbox"/>	WIMEPopenai1	...	Azure OpenAI
<input type="checkbox"/>	wimepproject	...	Azure AI project



# Create Generative AI Application – No Code.

1. Click on WIMEPAIHub or the name of the Azure AI hub or foundry resource created.
2. Click Launch AI foundry:

## Govern the environment for your team in AI Foundry



Your Azure AI hub provides enterprise-grade security, and a collaborative environment to build AI solutions. Centrally audit usage and cost, and set up connections to your company resources that all projects can use. [learn more about the Azure AI Foundry](#) ↗

Launch Azure AI Foundry

3. Will take you to a new tab, which is the user interface, we are going to use for our application development.
4. First create a Project name as below:

### You'll need a project to keep working

#### Name your project

Projects are an even better way to collaborate, connect data and services, and organize everything you need to build with AI. Your existing assets will still be available—just in an easier to manage container.

Project name \*

#### Hub (WIMEPAIHub)

Your new project will be added under your current hub, which provides security, governance controls, and shared configurations that all projects can use.

Create projectCancel

5. Wait for the project to created.
6. Next on the left Menu click on Chat playground
7. Click on Try Chat Playground as below

## Chat playground

Create and test a chatbot by interacting with it and seeing how it responds to various inputs.

[Try the Chat playground](#)

8. Now click Create deployment as below:

Don't have a deployment?

[+ Create a deployment](#)

9. Now Select GPT 4o as the model:

## Select a chat completion model

Choose a model to create a new deployment. For flows and other resources, create a deployment from their respective list. [Go to model catalog.](#)

**Models: 63** Collections Deployment options Inference tasks: Chat completion Show description

Search

**o1**  
Chat completion

**gpt-4o**  
Chat completion

**o1-preview**  
Chat completion

**o1-mini**  
Chat completion

**gpt-4o-mini**  
Chat completion

**gpt-4**  
Chat completion

**gpt-4-32k**  
Chat completion

< Prev

Next >

### gpt-4o

Task: Chat completion

gpt-4o offers a shift in how AI models interact with multimodal inputs. By seamlessly combining text, images, and audio, gpt-4o provides a richer, more engaging user experience.

Matching the intelligence of gpt-4 turbo, it is remarkably more efficient, delivering text at twice the speed and at half the cost. Additionally, GPT-4o exhibits the highest vision performance and excels in non-English languages compared to previous OpenAI models.

gpt-4o is engineered for speed and efficiency. Its advanced ability to handle complex queries with minimal resources can translate into cost savings and performance.

The introduction of gpt-4o opens numerous possibilities for businesses in various sectors:

- Enhanced customer service:** By integrating diverse data inputs, gpt-4o enables more dynamic and comprehensive customer support interactions.
- Advanced analytics:** Leverage gpt-4o's capability to process and analyze different types of data to enhance decision-making and uncover deeper insights.

Confirm

Cancel

10. Select gpt-4o and then confirm.

11. Now select the model details as below image:

## Deploy model gpt-4o

Deployment name \*



gpt-4o

Deployment type

Global Standard



Global Standard: Pay per API call with the highest rate limits. Learn more about [Global deployment types](#).

Data might be processed globally, outside of the resource's Azure geography, but data storage remains in the AI resource's Azure geography. Learn more about [data residency](#).

Deployment details

Collapse

Model version

2024-08-06



AI resource

WIMEPopenai1



450K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit

450K

Corresponding requests per minute (RPM) = 4.5K

Content filter

DefaultV2

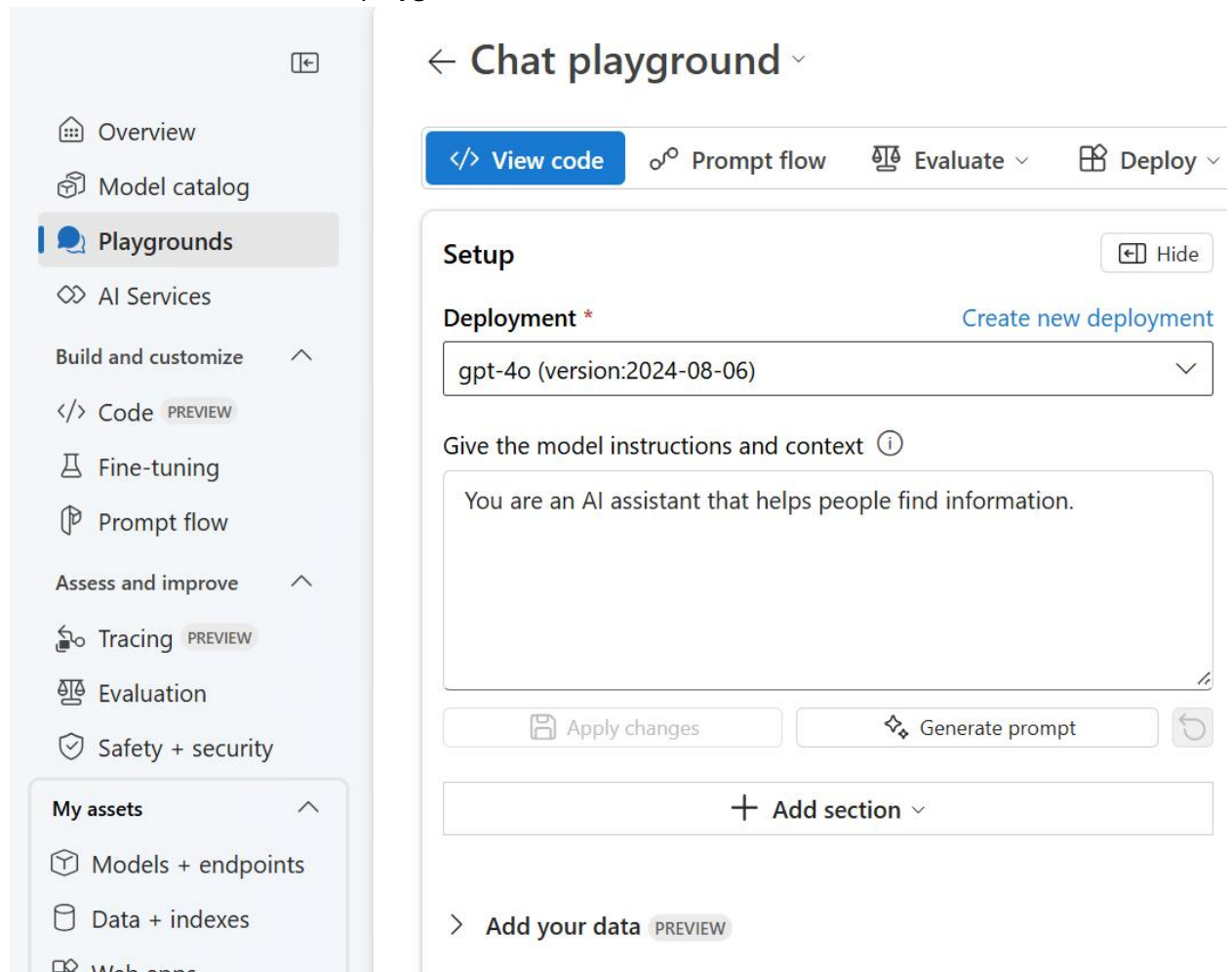


Your project will be connected to the selected resource  
A resource that supports the model has been pre-selected

Connect and deploy

Cancel

12. Select Deployment Type: Global deployment
13. And Make sure the Tokens per Minute Rate Limit to 450K, move the slider to the end.
14. Then Click Connect and Deploy and wait for the process to complete.
15. Now on the left menu select playgrounds and should see the below screen.



16. You can see the above gpt-4o in deployment.
17. Now we are going into Add your data, expand the data.

+ Add a new data source

18. Click on Add new data source:
19. Now follow the instructions on the guided step by step screen.
20. First select the data source as upload files.
21. Select the PDF file to upload like the below image:

Add your dataPREVIEW

1Source data

2Index configuration

3Search settings

4Review and finish

Select your data

Select the data you want the generative AI to reference so it can ground its responses on your specific data.

Your data will be ingested into an Index, which allows the Generative AI model to quickly and accurately find information for your specific use case.

Currently, only the file types .doc(x), .htm, .html, .md, .pdf, .ppt(x), .py, .txt, and .xls(x) are supported. Max file size limit is 16 MB.

Data source \* ⓘ

Upload files

Upload

☐ Overwrite if already exists

Upload list

2024-WI-Manufacturing-Report-v4-Oct-2-POST.pdf

✓ 5.17 MB/5.17 MB

...

ⓘ

An Azure AI Search resource and an Azure OpenAI connection will be required to index your data. [Create a new Azure AI Search resource](#) and create a connection to it to select it while creating an index.

Next

Create vector index

Cancel

22. Click Next
23. Now give a index name, then select the Azure AI Search resource

## Index settings

Configure your index

### Index storage \*

Azure AI Search

### Select Azure AI Search service \* ⓘ

Select Azure AI Search service

[Create a new Azure AI Search resource](#) 

### Vector index \* ⓘ


wimepaiindex


### Virtual machine \* ⓘ


☒ Auto select ☐ Select from recommended options ☐ Select from all options

*Selecting a virtual machine will incur additional costs.*

24. Select the AI Search service and click Connect if doesn't show in the list. If the list shows then select that resource.

<b>Name</b> wimepaisearch 	<b>Resource group</b> wimeptraining	<button>Add connection</button>
<b>Location</b> centralus	<b>SKU</b> basic	
<b>Subscription</b> MCAPS-Hybrid-REQ-39734-2022-babal	<b>Semantic search</b> free	

**Authentication**  
API key 

 Your hub will be granted access to this resource. Anyone with access to your project or hub will be able to use this

25. Click Add connection, once the connection is added, then only the drop down will show.

26. Now you select the AI Search

### Index settings

Configure your index

**Index storage \***

Azure AI Search

**Select Azure AI Search service \*** ⓘ

wimepaisearch

[Create a new Azure AI Search resource](#) ⓘ

**Vector index \*** ⓘ

wimepaiindex25

**Virtual machine \*** ⓘ

☒ Auto select ☐ Select from recommended options ☐ Select from all options

*Selecting a virtual machine will incur additional costs.*

Back

Next

Create vector index

Cancel

27. Make sure you choose the right AI search we created in the above deployment
28. Leave the other settings as is. Make sure vector index name is populated like above image.
29. Click Next
30. Leave the settings as is, nothing to change here



### Configure search settings

Adding vector search supports: Hybrid (vector + keyword search), Hybrid + Semantic (most accurate search results for generative AI applications), Vector, Semantic and Keyword retrieval. Hybrid will be set as default and can be changed at inference time in the playground. Not adding vector search supports: Keyword and Semantic retrieval. Keyword will be set as default and can be changed at inference time in the playground. Adding vector search requires an Azure OpenAI embedding model. [Learn more](#)

#### Vector settings

☒ Add vector search to this search resource

Azure OpenAI connection \* ⓘ

WIMEPopenai1

ⓘ This resource requires an embedding model. If you don't have one already, **text-embedding-ada-002 (Version 2)** will be deployed for you. Using vector embeddings will incur usage to your account. [View Azure OpenAI Service pricing](#)

Back

Next

Create vector index

Cancel

31. Click Next.

32. Validate all the settings are correct and Click Create Vector Index:

## Review and finish

Review the configurations you set for your index

## Vector index

wimepaiindex

## Index storage

## Azure AI Search

## Azure AI Search connection

lipsearch

### Include vector settings

Yes

## Schedule

OneTime

Compute

Serverless compute (Auto select)

[Back](#)

## Create vector index

Cancel

33. Wait for the process to complete might take 10 to 20 minutes:

[Add your data](#)
[PREVIEW](#)

Gain insights into your own data source. Your data is stored securely in your Azure subscription. [Learn more about how your data is protected.](#)



34. Status can also be seen in the screen as above.

35. Once the index is created, we can update the Model instruction as below screen:

**Deployment** \* [Create new deployment](#)

gpt-4o (version:2024-08-06) ▼

Give the model instructions and context ⓘ

You are an AI assistant that helps people find information. Only answer from data source provided.

Apply changes

Generate prompt

↺

36. Once completed we should be able to see the index listed as below

▼ Add your data PREVIEW

Gain insights into your own data source. Your data is stored securely in your Azure subscription. [Learn more about how your data is protected.](#)

**Index:**  
wimepaiindex25

**Search type:**  

Hybrid (vector + keyword) ▼

[Learn more about different search types](#) [↗](#)

Advanced settings >

Remove data source



37. Here is what we need to add: You are an AI assistant that helps people find information, Only answer from the data source provided.

38. Next on the Chat section try to ask questions. Now it's time to ask various different questions and observe the answers.

39. Questions: how do Wisconsin manufacturers see the business climate?

40. Here is the results:

Chat history

 | Response format 

Text

how do Wisconsin manufacturers see the business climate?

...

Wisconsin manufacturers are currently feeling less confident and increasingly uncertain about the state's business climate. Many believe that the economy is stagnant, with projected increases in gross revenues, profitability, and capital expenditures reaching their lowest levels since 2021 <sup>1</sup>. Despite this general uncertainty about the economic environment, a significant number of manufacturing executives in Wisconsin remain confident about the financial outlook for their own companies <sup>2</sup>.

2 references

1

2024-WI-Manufacturing-Report-v4-Oct-2-POST.pdf - Part 1

2

2024-WI-Manufacturing-Report-v4-Oct-2-POST.pdf - Part 2

41. As you can see the response is based on the document we uploaded.

42. So the output will have less hallucination, also we are adding new memory to the model in runtime.

43. There is no training happening to send the data into model's existing memory.

44. This provides trust to use the existing model and leverage company specific data to respond to queries or questions.

45. Based on the above process we can create and build so many different use cases. Here is also the basis to create AI Agents that can also automate process by taking decisions. Decisions can be leveraged using large language or small language models.

46. Try asking various questions, which is also called prompt.

47. We can also provide instructions to model to behave in a certain way.

## Conclusion

Now that we are able to create a simple generative ai application for manufacturing customers. There might be lots of documentation and processes in manufacturing so that we can enable them to provide answers quicker and provide instructions on specific tasks. It's huge time savings and can increase productivity and reduce downtime.