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

Dated: 01/01/2025

Audience:

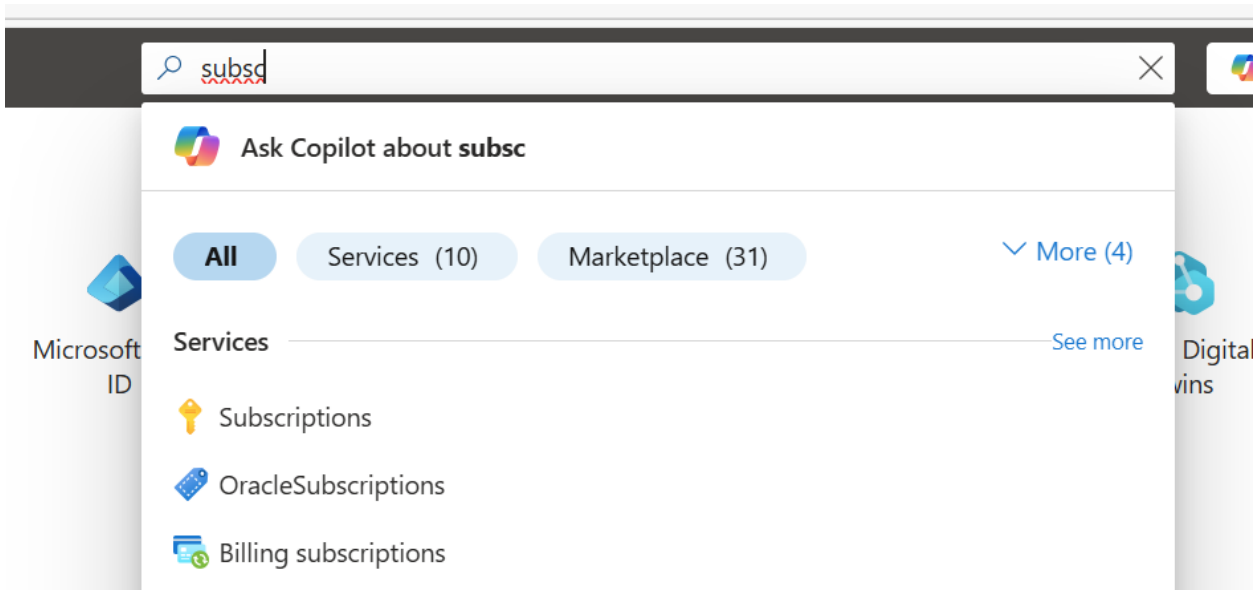
Low Code Developers

AI Engineers

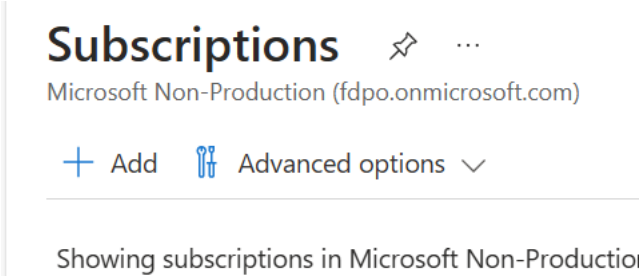
Business Enthusiast

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.



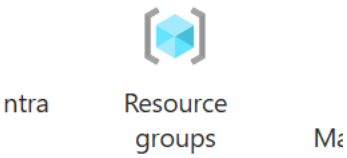
10. Follow the instructions to create a subscription with Microsoft Azure plan and select your billing profile.
11. For Plan type select Microsoft Azure plan.
12. Choose a billing profile created from the above sign up.
13. Give a name to the subscription and complete the process.
14. Once subscription 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:

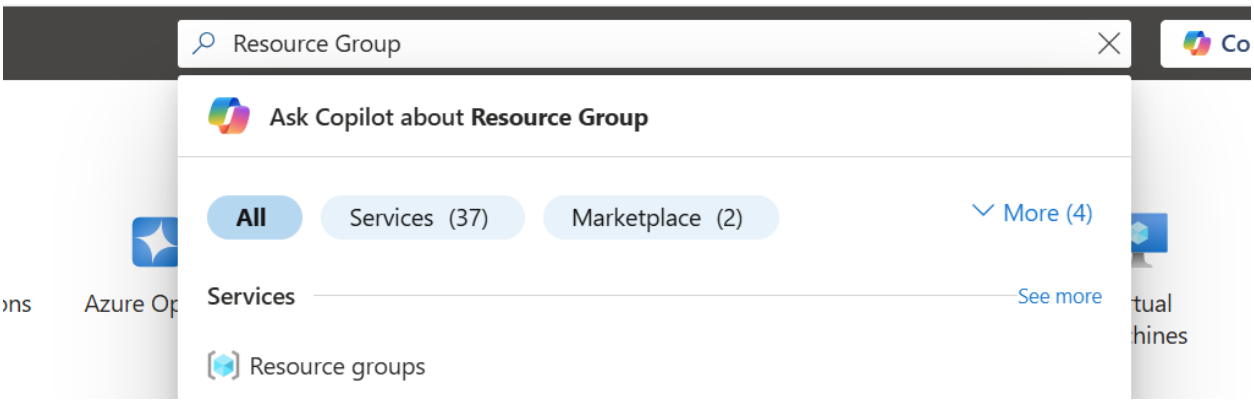
Now to create a resource group. Before creating a resource group pick a region to use like US central or East US etc. Resource group are logical separation provided in portal for us to organize resources or assets based on workload or functionality of workload.

Creating a Resource Group

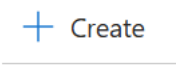
1) Log into Azure Portal https://portal.azure.com, use the above email address which was used to create azure account.
2) In the home page if you see Resource group like below then select:



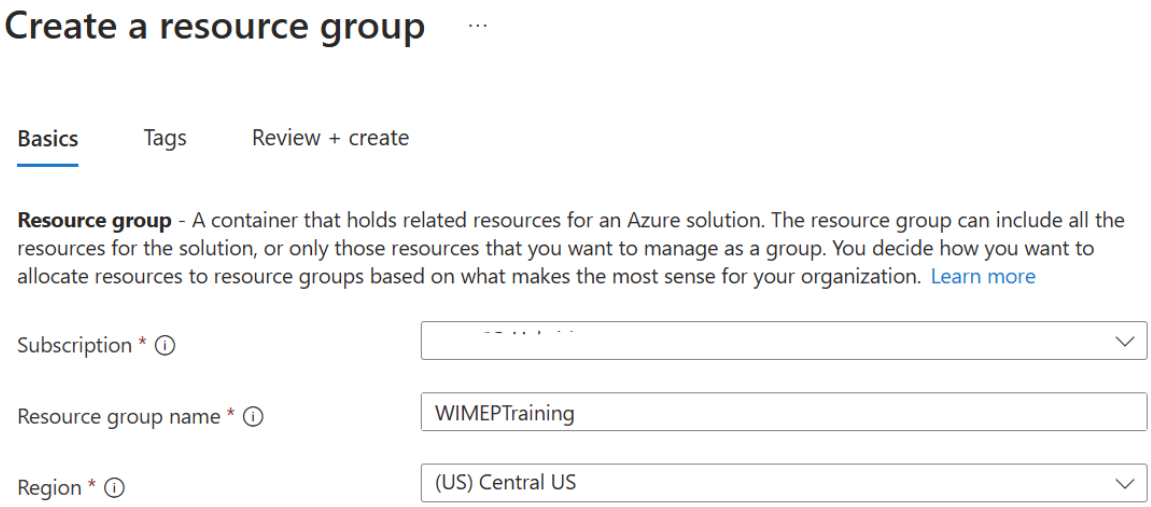
3) If there is no Resource group, then go to Search bar in the top center of the page and type Resource Group



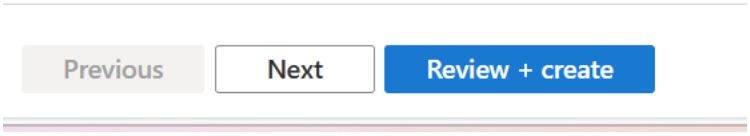
4) Select Resource group
5) Now click Create button like below:



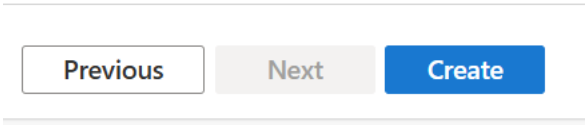
6) Fill the details like subscription, which you can choose from list created above.



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



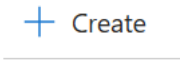
10) After validation it will show the page to create so click Create like below



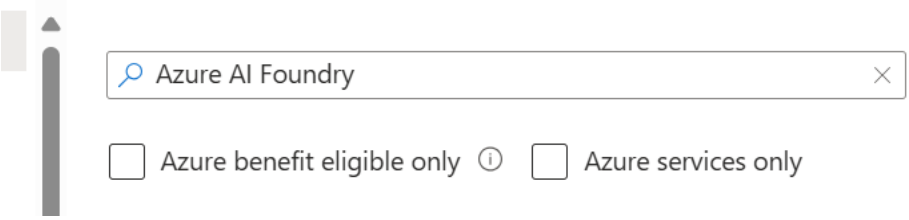
11) Wait for the resource to create.
12) Now Go to home page and select the Resource group created above from the recent activity section, or Search for resource group that we created in the top search bar. Click into the resource group and should see empty one.

Now it's time to create AI Foundry resources:

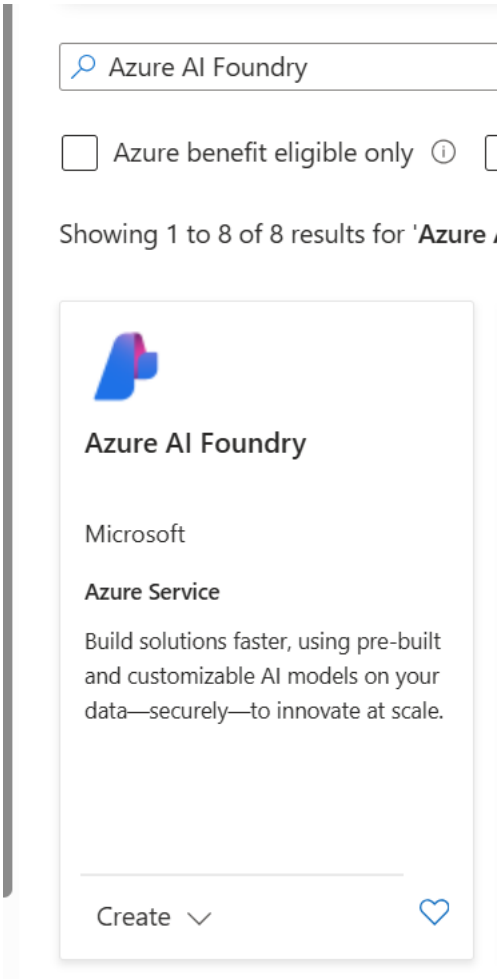
1) First we create Azure AI Foundry which is the interaction layer to create the gen ai application.
2) Click Create in the resource group selected like below:



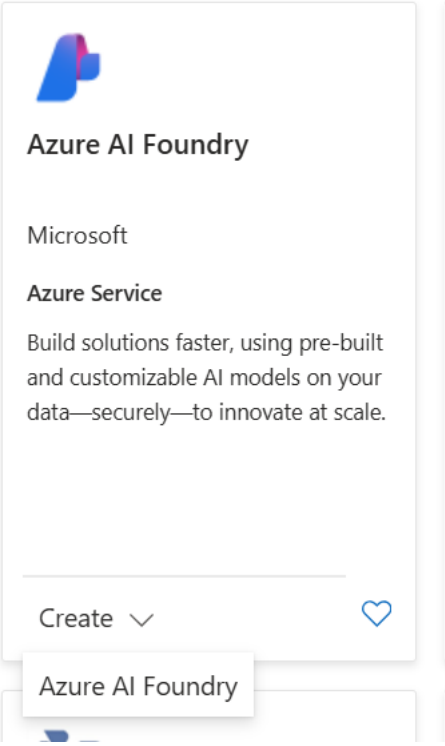
3) On the search bar shown below type in Azure AI Foundry and should see a one in the list to choose from



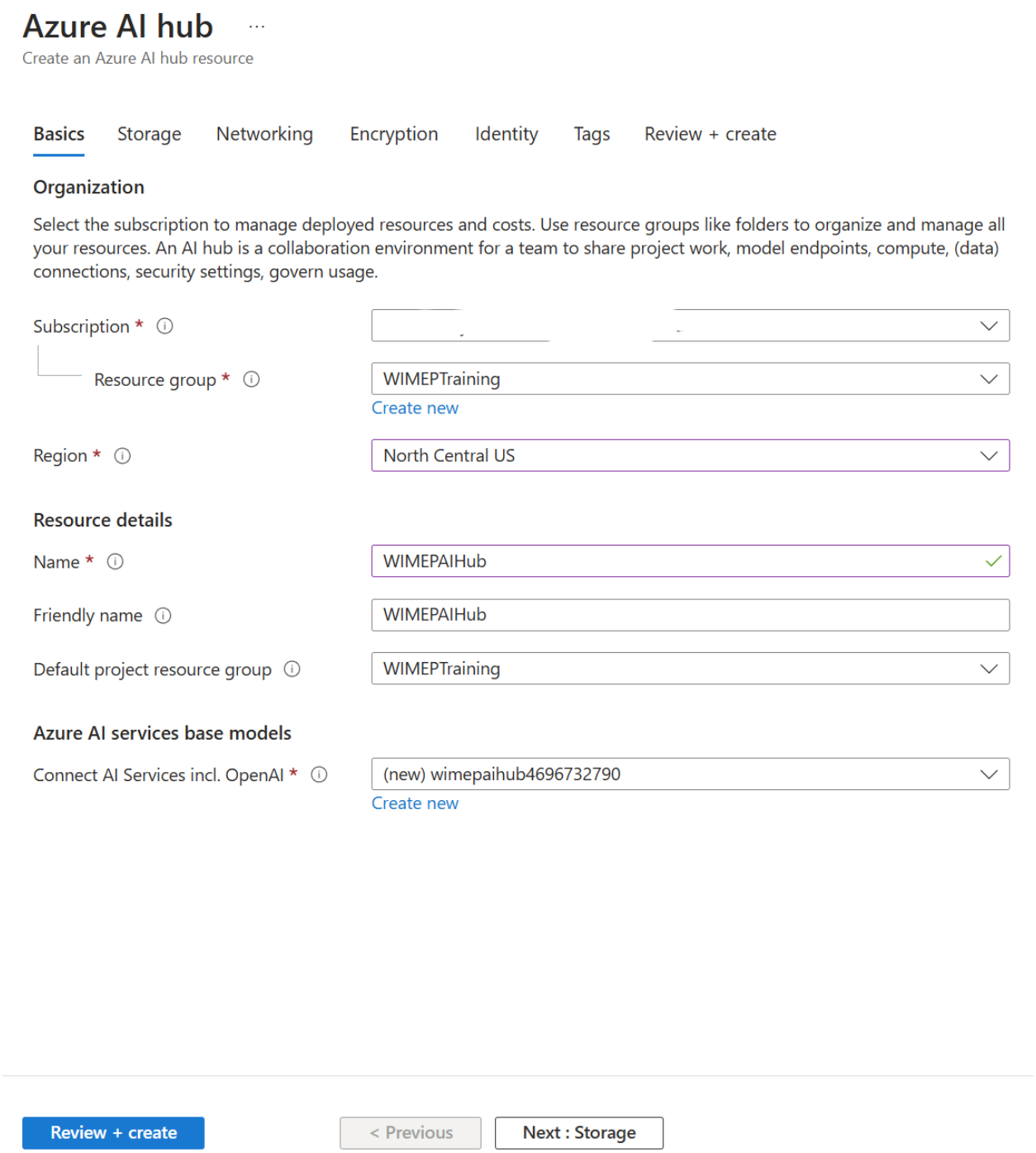
4) Or press enter or click search button to find the resource.
5) Here is the resource to create:



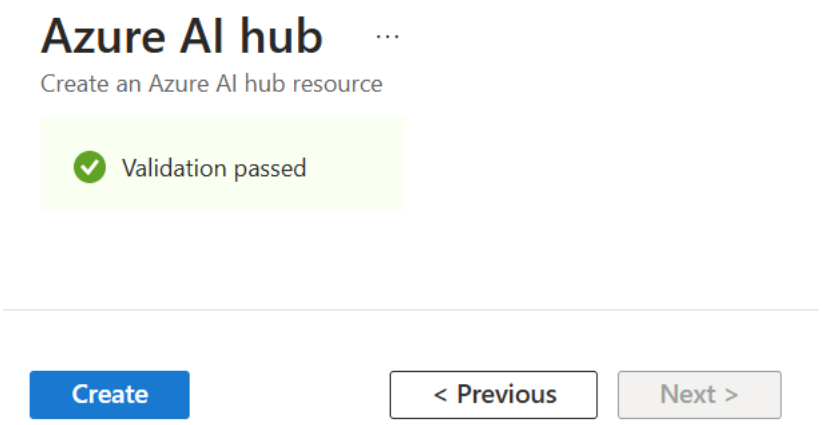
6) Now click create on the Azure AI foundry resource:



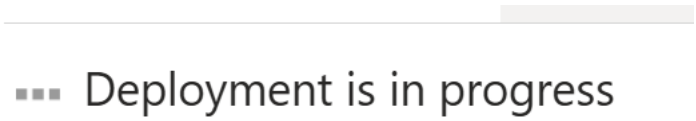
7) Click Azure AI Foundry
8) In the next screen, choose the subscription, should be autopopulated.
9) Then fill the remaining details like Resource group, region to use, name for Azure AI foundry like below and leave the other in tact.



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:



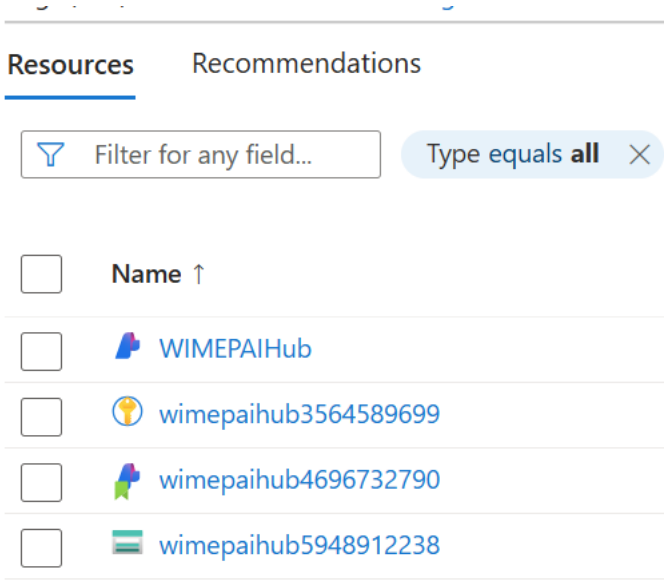
12) Validation should be passed, if you are using a corporate tenant you might not have permission to create, in that case talk to the administrator who can help you create.
13) Wait for the resource to create.



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

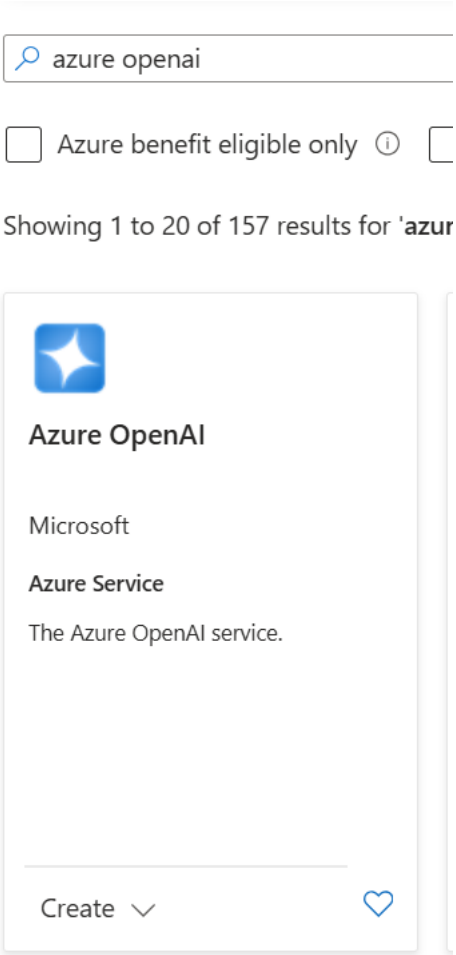
15) Once the resource is created go back to Resource group. You can click the home link shown below:

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.

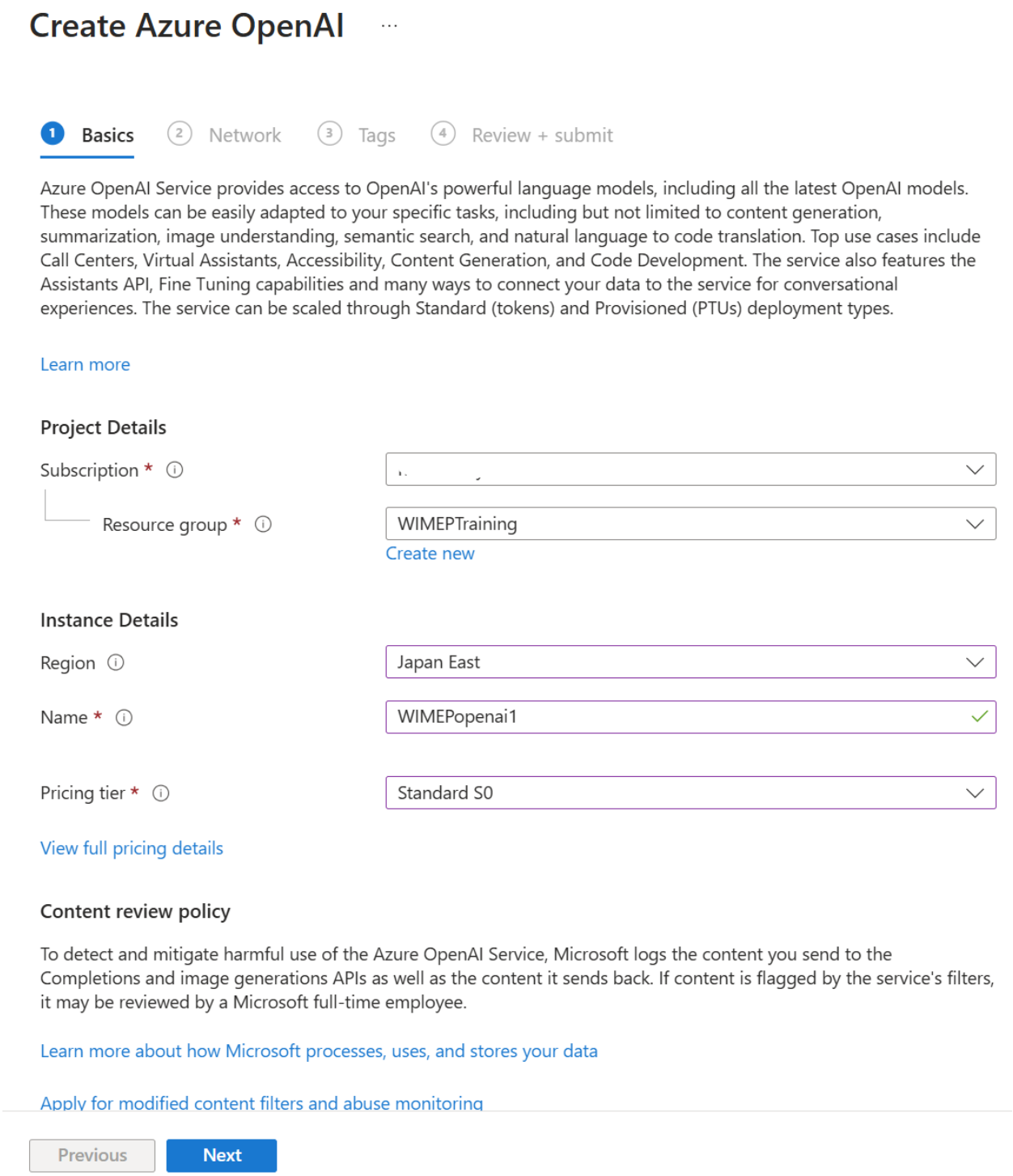


Now it's time to create Azure Open AI resource.

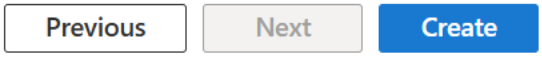
1) Click on the Create on the resource group: + Create
2) On the Search bar type: Azure Openai, and select azure openai in the list and press enter or click search button and here is what you should see:



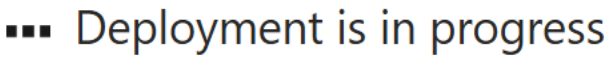
3) Click Create on the resource page
4) Below are sample screen shot shown how to fill the page:



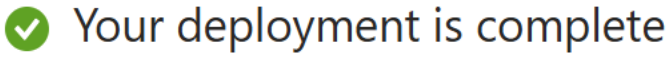
5) Provide a Name for the resource, has to be unique and also select Central US for region, in the above example I choose Japan east since I don't have quota in other regions. But for your training please Select Central US or a region in US.
6) Click Next
7) Keep Clicking Next on the other few pages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



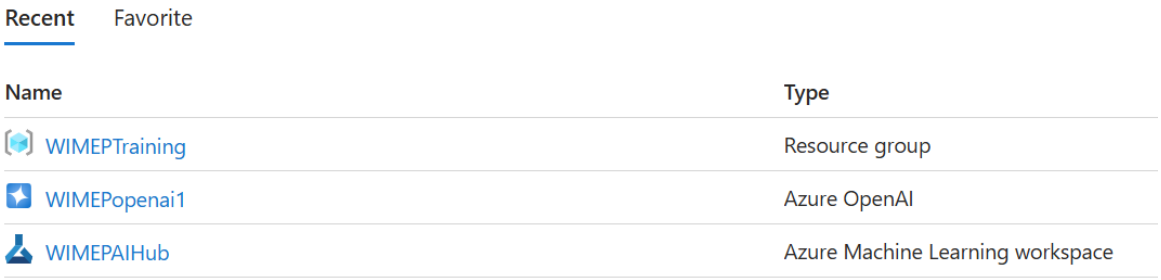
8) Wait for the resource to be created.



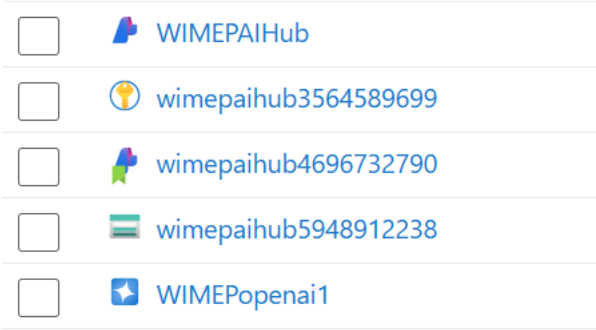
9) Resource creation usually takes like 2 to 5 minutes.
10) Once completed, here is what will be seen in the screen:



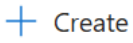
11) Now lets create Azure AI Search.
12) Go back to Home and select the Resource group:



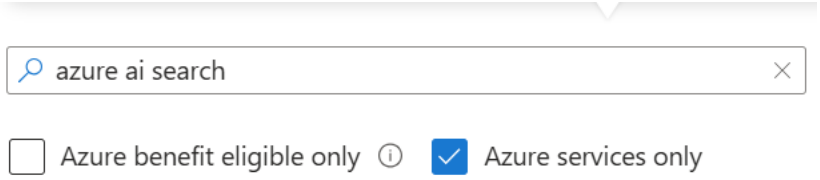
13) Click the WIMEPTraining Resource group and you will see all the resource that we created until now.



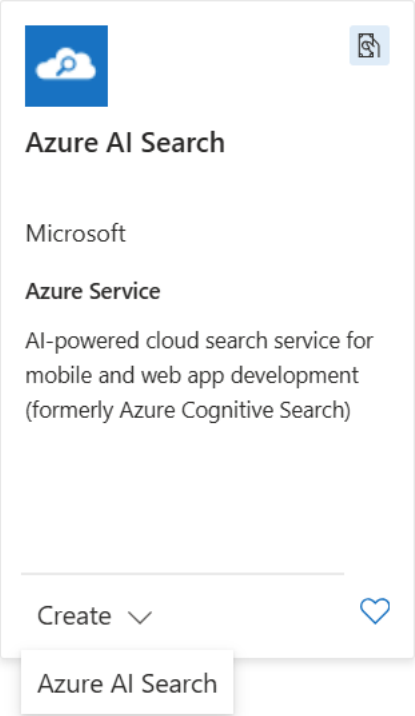
Create Azure AI Search Resource:

1) Now click Create on the top of the resource group page: 

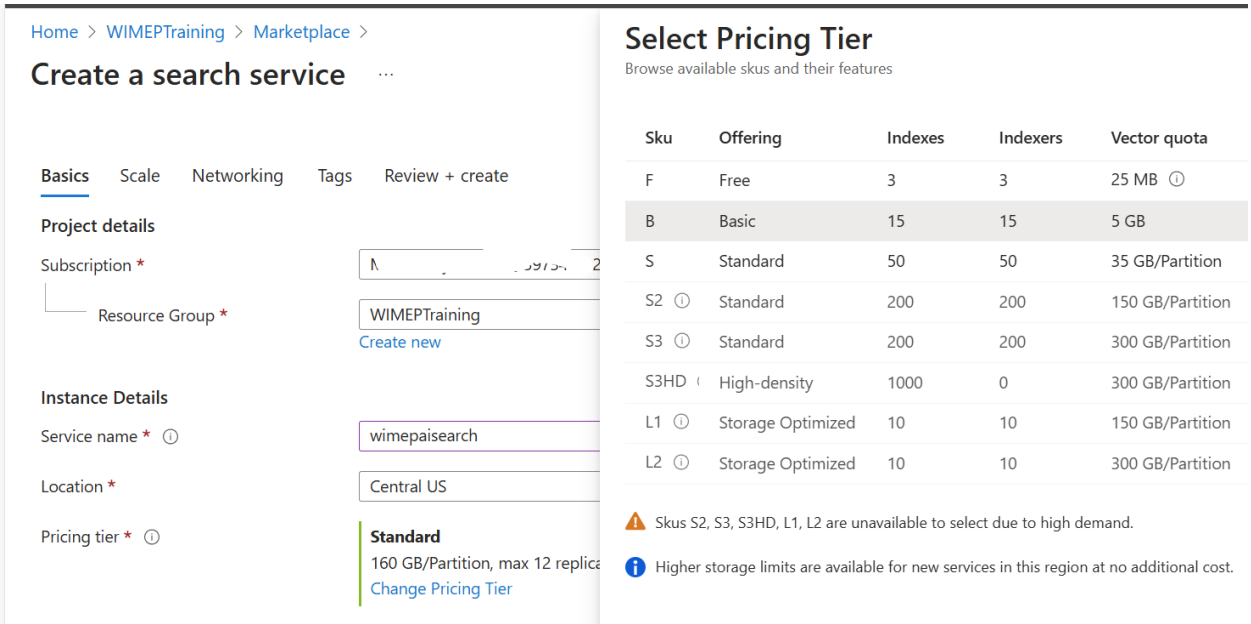
2) In the search bar type Azure AI Search.



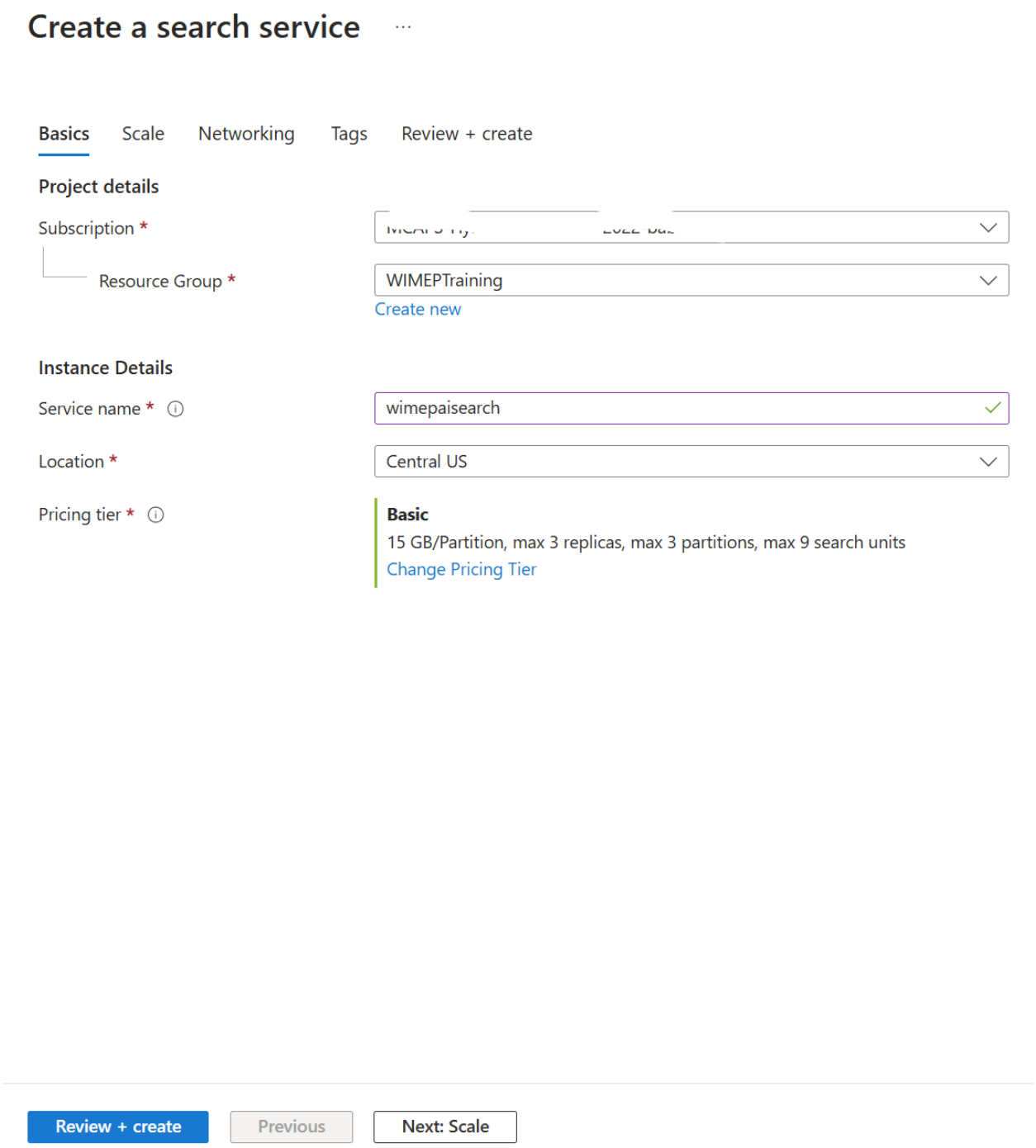
3) Scroll down until we see Azure AI Search as below image:



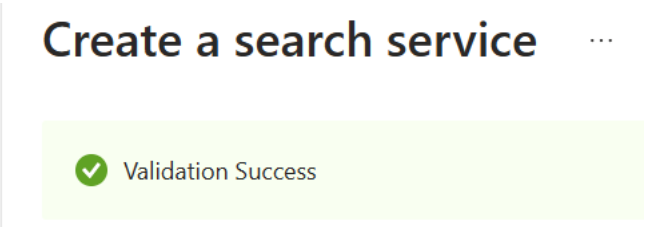
4) Then click Create and choose Azure AI Services. If you can see the AI Search on under the search bar click Azure Services only and should popup.

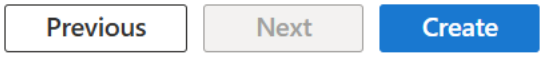


5) From the above image, please give a name for the search services under Name* section.
6) Then in the Pricing tier choose Basic.
7) Click the check box for the terms and conditions.
8) Then Click Review and Create

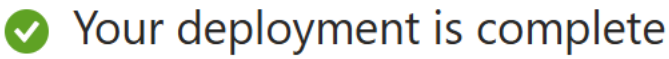


9) Wait for validation to complete and then click Create like below





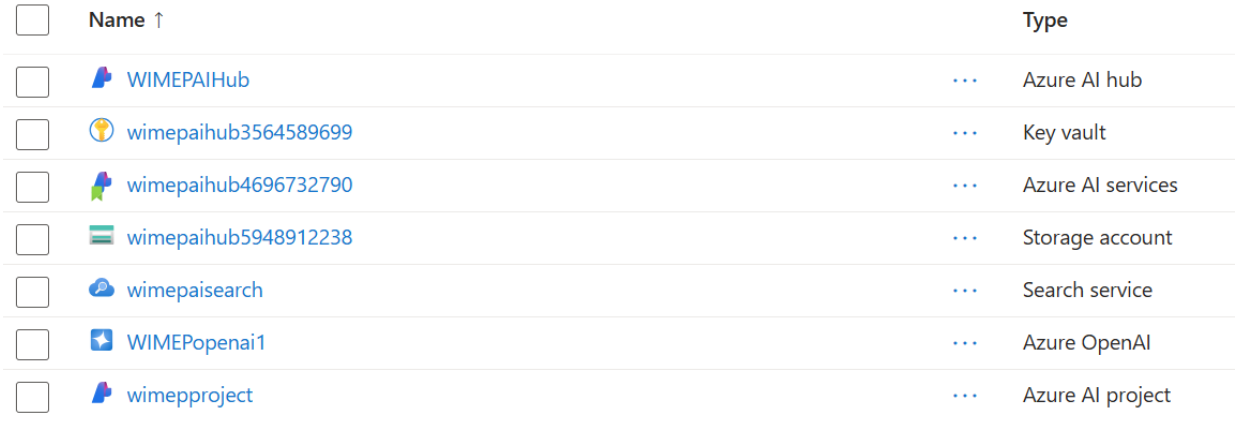
10) Wait for deployment to be completed



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.

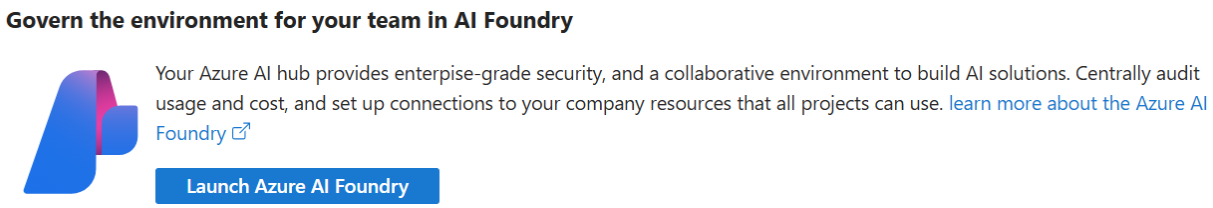
14) Now Go back to home and resource group and should see all the resources



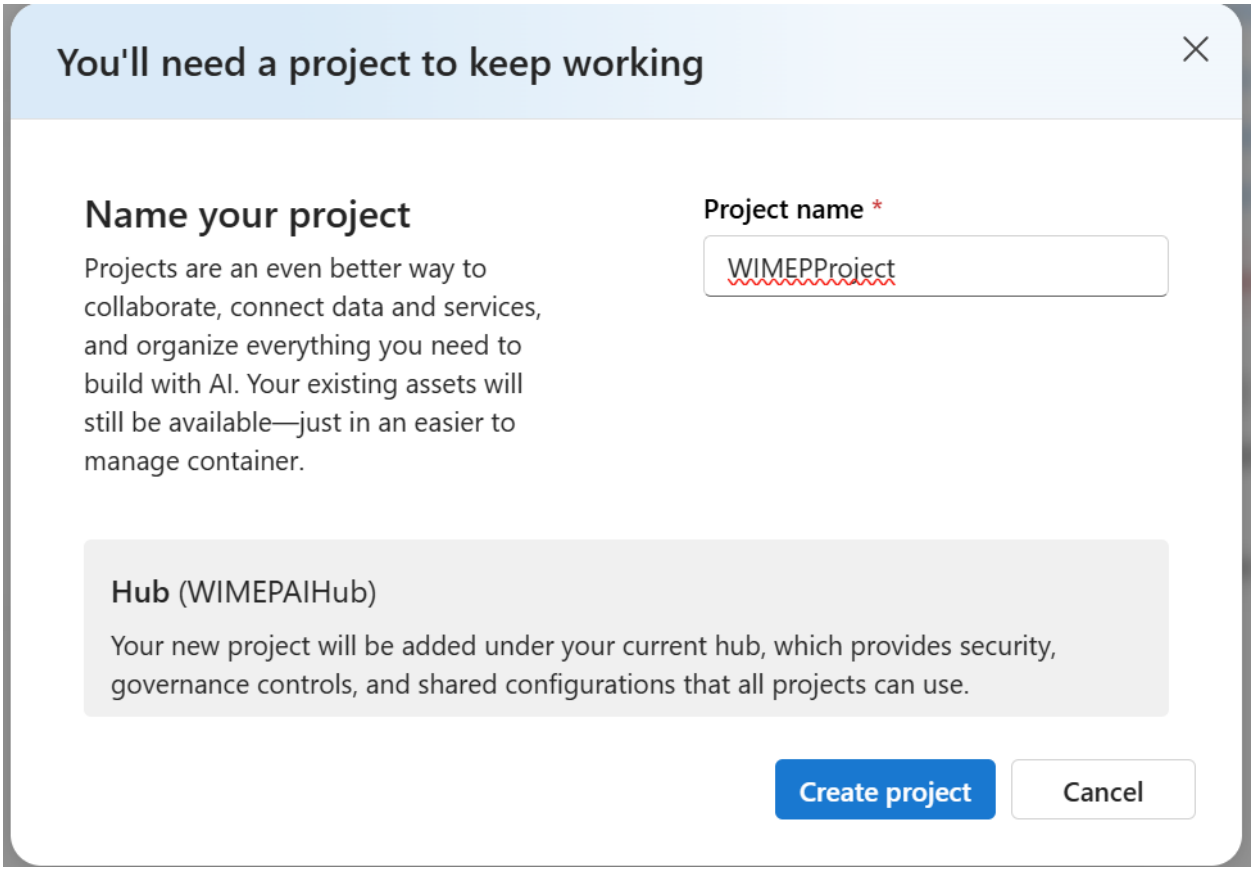
Name ↑	Type
WIMEPAIHub	Azure AI hub
wimepaihub3564589699	Key vault
wimepaihub4696732790	Azure AI services
wimepaihub5948912238	Storage account
wimepaisearch	Search service
WIMEPopenai1	Azure OpenAI
wimepproject	Azure AI project

Create Generative AI Application – No Code.

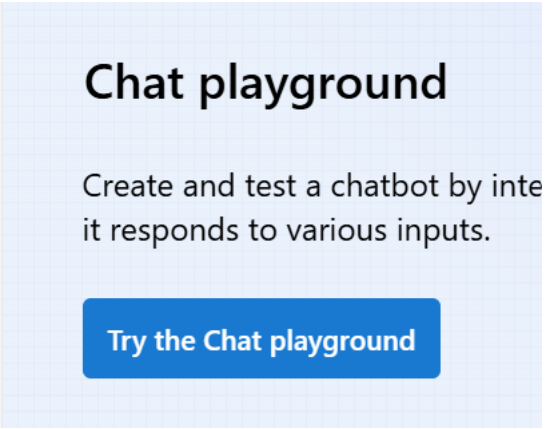
1. Click on WIMEPAIHub or the name of the Azure AI hub or foundry resource created.
2. Click Launch AI foundry:



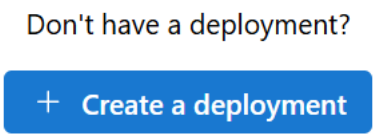
3. Will take you to a new tab, which is the user interface, we are going to use for our application development.
4. First create a Project name as below:



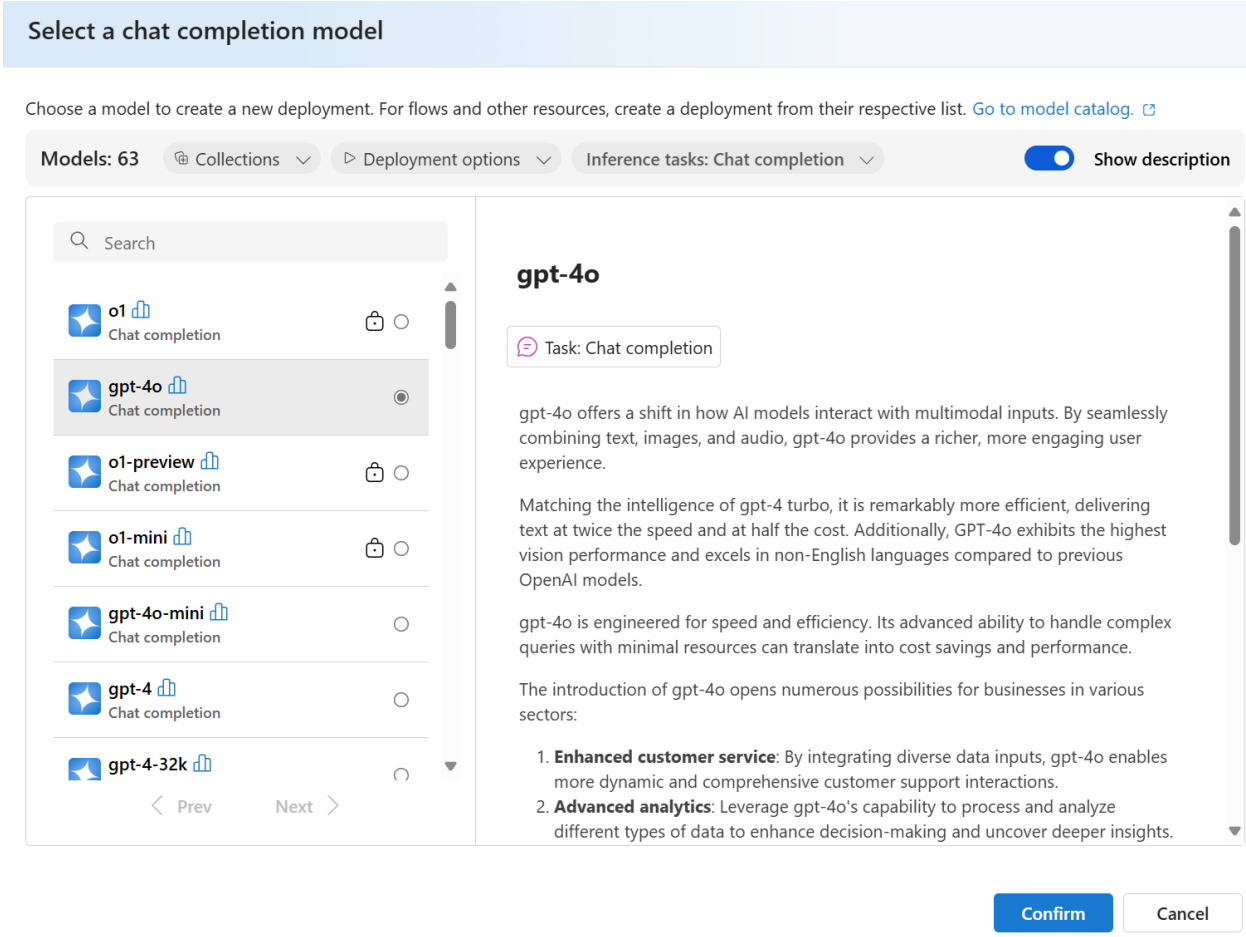
5. Wait for the project to created.
6. Next on the left Menu click on Chat playground
7. Click on Try Chat Playground as below



8. Now click Create deployment as below:



9. Now Select GPT 4o as the model:



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

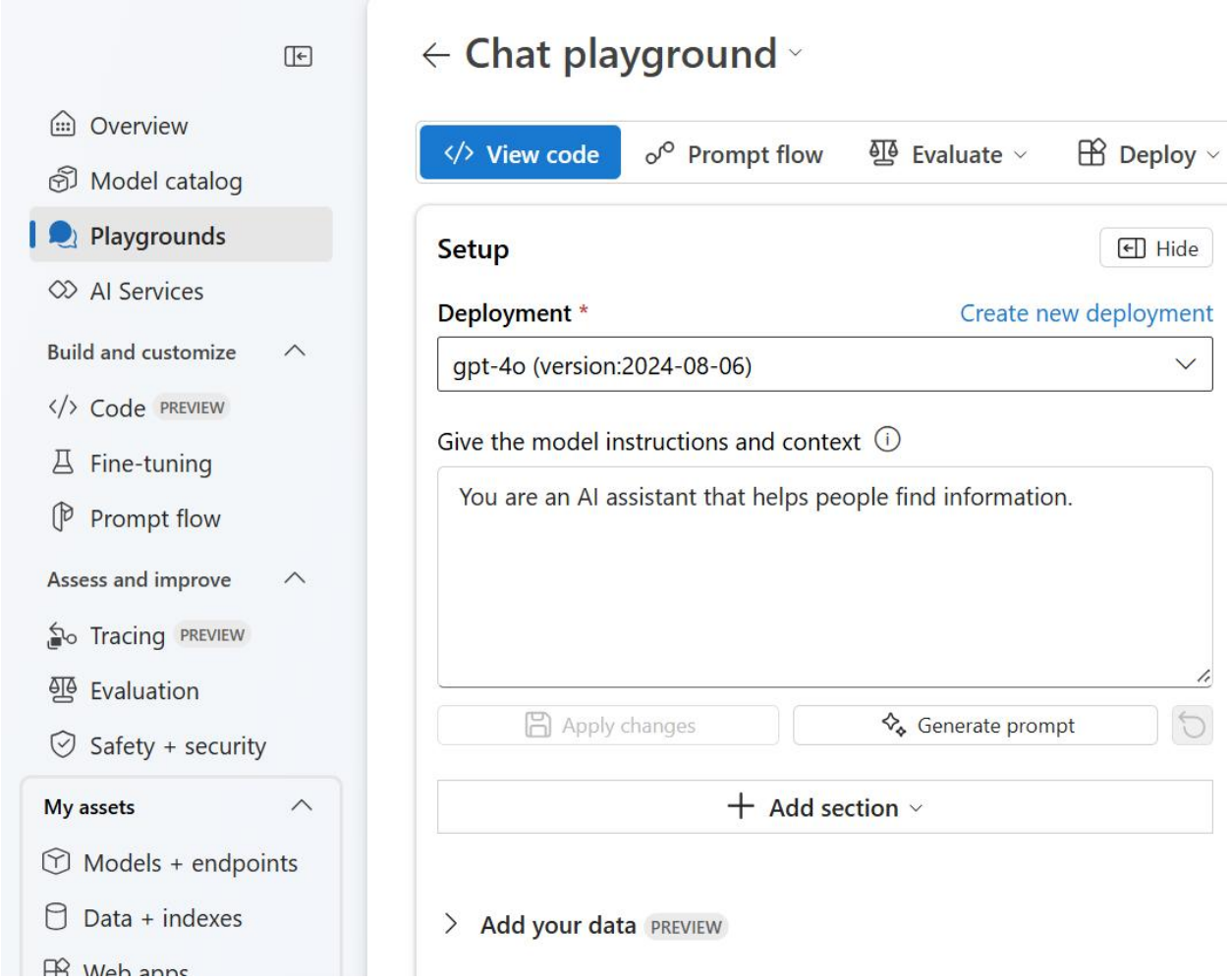
Content filter

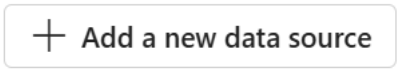
DefaultV2

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

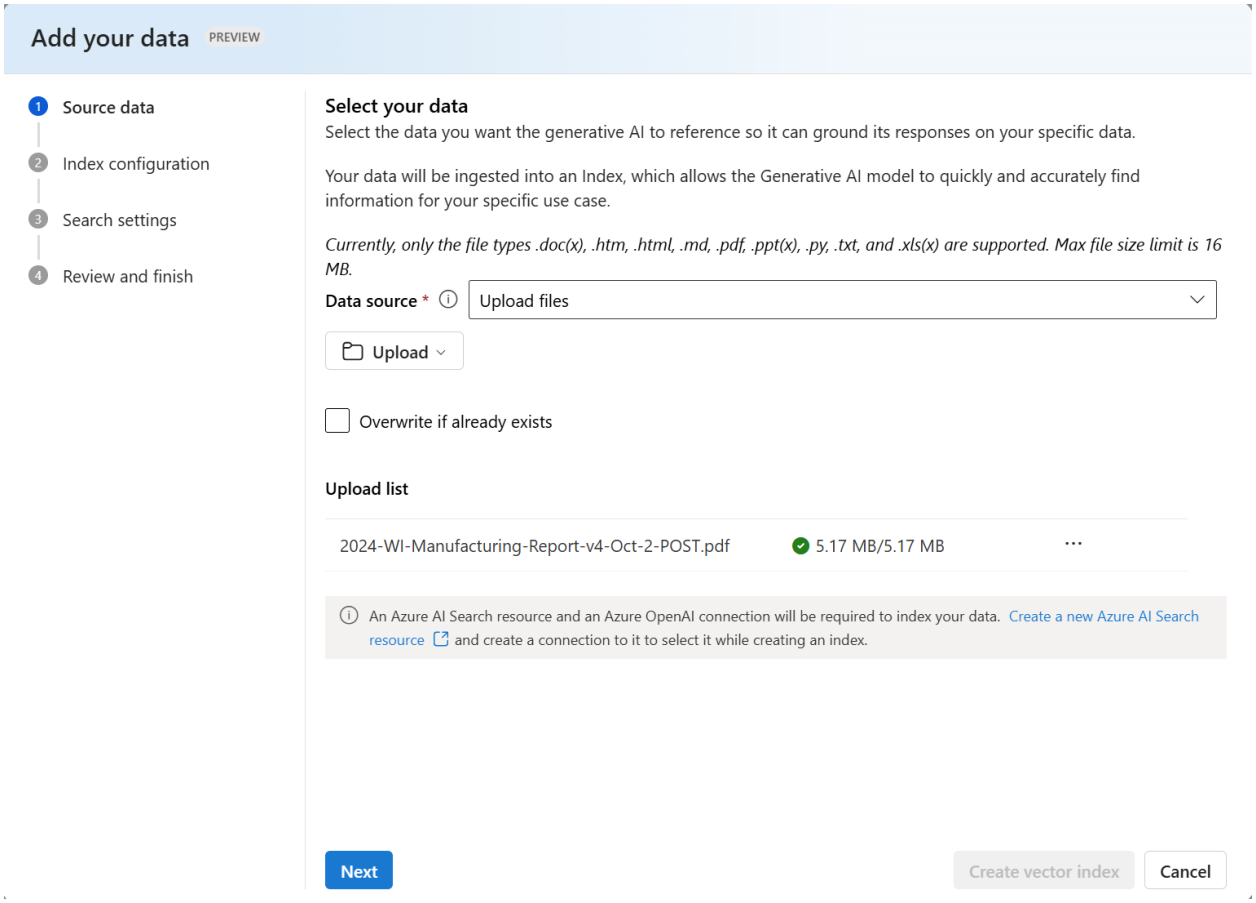
Connect and deploy | Cancel

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.

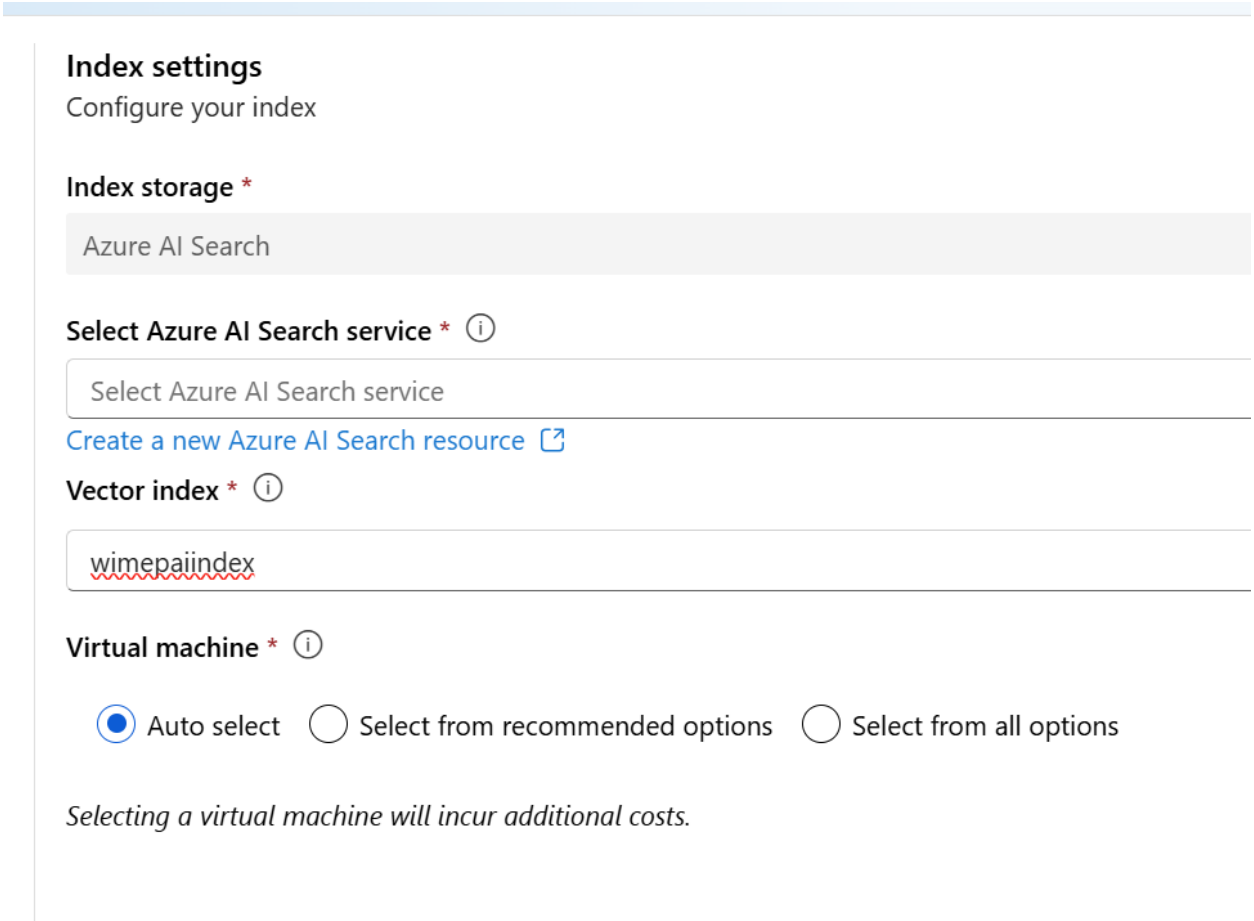


16. You can see the above gpt-4o in deployment.
17. Now we are going into Add your data, expand the data.
18. Click on Add new data source: 

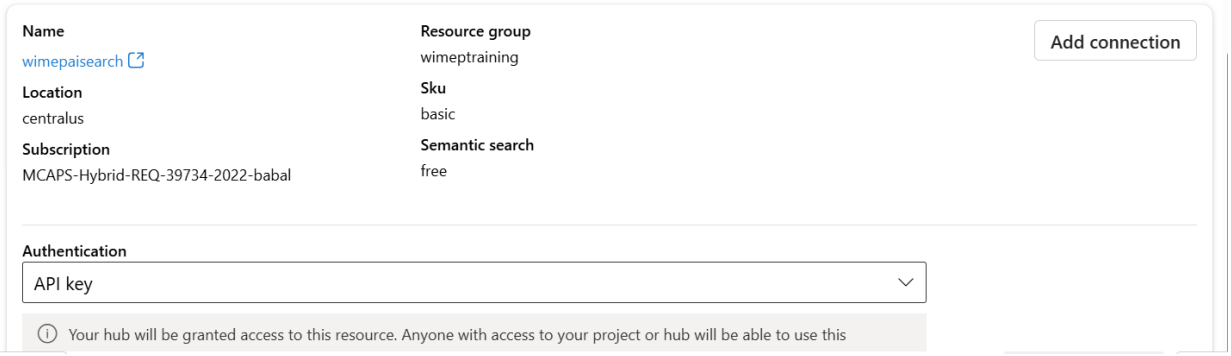
19. Now follow the instructions on the guided step by step screen.
20. First select the data source as upload files.
21. Select the PDF file to upload like the below image:



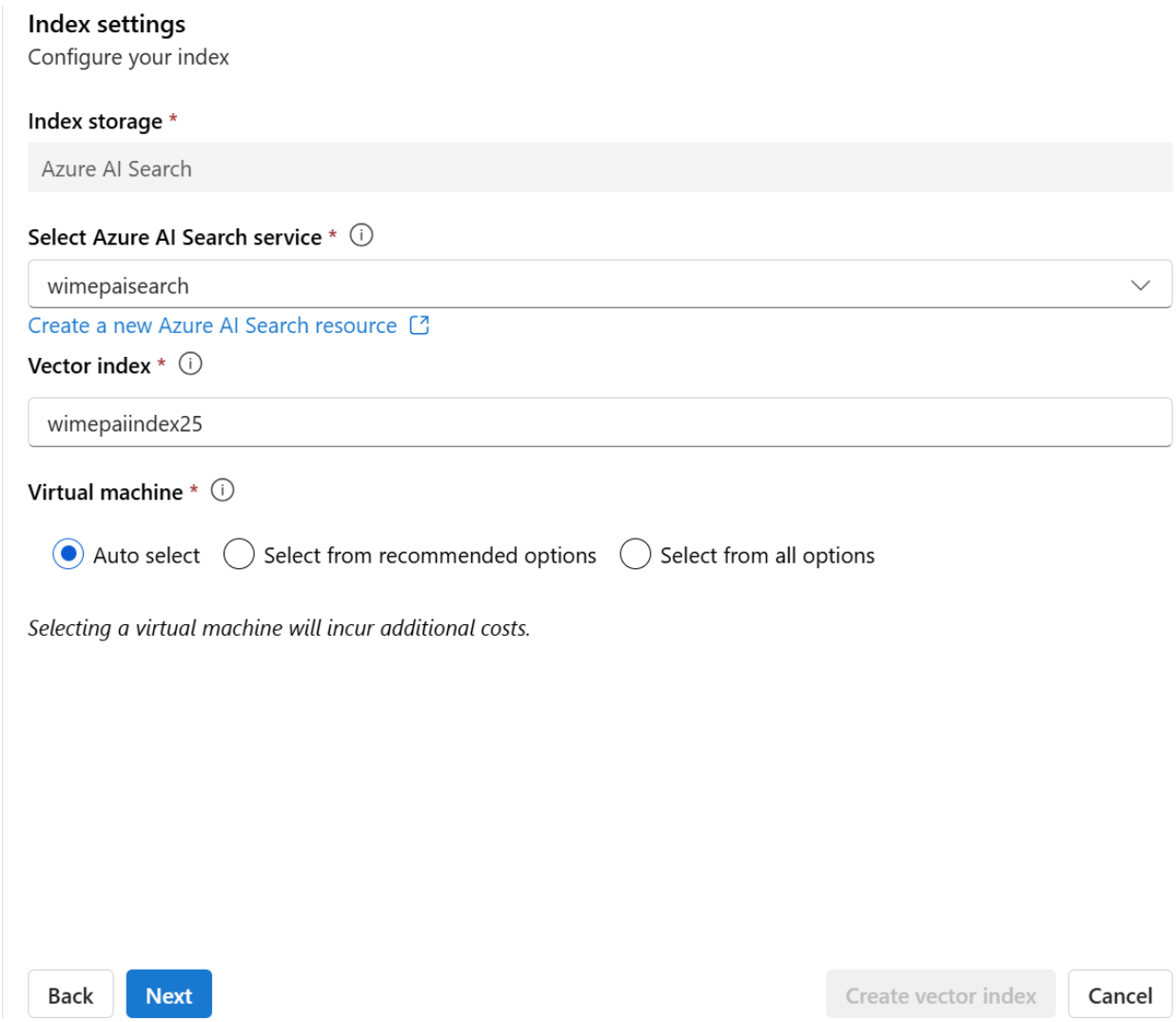
22. Click Next
23. Now give a index name, then select the Azure AI Search resource



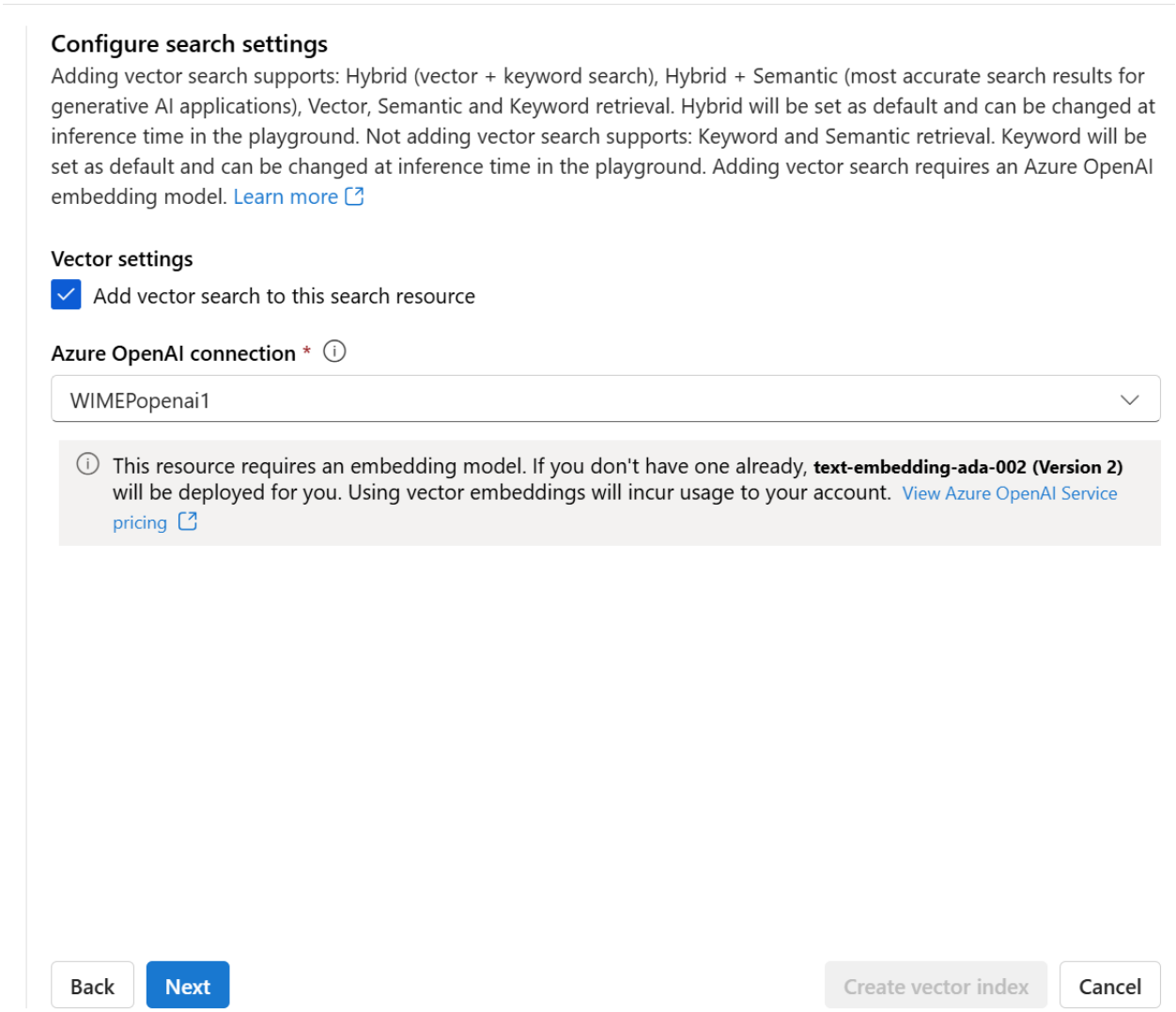
24. Select the AI Search service and click Connect if doesn't show in the list. If the list shows then select that resource.



25. Click Add connection, once the connection is added, then only the drop down will show.
26. Now you select the AI Search



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



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:

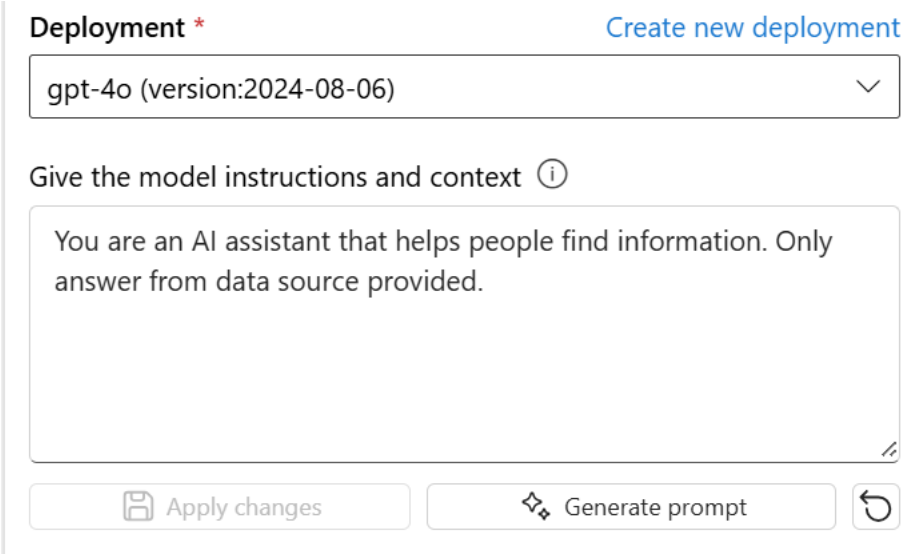
Add your data PREVIEW

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

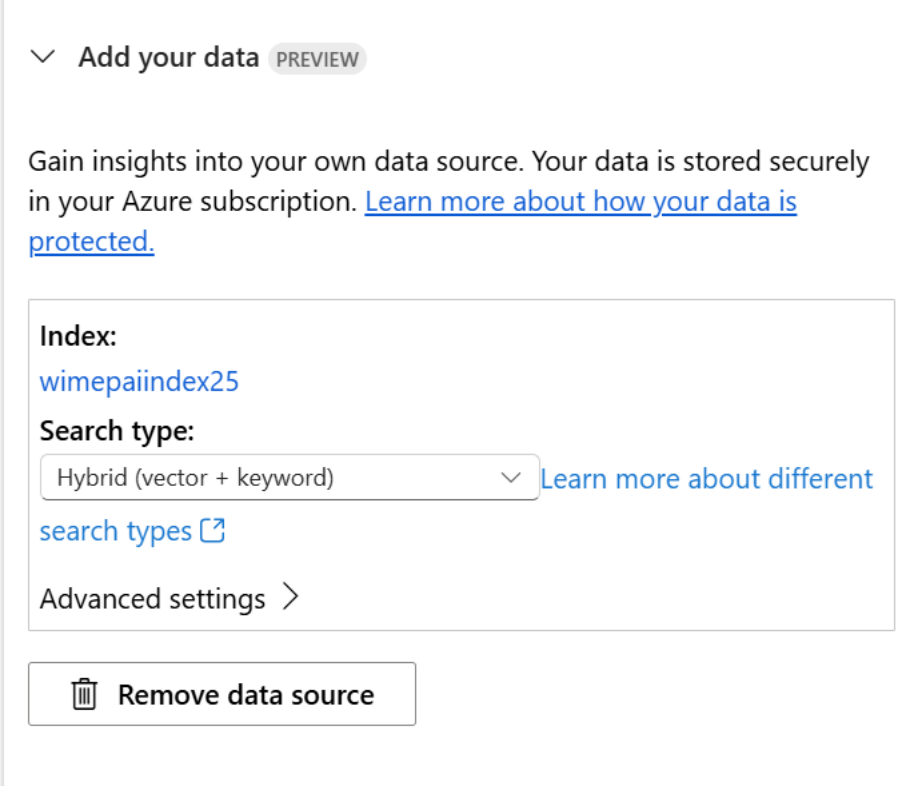
Loading...

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:

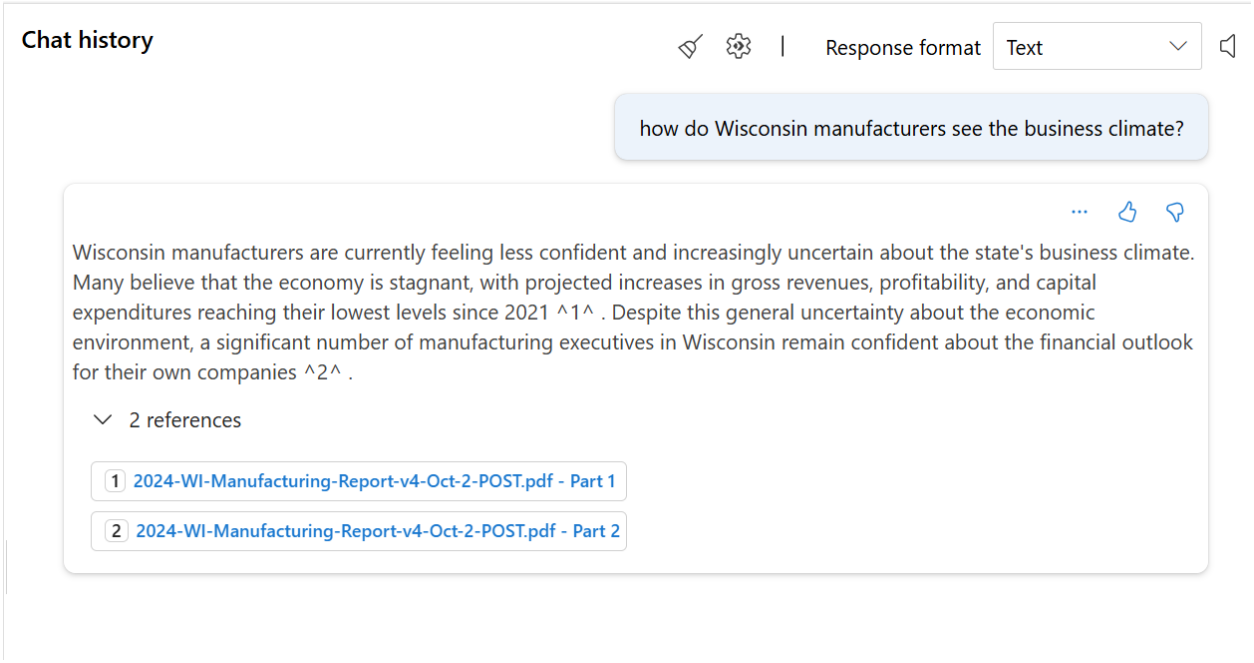


36. Once completed we should be able to see the index listed as below



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:



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.