



AI-050

Develop Generative AI Solutions with Azure OpenAI Service




Module 1

Get started with Azure OpenAI Service



Agenda



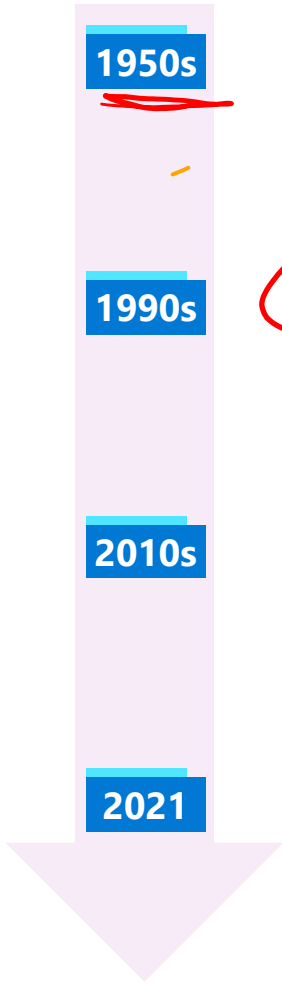
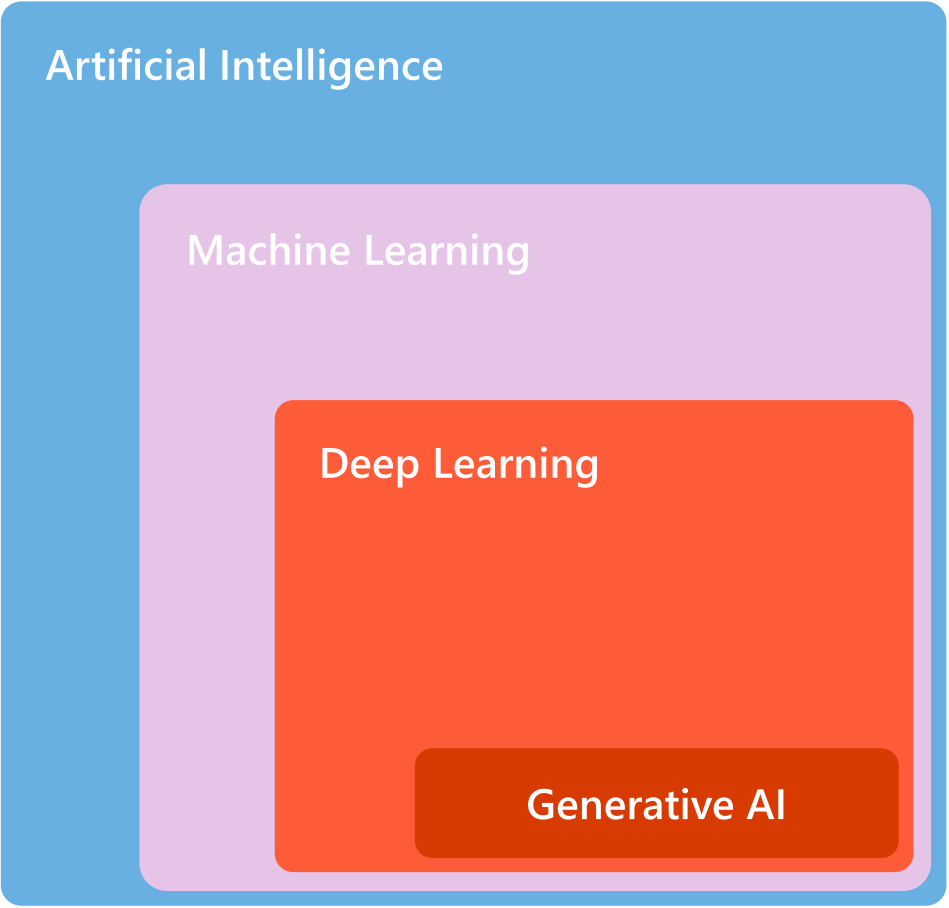
- What is generative AI?
- Provision a resource
- Deploying a model
- Using Azure OpenAI studio

What is generative AI?

LLM
SLM

GPT 3
3.5
4
4.0

4.0 mini



Artificial Intelligence

the field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence

Machine Learning

subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions

Deep Learning

a machine learning technique in which layers of neural networks are used to process data and make decisions

Generative AI

Create new written, visual, and auditory content given prompts or existing data.

?

Provision an Azure OpenAI resource in Azure

Deploy a model in Azure OpenAI Studio to use it

1. Apply for access to the Azure OpenAI service:
<https://aka.ms/oaiapply>
2. Create an **Azure OpenAI** resource in the Azure portal

Alternatively, use the Azure CLI

```
az cognitiveservices account create \  
-n MyOpenAIResource \  
-g MyResourceGroup \  
-l eastus \  
--kind OpenAI \  
--sku s0 \  
--subscription subscriptionID
```

Bash

Azure PowerShell PS .

Home > Azure AI services | Azure OpenAI >

Create Azure OpenAI ...

1 Basics 2 Network 3 Tags 4 Review + submit

Enable new business solutions with OpenAI's language generation capabilities powered by GPT-3 models. These models have been pretrained with trillions of words and can easily adapt to your scenario with a few short examples provided at inference. Apply them to numerous scenarios, from summarization to content and code generation.

[Learn more](#)

Project Details

Subscription * ⓘ

Resource group * ⓘ

[Create new](#)

Instance Details

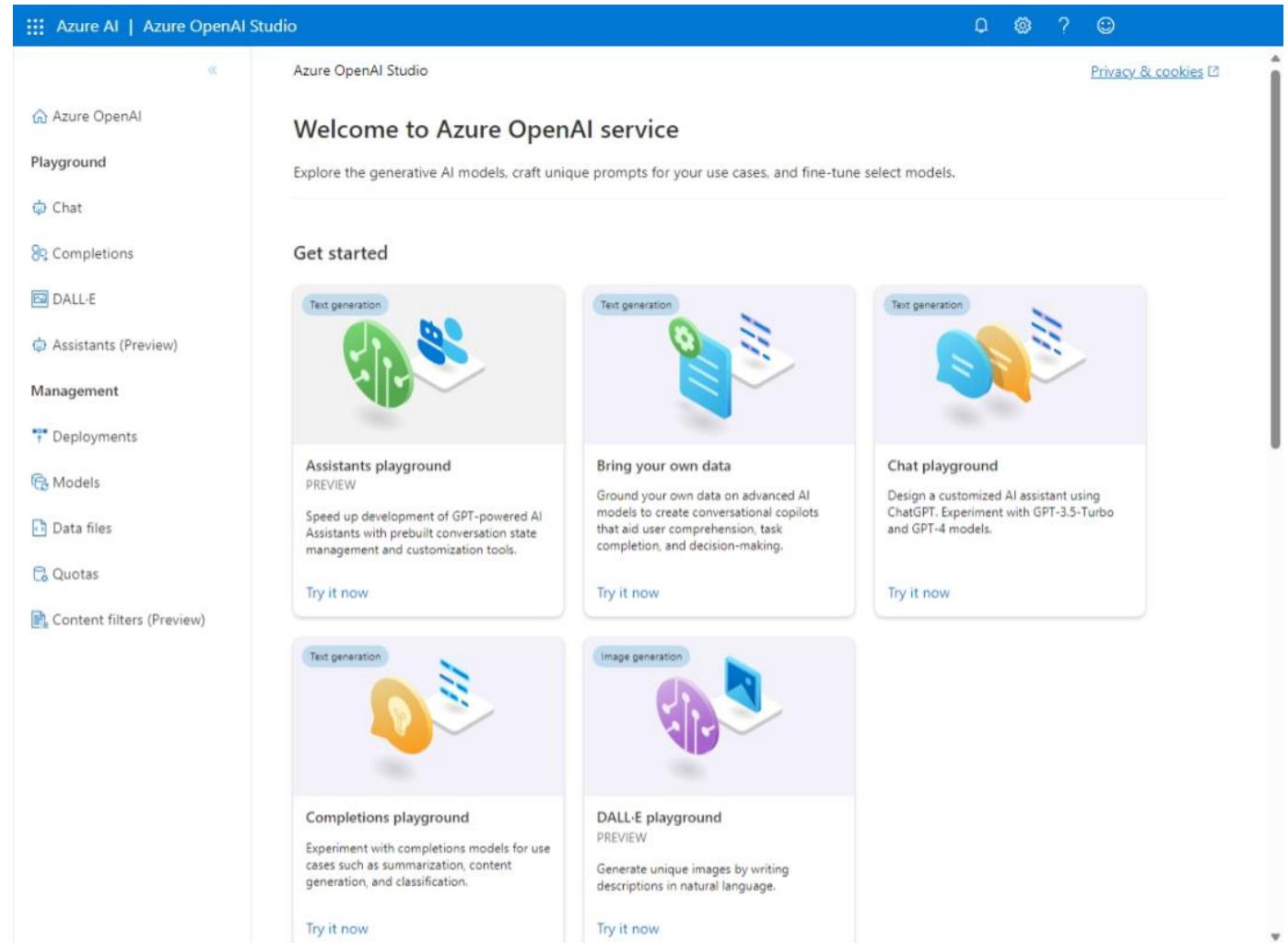
Region ⓘ

Name * ⓘ

Pricing tier * ⓘ

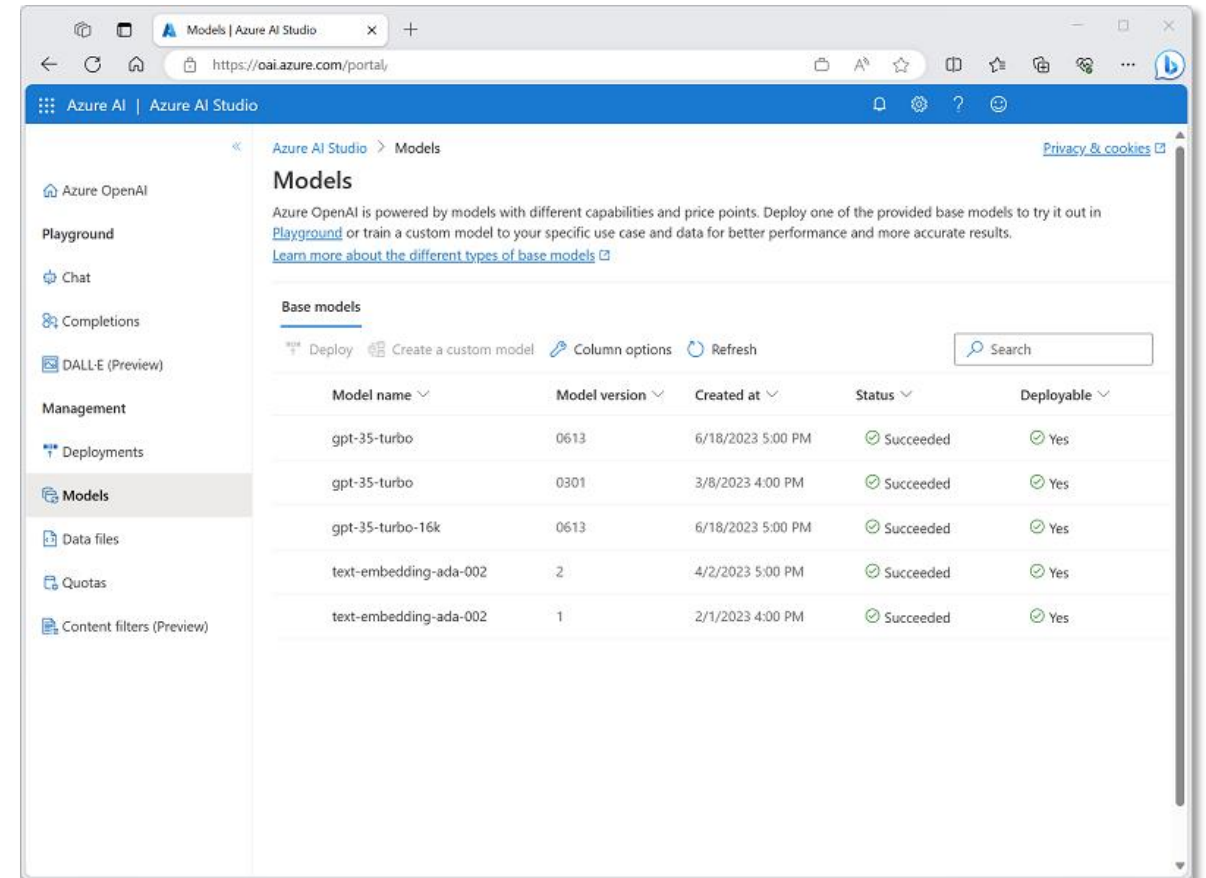
Azure OpenAI Studio

- Web portal for working with Azure OpenAI models:
<https://oai.azure.com/>
- View and deploy base models
- Connect your own data source
- Manage fine tuning and data files for custom models
- Test models in visual playgrounds:
 - **Chat** (GPT-3.5-Turbo and later models)
 - **Completions** (GPT-3 and earlier models)
 - **DALL-E** (Image generations)
 - **Assistants** (Custom and Copilot-like experiences)



Types of generative AI model

Model Family	Description
GPT-4	Newest, most capable chat-based models for language and code generation
GPT-3.5	Natural language and code-generation models
Embeddings	Models that use embeddings for specific tasks (similarity, text search, and code search)
DALL-E	Image-generation model

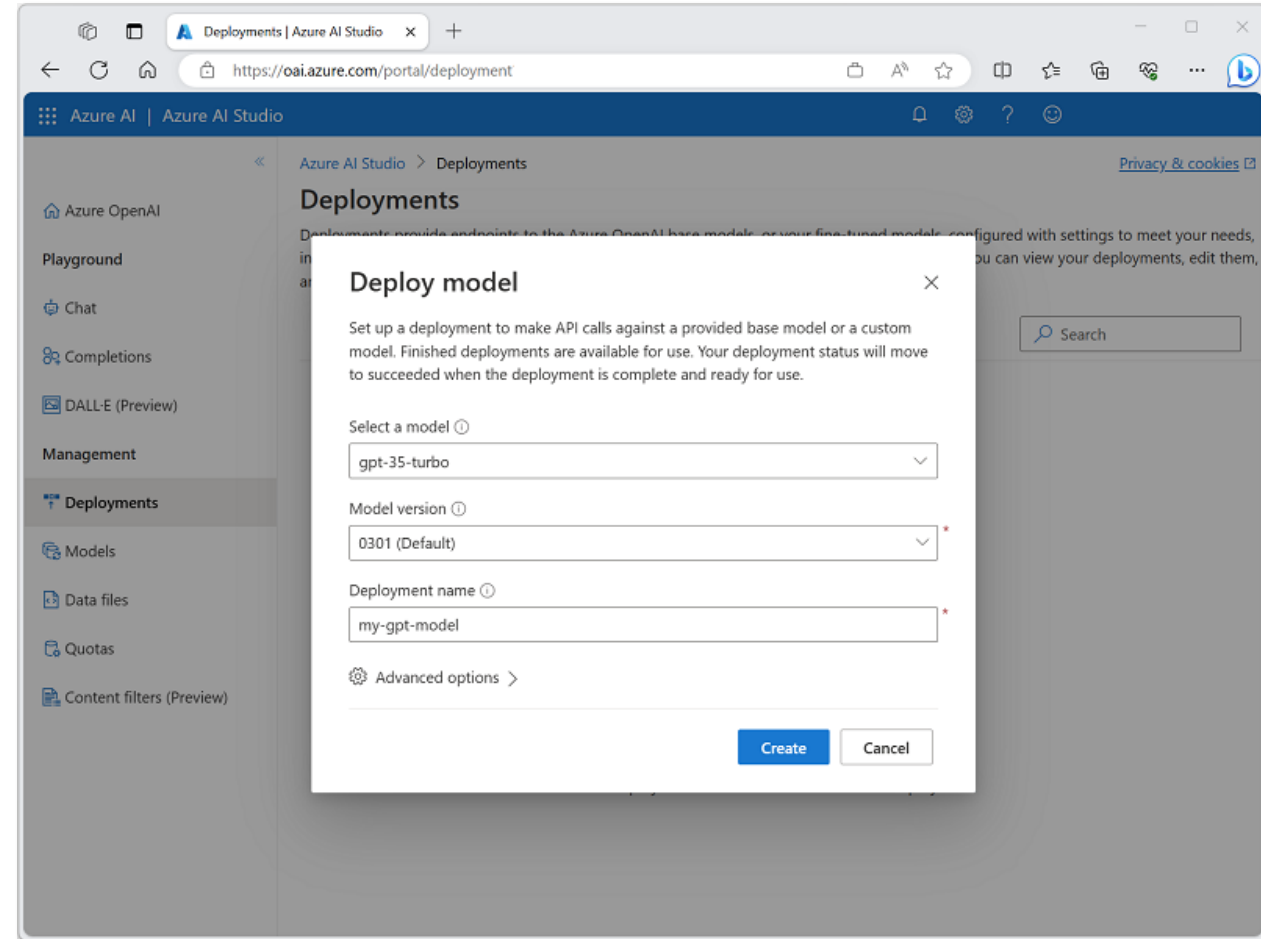


Deploying generative AI models

Deploy a model in Azure OpenAI Studio to use it

- You can deploy one or more instances of each available model
- The number of deployments depends on your quota, which you can see in the portal
- Alternatively, use the Azure CLI

```
az cognitiveservices account deployment create \  
-g myResourceGroupName \  
-n MyOpenAIResource \  
--deployment-name my-gpt-model \  
--model-name gpt-35-turbo \  
--model-version "0301" \  
--model-format OpenAI \  
--scale-settings-scale-type "Standard"
```



Using prompts to get completions from models

Task	Prompt	Completion
Classifying content	Tweet: I enjoyed the training course. Sentiment:	Positive
Generating new content	Write a poem about databases	Databases, oh databases, You keep our information safe, From the small to the large, You store our data in a place.
Transformation/Translation	English: Hello French:	Bonjour
Summarization	Scotland is [<i>long description of Scotland...</i>] Summarize the previous text	Scotland is [<i>summarized description...</i>]
Continuation	One way to grow tomatoes is to	start with seeds...
Question answering	How many moons does Earth have?	Earth has one moon.
Chat	<i>Setup, followed by messages...</i>	<i>A sequence of relevant responses</i>

Testing models in Azure OpenAI Studio playground

The screenshot displays the Azure OpenAI Studio Chat playground interface. The top navigation bar includes the Azure AI logo and the text 'Azure OpenAI Studio'. The left sidebar contains a navigation menu with options: Azure OpenAI, Playground, Chat (selected), Completions, DALL-E, Assistants (Preview), Management, Deployments, Models, Data files, Quotas, and Content filters (Preview). The main area is titled 'Chat playground' and features a 'Deploy to' button. Below the title, there are links for 'Import setup', 'Export setup', and 'Show panels'. The interface is divided into three main sections: Setup, Chat, and Configuration. The Setup section on the left includes a 'Prompt' tab, an 'Add your data' button, an 'Apply changes' button, and a 'Use a system message template' section. It also has a 'Using templates' panel with a 'Select a template' dropdown and a 'System message' text area containing the text 'You are an AI assistant that helps people find information.' The Chat section in the center features a 'Start chatting' button and a text input area with the placeholder 'Type user query here. (Shift + Enter for new line)'. The Configuration section on the right includes a 'Parameters' tab with sliders for 'Max response' (set to 800), 'Temperature' (set to 0.7), 'Top P' (set to 0.95), 'Stop sequence' (set to 'Stop sequences'), 'Frequency penalty' (set to 0), and 'Presence penalty' (set to 0). It also shows a 'Current token count' of 11/4000 and a 'Learn more' link.

Azure AI | Azure OpenAI Studio

Azure OpenAI Studio > Chat playground

Chat playground

Deploy to

Import setup Export setup Show panels

Setup

Prompt Add your data

Apply changes

Use a system message template

Using templates

Use a template to get started, or just start writing your own system message below. Want some tips? [Learn more](#)

Select a template

System message

You are an AI assistant that helps people find information.

Examples

Using examples

Add examples to show the chat what responses you want. It will try to mimic any responses you add here so make sure they match the rules you laid out in the system message.

Start chatting

Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.

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

Configuration

Deployment Parameters

Max response 800

Temperature 0.7

Top P 0.95

Stop sequence Stop sequences

Frequency penalty 0

Presence penalty 0

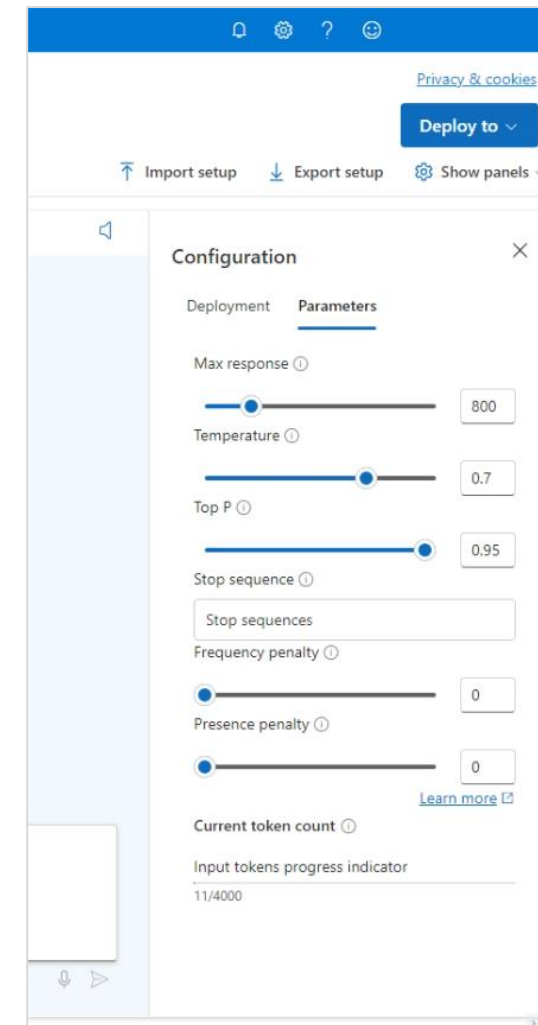
Learn more

Current token count 11/4000

Input tokens progress indicator

Model parameters and tokens in Azure OpenAI

- Control model behavior through parameters in studio pane or API call
- Parameters include:
 - **Max Response:** Limit on the number of tokens the response can include
 - **Temperature:** Controls randomness, with a higher number producing a less deterministic response
 - **Top P:** Controls randomness similarly but in a different way than temperature. If adjusting these two values, try one or the other but not both
 - **Frequency penalty:** New tokens are penalized for their existing frequency in the text so far, reducing the likelihood to repeat the same line
 - **Presence penalty:** New tokens are penalized whether they appear in the text so far, increasing likelihood of talking about new topics
- **Tokens** are text measurements, roughly four English characters long. Tokens are used for measuring model capacity, quotas, and prompt or response length



Exercise: Get started with Azure OpenAI Service



Use the hosted lab environment if provided, or view the lab instructions at the link below:

<https://aka.ms/mslearn-get-started-azure-openai>

Knowledge check



1 What do you need in order to test a generative AI model using the Azure OpenAI Service Studio?

- ☐ A deployed model name and Azure command line interface
- ☐ An Azure OpenAI resource and an Azure Cognitive Services resource
- ☒ An Azure OpenAI resource, a deployed model, and a playground

2 Which parameter could you adjust to change the randomness or creativeness of completions

- ☒ Temperature
- ☐ Frequency penalty
- ☐ Stop sequence

3 Which Azure OpenAI Studio playground is able to support conversational interactions that consist of a sequence of messages?

- ☐ Completions
- ☒ Chat
- ☐ Bot

Learning Recap

In this module, we:

Introduced generative AI and how it relates to AI and machine learning

Provisioned Azure OpenAI resources

Deployed OpenAI models

Generated completions in Azure OpenAI studio

Resources

Get started with Azure OpenAI Service

<https://aka.ms/mslearn-start-azure-openai>



On Prem
Kerberos

AD

sync.

SP

User
Devices
App

Entra ID (Azure AD)

Tenant

Root

M365

Thank you.

(Permission)
Role "Owner"

Policies

Sub \$

Management Group

Subscription \$

Resource Group

Resource

VM

Storage
Account

OpenAI
Account

Region

Bicep (TF)
json
Portal Templates
ARM
SDK
Azure