# Azure OpenAI Service Documentation

Learn how to use Azure OpenAI's powerful language models including the GPT-3, Codex and Embeddings model series for content generation, summarization, semantic search, and natural language to code translation.

OVERVIEW
**What is Azure OpenAI Service?**

QUICKSTART
**Quickstarts**

HOW-TO GUIDE
**Create a resource**

TUTORIAL
**Embeddings**

HOW-TO GUIDE
**Completions**

TRAINING
**Intro to Azure OpenAI training**

CONCEPT
**Azure OpenAI Models**

REFERENCE
**Support and help options**

## Additional resources

### Azure OpenAI

- Azure OpenAI Studio
- Region support
- Quotas and limits
- Apply for access to Azure OpenAI

### Video

- Combining OpenAI models with the power of Azure

## Reference

REST API

Terms of use

## Tools

Azure CLI

PowerShell

# What is Azure OpenAI Service?

Article • 07/18/2023

Azure OpenAI Service provides REST API access to OpenAI's powerful language models including the GPT-4, GPT-35-Turbo, and Embeddings model series. In addition, the new GPT-4 and gpt-35-turbo model series have now reached general availability. These models can be easily adapted to your specific task including but not limited to content generation, summarization, semantic search, and natural language to code translation. Users can access the service through REST APIs, Python SDK, or our web-based interface in the Azure OpenAI Studio.

## Features overview

| Feature | Azure OpenAI |
| --- | --- |
| Models available | **GPT-4 series**<br>**GPT-35-Turbo series**<br>Embeddings series<br>Learn more in our Models page. |
| Fine-tuning | Ada<br>Babbage<br>Curie<br>Cushman<br>Davinci<br>**Fine-tuning is currently unavailable to new customers**. |
| Price | Available here |
| Virtual network support & private link support | Yes, unless using Azure OpenAI on your data. |
| Managed Identity | Yes, via Azure Active Directory |
| UI experience | **Azure portal** for account & resource management,<br>**Azure OpenAI Service Studio** for model exploration and fine tuning |
| Model regional availability | Model availability |
| Content filtering | Prompts and completions are evaluated against our content policy with automated systems. High severity content will be filtered. |

# Responsible AI

At Microsoft, we're committed to the advancement of AI driven by principles that put people first. Generative models such as the ones available in Azure OpenAI have significant potential benefits, but without careful design and thoughtful mitigations, such models have the potential to generate incorrect or even harmful content. Microsoft has made significant investments to help guard against abuse and unintended harm, which includes requiring applicants to show well-defined use cases, incorporating Microsoft's principles for responsible AI use, building content filters to support customers, and providing responsible AI implementation guidance to onboarded customers.

# How do I get access to Azure OpenAI?

How do I get access to Azure OpenAI?

Access is currently limited as we navigate high demand, upcoming product improvements, and Microsoft's commitment to responsible AI. For now, we're working with customers with an existing partnership with Microsoft, lower risk use cases, and those committed to incorporating mitigations.

More specific information is included in the application form. We appreciate your patience as we work to responsibly enable broader access to Azure OpenAI.

Apply here for access:

Apply now

# Comparing Azure OpenAI and OpenAI

Azure OpenAI Service gives customers advanced language AI with OpenAI GPT-4, GPT-3, Codex, and DALL-E models with the security and enterprise promise of Azure. Azure OpenAI co-develops the APIs with OpenAI, ensuring compatibility and a smooth transition from one to the other.

With Azure OpenAI, customers get the security capabilities of Microsoft Azure while running the same models as OpenAI. Azure OpenAI offers private networking, regional availability, and responsible AI content filtering.

# Key concepts

## Prompts & completions

The completions endpoint is the core component of the API service. This API provides access to the model's text-in, text-out interface. Users simply need to provide an input **prompt** containing the English text command, and the model will generate a text **completion**.

Here's an example of a simple prompt and completion:

> **Prompt**: `""" count to 5 in a for loop """`
>
> **Completion**: `for i in range(1, 6): print(i)`

## Tokens

Azure OpenAI processes text by breaking it down into tokens. Tokens can be words or just chunks of characters. For example, the word “hamburger” gets broken up into the tokens “ham”, “bur” and “ger”, while a short and common word like “pear” is a single token. Many tokens start with a whitespace, for example “ hello” and “ bye”.

The total number of tokens processed in a given request depends on the length of your input, output and request parameters. The quantity of tokens being processed will also affect your response latency and throughput for the models.

## Resources

Azure OpenAI is a new product offering on Azure. You can get started with Azure OpenAI the same way as any other Azure product where you create a resource, or instance of the service, in your Azure Subscription. You can read more about Azure's resource management design.

## Deployments

Once you create an Azure OpenAI Resource, you must deploy a model before you can start making API calls and generating text. This action can be done using the Deployment APIs. These APIs allow you to specify the model you wish to use.

## Prompt engineering

GPT-3, GPT-3.5, and GPT-4 models from OpenAI are prompt-based. With prompt-based models, the user interacts with the model by entering a text prompt, to which the model responds with a text completion. This completion is the model’s continuation of the input text.

While these models are extremely powerful, their behavior is also very sensitive to the prompt. This makes prompt engineering an important skill to develop.

Prompt construction can be difficult. In practice, the prompt acts to configure the model weights to complete the desired task, but it's more of an art than a science, often requiring experience and intuition to craft a successful prompt.

## Models

The service provides users access to several different models. Each model provides a different capability and price point.

GPT-4 models are the latest available models. Due to high demand access to this model series is currently only available by request. To request access, existing Azure OpenAI customers can apply by filling out this form

The DALL-E models, currently in preview, generate images from text prompts that the user provides.

Learn more about each model on our models concept page.

# Next steps

Learn more about the underlying models that power Azure OpenAI.

# Azure OpenAI Service quotas and limits

Article • 08/18/2023

This article contains a quick reference and a detailed description of the quotas and limits for Azure OpenAI in Azure AI services.

## Quotas and limits reference

The following sections provide you with a quick guide to the default quotas and limits that apply to Azure OpenAI:

| Limit Name | Limit Value |
|---|---|
| OpenAI resources per region per Azure subscription | 30 |
| Default quota per model and region (in tokens-per-minute)[1] | Text-Davinci-003: 120 K<br>GPT-4: 20 K<br>GPT-4-32K: 60 K<br>All others: 240 K |
| Default DALL-E quota limits | 2 concurrent requests |
| Maximum prompt tokens per request | Varies per model. For more information, see Azure OpenAI Service models |
| Max fine-tuned model deployments | 2 |
| Total number of training jobs per resource | 100 |
| Max simultaneous running training jobs per resource | 1 |
| Max training jobs queued | 20 |
| Max Files per resource | 30 |
| Total size of all files per resource | 1 GB |
| Max training job time (job will fail if exceeded) | 720 hours |
| Max training job size (tokens in training file) x (# of epochs) | 2 Billion |
| Max size of all files per upload (Azure OpenAI on your data) | 16 MB |

[1] Default quota limits are subject to change.

## General best practices to remain within rate limits

To minimize issues related to rate limits, it's a good idea to use the following techniques:

- Implement retry logic in your application.
- Avoid sharp changes in the workload. Increase the workload gradually.
- Test different load increase patterns.
- Increase the quota assigned to your deployment. Move quota from another deployment, if necessary.

## How to request increases to the default quotas and limits

Quota increase requests can be submitted from the Quotas page of Azure OpenAI Studio. Please note that due to overwhelming demand, we are not currently approving new quota increase requests. Your request will be queued until it can be filled at a later time.

For other rate limits, please submit a service request.

# Next steps

Explore how to manage quota for your Azure OpenAI deployments. Learn more about the underlying models that power Azure OpenAI.

# Azure OpenAI Service models

Article • 09/05/2023

Azure OpenAI Service is powered by a diverse set of models with different capabilities and price points. Model availability varies by region. For GPT-3 and other models retiring in July 2024, see Azure OpenAI Service legacy models.

| Models | Description |
|---|---|
| GPT-4 | A set of models that improve on GPT-3.5 and can understand as well as generate natural language and code. |
| GPT-3.5 | A set of models that improve on GPT-3 and can understand as well as generate natural language and code. |
| Embeddings | A set of models that can convert text into numerical vector form to facilitate text similarity. |
| DALL-E (Preview) | A series of models in preview that can generate original images from natural language. |

## GPT-4

GPT-4 can solve difficult problems with greater accuracy than any of OpenAI's previous models. Like GPT-3.5 Turbo, GPT-4 is optimized for chat and works well for traditional completions tasks. Use the Chat Completions API to use GPT-4. To learn more about how to interact with GPT-4 and the Chat Completions API check out our in-depth how-to.

To request access to GPT-4, Azure OpenAI customers can apply by filling out this form

- `gpt-4`
- `gpt-4-32k`

The `gpt-4` model supports 8192 max input tokens and the `gpt-4-32k` model supports up to 32,768 tokens.

## GPT-3.5

GPT-3.5 models can understand and generate natural language or code. The most capable and cost effective model in the GPT-3.5 family is GPT-3.5 Turbo, which has been

optimized for chat and works well for traditional completions tasks as well. We recommend using GPT-3.5 Turbo over legacy GPT-3.5 and GPT-3 models.

- `gpt-35-turbo`
- `gpt-35-turbo-16k`

The `gpt-35-turbo` model supports 4096 max input tokens and the `gpt-35-turbo-16k` model supports up to 16,384 tokens.

Like GPT-4, use the Chat Completions API to use GPT-3.5 Turbo. To learn more about how to interact with GPT-3.5 Turbo and the Chat Completions API check out our in-depth how-to.

# Embeddings models

> **ⓘ Important**
>
> We strongly recommend using `text-embedding-ada-002 (Version 2)`. This model/version provides parity with OpenAI's `text-embedding-ada-002`. To learn more about the improvements offered by this model, please refer to **OpenAI's blog post**. Even if you are currently using Version 1 you should migrate to Version 2 to take advantage of the latest weights/updated token limit. Version 1 and Version 2 are not interchangeable, so document embedding and document search must be done using the same version of the model.

Currently, we offer three families of Embeddings models for different functionalities: The following list indicates the length of the numerical vector returned by the service, based on model capability:

| Base Model | Model(s) | Dimensions |
|---|---|---|
| Ada | models ending in -001 (Version 1) | 1024 |
| Ada | text-embedding-ada-002 (Version 2) | 1536 |

# DALL-E (Preview)

The DALL-E models, currently in preview, generate images from text prompts that the user provides.

# Model summary table and region availability

> **Important**
>
> Due to high demand:
>
> - South Central US is temporarily unavailable for creating new resources and deployments.

## GPT-4 models

These models can only be used with the Chat Completion API.

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| `gpt-4` [1,2] (0314) | | N/A | 8,192 | September 2021 |
| `gpt-4-32k` [1,2] (0314) | | N/A | 32,768 | September 2021 |
| `gpt-4` [13] (0613) | Australia East, Canada East, East US, East US 2, France Central, Japan East, Sweden Central, Switzerland North, UK South | N/A | 8,192 | September 2021 |
| `gpt-4-32k` [13] (0613) | Australia East, Canada East, East US, East US 2, France Central, Japan East, Sweden Central, Switzerland North, UK South | N/A | 32,768 | September 2021 |

[1] The model is only available by request.
[2] Version `0314` of gpt-4 and gpt-4-32k will be retired no earlier than July 5, 2024. See model updates for model upgrade behavior.
[3] We are rolling out availability of new regions to customers gradually to ensure a smooth experience. In East US and France Central, customers with existing deployments of GPT-4 can create additional deployments of GPT-4 version 0613. For customers new to GPT-4 on Azure OpenAI, please use one of the other available regions.

## GPT-3.5 models

GPT-3.5 Turbo is used with the Chat Completion API. GPT-3.5 Turbo (0301) can also be used with the Completions API. GPT3.5 Turbo (0613) only supports the Chat Completions API.

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| `gpt-35-turbo`[1] (0301) | East US, France Central, South Central US, UK South, West Europe | N/A | 4,096 | Sep 2021 |
| `gpt-35-turbo` (0613) | Australia East, Canada East, East US, East US 2, France Central, Japan East, North Central US, Sweden Central, Switzerland North, UK South | N/A | 4,096 | Sep 2021 |
| `gpt-35-turbo-16k` (0613) | Australia East, Canada East, East US, East US 2, France Central, Japan East, North Central US, Sweden Central, Switzerland North, UK South | N/A | 16,384 | Sep 2021 |

[1] Version `0301` of gpt-35-turbo will be retired no earlier than July 5, 2024. See model updates for model upgrade behavior.

## Embeddings models

These models can only be used with Embedding API requests.

> ⓘ Note
>
> We strongly recommend using `text-embedding-ada-002 (Version 2)`. This model/version provides parity with OpenAI's `text-embedding-ada-002`. To learn more about the improvements offered by this model, please refer to **OpenAI's blog post**. Even if you are currently using Version 1 you should migrate to Version 2 to take advantage of the latest weights/updated token limit. Version 1 and Version 2 are not interchangeable, so document embedding and document search must be done using the same version of the model.

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| text-embedding-ada-002 (version 2) | Canada East, East US, France Central, Japan East, North Central US, South Central US, Switzerland North, UK South, West Europe | N/A | 8,191 | Sep 2021 |
| text-embedding-ada-002 (version 1) | East US, South Central US, West Europe | N/A | 2,046 | Sep 2021 |

## DALL-E models (Preview)

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (characters) | Training Data (up to) |
|---|---|---|---|---|
| dalle2 | East US | N/A | 1000 | N/A |

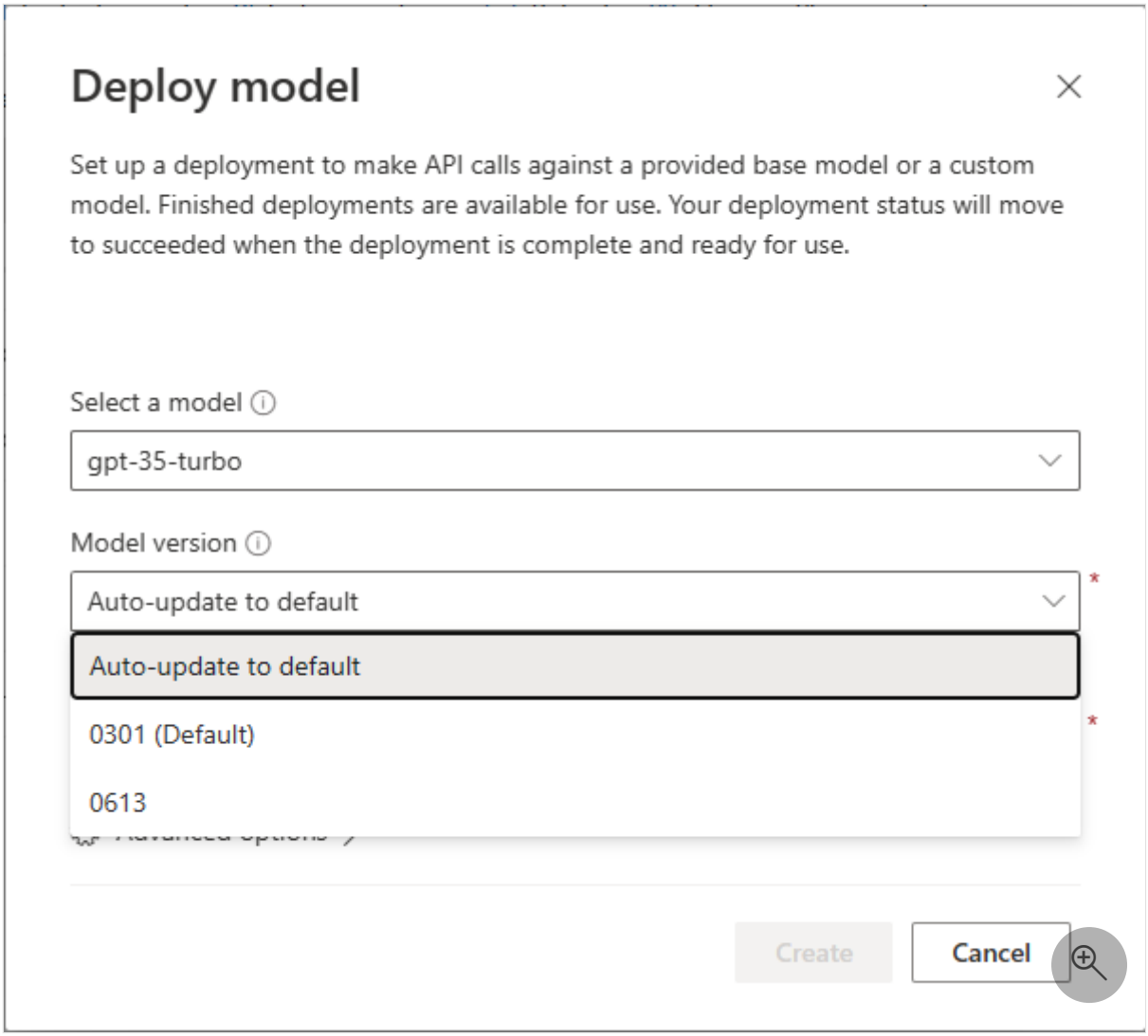
# Working with models

## Finding what models are available

You can get a list of models that are available for both inference and fine-tuning by your Azure OpenAI resource by using the Models List API.

## Model updates

Azure OpenAI now supports automatic updates for select model deployments. On models where automatic update support is available, a model version drop-down will be visible in Azure OpenAI Studio under **Create new deployment** and **Edit deployment**:




## Auto update to default

When **Auto-update to default** is selected your model deployment will be automatically updated within two weeks of a new version being released.

If you are still in the early testing phases for completion and chat completion based models, we recommend deploying models with **auto-update to default** set whenever it is available.

## Specific model version

As your use of Azure OpenAI evolves, and you start to build and integrate with applications you will likely want to manually control model updates so that you can first test and validate that model performance is remaining consistent for your use case prior to upgrade.

When you select a specific model version for a deployment this version will remain selected until you either choose to manually update yourself, or once you reach the

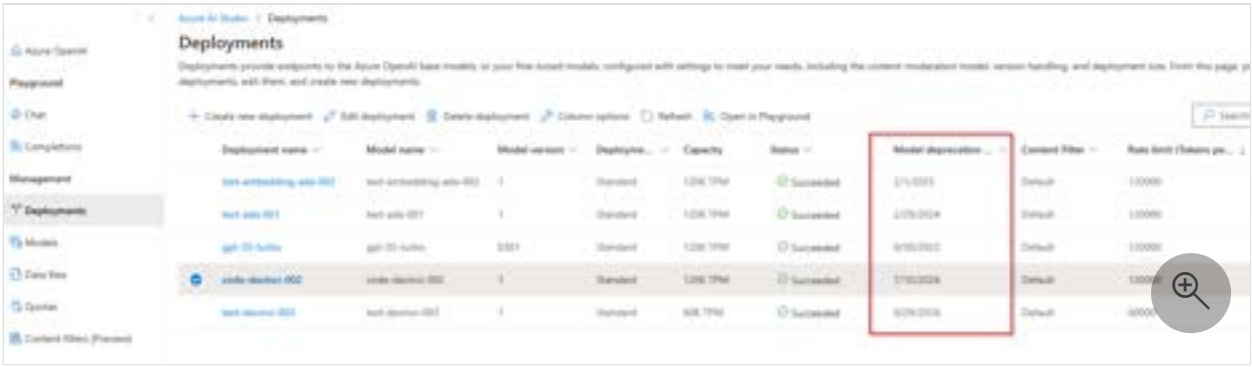
retirement date for the model. When the retirement date is reached the model will auto-upgrade to the default version at the time of retirement.

## GPT-35-Turbo 0301 and GPT-4 0314 retirement

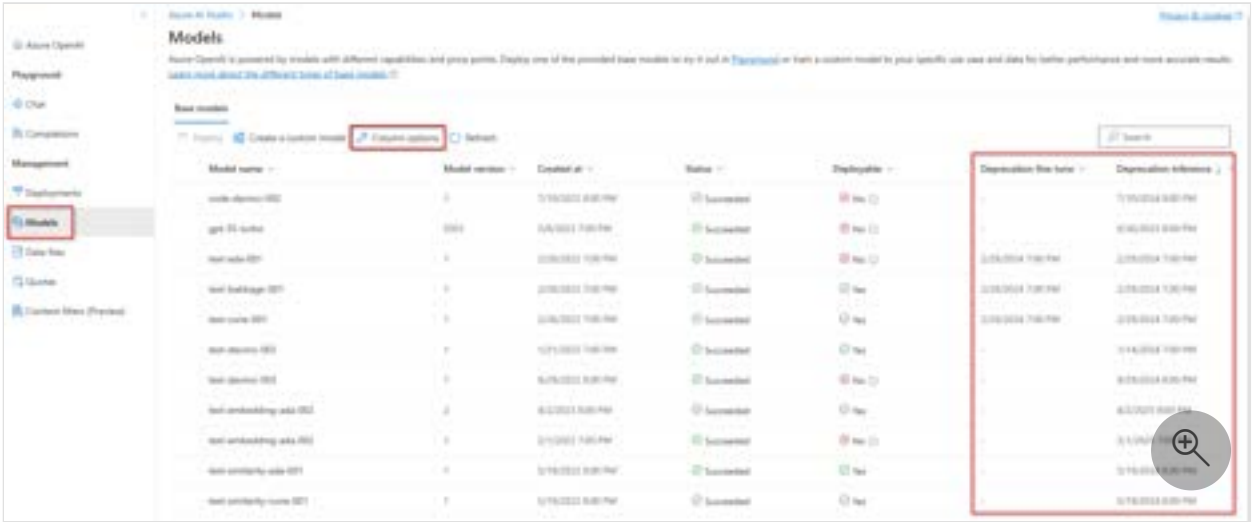
The `gpt-35-turbo` (`0301`) and both `gpt-4` (`0314`) models will be retired no earlier than July 5, 2024. Upon retirement, deployments will automatically be upgraded to the default version at the time of retirement. If you would like your deployment to stop accepting completion requests rather than upgrading, then you will be able to set the model upgrade option to expire through the API. We will publish guidelines on this by September 1.

## Viewing deprecation dates

For currently deployed models, from Azure OpenAI Studio select **Deployments**:




To view deprecation/expiration dates for all available models in a given region from Azure OpenAI Studio select **Models** > **Column options** > Select **Deprecation fine tune** and **Deprecation inference**:




## Update & deploy models via the API

HTTP

```
PUT
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{
resourceGroupName}/providers/Microsoft.CognitiveServices/accounts/{accountNa
me}/deployments/{deploymentName}?api-version=2023-05-01
```

### Path parameters

| Parameter | Type | Required? | Description |
|---|---|---|---|
| `acountname` | string | Required | The name of your Azure OpenAI Resource. |
| `deploymentName` | string | Required | The deployment name you chose when you deployed an existing model or the name you would like a new model deployment to have. |
| `resourceGroupName` | string | Required | The name of the associated resource group for this model deployment. |
| `subscriptionId` | string | Required | Subscription ID for the associated subscription. |
| `api-version` | string | Required | The API version to use for this operation. This follows the YYYY-MM-DD format. |

### Supported versions

- `2023-05-01` Swagger spec

### Request body

This is only a subset of the available request body parameters. For the full list of the parameters, you can refer to the REST API reference documentation.

| Parameter | Type | Description |
|---|---|---|
| versionUpgradeOption | String | Deployment model version upgrade options:<br>`OnceNewDefaultVersionAvailable`<br>`OnceCurrentVersionExpired`<br>`NoAutoUpgrade` |
| capacity | integer | This represents the amount of quota you are assigning to this deployment. A value of 1 equals 1,000 Tokens per Minute (TPM) |

## Example request

Bash

```
curl -X PUT https://management.azure.com/subscriptions/00000000-0000-0000-0000-000000000000/resourceGroups/resource-group-temp/providers/Microsoft.CognitiveServices/accounts/docs-openai-test-001/deployments/text-embedding-ada-002-test-1?api-version=2023-05-01 \
  -H "Content-Type: application/json" \
  -H 'Authorization: Bearer YOUR_AUTH_TOKEN' \
  -d '{"sku":{"name":"Standard","capacity":1},"properties": {"model": {"format": "OpenAI","name": "text-embedding-ada-002","version": "2"},"versionUpgradeOption":"OnceCurrentVersionExpired"}}'
```

> **Note**
>
> There are multiple ways to generate an authorization token. The easiest method for initial testing is to launch the Cloud Shell from the **Azure portal**. Then run **az account get-access-token**. You can use this token as your temporary authorization token for API testing.

## Example response

JSON


```
{
  "id": "/subscriptions/{subscription-id}/resourceGroups/resource-group-temp/providers/Microsoft.CognitiveServices/accounts/docs-openai-test-001/deployments/text-embedding-ada-002-test-1",
  "type": "Microsoft.CognitiveServices/accounts/deployments",
  "name": "text-embedding-ada-002-test-1",
  "sku": {
    "name": "Standard",
    "capacity": 1
  },
  "properties": {
    "model": {
      "format": "OpenAI",
      "name": "text-embedding-ada-002",
      "version": "2"
    },
    "versionUpgradeOption": "OnceCurrentVersionExpired",
    "capabilities": {
      "embeddings": "true",
      "embeddingsMaxInputs": "1"
    },
    "provisioningState": "Succeeded",
    "ratelimits": [
      {
        "key": "request",
```

```
          "renewalPeriod": 10,
          "count": 2
        },
        {
          "key": "token",
          "renewalPeriod": 60,
          "count": 1000
        }
      ]
    },
    "systemData": {
      "createdBy": "docs@contoso.com",
      "createdByType": "User",
      "createdAt": "2023-06-13T00:12:38.885937Z",
      "lastModifiedBy": "docs@contoso.com",
      "lastModifiedByType": "User",
      "lastModifiedAt": "2023-06-13T02:41:04.8410965Z"
    },
    "etag": "\"{GUID}\""
}
```


# Next steps

- Learn more about Azure OpenAI
- Learn more about fine-tuning Azure OpenAI models

# Azure OpenAI Service legacy models

Article • 07/18/2023

Azure OpenAI Service offers a variety of models for different use cases. The following models are not available for new deployments beginning July 6, 2023. Deployments created prior to July 6, 2023 remain available to customers until July 5, 2024. We recommend customers migrate to the replacement models prior to the July 5, 2024 retirement.

## GPT-3.5

The impacted GPT-3.5 models are the following. The replacement for the GPT-3.5 models is GPT-3.5 Turbo Instruct when that model becomes available.

- `text-davinci-002`
- `text-davinci-003`
- `code-davinci-002`

## GPT-3

The impacted GPT-3 models are the following. The replacement for the GPT-3 models is GPT-3.5 Turbo Instruct when that model becomes available.

- `text-ada-001`
- `text-babbage-001`
- `text-curie-001`
- `text-davinci-001`
- `code-cushman-001`

## Embedding models

The embedding models below will be retired effective July 5, 2024. Customers should migrate to `text-embedding-ada-002` (version 2).

- Similarity
- Text search
- Code search

Each family includes models across a range of capability. The following list indicates the length of the numerical vector returned by the service, based on model capability:

| Base Model | Model(s) | Dimensions |
|---|---|---|
| Ada | | 1024 |
| Babbage | | 2048 |
| Curie | | 4096 |
| Davinci | | 12288 |

## Similarity embedding

These models are good at capturing semantic similarity between two or more pieces of text.

| Use cases | Models |
|---|---|
| Clustering, regression, anomaly detection, visualization | `text-similarity-ada-001`<br>`text-similarity-babbage-001`<br>`text-similarity-curie-001`<br>`text-similarity-davinci-001` |

## Text search embedding

These models help measure whether long documents are relevant to a short search query. There are two input types supported by this family: `doc`, for embedding the documents to be retrieved, and `query`, for embedding the search query.

| Use cases | Models |
|---|---|
| Search, context relevance, information retrieval | `text-search-ada-doc-001`<br>`text-search-ada-query-001`<br>`text-search-babbage-doc-001`<br>`text-search-babbage-query-001`<br>`text-search-curie-doc-001`<br>`text-search-curie-query-001`<br>`text-search-davinci-doc-001`<br>`text-search-davinci-query-001` |

## Code search embedding

Similar to text search embedding models, there are two input types supported by this family: `code`, for embedding code snippets to be retrieved, and `text`, for embedding natural language search queries.

| Use cases | Models |
| --- | --- |
| Code search and relevance | `code-search-ada-code-001`<br>`code-search-ada-text-001`<br>`code-search-babbage-code-001`<br>`code-search-babbage-text-001` |

# Model summary table and region availability

Region availability is for customers with deployments of the models prior to July 6, 2023.

## GPT-3.5 models

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
| --- | --- | --- | --- | --- |
| text-davinci-002 | East US, South Central US, West Europe | N/A | 4,097 | Jun 2021 |
| text-davinci-003 | East US, West Europe | N/A | 4,097 | Jun 2021 |
| code-davinci-002 | East US, West Europe | N/A | 8,001 | Jun 2021 |

## GPT-3 models

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
| --- | --- | --- | --- | --- |
| ada | N/A | N/A | 2,049 | Oct 2019 |
| text-ada-001 | East US, South Central US, West Europe | N/A | 2,049 | Oct 2019 |
| babbage | N/A | N/A | 2,049 | Oct 2019 |
| text-babbage-001 | East US, South Central US, West Europe | N/A | 2,049 | Oct 2019 |
| curie | N/A | N/A | 2,049 | Oct 2019 |

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| text-curie-001 | East US, South Central US, West Europe | N/A | 2,049 | Oct 2019 |
| davinci | N/A | N/A | 2,049 | Oct 2019 |
| text-davinci-001 | South Central US, West Europe | N/A | | |

## Codex models

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| code-cushman-001 | South Central US, West Europe | N/A | 2,048 | |

## Embedding models

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| text-similarity-ada-001 | East US, South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-similarity-babbage-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-similarity-curie-001 | East US, South Central US, West Europe | N/A | 2046 | Aug 2020 |
| text-similarity-davinci-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-ada-doc-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-ada-query-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-babbage-doc-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-babbage-query-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |

| Model ID | Base model Regions | Fine-Tuning Regions | Max Request (tokens) | Training Data (up to) |
|---|---|---|---|---|
| text-search-curie-doc-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-curie-query-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-davinci-doc-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| text-search-davinci-query-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| code-search-ada-code-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| code-search-ada-text-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| code-search-babbage-code-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |
| code-search-babbage-text-001 | South Central US, West Europe | N/A | 2,046 | Aug 2020 |

# What's new in Azure OpenAI Service

Article • 08/25/2023

## August 2023

### Azure OpenAI on your own data (preview) updates

- You can now deploy Azure OpenAI on your data to Power Virtual Agents.
- Azure OpenAI on your data now supports private endpoints.
- Ability to filter access to sensitive documents.
- Automatically refresh your index on a schedule.
- Vector search and semantic search options.
- View your chat history in the deployed web app

## July 2023

### Support for function calling

- Azure OpenAI now supports function calling to enable you to work with functions in the chat completions API.

### Embedding input array increase

- Azure OpenAI now supports arrays with up to 16 inputs per API request with text-embedding-ada-002 Version 2.

### New Regions

- Azure OpenAI is now also available in the Canada East, East US 2, Japan East, and North Central US regions. Check the models page, for the latest information on model availability in each region.

## June 2023

### Use Azure OpenAI on your own data (preview)

- Azure OpenAI on your data is now available in preview, enabling you to chat with OpenAI models such as GPT-35-Turbo and GPT-4 and receive responses based on your data.

## New versions of gpt-35-turbo and gpt-4 models

- gpt-35-turbo (version 0613)
- gpt-35-turbo-16k (version 0613)
- gpt-4 (version 0613)
- gpt-4-32k (version 0613)

## UK South

- Azure OpenAI is now available in the UK South region. Check the models page, for the latest information on model availability in each region.

## Content filtering & annotations (Preview)

- How to configure content filters with Azure OpenAI Service.
- Enable annotations to view content filtering category and severity information as part of your GPT based Completion and Chat Completion calls.

## Quota

- Quota provides the flexibility to actively manage the allocation of rate limits across the deployments within your subscription.

# May 2023

## Java & JavaScript SDK support

- NEW Azure OpenAI preview SDKs offering support for JavaScript and Java.

## Azure OpenAI Chat Completion General Availability (GA)

- General availability support for:
  - Chat Completion API version `2023-05-15`.
  - GPT-35-Turbo models.

    - GPT-4 model series. Due to high demand access to this model series is currently only available by request. To request access, existing Azure OpenAI customers can apply by filling out this form

If you are currently using the `2023-03-15-preview` API, we recommend migrating to the GA `2023-05-15` API. If you are currently using API version `2022-12-01` this API remains GA, but does not include the latest Chat Completion capabilities.

> **ⓘ Important**
>
> Using the current versions of the GPT-35-Turbo models with the completion endpoint remains in preview.

### France Central

- Azure OpenAI is now available in the France Central region. Check the models page, for the latest information on model availability in each region.

## April 2023

- **DALL-E 2 public preview**. Azure OpenAI Service now supports image generation APIs powered by OpenAI's DALL-E 2 model. Get AI-generated images based on the descriptive text you provide. To learn more, check out the quickstart. To request access, existing Azure OpenAI customers can apply by filling out this form.
- **Inactive deployments of customized models will now be deleted after 15 days; models will remain available for redeployment.** If a customized (fine-tuned) model is deployed for more than fifteen (15) days during which no completions or chat completions calls are made to it, the deployment will automatically be deleted (and no further hosting charges will be incurred for that deployment). The underlying customized model will remain available and can be redeployed at any time. To learn more check out the how-to-article.

## March 2023

- **GPT-4 series models are now available in preview on Azure OpenAI**. To request access, existing Azure OpenAI customers can apply by filling out this form. These models are currently available in the East US and South Central US regions.

- **New Chat Completion API for GPT-35-Turbo and GPT-4 models released in preview on 3/21**. To learn more checkout the updated quickstarts and how-to article.
- **GPT-35-Turbo preview**. To learn more checkout the how-to article.
- Increased training limits for fine-tuning: The max training job size (tokens in training file) x (# of epochs) is 2 Billion tokens for all models. We have also increased the max training job from 120 to 720 hours.
- Adding additional use cases to your existing access.  Previously, the process for adding new use cases required customers to reapply to the service. Now, we're releasing a new process that allows you to quickly add new use cases to your use of the service. This process follows the established Limited Access process within Azure AI services. Existing customers can attest to any and all new use cases here. Please note that this is required anytime you would like to use the service for a new use case you did not originally apply for.

# February 2023

## New Features

- .NET SDK(inference) preview release | Samples
- Terraform SDK update to support Azure OpenAI management operations.
- Inserting text at the end of a completion is now supported with the `suffix` parameter.

## Updates

- Content filtering is on by default.

New articles on:

- Monitoring an Azure OpenAI Service
- Plan and manage costs for Azure OpenAI

New training course:

- Intro to Azure OpenAI

# January 2023

## New Features

- **Service GA**. Azure OpenAI Service is now generally available.
- **New models**: Addition of the latest text model, text-davinci-003 (East US, West Europe), text-ada-embeddings-002 (East US, South Central US, West Europe)

# December 2022

## New features

- **The latest models from OpenAI.** Azure OpenAI provides access to all the latest models including the GPT-3.5 series.
- **New API version (2022-12-01).** This update includes several requested enhancements including token usage information in the API response, improved error messages for files, alignment with OpenAI on fine-tuning creation data structure, and support for the suffix parameter to allow custom naming of fine-tuned jobs.
- **Higher request per second limits.** 50 for non-Davinci models. 20 for Davinci models.
- **Faster fine-tune deployments.** Deploy an Ada and Curie fine-tuned models in under 10 minutes.
- **Higher training limits:** 40M training tokens for Ada, Babbage, and Curie. 10M for Davinci.
- **Process for requesting modifications to the abuse & miss-use data logging & human review.** Today, the service logs request/response data for the purposes of abuse and misuse detection to ensure that these powerful models aren't abused. However, many customers have strict data privacy and security requirements that require greater control over their data. To support these use cases, we're releasing a new process for customers to modify the content filtering policies or turn off the abuse logging for low-risk use cases. This process follows the established Limited Access process within Azure AI services and existing OpenAI customers can apply here.
- **Customer managed key (CMK) encryption.** CMK provides customers greater control over managing their data in Azure OpenAI by providing their own encryption keys used for storing training data and customized models. Customer-

managed keys (CMK), also known as bring your own key (BYOK), offer greater flexibility to create, rotate, disable, and revoke access controls. You can also audit the encryption keys used to protect your data. Learn more from our encryption at rest documentation.

- **Lockbox support**
- **SOC-2 compliance**
- **Logging and diagnostics** through Azure Resource Health, Cost Analysis, and Metrics & Diagnostic settings.
- **Studio improvements.** Numerous usability improvements to the Studio workflow including Azure AD role support to control who in the team has access to create fine-tuned models and deploy.

## Changes (breaking)

**Fine-tuning** create API request has been updated to match OpenAI's schema.

**Preview API versions:**

JSON

```
{
    "training_file": "file-XGinujblHPwGLSztz8cPS8XY",
    "hyperparams": {
        "batch_size": 4,
        "learning_rate_multiplier": 0.1,
        "n_epochs": 4,
        "prompt_loss_weight": 0.1,
    }
}
```

**API version 2022-12-01:**

JSON

```
{
    "training_file": "file-XGinujblHPwGLSztz8cPS8XY",
    "batch_size": 4,
    "learning_rate_multiplier": 0.1,
    "n_epochs": 4,
    "prompt_loss_weight": 0.1,
}
```

- Immersive Reader
- Language Understanding (LUIS)
- QnA Maker
- Language service
- Translator

**Speech**

- Speech service

**Decision**

- Anomaly Detector
- Content Moderator
- Metrics Advisor
- Personalizer

**Azure OpenAI**

- Azure OpenAI

# Post a question to Stack Overflow

For answers on your developer questions from the largest community developer ecosystem, ask your question on Stack Overflow.

If you do submit a new question to Stack Overflow, please use one or more of the following tags when you create the question:

- Azure AI services

**Vision**

- Azure AI Vision
- Custom Vision
- Face
- Document Intelligence
- Video Indexer

**Language**

- Immersive Reader
- Language Understanding (LUIS)

- QnA Maker
- Language service
- Translator

**Speech**

- Speech service

**Decision**

- Anomaly Detector
- Content Moderator
- Metrics Advisor
- Personalizer

**Azure OpenAI**

- Azure OpenAI

# Submit feedback

To request new features, post them on https://feedback.azure.com. Share your ideas for making Azure AI services and its APIs work better for the applications you develop.

- Azure AI services

**Vision**

- Azure AI Vision
- Custom Vision
- Face
- Document Intelligence
- Video Indexer

**Language**

- Immersive Reader
- Language Understanding (LUIS)
- QnA Maker
- Language service
- Translator

**Speech**

- Speech service

**Decision**

- Anomaly Detector
- Content Moderator
- Metrics Advisor
- Personalizer

## Stay informed

Staying informed about features in a new release or news on the Azure blog can help you find the difference between a programming error, a service bug, or a feature not yet available in Azure AI services.

- Learn more about product updates, roadmap, and announcements in Azure Updates.
- News about Azure AI services is shared in the Azure blog.
- Join the conversation on Reddit about Azure AI services.

## Next steps

What are Azure AI services?