**[Azure OpenAI Service models](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models)**

* Article
* 12/11/2023
* 8 contributors

Feedback

**In this article**

1. [GPT-4 and GPT-4 Turbo Preview](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#gpt-4-and-gpt-4-turbo-preview)
2. [GPT-3.5](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#gpt-35)
3. [Embeddings](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#embeddings)
4. [DALL-E (Preview)](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#dall-e-preview)
5. [Model summary table and region availability](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#model-summary-table-and-region-availability)

Show 3 more

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](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/legacy-models).

Expand table

| **Models** | **Description** |
| --- | --- |
| [GPT-4](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#gpt-4-and-gpt-4-turbo-preview) | A set of models that improve on GPT-3.5 and can understand and generate natural language and code. |
| [GPT-3.5](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#gpt-35) | A set of models that improve on GPT-3 and can understand and generate natural language and code. |
| [Embeddings](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#embeddings-models) | A set of models that can convert text into numerical vector form to facilitate text similarity. |
| [DALL-E](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#dall-e-models-preview) (Preview) | A series of models in preview that can generate original images from natural language. |
| [Whisper](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#whisper-models-preview) (Preview) | A series of models in preview that can transcribe and translate speech to text. |

**GPT-4 and GPT-4 Turbo Preview**

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](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/chatgpt).

* gpt-4
* gpt-4-32k
* gpt-4-vision

You can see the token context length supported by each model in the [model summary table](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#model-summary-table-and-region-availability).

**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. GPT-3.5 Turbo is available for use with the Chat Completions API. GPT-3.5 Turbo Instruct has similar capabilities to text-davinci-003 using the Completions API instead of the Chat Completions API. We recommend using GPT-3.5 Turbo and GPT-3.5 Turbo Instruct over [legacy GPT-3.5 and GPT-3 models](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/legacy-models).

* gpt-35-turbo
* gpt-35-turbo-16k
* gpt-35-turbo-instruct

You can see the token context length supported by each model in the [model summary table](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#model-summary-table-and-region-availability).

To learn more about how to interact with GPT-3.5 Turbo and the Chat Completions API check out our [in-depth how-to](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/chatgpt).

**Embeddings**

**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**](https://openai.com/blog/new-and-improved-embedding-model). 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.

The previous embeddings models have been consolidated into the following new replacement model:

text-embedding-ada-002

**DALL-E (Preview)**

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

**Whisper (Preview)**

The Whisper models, currently in preview, can be used for speech to text.

You can also use the Whisper model via Azure AI Speech [batch transcription](https://learn.microsoft.com/en-us/azure/ai-services/speech-service/batch-transcription-create) API. Check out [What is the Whisper model?](https://learn.microsoft.com/en-us/azure/ai-services/speech-service/whisper-overview) to learn more about when to use Azure AI Speech vs. Azure OpenAI Service.

**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 and GPT-4 Turbo Preview models**

GPT-4, GPT-4-32k, and GPT-4 Turbo with Vision are now available to all Azure OpenAI Service customers. Availability varies by region. If you don't see GPT-4 in your region, please check back later.

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

GPT-4 version 0314 is the first version of the model released. Version 0613 is the second version of the model and adds function calling support.

See [model versions](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/model-versions) to learn about how Azure OpenAI Service handles model version upgrades, and [working with models](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/working-with-models) to learn how to view and configure the model version settings of your GPT-4 deployments.

**Note**

Version 0314 of gpt-4 and gpt-4-32k will be retired no earlier than July 5, 2024. See [**model updates**](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/working-with-models#model-updates) for model upgrade behavior.

Expand table

| **Model ID** | **Max Request (tokens)** | **Training Data (up to)** |
| --- | --- | --- |
| gpt-4 (0314) | 8,192 | Sep 2021 |
| gpt-4-32k(0314) | 32,768 | Sep 2021 |
| gpt-4 (0613) | 8,192 | Sep 2021 |
| gpt-4-32k (0613) | 32,768 | Sep 2021 |
| gpt-4 (1106-preview)**1** **GPT-4 Turbo Preview** | Input: 128,000 Output: 4096 | Apr 2023 |
| gpt-4 (vision-preview)**2** **GPT-4 Turbo with Vision Preview** | Input: 128,000 Output: 4096 | Apr 2023 |

**1** GPT-4 Turbo Preview = gpt-4 (1106-preview). To deploy this model, under **Deployments** select model **gpt-4**. For **Model version** select **1106-preview**.

**2** GPT-4 Turbo with Vision Preview = gpt-4 (vision-preview). To deploy this model, under **Deployments** select model **gpt-4**. For **Model version** select **vision-preview**.

**Caution**

We don't recommend using these models in production. We will upgrade all deployments of these models to a future stable version. Models designated preview do not follow the standard Azure OpenAI model lifecycle.

**Note**

Regions where GPT-4 (0314) & (0613) are listed as available have access to both the 8K and 32K versions of the model

**GPT-4 and GPT-4 Turbo Preview model availability**

Expand table

| **Model Availability** | **gpt-4 (0314)** | **gpt-4 (0613)** | **gpt-4 (1106-preview)** | **gpt-4 (vision-preview)** |
| --- | --- | --- | --- | --- |
| Available to all subscriptions with Azure OpenAI access |  | Australia East Canada East France Central Sweden Central Switzerland North | Australia East Canada East East US 2 France Central Norway East South India Sweden Central UK South West US | Switzerland North West US |
| Available to subscriptions with current access to the model version in the region | East US France Central South Central US UK South | East US East US 2 Japan East UK South |  | Australia East Sweden Central |

**GPT-3.5 models**

GPT-3.5 Turbo is used with the Chat Completion API. GPT-3.5 Turbo version 0301 can also be used with the Completions API. GPT-3.5 Turbo versions 0613 and 1106 only support the Chat Completions API.

GPT-3.5 Turbo version 0301 is the first version of the model released. Version 0613 is the second version of the model and adds function calling support.

See [model versions](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/model-versions) to learn about how Azure OpenAI Service handles model version upgrades, and [working with models](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/working-with-models) to learn how to view and configure the model version settings of your GPT-3.5 Turbo deployments.

**Note**

Version 0301 of gpt-35-turbo will be retired no earlier than July 5, 2024. See [**model updates**](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/working-with-models#model-updates) for model upgrade behavior.

**GPT-3.5-Turbo model availability**

Expand table

| **Model ID** | **Model Availability** | **Max Request (tokens)** | **Training Data (up to)** |
| --- | --- | --- | --- |
| gpt-35-turbo**1** (0301) | East US France Central South Central US UK South West Europe | 4096 | 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 | 4096 | 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 | 16,384 | Sep 2021 |
| gpt-35-turbo-instruct (0914) | East US Sweden Central | 4097 | Sep 2021 |
| gpt-35-turbo (1106) | Australia East Canada East France Central South India Sweden Central UK South West US | Input: 16,385 Output: 4,096 | Sep 2021 |

**1** This model will accept requests > 4096 tokens. It is not recommended to exceed the 4096 input token limit as the newer version of the model are capped at 4096 tokens. If you encounter issues when exceeding 4096 input tokens with this model this configuration is not officially supported.

**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**](https://openai.com/blog/new-and-improved-embedding-model). 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.

Expand table

| **Model ID** | **Model Availability** | **Max Request (tokens)** | **Training Data (up to)** | **Output Dimensions** |
| --- | --- | --- | --- | --- |
| text-embedding-ada-002 (version 2) | Australia East Canada East East US East US2 France Central Japan East North Central US South Central US Sweden Central Switzerland North UK South West Europe | 8,191 | Sep 2021 | 1536 |
| text-embedding-ada-002 (version 1) | East US South Central US West Europe | 2,046 | Sep 2021 | 1536 |

**DALL-E models (Preview)**

Expand table

| **Model ID** | **Feature Availability** | **Max Request (characters)** |
| --- | --- | --- |
| dalle2 | East US | 1000 |
| dalle3 | Sweden Central | 4000 |

**Fine-tuning models**

babbage-002 and davinci-002 are not trained to follow instructions. Querying these base models should only be done as a point of reference to a fine-tuned version to evaluate the progress of your training.

gpt-35-turbo-0613 - fine-tuning of this model is limited to a subset of regions, and is not available in every region the base model is available.

Expand table

| **Model ID** | **Fine-Tuning Regions** | **Max Request (tokens)** | **Training Data (up to)** |
| --- | --- | --- | --- |
| babbage-002 | North Central US Sweden Central | 16,384 | Sep 2021 |
| davinci-002 | North Central US Sweden Central | 16,384 | Sep 2021 |
| gpt-35-turbo (0613) | North Central US Sweden Central | 4096 | Sep 2021 |

**Whisper models (Preview)**

Expand table

| **Model ID** | **Model Availability** | **Max Request (audio file size)** |
| --- | --- | --- |
| whisper | North Central US West Europe | 25 MB |

**Next steps**

* [Learn more about working with Azure OpenAI models](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/working-with-models)
* [Learn more about Azure OpenAI](https://learn.microsoft.com/en-us/azure/ai-services/openai/overview)
* [Learn more about fine-tuning Azure OpenAI models](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/fine-tuning)