

What's new in Azure OpenAI Service

Article • 11/12/2024

This article provides a summary of the latest releases and major documentation updates for Azure OpenAI.

December 2024

NEW data zone provisioned deployment type

Data zone provisioned deployments are available in the same Azure OpenAI resource as all other Azure OpenAI deployment types but allow you to leverage Azure global infrastructure to dynamically route traffic to the data center within the Microsoft defined data zone with the best availability for each request. Data zone provisioned deployments provide reserved model processing capacity for high and predictable throughput using Azure infrastructure within Microsoft specified data zones. Data zone provisioned deployments are supported on `gpt-4o-2024-08-06`, `gpt-4o-2024-05-13`, and `gpt-4o-mini-2024-07-18` models.

For more information, see the [deployment types guide](#).

November 2024

Vision Fine-tuning GA

Vision fine-tuning with GPT-4o (2024-08-06) is now Generally Available (GA).

[Vision fine-tuning](#) allows you to add images to your JSONL training data. Just as you can send one or many image inputs to chat completions, you can include those same message types within your training data. Images can be provided either as URLs or as base64 encoded images.

For fine-tuning model region availability, see the [models page](#).

NEW AI abuse monitoring

We are introducing new forms of abuse monitoring that leverage LLMs to improve efficiency of detection of potentially abusive use of the Azure OpenAI service and to

enable abuse monitoring without the need for human review of prompts and completions. Learn more, see [Abuse monitoring](#).

Prompts and completions that are flagged through content classification and/or identified to be part of a potentially abusive pattern of use are subjected to an additional review process to help confirm the system's analysis and inform actioning decisions. Our abuse monitoring systems have been expanded to enable review by LLM by default and by humans when necessary and appropriate.

October 2024

NEW data zone standard deployment type

Data zone standard deployments are available in the same Azure OpenAI resource as all other Azure OpenAI deployment types but allow you to leverage Azure global infrastructure to dynamically route traffic to the data center within the Microsoft defined data zone with the best availability for each request. Data zone standard provides higher default quotas than our Azure geography-based deployment types. Data zone standard deployments are supported on `gpt-4o-2024-08-06`, `gpt-4o-2024-05-13`, and `gpt-4o-mini-2024-07-18` models.

For more information, see the [deployment types guide](#).

Global Batch GA

Azure OpenAI global batch is now generally available.

The Azure OpenAI Batch API is designed to handle large-scale and high-volume processing tasks efficiently. Process asynchronous groups of requests with separate quota, with 24-hour target turnaround, at [50% less cost than global standard](#). With batch processing, rather than send one request at a time you send a large number of requests in a single file. Global batch requests have a separate enqueued token quota avoiding any disruption of your online workloads.

Key use cases include:

- **Large-Scale Data Processing:** Quickly analyze extensive datasets in parallel.
- **Content Generation:** Create large volumes of text, such as product descriptions or articles.

- **Document Review and Summarization:** Automate the review and summarization of lengthy documents.
- **Customer Support Automation:** Handle numerous queries simultaneously for faster responses.
- **Data Extraction and Analysis:** Extract and analyze information from vast amounts of unstructured data.
- **Natural Language Processing (NLP) Tasks:** Perform tasks like sentiment analysis or translation on large datasets.
- **Marketing and Personalization:** Generate personalized content and recommendations at scale.

For more information on [getting started with global batch deployments](#).

o1-preview and o1-mini models limited access

The `o1-preview` and `o1-mini` models are now available for API access and model deployment. **Registration is required, and access will be granted based on Microsoft's eligibility criteria.**

Request access: [limited access model application](#) 

Customers who were already approved and have access to the model through the early access playground don't need to apply again, you'll automatically be granted API access. Once access has been granted, you'll need to create a deployment for each model.

API support:

Support for the **o1 series** models was added in API version `2024-09-01-preview`.

The `max_tokens` parameter has been deprecated and replaced with the new `max_completion_tokens` parameter. **o1 series** models will only work with the `max_completion_tokens` parameter.

Region availability:

Models are available for standard and global standard deployment in East US2 and Sweden Central for approved customers.

New GPT-4o Realtime API for speech and audio public preview

Azure OpenAI GPT-4o audio is part of the GPT-4o model family that supports low-latency, "speech in, speech out" conversational interactions. The GPT-4o audio `realtime` API is designed to handle real-time, low-latency conversational interactions, making it a great fit for use cases involving live interactions between a user and a model, such as customer support agents, voice assistants, and real-time translators.

The `gpt-4o-realtime-preview` model is available for global deployments in [East US 2](#) and [Sweden Central](#) regions.

For more information, see the [GPT-4o real-time audio documentation](#).

Global batch support updates

Global batch now supports GPT-4o (2024-08-06). See the [global batch getting started guide](#) for more information.

September 2024

Azure OpenAI Studio UX updates

On September 19, when you access the [Azure OpenAI Studio](#) you'll begin to no longer see the legacy AI Foundry portal by default. If needed you'll still be able to go back to the previous experience by using the **Switch to the old look** toggle in the top bar of the UI for the next couple of weeks. If you switch back to legacy AI Foundry portal, it helps if you fill out the feedback form to let us know why. We're actively monitoring this feedback to improve the new experience.

GPT-4o 2024-08-06 provisioned deployments

GPT-4o 2024-08-06 is now available for provisioned deployments in East US, East US 2, North Central US, and Sweden Central. It's also available for global provisioned deployments.

For the latest information on model availability, see the [models page](#).

NEW Global provisioned deployment type

Global deployments are available in the same Azure OpenAI resources as non-global deployment types but allow you to leverage Azure's global infrastructure to dynamically route traffic to the data center with best availability for each request. Global provisioned

deployments provide reserved model processing capacity for high and predictable throughput using Azure global infrastructure. Global provisioned deployments are supported on `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18` models.

For more information, see the [deployment types guide](#).

NEW o1-preview and o1-mini models available for limited access

The Azure OpenAI `o1-preview` and `o1-mini` models are designed to tackle reasoning and problem-solving tasks with increased focus and capability. These models spend more time processing and understanding the user's request, making them exceptionally strong in areas like science, coding, and math compared to previous iterations.

Key capabilities of the o1 series

- **Complex Code Generation:** Capable of generating algorithms and handling advanced coding tasks to support developers.
- **Advanced Problem Solving:** Ideal for comprehensive brainstorming sessions and addressing multifaceted challenges.
- **Complex Document Comparison:** Perfect for analyzing contracts, case files, or legal documents to identify subtle differences.
- **Instruction Following and Workflow Management:** Particularly effective for managing workflows requiring shorter contexts.

Model variants

- `o1-preview`: `o1-preview` is the more capable of the `o1` series models.
- `o1-mini`: `o1-mini` is the faster and cheaper of the `o1` series models.

Model version: `2024-09-12`

Request access: [limited access model application](#)

Limitations

The `o1` series models are currently in preview and don't include some features available in other models, such as image understanding and structured outputs which are available in the latest GPT-4o model. For many tasks, the generally available GPT-4o models might still be more suitable.

Safety

OpenAI has incorporated additional safety measures into the `o1` models, including new techniques to help the models refuse unsafe requests. These advancements make the `o1` series some of the most robust models available.

Availability

The `o1-preview` and `o1-mini` are available in the East US2 region for limited access through the [AI Foundry portal](#) early access playground. Data processing for the `o1` models might occur in a different region than where they are available for use.

To try the `o1-preview` and `o1-mini` models in the early access playground **registration is required, and access will be granted based on Microsoft's eligibility criteria.**

Request access: [limited access model application](#)

Once access has been granted, you will need to:

1. Navigate to <https://ai.azure.com/resources> and select a resource in the `eastus2` region. If you don't have an Azure OpenAI resource in this region you'll need to [create one](#).
2. Once the `eastus2` Azure OpenAI resource is selected, in the upper left-hand panel under **Playgrounds** select **Early access playground (preview)**.

August 2024

GPT-4o 2024-08-06 structured outputs

- Available for standard and global deployments in [all US regions and Sweden Central](#).
- This model adds support for [structured outputs](#).

GPT-4o mini provisioned deployments

GPT-4o mini is now available for provisioned deployments in Canada East, East US, East US2, North Central US, and Sweden Central.

For the latest information on model availability, see the [models page](#).

GPT-4o fine-tuning (Public Preview)

GPT-4o fine-tuning is now available for Azure OpenAI in public preview in North Central US and Sweden Central.

For more information, see our [blog post](#).

New preview API release

API version `2024-07-01-preview` is the latest dataplane authoring & inference API release. It replaces API version `2024-05-01-preview` and adds support for:

- [Batch API support added](#)
- [Vector store chunking strategy parameters](#)
- `max_num_results` that the file search tool should output.

For more information see our [reference documentation](#)

GPT-4o mini regional availability

- GPT-4o mini is available for standard and global standard deployment in the East US and Sweden Central regions.
- GPT-4o mini is available for global batch deployment in East US, Sweden Central, and West US regions.

Evaluations guide

- New blog post on [getting started with model evaluations](#). We recommend using this guide as part of the [model upgrade and retirement process](#).

Latest GPT-4o model available in the early access playground (preview)

On August 6, 2024, OpenAI [announced](#) the latest version of their flagship GPT-4o model version `2024-08-06`. GPT-4o `2024-08-06` has all the capabilities of the previous version as well as:

- An enhanced ability to support complex structured outputs.
- Max output tokens have been increased from 4,096 to 16,384.

Azure customers can test out GPT-4o `2024-08-06` today in the new AI Foundry early access playground (preview).

Unlike the previous early access playground, the AI Foundry portal early access playground (preview) doesn't require you to have a resource in a specific region.

ⓘ Note

Prompts and completions made through the early access playground (preview) might be processed in any Azure OpenAI region, and are currently subject to a 10 request per minute per Azure subscription limit. This limit might change in the future.

Azure OpenAI Service abuse monitoring is enabled for all early access playground users even if approved for modification; default content filters are enabled and cannot be modified.

To test out GPT-4o `2024-08-06`, sign-in to the Azure AI early access playground (preview) using this [link](#).

Global batch deployments are now available

The Azure OpenAI Batch API is designed to handle large-scale and high-volume processing tasks efficiently. Process asynchronous groups of requests with separate quota, with 24-hour target turnaround, at [50% less cost than global standard](#). With batch processing, rather than send one request at a time you send a large number of requests in a single file. Global batch requests have a separate enqueued token quota avoiding any disruption of your online workloads.

Key use cases include:

- **Large-Scale Data Processing:** Quickly analyze extensive datasets in parallel.
- **Content Generation:** Create large volumes of text, such as product descriptions or articles.
- **Document Review and Summarization:** Automate the review and summarization of lengthy documents.
- **Customer Support Automation:** Handle numerous queries simultaneously for faster responses.
- **Data Extraction and Analysis:** Extract and analyze information from vast amounts of unstructured data.

- **Natural Language Processing (NLP) Tasks:** Perform tasks like sentiment analysis or translation on large datasets.
- **Marketing and Personalization:** Generate personalized content and recommendations at scale.

For more information on [getting started with global batch deployments](#).

July 2024

GPT-4o mini is now available for fine-tuning

GPT-4o mini fine-tuning is [now available in public preview](#) in Sweden Central and in North Central US.

Assistants File Search tool is now billed

The [file search](#) tool for Assistants now has additional charges for usage. See the [pricing page](#) [↗] for more information.

GPT-4o mini model available for deployment

GPT-4o mini is the latest Azure OpenAI model first [announced on July 18, 2024](#) [↗]:

"GPT-4o mini allows customers to deliver stunning applications at a lower cost with blazing speed. GPT-4o mini is significantly smarter than GPT-3.5 Turbo—scoring 82% on Measuring Massive Multitask Language Understanding (MMLU) compared to 70%—and is more than 60% cheaper.¹ The model delivers an expanded 128K context window and integrates the improved multilingual capabilities of GPT-4o, bringing greater quality to languages from around the world."

The model is currently available for both [standard and global standard deployment](#) in the East US region.

For information on model quota, consult the [quota and limits page](#) and for the latest info on model availability refer to the [models page](#).

New Responsible AI default content filtering policy

The new default content filtering policy `DefaultV2` delivers the latest safety and security mitigations for the GPT model series (text), including:

- Prompt Shields for jailbreak attacks on user prompts (filter),
- Protected material detection for text (filter) on model completions
- Protected material detection for code (annotate) on model completions

While there are no changes to content filters for existing resources and deployments (default or custom content filtering configurations remain unchanged), new resources and GPT deployments will automatically inherit the new content filtering policy `DefaultV2`. Customers have the option to switch between safety defaults and create custom content filtering configurations.

Refer to our [Default safety policy documentation](#) for more information.

New GA API release

API version `2024-06-01` is the latest GA data plane inference API release. It replaces API version `2024-02-01` and adds support for:

- embeddings `encoding_format` & `dimensions` parameters.
- chat completions `logprobs` & `top_logprobs` parameters.

Refer to our [data plane inference reference documentation](#) for more information.

Expansion of regions available for global standard deployments of gpt-4o

GPT-4o is now available for [global standard deployments](#) in:

- australiaeast
- brazilsouth
- canadaeast
- eastus
- eastus2
- francecentral git
- germanywestcentral
- japaneast
- koreacentral
- northcentralus
- norwayeast
- polandcentral
- southafricanorth
- southcentralus
- southindia

- swedencentral
- switzerlandnorth
- uksouth
- westeurope
- westus
- westus3

For information on global standard quota, consult the [quota and limits page](#).

June 2024

Retirement date updates

- Updated `gpt-35-turbo` 0301 retirement date to no earlier than October 1, 2024.
- Updated `gpt-35-turbo` & `gpt-35-turbo-16k` 0613 retirement date to October 1, 2024.
- Updated `gpt-4` & `gpt-4-32k` 0314 deprecation date to October 1, 2024, and retirement date to June 6, 2025.

Refer to our [model retirement guide](#) for the latest information on model deprecation and retirement.

Token based billing for fine-tuning

- Azure OpenAI fine-tuning billing is now based on the number of tokens in your training file – instead of the total elapsed training time. This can result in a significant cost reduction for some training runs, and makes estimating fine-tuning costs much easier. To learn more, you can consult the [official announcement](#) [↗].

GPT-4o released in new regions

- GPT-4o is now also available in:
 - Sweden Central for standard regional deployment.
 - Australia East, Canada East, Japan East, Korea Central, Sweden Central, Switzerland North, & West US 3 for provisioned deployment.

For the latest information on model availability, see the [models page](#).

Customer-managed key (CMK) support for Assistants

Threads and Files in Assistants now supports CMK in the following region:

- West US 3


May 2024

GPT-4o provisioned deployments

`gpt-4o` Version: `2024-05-13` is available for both standard and provisioned deployments. Provisioned and standard model deployments accept both text and image/vision inference requests. For information on model regional availability, consult the model matrix for [provisioned deployments](#).

Assistants v2 (preview)

A refresh of the Assistants API is now publicly available. It contains the following updates:

- [File search tool and vector storage](#) 
- [Max completion and max prompt token support](#) for managing token usage.
- `tool_choice` [parameter](#) for forcing the Assistant to use a specified tool. You can now create messages with the [assistant](#) role to create custom conversation histories in Threads.
- Support for `temperature`, `top_p`, `response_format` [parameters](#).
- Streaming and polling support. You can use the helper functions in our Python SDK to create runs and stream responses. We have also added polling SDK helpers to share object status updates without the need for polling.
- Experiment with [Logic Apps and Function Calling using Azure OpenAI Studio](#). Import your REST APIs implemented in Logic Apps as functions and the studio invokes the function (as a Logic Apps workflow) automatically based on the user prompt.
- AutoGen by Microsoft Research provides a multi-agent conversation framework to enable convenient building of Large Language Model (LLM) workflows across a wide range of applications. Azure OpenAI assistants are now integrated into AutoGen via `GPTAssistantAgent`, a new experimental agent that lets you seamlessly add Assistants into AutoGen-based multi-agent workflows. This enables multiple Azure OpenAI assistants that could be task or domain specialized to collaborate and tackle complex tasks.
- Support for fine-tuned `gpt-3.5-turbo-0125` [models](#) in the following regions:
 - East US 2

- Sweden Central
- Expanded [regional support](#) for:
 - Japan East
 - UK South
 - West US
 - West US 3
 - Norway east

For more information, see the [blog post](#) about assistants.

GPT-4o model general availability (GA)

GPT-4o ("o is for "omni") is the latest model from OpenAI launched on May 13, 2024.

- GPT-4o integrates text, and images in a single model, enabling it to handle multiple data types simultaneously. This multimodal approach enhances accuracy and responsiveness in human-computer interactions.
- GPT-4o matches GPT-4 Turbo in English text and coding tasks while offering superior performance in non-English languages and in vision tasks, setting new benchmarks for AI capabilities.

For information on model regional availability, see the [models page](#).

Global standard deployment type (preview)

Global deployments are available in the same Azure OpenAI resources as non-global offers but allow you to leverage Azure's global infrastructure to dynamically route traffic to the data center with best availability for each request. Global standard provides the highest default quota for new models and eliminates the need to load balance across multiple resources.

For more information, see the [deployment types guide](#).

Fine-tuning updates

- GPT-4 fine-tuning is [now available in public preview](#).
- Added support for [seed](#), [events](#), [full validation statistics](#), and [checkpoints](#) as part of the `2024-05-01-preview` API release.

DALL-E and GPT-4 Turbo Vision GA configurable content filters

Create custom content filters for your DALL-E 2 and 3, GPT-4 Turbo with Vision GA (`turbo-2024-04-09`), and GPT-4o deployments. [Content filtering](#)

Asynchronous Filter available for all Azure OpenAI customers

Running filters asynchronously for improved latency in streaming scenarios is now available for all Azure OpenAI customers. [Content filtering](#)

Prompt Shields

Prompt Shields protect applications powered by Azure OpenAI models from two types of attacks: direct (jailbreak) and indirect attacks. Indirect Attacks (also known as Indirect Prompt Attacks or Cross-Domain Prompt Injection Attacks) are a type of attack on systems powered by Generative AI models that might occur when an application processes information that wasn't directly authored by either the developer of the application or the user. [Content filtering](#)

2024-05-01-preview API release

- For more information, see the [API version lifecycle](#).

GPT-4 Turbo model general availability (GA)

The latest GA release of GPT-4 Turbo is:

- `gpt-4` **Version:** `turbo-2024-04-09`

This is the replacement for the following preview models:

- `gpt-4` **Version:** `1106-Preview`
- `gpt-4` **Version:** `0125-Preview`
- `gpt-4` **Version:** `vision-preview`

Differences between OpenAI and Azure OpenAI GPT-4 Turbo GA Models

- OpenAI's version of the latest `0409` turbo model supports JSON mode and function calling for all inference requests.

- Azure OpenAI's version of the latest `turbo-2024-04-09` currently doesn't support the use of JSON mode and function calling when making inference requests with image (vision) input. Text based input requests (requests without `image_url` and inline images) do support JSON mode and function calling.

Differences from gpt-4 vision-preview

- Azure AI specific Vision enhancements integration with GPT-4 Turbo with Vision isn't supported for `gpt-4` **Version:** `turbo-2024-04-09`. This includes Optical Character Recognition (OCR), object grounding, video prompts, and improved handling of your data with images.

Important

Vision enhancements preview features including Optical Character Recognition (OCR), object grounding, video prompts will be retired and no longer available once `gpt-4` **Version:** `vision-preview` is upgraded to `turbo-2024-04-09`. If you are currently relying on any of these preview features, this automatic model upgrade will be a breaking change.

GPT-4 Turbo provisioned managed availability

- `gpt-4` **Version:** `turbo-2024-04-09` is available for both standard and provisioned deployments. Currently the provisioned version of this model **doesn't support image/vision inference requests**. Provisioned deployments of this model only accept text input. Standard model deployments accept both text and image/vision inference requests.

Deploying GPT-4 Turbo with Vision GA

To deploy the GA model from the AI Foundry portal, select `GPT-4` and then choose the `turbo-2024-04-09` version from the dropdown menu. The default quota for the `gpt-4-turbo-2024-04-09` model will be the same as current quota for GPT-4-Turbo. See the [regional quota limits](#).

April 2024

Fine-tuning is now supported in two new regions East US 2 and Switzerland West

Fine-tuning is now available with support for:

East US 2

- `gpt-35-turbo` (0613)
- `gpt-35-turbo` (1106)
- `gpt-35-turbo` (0125)

Switzerland West

- `babbage-002`
- `davinci-002`
- `gpt-35-turbo` (0613)
- `gpt-35-turbo` (1106)
- `gpt-35-turbo` (0125)

Check the [models page](#), for the latest information on model availability and fine-tuning support in each region.

Multi-turn chat training examples

Fine-tuning now supports [multi-turn chat training examples](#).

GPT-4 (0125) is available for Azure OpenAI On Your Data

You can now use the GPT-4 (0125) model in [available regions](#) with Azure OpenAI On Your Data.

March 2024

Risks & Safety monitoring in Azure OpenAI Studio

Azure OpenAI Studio now provides a Risks & Safety dashboard for each of your deployments that uses a content filter configuration. Use it to check the results of the filtering activity. Then you can adjust your filter configuration to better serve your business needs and meet Responsible AI principles.

Azure OpenAI On Your Data updates

- You can now connect to an Elasticsearch vector database to be used with [Azure OpenAI On Your Data](#).
- You can use the [chunk size parameter](#) during data ingestion to set the maximum number of tokens of any given chunk of data in your index.

2024-02-01 general availability (GA) API released

This is the latest GA API release and is the replacement for the previous `2023-05-15` GA release. This release adds support for the latest Azure OpenAI GA features like Whisper, DALL-E-3, fine-tuning, on your data, and more.

Features that are in preview such as Assistants, text to speech (TTS), and some of the "on your data" datasources, require a preview API version. For more information, check out our [API version lifecycle guide](#).

Whisper general availability (GA)

The Whisper speech to text model is now GA for both REST and Python. Client library SDKs are currently still in public preview.

Try out Whisper by following a [quickstart](#).

DALL-E 3 general availability (GA)

DALL-E 3 image generation model is now GA for both REST and Python. Client library SDKs are currently still in public preview.

Try out DALL-E 3 by following a [quickstart](#).

New regional support for DALL-E 3

You can now access DALL-E 3 with an Azure OpenAI resource in the `East US` or `AustraliaEast` Azure region, in addition to `SwedenCentral`.

Model deprecations and retirements

We have added a page to track [model deprecations and retirements](#) in Azure OpenAI Service. This page provides information about the models that are currently available, deprecated, and retired.

2024-03-01-preview API released

`2024-03-01-preview` has all the same functionality as `2024-02-15-preview` and adds two new parameters for embeddings:

- `encoding_format` allows you to specify the format to generate embeddings in `float`, or `base64`. The default is `float`.
- `dimensions` allows you set the number of output embeddings. This parameter is only supported with the new third generation embeddings models: `text-embedding-3-large`, `text-embedding-3-small`. Typically larger embeddings are more expensive from a compute, memory, and storage perspective. Being able to adjust the number of dimensions allows more control over overall cost and performance. The `dimensions` parameter isn't supported in all versions of the OpenAI 1.x Python library, to take advantage of this parameter we recommend upgrading to the latest version: `pip install openai --upgrade`.

If you're currently using a preview API version to take advantage of the latest features, we recommend consulting the [API version lifecycle](#) article to track how long your current API version will be supported.

Update to GPT-4-1106-Preview upgrade plans

The deployment upgrade of `gpt-4` 1106-Preview to `gpt-4` 0125-Preview scheduled for March 8, 2024 is no longer taking place. Deployments of `gpt-4` versions 1106-Preview and 0125-Preview set to "Auto-update to default" and "Upgrade when expired" will start to be upgraded after a stable version of the model is released.

For more information on the upgrade process refer to the [models page](#).

February 2024

GPT-3.5-turbo-0125 model available

This model has various improvements, including higher accuracy at responding in requested formats and a fix for a bug which caused a text encoding issue for non-English language function calls.

For information on model regional availability and upgrades refer to the [models page](#).

Third generation embeddings models available

- `text-embedding-3-large`
- `text-embedding-3-small`

In testing, OpenAI reports both the large and small third generation embeddings models offer better average multi-language retrieval performance with the [MIRACL](#) benchmark while still maintaining better performance for English tasks with the [MTEB](#) benchmark than the second generation `text-embedding-ada-002` model.

For information on model regional availability and upgrades refer to the [models page](#).

GPT-3.5 Turbo quota consolidation

To simplify migration between different versions of the GPT-3.5-Turbo models (including 16k), we'll be consolidating all GPT-3.5-Turbo quota into a single quota value.

- Any customers who have increased quota approved will have combined total quota that reflects the previous increases.
- Any customer whose current total usage across model versions is less than the default will get a new combined total quota by default.

GPT-4-0125-preview model available

The `gpt-4` model version `0125-preview` is now available on Azure OpenAI Service in the East US, North Central US, and South Central US regions. Customers with deployments of `gpt-4` version `1106-preview` will be automatically upgraded to `0125-preview` in the coming weeks.

For information on model regional availability and upgrades refer to the [models page](#).

Assistants API public preview

Azure OpenAI now supports the API that powers OpenAI's GPTs. Azure OpenAI Assistants (Preview) allows you to create AI assistants tailored to your needs through custom instructions and advanced tools like code interpreter, and custom functions. To learn more, see:

- [Quickstart](#)

- [Concepts](#)
- [In-depth Python how-to](#)
- [Code Interpreter](#)
- [Function calling](#)
- [Assistants model & region availability](#)
- [Assistants Python & REST reference](#)
- [Assistants Samples](#) ↗

OpenAI text to speech voices public preview

Azure OpenAI Service now supports text to speech APIs with OpenAI's voices. Get AI-generated speech from the text you provide. To learn more, see the [overview guide](#) and try the [quickstart](#).

ⓘ Note

Azure AI Speech also supports OpenAI text to speech voices. To learn more, see [OpenAI text to speech voices via Azure OpenAI Service or via Azure AI Speech guide](#).

New Fine-tuning capabilities and model support

- [Continuous fine-tuning](#) ↗
- [Fine-tuning & function calling](#)
- [gpt-35-turbo 1106 support](#)

New regional support for Azure OpenAI On Your Data

You can now use Azure OpenAI On Your Data in the following Azure region:

- South Africa North

Azure OpenAI On Your Data general availability

- [Azure OpenAI On Your Data](#) is now generally available.

December 2023

Azure OpenAI On Your Data

- Full VPN and private endpoint support for Azure OpenAI On Your Data, including security support for: storage accounts, Azure OpenAI resources, and Azure AI Search service resources.
- New article for using [Azure OpenAI On Your Data configuration](#) by protecting data with virtual networks and private endpoints.

GPT-4 Turbo with Vision now available

GPT-4 Turbo with Vision on Azure OpenAI service is now in public preview. GPT-4 Turbo with Vision is a large multimodal model (LMM) developed by OpenAI that can analyze images and provide textual responses to questions about them. It incorporates both natural language processing and visual understanding. With enhanced mode, you can use the [Azure AI Vision](#) features to generate additional insights from the images.

- Explore the capabilities of GPT-4 Turbo with Vision in a no-code experience using the [Azure OpenAI Playground](#) [↗](#). Learn more in the [Quickstart guide](#).
- Vision enhancement using GPT-4 Turbo with Vision is now available in the [Azure OpenAI Playground](#) [↗](#) and includes support for Optical Character Recognition, object grounding, image support for "add your data," and support for video prompt.
- Make calls to the chat API directly using the [REST API](#) [↗](#).
- Region availability is currently limited to `SwitzerlandNorth`, `SwedenCentral`, `WestUS`, and `AustraliaEast`.
- Learn more about the known limitations of GPT-4 Turbo with Vision and other [frequently asked questions](#).

November 2023

New data source support in Azure OpenAI On Your Data

- You can now use [Azure Cosmos DB for MongoDB vCore](#) and URLs/web addresses as data sources to ingest your data and chat with a supported Azure OpenAI model.

GPT-4 Turbo Preview & GPT-3.5-Turbo-1106 released

Both models are the latest release from OpenAI with improved instruction following, [JSON mode](#), [reproducible output](#), and parallel function calling.

- **GPT-4 Turbo Preview** has a max context window of 128,000 tokens and can generate 4,096 output tokens. It has the latest training data with knowledge up to April 2023. This model is in preview and isn't recommended for production use. All deployments of this preview model will be automatically updated in place once the stable release becomes available.
- **GPT-3.5-Turbo-1106** has a max context window of 16,385 tokens and can generate 4,096 output tokens.

For information on model regional availability consult the [models page](#).

The models have their own unique per region [quota allocations](#).

DALL-E 3 public preview

DALL-E 3 is the latest image generation model from OpenAI. It features enhanced image quality, more complex scenes, and improved performance when rendering text in images. It also comes with more aspect ratio options. DALL-E 3 is available through OpenAI Studio and through the REST API. Your OpenAI resource must be in the `SwedenCentral` Azure region.

DALL-E 3 includes built-in prompt rewriting to enhance images, reduce bias, and increase natural variation.

Try out DALL-E 3 by following a [quickstart](#).

Responsible AI

- **Expanded customer configurability:** All Azure OpenAI customers can now configure all severity levels (low, medium, high) for the categories hate, violence, sexual and self-harm, including filtering only high severity content. [Configure content filters](#)
- **Content Credentials in all DALL-E models:** AI-generated images from all DALL-E models now include a digital credential that discloses the content as AI-generated. Applications that display image assets can leverage the open source [Content Authenticity Initiative SDK](#) to display credentials in their AI generated images. [Content Credentials in Azure OpenAI](#)
- **New RAI models**
 - **Jailbreak risk detection:** Jailbreak attacks are user prompts designed to provoke the Generative AI model into exhibiting behaviors it was trained to avoid or to break the rules set in the System Message. The jailbreak risk detection model is

optional (default off), and available in annotate and filter model. It runs on user prompts.

- **Protected material text:** Protected material text describes known text content (for example, song lyrics, articles, recipes, and selected web content) that can be outputted by large language models. The protected material text model is optional (default off), and available in annotate and filter model. It runs on LLM completions.
- **Protected material code:** Protected material code describes source code that matches a set of source code from public repositories, which can be outputted by large language models without proper citation of source repositories. The protected material code model is optional (default off), and available in annotate and filter model. It runs on LLM completions.

[Configure content filters](#)

- **Blocklists:** Customers can now quickly customize content filter behavior for prompts and completions further by creating a custom blocklist in their filters. The custom blocklist allows the filter to take action on a customized list of patterns, such as specific terms or regex patterns. In addition to custom blocklists, we provide a Microsoft profanity blocklist (English). [Use blocklists](#)

October 2023

New fine-tuning models (preview)

- `gpt-35-turbo-0613` is [now available for fine-tuning](#).
- `babbage-002` and `davinci-002` are [now available for fine-tuning](#). These models replace the legacy ada, babbage, curie, and davinci base models that were previously available for fine-tuning.
- Fine-tuning availability is limited to certain regions. Check the [models page](#), for the latest information on model availability in each region.
- Fine-tuned models have different [quota limits](#) than regular models.
- [Tutorial: fine-tuning GPT-3.5-Turbo](#)

Azure OpenAI On Your Data

- New [custom parameters](#) for determining the number of retrieved documents and strictness.

- The strictness setting sets the threshold to categorize documents as relevant to your queries.
- The retrieved documents setting specifies the number of top-scoring documents from your data index used to generate responses.
- You can see data ingestion/upload status in the Azure OpenAI Studio.
- Support for private endpoints & VPNs for blob containers.

September 2023

GPT-4

GPT-4 and GPT-4-32k are now available to all Azure OpenAI Service customers. Customers no longer need to apply for the waitlist to use GPT-4 and GPT-4-32k (the Limited Access registration requirements continue to apply for all Azure OpenAI models). Availability might vary by region. Check the [models page](#), for the latest information on model availability in each region.

GPT-3.5 Turbo Instruct

Azure OpenAI Service now supports the GPT-3.5 Turbo Instruct model. This model has performance comparable to `text-davinci-003` and is available to use with the Completions API. Check the [models page](#), for the latest information on model availability in each region.

Whisper public preview

Azure OpenAI Service now supports speech to text APIs powered by OpenAI's Whisper model. Get AI-generated text based on the speech audio you provide. To learn more, check out the [quickstart](#).

ⓘ Note

Azure AI Speech also supports OpenAI's Whisper model via the batch transcription API. To learn more, check out the [Create a batch transcription](#) guide. Check out [What is the Whisper model?](#) to learn more about when to use Azure AI Speech vs. Azure OpenAI Service.

New Regions

- Azure OpenAI is now also available in the Sweden Central, and Switzerland North regions. Check the [models page](#), for the latest information on model availability in each region.

Regional quota limits increases

- Increases to the max default quota limits for certain models and regions. Migrating workloads to [these models and regions](#) will allow you to take advantage of higher Tokens per minute (TPM).

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.

If you're currently using the `2023-03-15-preview` API, we recommend migrating to the GA `2023-05-15` API. If you're currently using API version `2022-12-01` this API remains GA, but doesn't 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](#).
- **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 didn't 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,
}
```

Content filtering is temporarily off by default. Azure content moderation works differently than Azure OpenAI. Azure OpenAI runs content filters during the generation call to detect harmful or abusive content and filters them from the response. [Learn More](#)

These models will be re-enabled in Q1 2023 and be on by default.

Customer actions

- [Contact Azure Support](#)  if you would like these turned on for your subscription.
- [Apply for filtering modifications](#) , if you would like to have them remain off. (This option will be for low-risk use cases only.)

Next steps

Learn more about the [underlying models that power Azure OpenAI](#).

Feedback

Was this page helpful?

 Yes

 No

[Provide product feedback](#)  | [Get help at Microsoft Q&A](#)