

Azure OpenAI Service frequently asked questions

FAQ

If you can't find answers to your questions in this document, and still need help check the [Azure AI services support options guide](#). Azure OpenAI is part of Azure AI services.

Data and Privacy

Do you use my company data to train any of the models?

Azure OpenAI doesn't use customer data to retrain models. For more information, see the [Azure OpenAI data, privacy, and security guide](#).

General

Does Azure OpenAI support custom API headers? We append additional custom headers to our API requests and are seeing HTTP 431 failure errors.

Our current APIs allow up to 10 custom headers, which are passed through the pipeline, and returned. We have noticed some customers now exceed this header count resulting in HTTP 431 errors. There is no solution for this error, other than to reduce header volume. In future API versions we will no longer pass through custom headers. We recommend customers not depend on custom headers in future system architectures.

Does Azure OpenAI work with the latest Python library released by OpenAI (version >=1.0)?

Azure OpenAI is supported by the latest release of the [OpenAI Python library \(version >= 1.0\)](#). However, it's important to note migration of your codebase using `openai migrate` is not supported and will not work with code that targets Azure OpenAI.

I can't find GPT-4 Turbo Preview, where is it?

GPT-4 Turbo Preview is the `gpt-4` (1106-preview) model. To deploy this model, under **Deployments** select model **gpt-4**. For **Model version** select **1106-preview**. To check which regions this model is available, refer to the [models page](#).

Does Azure OpenAI support GPT-4?

Azure OpenAI supports the latest GPT-4 models. It supports both GPT-4 and GPT-4-32K.

How do the capabilities of Azure OpenAI compare to OpenAI?

Azure OpenAI Service gives customers advanced language AI with OpenAI GPT-3, Codex, and DALL-E models with the security and enterprise promise of Azure. Azure OpenAI codevelops 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.


Does Azure OpenAI support VNETs and Private Endpoints?

Yes, as part of Azure AI services, Azure OpenAI supports VNETs and Private Endpoints. To learn more, consult the [Azure AI services virtual networking guidance](#).

Do the GPT-4 models currently support image input?

No, GPT-4 is designed by OpenAI to be multimodal, but currently only text input and output are supported.

How do I apply for new use cases?

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](#) . Note that this is required anytime you would like to use the service for a new use case you didn't originally apply for.

I'm trying to use embeddings and received the error "InvalidRequestError: Too many inputs. The max number of inputs is 16." How do I fix this?

This error typically occurs when you try to send a batch of text to embed in a single API request as an array. Currently Azure OpenAI only supports arrays of embeddings with multiple inputs for the `text-embedding-ada-002` Version 2 model. This model version supports an array consisting of up to 16 inputs per API request. The array can be up to 8,191 tokens in length when using the text-embedding-ada-002 (Version 2) model.

Where can I read about better ways to use Azure OpenAI to get the responses I want from the service?

Check out our [introduction to prompt engineering](#). While these models are powerful, their behavior is also very sensitive to the prompts they receive from the user. This makes prompt construction an important skill to develop. After you've completed the introduction, check out our article on [system messages](#).

My guest account has been given access to an Azure OpenAI resource, but I'm unable to access that resource in the Azure AI Foundry portal. How do I enable access?

This is expected behavior when using the default sign-in experience for the [Azure AI Foundry](#).

To access Azure AI Foundry from a guest account that has been granted access to an Azure OpenAI resource:

1. Open a private browser session and then navigate to <https://ai.azure.com>.
2. Rather than immediately entering your guest account credentials instead select **Sign-in options**
3. Now select **Sign in to an organization**
4. Enter the domain name of the organization that granted your guest account access to the Azure OpenAI resource.
5. Now sign-in with your guest account credentials.

You should now be able to access the resource via the Azure AI Foundry portal.

Alternatively if you're signed into the [Azure portal](#) from the Azure OpenAI resource's Overview pane you can select **Go to Azure AI Foundry** to automatically sign in with the appropriate organizational context.

When I ask GPT-4 which model it's running, it tells me it's running GPT-3. Why does this happen?

Azure OpenAI models (including GPT-4) being unable to correctly identify what model is running is expected behavior.

Why does this happen?

Ultimately, the model is performing next **token** prediction in response to your question. The model doesn't have any native ability to query what model version is currently being run to answer your question. To answer this question, you can always go to **Azure AI**

Foundry > Management > Deployments > and consult the model name column to confirm what model is currently associated with a given deployment name.

The questions, "What model are you running?" or "What is the latest model from OpenAI?" produce similar quality results to asking the model what the weather will be today. It might return the correct result, but purely by chance. On its own, the model has no real-world information other than what was part of its training/training data. In the case of GPT-4, as of August 2023 the underlying training data goes only up to September 2021. GPT-4 wasn't released until March 2023, so barring OpenAI releasing a new version with updated training data, or a new version that is fine-tuned to answer those specific questions, it's expected behavior for GPT-4 to respond that GPT-3 is the latest model release from OpenAI.

If you wanted to help a GPT based model to accurately respond to the question "what model are you running?", you would need to provide that information to the model through techniques like [prompt engineering of the model's system message](#), [Retrieval Augmented Generation \(RAG\)](#) which is the technique used by [Azure OpenAI on your data](#) where up-to-date information is injected to the system message at query time, or via [fine-tuning](#) where you could fine-tune specific versions of the model to answer that question in a certain way based on model version.

To learn more about how GPT models are trained and work we recommend watching [Andrej Karpathy's talk from Build 2023 on the state of GPT](#) [↗](#).

I asked the model when its knowledge cutoff is and it gave me a different answer than what is on the Azure OpenAI model's page. Why does this happen?

This is expected behavior. The models aren't able to answer questions about themselves. If you want to know when the knowledge cutoff for the model's training data is, consult the [models page](#).

I asked the model a question about something that happened recently before the knowledge cutoff and it got the answer wrong. Why does this happen?

This is expected behavior. First there's no guarantee that every recent event was part of the model's training data. And even when information was part of the training data, without using additional techniques like Retrieval Augmented Generation (RAG) to help ground the model's responses there's always a chance of ungrounded responses occurring. Both Azure OpenAI's [use your data feature](#) and [Bing Chat](#) use Azure OpenAI models combined with Retrieval Augmented Generation to help further ground model responses.

The frequency that a given piece of information appeared in the training data can also impact the likelihood that the model will respond in a certain way.

Asking the latest GPT-4 Turbo Preview model about something that changed more recently like "Who is the prime minister of New Zealand?", is likely to result in the fabricated response `Jacinda Ardern`. However, asking the model "When did `Jacinda Ardern` step down as prime minister?" Tends to yield an accurate response which demonstrates training data knowledge going to at least January of 2023.

So while it is possible to probe the model with questions to guess its training data knowledge cutoff, the [model's page](#) is the best place to check a model's knowledge cutoff.

Where do I access pricing information for legacy models, which are no longer available for new deployments?

Legacy pricing information is available via a [downloadable PDF file](#). For all other models, consult the [official pricing page](#).

How do I fix InternalServerError - 500 - Failed to create completion as the model generated invalid Unicode output?

You can minimize the occurrence of these errors by reducing the temperature of your prompts to less than 1 and ensuring you're using a client with retry logic. Reattempting the request often results in a successful response.

We noticed charges associated with API calls that failed to complete with status code 400. Why are failed API calls generating a charge?

If the service performs processing, you will be charged even if the status code is not successful (not 200). Common examples of this are, a 400 error due to a content filter or input limit, or a 408 error due to a timeout. Charges will also occur when a `status 200` is received with a `finish_reason` of `content_filter`. In this case the prompt did not have any issues, but the completion generated by the model was detected to violate the content filtering rules, which result in the completion being filtered. If the service doesn't perform processing, you won't be charged. For example, a 401 error due to authentication or a 429 error due to exceeding the Rate Limit.

Getting access to Azure OpenAI Service

How do I get access to Azure OpenAI?

A Limited Access registration form is not required to access most Azure OpenAI models. Learn more on the [Azure OpenAI Limited Access page](#).

Learning more and where to ask questions

Where can I read about the latest updates to Azure OpenAI?

For monthly updates, see our [what's new page](#).

Where can I get training to get started learning and build my skills around Azure OpenAI?

Check out our [introduction to Azure OpenAI training course](#).

Where can I post questions and see answers to other common questions?

- We recommend posting questions on [Microsoft Q&A](#).
- Alternatively, you can post questions on [Stack Overflow](#) [↗](#).

Where do I go for Azure OpenAI customer support?

Azure OpenAI is part of Azure AI services. You can learn about all the support options for Azure AI services in the [support and help options guide](#).

Models and fine-tuning

What models are available?

Consult the Azure OpenAI [model availability guide](#).

Where can I find out what region a model is available in?

Consult the Azure OpenAI [model availability guide](#) for region availability.

What are the SLAs (Service Level Agreements) in Azure OpenAI?

We do offer an Availability SLA for all resources and a Latency SLA for Provisioned-Managed Deployments. For more information about the SLA for Azure OpenAI Service, see the [Service Level Agreements \(SLA\) for Online Services page](#).

How do I enable fine-tuning? Create a custom model is greyed out in Azure AI Foundry portal.

In order to successfully access fine-tuning, you need Cognitive Services OpenAI Contributor assigned. Even someone with high-level Service Administrator permissions would still need this account explicitly set in order to access fine-tuning. For more information, please review the [role-based access control guidance](#).

What is the difference between a base model and a fine-tuned model?

A base model is a model that hasn't been customized or fine-tuned for a specific use case. Fine-tuned models are customized versions of base models where a model's weights are trained on a unique set of prompts. Fine-tuned models let you achieve better results on a wider number of tasks without needing to provide detailed examples for in-context learning as part of your completion prompt. To learn more, review our [fine-tuning guide](#).

What is the maximum number of fine-tuned models I can create?

100

Why was my fine-tuned model deployment deleted?

If a customized (fine-tuned) model is deployed for more than 15 days during which no completions or chat completions calls are made to it, the deployment is automatically deleted (and no further hosting charges are incurred for that deployment). The underlying customized model remains available and can be redeployed at any time. To learn more, check out the [how-to-article](#).

How do I deploy a model with the REST API?

There are currently two different REST APIs that allow model deployment. For the latest model deployment features such as the ability to specify a model version during deployment for models like text-embedding-ada-002 Version 2, use the [Deployments - Create Or Update](#) REST API call.

Can I use quota to increase the max token limit of a model?

No, quota Tokens-Per-Minute (TPM) allocation isn't related to the max input token limit of a model. Model input token limits are defined in the [models table](#) and aren't impacted by changes made to TPM.

GPT-4 Turbo with Vision

Can I fine-tune the image capabilities in GPT-4?

No, we don't support fine-tuning the image capabilities of GPT-4 at this time.

Can I use GPT-4 to generate images?

No, you can use `dall-e-3` to generate images and `gpt-4-vision-preview` to understand images.

What type of files can I upload?

We currently support PNG (.png), JPEG (.jpeg and .jpg), WEBP (.webp), and nonanimated GIF (.gif).

Is there a limit to the size of the image I can upload?

Yes, we restrict image uploads to 20 MB per image.

Can I delete an image I uploaded?

No, we'll delete the image for you automatically after it has been processed by the model.

How do the rate limits for GPT-4 Turbo with Vision work?

We process images at the token level, so each image we process counts towards your tokens per minute (TPM) limit. See the [Image tokens section](#) of the Overview for details on the formula used to determine token count per image.

Can GPT-4 Turbo with Vision understand image metadata?

No, the model doesn't receive image metadata.

What happens if my image is unclear?

If an image is ambiguous or unclear, the model will do its best to interpret it. However, the results might be less accurate. A good rule of thumb is that if an average human can't see the info in an image at the resolutions used in low/high res mode, then the model can't either.

What are the known limitations of GPT-4 Turbo with Vision?

See the [limitations](#) section of the GPT-4 Turbo with Vision concepts guide.

I keep getting truncated responses when I use GPT-4 Turbo vision models. Why is this happening?

By default GPT-4 `vision-preview` and GPT-4 `turbo-2024-04-09` have a `max_tokens` value of 16. Depending on your request this value is often too low and can lead to truncated responses. To resolve this issue, pass a larger `max_tokens` value as part of your chat completions API requests. GPT-4o defaults to 4096 `max_tokens`.

Assistants

Do you store any data used in the Assistants API?

Yes. Unlike Chat Completions API, Azure OpenAI Assistants is a stateful API, meaning it retains data. There are two types of data stored in the Assistants API:

- Stateful entities: Threads, messages, and runs created during Assistants use.
- Files: Uploaded during Assistants setup or as part of a message.

Where is this data stored?

Data is stored in a secure, Microsoft-managed storage account that is logically separated.

How long is this data stored?

All used data persists in this system unless you explicitly delete this data. Use the [delete function](#) with the thread ID of the thread you want to delete. Clearing the Run in the Assistants Playground does not delete threads, however deleting them using delete function will not list them in the thread page.

Can I bring my own data store to use with Assistants?

No. Currently Assistants supports only local files uploaded to the Assistants-managed storage. You cannot use your private storage account with Assistants.

Does Assistants support customer-managed key encryption (CMK)?

Today we support CMK for Threads and Files in Assistants. See the [What's new page](#) for available regions for this feature.

Is my data used by Microsoft for training models?

No. Data is not used for Microsoft not used for training models. See the [Responsible AI documentation](#) for more information.

Where is data stored geographically?

Azure OpenAI Assistants endpoints are regional, and data is stored in the same region as the endpoint. For more information, see the [Azure data residency documentation](#).

How am I charged for Assistants?

- Inference cost (input and output) of the base model you're using for each Assistant (for example gpt-4-0125). If you've created multiple Assistants, you will be charged for the base model attached to each Assistant.
- If you've enabled the Code Interpreter tool. For example if your assistant calls Code Interpreter simultaneously in two different threads, this would create two Code Interpreter sessions, each of which would be charged. Each session is active by default for one hour, which means that you would only pay this fee once if your user keeps giving instructions to Code Interpreter in the same thread for up to one hour.
- File search is billed based on the vector storage used.

For more information, see the [pricing page](#).

Is there any additional pricing or quota for using Assistants?

No. All [quotas](#) apply to using models with Assistants.

Does the Assistants API support non-Azure OpenAI models?

Assistants API only supports Azure OpenAI models.

Is the Assistants API generally available?

The Assistants API is currently in public preview. Stay informed of our latest product updates by regularly visiting our [What's New](#) page.

What are some examples or other resources I can use to learn about Assistants?

See the [Conceptual](#), [quickstart](#), [how-to](#) articles for information on getting started and using Assistants. You can also check out Azure OpenAI Assistants code samples on [GitHub](#) [↗](#).

Web app

How can I customize my published web app?

You can customize your published web app in the Azure portal. The source code for the published web app is [available on GitHub](#) [↗](#), where you can find information on changing the app frontend, as well as instructions for building and deploying the app.

Will my web app be overwritten when I deploy the app again from the Azure AI Foundry portal?

Your app code won't be overwritten when you update your app. The app will be updated to use the Azure OpenAI resource, Azure AI Search index (if you're using Azure OpenAI on your data), and model settings selected in the Azure AI Foundry portal without any change to the appearance or functionality.

Using your data

What is Azure OpenAI on your data?

Azure OpenAI on your data is a feature of the Azure OpenAI Services that helps organizations to generate customized insights, content, and searches using their designated data sources. It works with the capabilities of the OpenAI models in Azure OpenAI to provide more accurate and relevant responses to user queries in natural language. Azure OpenAI on your data can be integrated with customer's existing applications and workflows, offers insights into key performance indicators, and can interact with users seamlessly.

How can I access Azure OpenAI on your data?

All Azure OpenAI customers can use Azure OpenAI on your data via the Azure AI Foundry portal and Rest API.

What data sources does Azure OpenAI on your data support?

Azure OpenAI on your data supports ingestion from Azure AI Search, Azure Blob Storage, and uploading local files. You can learn more about Azure OpenAI on your data from the [conceptual article](#) and [quickstart](#).

How much does it cost to use Azure OpenAI on your data?

When using Azure OpenAI on your data, you incur costs when you use Azure AI Search, Azure Blob Storage, Azure Web App Service, semantic search and OpenAI models. There's no additional cost for using the "your data" feature in the Azure AI Foundry portal.

How can I customize or automate the index creation process?

You can prepare the index yourself using a [script provided on GitHub](#). Using this script will create an Azure AI Search index with all the information needed to better use your data, with your documents broken down into manageable chunks. See the README file with the data preparation code for details on how to run it.

How can I update my index?

You can [schedule an automatic index refresh](#), or upload additional data to your Azure Blob Container and use it as your data source when you create a new index. The new index will include all of the data in your container.

What file types does Azure OpenAI on your data support?

See [Using your data](#) for more information on supported file types.

Is responsible AI supported by Azure OpenAI on your data?

Yes, [Azure OpenAI on your data](#) is part of the Azure OpenAI Service and works with the [models](#) available in Azure OpenAI. The [content filtering](#) and abuse monitoring features of Azure OpenAI still apply. For more information, see the [overview of Responsible AI practices for Azure OpenAI models](#) and the [Transparency Note for Azure OpenAI](#) for extra guidance on using Azure OpenAI on your data responsibly.

Is there a token limit on the system message?

Yes, the token limit on the system message is 400. If the system message is more than 400 tokens, the rest of the tokens beyond the first 400 will be ignored. This limitation only applies to the Azure OpenAI [on your data feature](#).

Does Azure OpenAI on your data support function calling?

Azure OpenAI on your data currently doesn't support function calling.

Does the query language and the data source language need to be the same?

You must send queries in the same language of your data. Your data can be in any of the languages supported by [Azure AI Search](#).

If Semantic Search is enabled for my Azure AI Search resource, will it be automatically applied to Azure OpenAI on your data in the Azure AI Foundry portal?

When you select "Azure AI Search" as the data source, you can choose to apply semantic search. If you select "Azure Blob Container" or "Upload files" as the data source, you can create the index as usual. Afterwards you would reingest the data using the "Azure AI Search" option to select the same index and apply Semantic Search. You will then be ready to chat on your data with semantic search applied.

How can I add vector embeddings when indexing my data?

When you select "Azure Blob Container", "Azure AI Search", or "Upload files" as the data source, you can also select an Ada embedding model deployment to use when ingesting your data. This will create an Azure AI Search index with vector embeddings.

Why is index creation failing after I added an embedding model?

Index creation can fail when adding embeddings to your index if the rate limit on your Ada embedding model deployment is too low, or if you have a very large set of documents. You can use this [script provided on GitHub](#) to create the index with embeddings manually.

Customer Copyright Commitment

How do I obtain coverage under the Customer Copyright Commitment?

The Customer Copyright Commitment is a provision to be included in the December 1, 2023, Microsoft Product Terms that describes Microsoft's obligation to defend customers against certain third-party intellectual property claims relating to Output Content. If the subject of the claim is Output Content generated from the Azure OpenAI Service (or any other Covered Product that allows customers to configure the safety systems), then to receive coverage, customer must have implemented all mitigations required by the Azure OpenAI Service documentation in the offering that delivered the Output Content. The required mitigations are documented [here](#) and updated on an ongoing basis. For new services, features, models, or use cases, new CCC requirements will be posted and take effect at or following the launch of such service, feature, model, or use case. Otherwise, customers will have six months from the time of publication to implement new mitigations to maintain coverage under the CCC. If a customer tenders a claim, the customer will be required to demonstrate compliance with the relevant requirements. These mitigations are required for Covered Products that allow customers to configure the safety systems, including Azure OpenAI Service; they do not impact coverage for customers using other Covered Products.

Next steps

- [Azure OpenAI quotas and limits](#)
- [Azure OpenAI what's new](#)

- [Azure OpenAI quickstarts](#)

Feedback

Was this page helpful?



Yes



No

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