Azure OpenAI Service REST API reference

Article • 02/02/2024

This article provides details on the inference REST API endpoints for Azure OpenAI.

Authentication

Azure OpenAI provides two methods for authentication. you can use either API Keys or Microsoft Entra ID.

- **API Key authentication**: For this type of authentication, all API requests must include the API Key in the `api-key` HTTP header. The Quickstart provides guidance for how to make calls with this type of authentication.
- **Microsoft Entra ID authentication**: You can authenticate an API call using a Microsoft Entra token. Authentication tokens are included in a request as the `Authorization` header. The token provided must be preceded by `Bearer`, for example `Bearer YOUR_AUTH_TOKEN`. You can read our how-to guide on authenticating with Microsoft Entra ID.

REST API versioning

The service APIs are versioned using the `api-version` query parameter. All versions follow the YYYY-MM-DD date structure. For example:

```
HTTP

POST https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-15
```

Completions

With the Completions operation, the model generates one or more predicted completions based on a provided prompt. The service can also return the probabilities of alternative tokens at each position.

Create a completion

```
HTTP

POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/completions?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`deployment-id`	string	Required	The deployment name you chose when you deployed the model.
`api-version`	string	Required	The API version to use for this operation. This follows the YYYY-MM-DD format.

Supported versions

- `2022-12-01` Swagger spec
- `2023-03-15-preview` (retiring 2024-04-02) Swagger spec
- `2023-05-15` Swagger spec
- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-09-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-12-01-preview` Swagger spec

Request body

Expand table

Parameter	Type	Required?	Default	Description
`prompt`	string or array	Optional	`<|endoftext|>`	The prompt(s) to generate completions for, encoded as a string, or array of strings. Note that `<|endoftext|>` is the document separator that the model sees during training, so if a prompt isn't specified the model generates as if from the beginning of a new document.
`max_tokens`	integer	Optional	16	The maximum number of tokens to generate in the completion. The token count of your prompt plus max_tokens can't exceed the model's context length. Most models have a context length of 2048 tokens (except for the newest models, which support 4096).

`temperature`	number	Optional	1	What sampling temperature to use, between 0 and 2. Higher values means the model takes more risks. Try 0.9 for more creative applications, and 0 (`argmax sampling`) for ones with a well-defined answer. We generally recommend altering this or top_p but not both.
`top_p`	number	Optional	1	An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered. We generally recommend altering this or temperature but not both.
`logit_bias`	map	Optional	null	Modify the likelihood of specified tokens appearing in the completion. Accepts a json object that maps tokens (specified by their token ID in the GPT tokenizer) to an associated bias value from -100 to 100. You can use this tokenizer tool (which works for both GPT-2 and GPT-3) to convert text to token IDs. Mathematically, the bias is added to the logits generated by the model prior to sampling. The exact effect varies per model, but values between -1 and 1 should decrease or increase likelihood of selection; values like -100 or 100 should result in a ban or exclusive selection of the relevant token. As an example, you can pass {"50256": -100} to prevent the <|endoftext|> token from being generated.
`user`	string	Optional		A unique identifier representing your end-user, which can help monitoring and detecting abuse
`n`	integer	Optional	1	How many completions to generate for each prompt. Note: Because this parameter generates many completions, it can quickly consume your token quota. Use carefully and ensure that you have reasonable settings for max_tokens and stop.
`stream`	boolean	Optional	False	Whether to stream back partial progress. If set, tokens are sent as data-only server-sent events as they become available, with the stream terminated by a data: [DONE] message.
`logprobs`	integer	Optional	null	Include the log probabilities on the logprobs most likely tokens, as well the chosen tokens. For example, if logprobs is 10, the API will return a list of the 10 most likely tokens. the API will always return the logprob of the sampled token, so there might be up to logprobs+1 elements in the response. This parameter cannot be used with `gpt-35-turbo`.
`suffix`	string	Optional	null	The suffix that comes after a completion of inserted text.
`echo`	boolean	Optional	False	Echo back the prompt in addition to the completion. This parameter cannot be used with `gpt-35-turbo`.
`stop`	string or array	Optional	null	Up to four sequences where the API will stop generating further tokens. The returned text won't contain the stop sequence. For GPT-4 Turbo with Vision, up to two sequences are supported.
`presence_penalty`	number	Optional	0	Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.
`frequency_penalty`	number	Optional	0	Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.
`best_of`	integer	Optional	1	Generates best_of completions server-side and returns the "best" (the one with the lowest log probability per token). Results can't be streamed. When used with n, best_of controls the number of candidate completions and n specifies how many to return – best_of must be greater than n. Note: Because this parameter generates many completions, it can quickly consume your token quota. Use carefully and ensure that you have reasonable settings for max_tokens and stop. This parameter cannot be used with `gpt-35-turbo`.

Example request

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-15\
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d "{
  \"prompt\": \"Once upon a time\",
  \"max_tokens\": 5
}"
```

Example response

JSON

```
{
    "id": "cmpl-4kGh7iXtjW4lc9eGhff6Hp8C7btdQ",
    "object": "text_completion",
    "created": 1646932609,
    "model": "ada",
    "choices": [
        {
            "text": ", a dark line crossed",
            "index": 0,
            "logprobs": null,
            "finish_reason": "length"
        }
    ]
}
```

In the example response, `finish_reason` equals `stop`. If `finish_reason` equals `content_filter` consult our content filtering guide to understand why this is occurring.

Embeddings

Get a vector representation of a given input that can be easily consumed by machine learning models and other algorithms.

> **Note**
>
> OpenAI currently allows a larger number of array inputs with `text-embedding-ada-002`. Azure OpenAI currently supports input arrays up to 16 for `text-embedding-ada-002 (Version 2)`. Both require the max input token limit per API request to remain under 8191 for this model.

Create an embedding

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/embeddings?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`deployment-id`	string	Required	The name of your model deployment. You're required to first deploy a model before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This follows the YYYY-MM-DD format.

Supported versions

- `2022-12-01` Swagger spec
- `2023-03-15-preview` (retiring 2024-04-02) Swagger spec
- `2023-05-15` Swagger spec
- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-09-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-12-01-preview` Swagger spec

Request body

Expand table

Parameter	Type	Required?	Default	Description
`input`	string or array	Yes	N/A	Input text to get embeddings for, encoded as an array or string. The number of input tokens varies depending on what model you are using. Only `text-embedding-ada-002 (Version 2)` supports array input.
`user`	string	No	Null	A unique identifier representing your end-user. This will help Azure OpenAI monitor and detect abuse. **Do not pass PII identifiers instead use pseudoanonymized values such as GUIDs**

Example request

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/embeddings?api-version=2023-05-15 \
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d "{\"input\": \"The food was delicious and the waiter...\"}"
```

Example response

JSON

```
{
  "object": "list",
  "data": [
    {
      "object": "embedding",
      "embedding": [
        0.018990106880664825,
```

```
        -0.0073809814639389515,
        .... (1024 floats total for ada)
        0.021276434883475304,
      ],
      "index": 0
    }
  ],
  "model": "text-similarity-babbage:001"
}
```

Chat completions

Create completions for chat messages with the GPT-35-Turbo and GPT-4 models.

Create chat completions

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/chat/completions?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`deployment-id`	string	Required	The name of your model deployment. You're required to first deploy a model before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This follows the YYYY-MM-DD or YYYY-MM-DD-preview format.

Supported versions

- `2023-03-15-preview` (retiring 2024-04-02) Swagger spec
- `2023-05-15` Swagger spec
- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-09-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-12-01-preview` (required for Vision scenarios) Swagger spec

Request body

The request body consists of a series of messages. The model will generate a response to the last message, using earlier messages as context.

Expand table

Parameter	Type	Required?	Default	Description
`messages`	array	Yes	N/A	The series of messages associated with this chat completion request. It should include previous messages in the conversation. Each message has a `role` and `content`.
`role`	string	Yes	N/A	Indicates who is giving the current message. Can be `system`, `user`, `assistant`, `tool`, or `function`.

`content`	string
	or
	array
		Yes	N/A	The content of the message. It must be a string, unless in a Vision-enabled scenario. If it's part of the `user` message, using the GPT-4 Turbo with Vision model, with the latest API version, then `content` must be an array of structures, where each item represents either text or an image:
				• `text`: input text is represented as a structure with the following properties:
				  ◦ `type` = "text"
				  ◦ `text` = the input text
				• `images`: an input image is represented as a structure with the following properties:
				  ◦ `type` = "image_url"
				  ◦ `image_url` = a structure with the following properties:
				  ▪ `url` = the image URL
				  ▪ (optional) `detail` = "high", "low", or "auto"

`contentPart`	object	No	N/A	Part of a user's multi-modal message. It can be either text type or image type. If text, it will be a text string. If image, it will be a `contentPartImage` object.
`contentPartImage`	object	No	N/A	Represents a user-uploaded image. It has a `url` property, which is either a URL of the image or the base 64 encoded image data. It also has a `detail` property which can be `auto`, `low`, or `high`.
`enhancements`	object	No	N/A	Represents the Vision enhancement features requested for the chat. It has a `grounding` and `ocr` property, which each have a boolean `enabled` property. Use these to request the OCR service and/or the object detection/grounding service.
`dataSources`	object	No	N/A	Represents additional resource data. Computer Vision resource data is needed for Vision enhancement. It has a

`type` property which should be `"AzureComputerVision"` and a `parameters` property which has an `endpoint` and `key` property. These strings should be set to the endpoint URL and access key of your Computer Vision resource.

Example request

Text-only chat

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/chat/completions?api-version=2023-05-15 \
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d '{"messages":[{"role": "system", "content": "You are a helpful assistant."},{"role": "user", "content": "Does Azure OpenAI support customer managed keys?"},{"role": "assistant", "content": "Yes, customer managed keys are supported by Azure OpenAI."},{"role": "user", "content": "Do other Azure AI services support this too?"}]}'
```

Chat with vision

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/chat/completions?api-version=2023-12-01-preview \
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d '{"messages":[{"role":"system","content":"You are a helpful assistant."},{"role":"user","content":[{"type":"text","text":"Describe this picture:"},{ "type": "image_url", "image_url": { "url": "https://learn.microsoft.com/azure/ai-services/computer-vision/media/quickstarts/presentation.png", "detail": "high" } }]}]}'
```

Enhanced chat with vision

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/extensions/chat/completions?api-version=2023-12-01-preview \
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d '{"enhancements":{"ocr":{"enabled":true},"grounding":{"enabled":true}},"dataSources":[{"type":"AzureComputerVision","parameters":{"endpoint":" <Computer Vision Resource Endpoint> ","key":"<Computer Vision Resource Key>"}}],"messages":[{"role":"system","content":"You are a helpful assistant."},{"role":"user","content":[{"type":"text","text":"Describe this picture:"},{"type":"image_url","image_url":"https://learn.microsoft.com/azure/ai-services/computer-vision/media/quickstarts/presentation.png"}]}]}'
```

Example response

Console

```
{
    "id": "chatcmpl-6v7mkQj980V1yBec6ETrKPRqFjNw9",
    "object": "chat.completion",
    "created": 1679072642,
    "model": "gpt-35-turbo",
    "usage":
    {
        "prompt_tokens": 58,
        "completion_tokens": 68,
        "total_tokens": 126
    },
    "choices":
    [
        {
            "message":
            {
                "role": "assistant",
                "content": "Yes, other Azure AI services also support customer managed keys.
                    Azure AI services offer multiple options for customers to manage keys, such as
                    using Azure Key Vault, customer-managed keys in Azure Key Vault or
                    customer-managed keys through Azure Storage service. This helps customers ensure
                    that their data is secure and access to their services is controlled."
            },
            "finish_reason": "stop",
            "index": 0
        }
    ]
}
```

Output formatting adjusted for ease of reading, actual output is a single block of text without line breaks.

In the example response, `finish_reason` equals `stop`. If `finish_reason` equals `content_filter` consult our content filtering guide to understand why this is occurring.

> **Important**

The `functions` and `function_call` parameters have been deprecated with the release of the 2023-12-01-preview version of the API. The replacement for `functions` is the `tools` parameter. The replacement for `function_call` is the `tool_choice` parameter. Parallel function calling which was introduced as part of the 2023-12-01-preview is only supported with `gpt-35-turbo` (1106) and `gpt-4` (1106-preview) also known as GPT-4 Turbo Preview.

Expand table

Parameter	Type	Required?	Default	Description
`messages`	array	Required		The collection of context messages associated with this chat completions request. Typical usage begins with a chat message for the System role that provides instructions for the behavior of the assistant, followed by alternating messages between the User and Assistant roles.
`temperature`	number	Optional	1	What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic. We generally recommend altering this or `top_p` but not both.
`n`	integer	Optional	1	How many chat completion choices to generate for each input message.
`stream`	boolean	Optional	false	If set, partial message deltas will be sent, like in ChatGPT. Tokens will be sent as data-only server-sent events as they become available, with the stream terminated by a `data: [DONE]` message."
`stop`	string or array	Optional	null	Up to 4 sequences where the API will stop generating further tokens.
`max_tokens`	integer	Optional	inf	The maximum number of tokens allowed for the generated answer. By default, the number of tokens the model can return will be (4096 - prompt tokens).
`presence_penalty`	number	Optional	0	Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.
`frequency_penalty`	number	Optional	0	Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.
`logit_bias`	object	Optional	null	Modify the likelihood of specified tokens appearing in the completion. Accepts a json object that maps tokens (specified by their token ID in the tokenizer) to an associated bias value from -100 to 100. Mathematically, the bias is added to the logits generated by the model prior to sampling. The exact effect varies per model, but values between -1 and 1 should decrease or increase likelihood of selection; values like -100 or 100 should result in a ban or exclusive selection of the relevant token.
`user`	string	Optional		A unique identifier representing your end-user, which can help Azure OpenAI to monitor and detect abuse.
`function_call`		Optional		`[Deprecated in 2023-12-01-preview replacement paremeter is tools_choice]` Controls how the model responds to function calls. "none" means the model doesn't call a function, and responds to the end-user. "auto" means the model can pick between an end-user or calling a function. Specifying a particular function via {"name": "my_function"} forces the model to call that function. "none" is the default when no functions are present. "auto" is the default if functions are present. This parameter requires API version 2023-07-01-preview
`functions`	FunctionDefinition[]	Optional		`[Deprecated in 2023-12-01-preview replacement paremeter is tools]` A list of functions the model can generate JSON inputs for. This parameter requires API version 2023-07-01-preview
`tools`	string (The type of the tool. Only function is supported.)	Optional		A list of tools the model can call. Currently, only functions are supported as a tool. Use this to provide a list of functions the model can generate JSON inputs for. This parameter requires API version 2023-12-01-preview
`tool_choice`	string or object	Optional	none is the default when no functions are present. auto is the default if functions are present.	Controls which (if any) function is called by the model. none means the model won't call a function and instead generates a message. auto means the model can pick between generating a message or calling a function. Specifying a particular function via {"type: "function", "function": {"name": "my_function"}} forces the model to call that function. This parameter requires API version 2023-12-01-preview

ChatMessage

A single, role-attributed message within a chat completion interaction.

Expand table

Name	Type	Description
content	string	The text associated with this message payload.

function_call	FunctionCall	The name and arguments of a function that should be called, as generated by the model.
name	string	The `name` of the author of this message. `name` is required if role is `function`, and it should be the name of the function whose response is in the `content`. Can contain a-z, A-Z, 0-9, and underscores, with a maximum length of 64 characters.
role	ChatRole	The role associated with this message payload

ChatRole

A description of the intended purpose of a message within a chat completions interaction.

Expand table

Name	Type	Description
assistant	string	The role that provides responses to system-instructed, user-prompted input.
function	string	The role that provides function results for chat completions.
system	string	The role that instructs or sets the behavior of the assistant.
user	string	The role that provides input for chat completions.

Function

This is used with the `tools` parameter that was added in API version 2023-12-01-preview.

Expand table

Name	Type	Description
description	string	A description of what the function does, used by the model to choose when and how to call the function
name	string	The name of the function to be called. Must be a-z, A-Z, 0-9, or contain underscores and dashes, with a maximum length of 64
parameters	object	The parameters the functions accepts, described as a JSON Schema object. See the JSON Schema reference for documentation about the format."

FunctionCall-Deprecated

The name and arguments of a function that should be called, as generated by the model. This requires API version 2023-07-01-preview

Expand table

Name	Type	Description
arguments	string	The arguments to call the function with, as generated by the model in JSON format. The model doesn't always generate valid JSON, and might fabricate parameters not defined by your function schema. Validate the arguments in your code before calling your function.
name	string	The name of the function to call.

FunctionDefinition-Deprecated

The definition of a caller-specified function that chat completions can invoke in response to matching user input. This requires API version 2023-07-01-preview

Expand table

Name	Type	Description
description	string	A description of what the function does. The model uses this description when selecting the function and interpreting its parameters.
name	string	The name of the function to be called.
parameters		The parameters the functions accepts, described as a JSON Schema object.

Completions extensions

Extensions for chat completions, for example Azure OpenAI on your data.

Use chat completions extensions

HTTP

```
POST {your-resource-name}/openai/deployments/{deployment-id}/extensions/chat/completions?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`deployment-id`	string	Required	The name of your model deployment. You're required to first deploy a model before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This follows the YYYY-MM-DD format.

Supported versions

- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-09-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-12-01-preview` Swagger spec

Example request

You can make requests using Azure AI Search, Azure Cosmos DB for MongoDB vCore, Azure Machine Learning, Pinecone, and Elasticsearch.

Azure AI Search

Console

```
curl -i -X POST YOUR_RESOURCE_NAME/openai/deployments/YOUR_DEPLOYMENT_NAME/extensions/chat/completions?api-version=2023-06-01-preview \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d \
'
{
    "temperature": 0,
    "max_tokens": 1000,
    "top_p": 1.0,
    "dataSources": [
        {
            "type": "AzureCognitiveSearch",
            "parameters": {
                "endpoint": "YOUR_AZURE_COGNITIVE_SEARCH_ENDPOINT",
                "key": "YOUR_AZURE_COGNITIVE_SEARCH_KEY",
                "indexName": "YOUR_AZURE_COGNITIVE_SEARCH_INDEX_NAME"
            }
        }
    ],
    "messages": [
        {
            "role": "user",
            "content": "What are the differences between Azure Machine Learning and Azure AI services?"
        }
    ]
}
'
```

Azure Cosmos DB for MongoDB vCore

JSON

```
curl -i -X POST YOUR_RESOURCE_NAME/openai/deployments/YOUR_DEPLOYMENT_NAME/extensions/chat/completions?api-version=2023-06-01-preview \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d \
'
{
    "temperature": 0,
    "top_p": 1.0,
    "max_tokens": 800,
    "stream": false,
    "messages": [
        {
            "role": "user",
            "content": "What is the company insurance plan?"
        }
    ],
    "dataSources": [
        {
            "type": "AzureCosmosDB",
            "parameters": {
                "authentication": {
                    "type": "ConnectionString",
                    "connectionString": "mongodb+srv://onyourdatatest:{password}$@{cluster-name}.mongocluster.cosmos.azure.com/?tls=true&authMechanism=SCRAM-SHA-256&retrywrites=false&maxIdleTimeMS=120000"
                },
                "databaseName": "vectordb",
```

```
                "containerName": "azuredocs",
                "indexName": "azuredocindex",
                "embeddingDependency": {
                    "type": "DeploymentName",
                    "deploymentName": "{embedding deployment name}"
                },
                "fieldsMapping": {
                    "vectorFields": [
                        "contentvector"
                    ]
                }
            }
        }
    ]
}
'
```

Elasticsearch

Console

```
curl -i -X POST YOUR_RESOURCE_NAME/openai/deployments/YOUR_DEPLOYMENT_NAME/extensions/chat/completions?api-version=2023-
12-01-preview \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d \
{
  "messages": [
    {
      "role": "system",
      "content": "you are a helpful assistant that talks like a pirate"
    },
    {
      "role": "user",
      "content": "can you tell me how to care for a parrot?"
    }
  ],
  "dataSources": [
    {
      "type": "Elasticsearch",
      "parameters": {
        "endpoint": "{search endpoint}",
        "indexName": "{index name}",
        "authentication": {
          "type": "KeyAndKeyId",
          "key": "{key}",
          "keyId": "{key id}"
        }
      }
    }
  ]
}
```

Azure Machine Learning

Console

```
curl -i -X POST YOUR_RESOURCE_NAME/openai/deployments/YOUR_DEPLOYMENT_NAME/extensions/chat/completions?api-version=2023-
12-01-preview \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d \
'
{
  "messages": [
    {
      "role": "system",
      "content": "you are a helpful assistant that talks like a pirate"
    },
    {
      "role": "user",
      "content": "can you tell me how to care for a parrot?"
    }
  ],
  "dataSources": [
    {
      "type": "AzureMLIndex",
      "parameters": {
        "projectResourceId": "/subscriptions/{subscription-id}/resourceGroups/{resource-group-name}/providers/Microsoft-
.MachineLearningServices/workspaces/{workspace-id}",
        "name": "my-project",
        "version": "5"
      }
    }
  ]
}
'
```

Pinecone

Console

```
curl -i -X POST YOUR_RESOURCE_NAME/openai/deployments/YOUR_DEPLOYMENT_NAME/extensions/chat/completions?api-version=2023-12-01-preview \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d \
'
{
  "messages": [
    {
      "role": "system",
      "content": "you are a helpful assistant that talks like a pirate"
    },
    {
      "role": "user",
      "content": "can you tell me how to care for a parrot?"
    }
  ],
  "dataSources": [
    {
      "type": "Pinecone",
      "parameters": {
        "authentication": {
          "type": "APIKey",
          "apiKey": "{api key}"
        },
        "environment": "{environment name}",
        "indexName": "{index name}",
        "embeddingDependency": {
          "type": "DeploymentName",
          "deploymentName": "{embedding deployment name}"
        },
        "fieldsMapping": {
          "titleField": "title",
          "urlField": "url",
          "filepathField": "filepath",
          "contentFields": [
            "content"
          ],
          "contentFieldsSeparator": "\n"
        }
      }
    }
  ]
}
'
```

Example response

JSON


```
{
    "id": "12345678-1a2b-3c4e5f-a123-12345678abcd",
    "model": "",
    "created": 1684304924,
    "object": "chat.completion",
    "choices": [
        {
            "index": 0,
            "messages": [
                {
                    "role": "tool",
                    "content": "{\"citations\": [{\"content\": \"\\nAzure AI services are cloud-based artificial intel-
ligence (AI) services...\", \"id\": null, \"title\": \"What is Azure AI services\", \"filepath\": null, \"url\": null,
\"metadata\": {\"chunking\": \"orignal document size=250. Scores=0.4314117431640625 and 1.72564697265625.Org Highlight
count=4.\"}, \"chunk_id\": \"0\"}], \"intent\": \"[\\\"Learn about Azure AI services.\\\"]\"}",
                    "end_turn": false
                },
                {
                    "role": "assistant",
                    "content": " \nAzure AI services are cloud-based artificial intelligence (AI) services that help de-
velopers build cognitive intelligence into applications without having direct AI or data science skills or knowledge.
[doc1]. Azure Machine Learning is a cloud service for accelerating and managing the machine learning project lifecycle.
[doc1].",
                    "end_turn": true
                }
            ]
        }
    ]
}
```


Expand table

Parameters	Type	Required?	Default	Description
`messages`	array	Required	null	The messages to generate chat completions for, in the chat format.
`dataSources`	array	Required		The data sources to be used for the Azure OpenAI on your data feature.
`temperature`	number	Optional	0	What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random,

				while lower values like 0.2 will make it more focused and deterministic. We generally recommend altering this or `top_p` but not both.
`top_p`	number	Optional	1	An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with `top_p` probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered. We generally recommend altering this or temperature but not both.
`stream`	boolean	Optional	false	If set, partial message deltas are sent, like in ChatGPT. Tokens are sent as data-only server-sent events as they become available, with the stream terminated by a message `"messages": [{"delta": {"content": "[DONE]"}, "index": 2, "end_turn": true}]`
`stop`	string or array	Optional	null	Up to two sequences where the API will stop generating further tokens.
`max_tokens`	integer	Optional	1000	The maximum number of tokens allowed for the generated answer. By default, the number of tokens the model can return is `4096 - prompt_tokens`.

The following parameters can be used inside of the `parameters` field inside of `dataSources`.

Expand table

Parameters	Type	Required?	Default	Description
`type`	string	Required	null	The data source to be used for the Azure OpenAI on your data feature. For Azure AI Search the value is `AzureCognitiveSearch`. For Azure Cosmos DB for MongoDB vCore, the value is `AzureCosmosDB`. For Elasticsearch the value is `Elasticsearch`. For Azure Machine Learning, the value is `AzureMLIndex`. For Pinecone, the value is `Pinecone`.
`indexName`	string	Required	null	The search index to be used.
`inScope`	boolean	Optional	true	If set, this value limits responses specific to the grounding data content.
`topNDocuments`	number	Optional	5	Specifies the number of top-scoring documents from your data index used to generate responses. You might want to increase the value when you have short documents or want to provide more context. This is the *retrieved documents* parameter in Azure OpenAI studio.
`semanticConfiguration`	string	Optional	null	The semantic search configuration. Only required when `queryType` is set to `semantic` or `vectorSemanticHybrid`.
`roleInformation`	string	Optional	null	Gives the model instructions about how it should behave and the context it should reference when generating a response. Corresponds to the "System Message" in Azure OpenAI Studio. See Using your data for more information. There's a 100 token limit, which counts towards the overall token limit.
`filter`	string	Optional	null	The filter pattern used for restricting access to sensitive documents
`embeddingEndpoint`	string	Optional	null	The endpoint URL for an Ada embedding model deployment, generally of the format `https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/embeddings?api-version=2023-05-15`. Use with the `embeddingKey` parameter for vector search outside of private networks and private endpoints.
`embeddingKey`	string	Optional	null	The API key for an Ada embedding model deployment. Use with `embeddingEndpoint` for vector search outside of private networks and private endpoints.
`embeddingDeploymentName`	string	Optional	null	The Ada embedding model deployment name within the same Azure OpenAI resource. Used instead of `embeddingEndpoint` and `embeddingKey` for vector search. Should only be used when both the `embeddingEndpoint` and `embeddingKey` parameters are defined. When this parameter is provided, Azure OpenAI on your data use an internal call to evaluate the Ada embedding model, rather than calling the Azure OpenAI endpoint. This enables you to use vector search in private networks and private endpoints. Billing remains the same whether this parameter is defined or not. Available in regions where embedding models are available starting in API versions `2023-06-01-preview` and later.
`strictness`	number	Optional	3	Sets the threshold to categorize documents as relevant to your queries. Raising the value means a higher threshold for relevance and filters out more less-relevant documents for responses. Setting this value too high might cause the model to fail to generate responses due to limited available documents.

Azure AI Search parameters

The following parameters are used for Azure AI Search.

Expand table

Parameters	Type	Required?	Default	Description
`endpoint`	string	Required	null	Azure AI Search only. The data source endpoint.
`key`	string	Required	null	Azure AI Search only. One of the Azure AI Search admin keys for your service.
`queryType`	string	Optional	simple	Indicates which query option is used for Azure AI Search. Available types: `simple`, `semantic`, `vector`, `vectorSimpleHybrid`, `vectorSemanticHybrid`.
`fieldsMapping`	dictionary	Optional for Azure AI Search.	null	defines which fields you want to map when you add your data source.

The following parameters are used inside of the `authentication` field, which enables you to use Azure OpenAI without public network access.

Expand table

Parameters	Type	Required?	Default	Description
`type`	string	Required	null	The authentication type.
`managedIdentityResourceId`	string	Required	null	The resource ID of the user-assigned managed identity to use for authentication.

JSON

```
"authentication": {
  "type": "UserAssignedManagedIdentity",
  "managedIdentityResourceId": "/subscriptions/{subscription-id}/resourceGroups/{resource-group}/providers/Microsoft.-
ManagedIdentity/userAssignedIdentities/{resource-name}"
},
```

The following parameters are used inside of the `fieldsMapping` field.

Expand table

Parameters	Type	Required?	Default	Description
`titleField`	string	Optional	null	The field in your index that contains the original title of each document.
`urlField`	string	Optional	null	The field in your index that contains the original URL of each document.
`filepathField`	string	Optional	null	The field in your index that contains the original file name of each document.
`contentFields`	dictionary	Optional	null	The fields in your index that contain the main text content of each document.
`contentFieldsSeparator`	string	Optional	null	The separator for the content fields. Use `\n` by default.

JSON

```
"fieldsMapping": {
  "titleField": "myTitleField",
  "urlField": "myUrlField",
  "filepathField": "myFilePathField",
  "contentFields": [
    "myContentField"
  ],
  "contentFieldsSeparator": "\n"
}
```

The following parameters are used inside of the optional `embeddingDependency` parameter, which contains details of a vectorization source that is based on an internal embeddings model deployment name in the same Azure OpenAI resource.

Expand table

Parameters	Type	Required?	Default	Description
`deploymentName`	string	Optional	null	The type of vectorization source to use.
`type`	string	Optional	null	The embedding model deployment name, located within the same Azure OpenAI resource. This enables you to use vector search without an Azure OpenAI API key and without Azure OpenAI public network access.

JSON

```
"embeddingDependency": {
  "type": "DeploymentName",
  "deploymentName": "{embedding deployment name}"
},
```

Azure Cosmos DB for MongoDB vCore parameters

The following parameters are used for Azure Cosmos DB for MongoDB vCore.

Expand table

Parameters	Type	Required?	Default	Description
`type` (found inside of `authentication`)	string	Required	null	Azure Cosmos DB for MongoDB vCore only. The authentication to be used For. Azure Cosmos Mongo vCore, the value is `ConnectionString`
`connectionString`	string	Required	null	Azure Cosmos DB for MongoDB vCore only. The connection string to be used for authenticate Azure Cosmos Mongo vCore Account.
`databaseName`	string	Required	null	Azure Cosmos DB for MongoDB vCore only. The Azure Cosmos Mongo vCore database name.
`containerName`	string	Required	null	Azure Cosmos DB for MongoDB vCore only. The Azure Cosmos Mongo vCore

				container name in the database.
`type` (found inside of `embeddingDependencyType`)	string	Required	null	Indicates the embedding model dependency.
`deploymentName` (found inside of `embeddingDependencyType`)	string	Required	null	The embedding model deployment name.
`fieldsMapping`	dictionary	Required for Azure Cosmos DB for MongoDB vCore.	null	Index data column mapping. When you use Azure Cosmos DB for MongoDB vCore, the value `vectorFields` is required, which indicates the fields that store vectors.

The following parameters are used inside of the optional `embeddingDependency` parameter, which contains details of a vectorization source that is based on an internal embeddings model deployment name in the same Azure OpenAI resource.

Expand table

Parameters	Type	Required?	Default	Description
`deploymentName`	string	Optional	null	The type of vectorization source to use.
`type`	string	Optional	null	The embedding model deployment name, located within the same Azure OpenAI resource. This enables you to use vector search without an Azure OpenAI API key and without Azure OpenAI public network access.

JSON

```
"embeddingDependency": {
  "type": "DeploymentName",
  "deploymentName": "{embedding deployment name}"
},
```

Elasticsearch parameters

The following parameters are used for Elasticsearch.

Expand table

Parameters	Type	Required?	Default	Description
`endpoint`	string	Required	null	The endpoint for connecting to Elasticsearch.
`indexName`	string	Required	null	The name of the Elasticsearch index.
`type` (found inside of `authentication`)	string	Required	null	The authentication to be used. For Elasticsearch, the value is `KeyAndKeyId`.
`key` (found inside of `authentication`)	string	Required	null	The key used to connect to Elasticsearch.
`keyId` (found inside of `authentication`)	string	Required	null	The key ID to be used. For Elasticsearch.

The following parameters are used inside of the `fieldsMapping` field.

Expand table

Parameters	Type	Required?	Default	Description
`titleField`	string	Optional	null	The field in your index that contains the original title of each document.
`urlField`	string	Optional	null	The field in your index that contains the original URL of each document.
`filepathField`	string	Optional	null	The field in your index that contains the original file name of each document.
`contentFields`	dictionary	Optional	null	The fields in your index that contain the main text content of each document.
`contentFieldsSeparator`	string	Optional	null	The separator for the content fields. Use `\n` by default.
`vectorFields`	dictionary	Optional	null	The names of fields that represent vector data

JSON

```
"fieldsMapping": {
  "titleField": "myTitleField",
  "urlField": "myUrlField",
  "filepathField": "myFilePathField",
  "contentFields": [
    "myContentField"
  ],
  "contentFieldsSeparator": "\n",
  "vectorFields": [
    "myVectorField"
  ]
}
```

The following parameters are used inside of the optional `embeddingDependency` parameter, which contains details of a vectorization source that is based on an internal embeddings model deployment name in the same Azure OpenAI resource.

Expand table

Parameters	Type	Required?	Default	Description
`deploymentName`	string	Optional	null	The type of vectorization source to use.
`type`	string	Optional	null	The embedding model deployment name, located within the same Azure OpenAI resource. This enables you to use vector search without an Azure OpenAI API key and without Azure OpenAI public network access.

JSON

```
"embeddingDependency": {
  "type": "DeploymentName",
  "deploymentName": "{embedding deployment name}"
},
```

Azure Machine Learning parameters

The following parameters are used for Azure Machine Learning.

Expand table

Parameters	Type	Required?	Default	Description
`projectResourceId`	string	Required	null	The project resource ID.
`name`	string	Required	null	The name of the Azure Machine Learning project name.
`version` (found inside of `authentication`)	string	Required	null	The version of the Azure Machine Learning vector index.

The following parameters are used inside of the optional `embeddingDependency` parameter, which contains details of a vectorization source that is based on an internal embeddings model deployment name in the same Azure OpenAI resource.

Expand table

Parameters	Type	Required?	Default	Description
`deploymentName`	string	Optional	null	The type of vectorization source to use.
`type`	string	Optional	null	The embedding model deployment name, located within the same Azure OpenAI resource. This enables you to use vector search without an Azure OpenAI API key and without Azure OpenAI public network access.

JSON

```
"embeddingDependency": {
  "type": "DeploymentName",
  "deploymentName": "{embedding deployment name}"
},
```

Pinecone parameters

The following parameters are used for Pinecone.

Expand table

Parameters	Type	Required?	Default	Description
`type` (found inside of `authentication`)	string	Required	null	The authentication to be used. For Pinecone, the value is `APIKey`.
`apiKey` (found inside of `authentication`)	string	Required	null	The API key for Pinecone.
`environment`	string	Required	null	The name of the Pinecone environment.
`indexName`	string	Required	null	The name of the Pinecone index.
`embeddingDependency`	string	Required	null	The embedding dependency for vector search.
`type` (found inside of `embeddingDependency`)	string	Required	null	The type of dependency. For Pinecone the value is `DeploymentName`.
`deploymentName` (found inside of `embeddingDependency`)	string	Required	null	The name of the deployment.
`titleField` (found inside of `fieldsMapping`)	string	Required	null	The name of the index field to use as a title.
`urlField` (found inside of `fieldsMapping`)	string	Required	null	The name of the index field to use as a URL.
`filepathField` (found inside of `fieldsMapping`)	string	Required	null	The name of the index field to use as a file path.
`contentFields` (found inside of `fieldsMapping`)	string	Required	null	The name of the index fields that should be treated as content.

`vectorFields`	dictionary	Optional	null	The names of fields that represent vector data
`contentFieldsSeparator` (found inside of `fieldsMapping`)	string	Required	null	The separator for your content fields. Use `\n` by default.

The following parameters are used inside of the optional `embeddingDependency` parameter, which contains details of a vectorization source that is based on an internal embeddings model deployment name in the same Azure OpenAI resource.

Expand table

Parameters	Type	Required?	Default	Description
`deploymentName`	string	Optional	null	The type of vectorization source to use.
`type`	string	Optional	null	The embedding model deployment name, located within the same Azure OpenAI resource. This enables you to use vector search without an Azure OpenAI API key and without Azure OpenAI public network access.

JSON

```
"embeddingDependency": {
  "type": "DeploymentName",
  "deploymentName": "{embedding deployment name}"
},
```

Start an ingestion job

Tip

The `JOB_NAME` you choose will be used as the index name. Be aware of the **constraints** for the *index name*.

Console

```
curl -i -X PUT https://YOUR_RESOURCE_NAME.openai.azure.com/openai/extensions/on-your-data/ingestion-jobs/JOB_NAME?api-version=2023-10-01-preview \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-H "searchServiceEndpoint: https://YOUR_AZURE_COGNITIVE_SEARCH_NAME.search.windows.net" \
-H "searchServiceAdminKey: YOUR_SEARCH_SERVICE_ADMIN_KEY" \
-H  "storageConnectionString: YOUR_STORAGE_CONNECTION_STRING" \
-H "storageContainer: YOUR_INPUT_CONTAINER" \
-d '{ "dataRefreshIntervalInMinutes": 10 }'
```

Example response

JSON

```
{
    "id": "test-1",
    "dataRefreshIntervalInMinutes": 10,
    "completionAction": "cleanUpAssets",
    "status": "running",
    "warnings": [],
    "progress": {
        "stageProgress": [
            {
                "name": "Preprocessing",
                "totalItems": 100,
                "processedItems": 100
            },
            {
                "name": "Indexing",
                "totalItems": 350,
                "processedItems": 40
            }
        ]
    }
}
```

Header Parameters

Expand table

Parameters	Type	Required?	Default	Description
`searchServiceEndpoint`	string	Required	null	The endpoint of the search resource in which the data will be ingested.
`searchServiceAdminKey`	string	Optional	null	If provided, the key is used to authenticate with the `searchServiceEndpoint`. If not provided, the system-assigned identity of the Azure OpenAI resource will be used. In this case, the system-assigned identity must have "Search Service Contributor" role assignment on the search resource.
`storageConnectionString`	string	Required	null	The connection string for the storage account where the input data is located. An account key has to be

				provided in the connection string. It should look something like `DefaultEndpointsProtocol=https;AccountName=<your storage account>;AccountKey=<your account key>`
`storageContainer`	string	Required	null	The name of the container where the input data is located.
`embeddingEndpoint`	string	Optional	null	Not required if you use semantic or only keyword search. It's required if you use vector, hybrid, or hybrid + semantic search
`embeddingKey`	string	Optional	null	The key of the embedding endpoint. This is required if the embedding endpoint isn't empty.
`url`	string	Optional	null	If URL isn't null, the provided url is crawled into the provided storage container and then ingested accordingly.

Body Parameters

Expand table

Parameters	Type	Required?	Default	Description
`dataRefreshIntervalInMinutes`	string	Required	0	The data refresh interval in minutes. If you want to run a single ingestion job without a schedule, set this parameter to `0`.
`completionAction`	string	Optional	`cleanUpAssets`	What should happen to the assets created during the ingestion process upon job completion. Valid values are `cleanUpAssets` or `keepAllAssets`. `keepAllAssets` leaves all the intermediate assets for users interested in reviewing the intermediate results, which can be helpful for debugging assets. `cleanUpAssets` removes the assets after job completion.
`chunkSize`	int	Optional	1024	This number defines the maximum number of tokens in each chunk produced by the ingestion flow.

List ingestion jobs

Console

```
curl -i -X GET https://YOUR_RESOURCE_NAME.openai.azure.com/openai/extensions/on-your-data/ingestion-jobs?api-ver-
sion=2023-10-01-preview \
-H "api-key: YOUR_API_KEY"
```

Example response

JSON

```
{
    "value": [
        {
            "id": "test-1",
            "dataRefreshIntervalInMinutes": 10,
            "completionAction": "cleanUpAssets",
            "status": "succeeded",
            "warnings": []
        },
        {
            "id": "test-2",
            "dataRefreshIntervalInMinutes": 10,
            "completionAction": "cleanUpAssets",
            "status": "failed",
            "error": {
                "code": "BadRequest",
                "message": "Could not execute skill because the Web Api request failed."
            },
            "warnings": []
        }
    ]
}
```

Get the status of an ingestion job

Console

```
curl -i -X GET https://YOUR_RESOURCE_NAME.openai.azure.com/openai/extensions/on-your-data/ingestion-jobs/YOUR_JOB_NAME?
api-version=2023-10-01-preview \
-H "api-key: YOUR_API_KEY"
```

Example response body

JSON

```
{
    "id": "test-1",
    "dataRefreshIntervalInMinutes": 10,
```

```
    "completionAction": "cleanUpAssets",
    "status": "succeeded",
    "warnings": []
}
```

Image generation

Request a generated image (DALL-E 3)

Generate and retrieve a batch of images from a text caption.

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/images/generations?api-version=
{api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`deployment-id`	string	Required	The name of your DALL-E 3 model deployment such as *MyDalle3*. You're required to first deploy a DALL-E 3 model before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This follows the YYYY-MM-DD format.

Supported versions

- `2023-12-01-preview` Swagger spec

Request body

Expand table

Parameter	Type	Required?	Default	Description
`prompt`	string	Required		A text description of the desired image(s). The maximum length is 4000 characters.
`n`	integer	Optional	1	The number of images to generate. Only `n=1` is supported for DALL-E 3.
`size`	string	Optional	`1024x1024`	The size of the generated images. Must be one of `1792x1024`, `1024x1024`, or `1024x1792`.
`quality`	string	Optional	`standard`	The quality of the generated images. Must be `hd` or `standard`.
`response_format`	string	Optional	`url`	The format in which the generated images are returned Must be `url` (a URL pointing to the image) or `b64_json` (the base 64 byte code in JSON format).
`style`	string	Optional	`vivid`	The style of the generated images. Must be `natural` or `vivid` (for hyper-realistic / dramatic images).

Example request

Console

```
curl -X POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/images/generations?api-
version=2023-12-01-preview \
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d '{
    "prompt": "An avocado chair",
    "size": "1024x1024",
    "n": 3,
    "quality": "hd",
    "style": "vivid"
  }'
```

Example response

The operation returns a `202` status code and an `GenerateImagesResponse` JSON object containing the ID and status of the operation.

JSON

```
{
    "created": 1698116662,
    "data": [
        {
            "url": "url to the image",
```

```
            "revised_prompt": "the actual prompt that was used"
        },
        {
            "url": "url to the image"
        },
        ...
    ]
}
```

Request a generated image (DALL-E 2)

Generate a batch of images from a text caption.

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/images/generations:submit?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`api-version`	string	Required	The API version to use for this operation. This follows the YYYY-MM-DD format.

Supported versions

- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-12-01-preview` Swagger spec

Request body

Expand table

Parameter	Type	Required?	Default	Description
`prompt`	string	Required		A text description of the desired image(s). The maximum length is 1000 characters.
`n`	integer	Optional	1	The number of images to generate. Must be between 1 and 5.
`size`	string	Optional	1024x1024	The size of the generated images. Must be one of `256x256`, `512x512`, or `1024x1024`.

Example request

Console

```
curl -X POST https://YOUR_RESOURCE_NAME.openai.azure.com/openai/images/generations:submit?api-version=2023-06-01-preview
\
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d '{
"prompt": "An avocado chair",
"size": "512x512",
"n": 3
}'
```

Example response

The operation returns a `202` status code and an `GenerateImagesResponse` JSON object containing the ID and status of the operation.

JSON

```
{
  "id": "f508bcf2-e651-4b4b-85a7-58ad77981ffa",
  "status": "notRunning"
}
```

Get a generated image result (DALL-E 2)

Use this API to retrieve the results of an image generation operation. Image generation is currently only available with `api-version=2023-06-01-preview`.

HTTP

```
GET https://{your-resource-name}.openai.azure.com/openai/operations/images/{operation-id}?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`operation-id`	string	Required	The GUID that identifies the original image generation request.

Supported versions

- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec

Example request

Console

```
curl -X GET "https://{your-resource-name}.openai.azure.com/openai/operations/images/{operation-id}?api-version=2023-06-01-preview"
-H "Content-Type: application/json"
-H "Api-Key: {api key}"
```

Example response

Upon success the operation returns a `200` status code and an `OperationResponse` JSON object. The `status` field can be `"notRunning"` (task is queued but hasn't started yet), `"running"`, `"succeeded"`, `"canceled"` (task has timed out), `"failed"`, or `"deleted"`. A `succeeded` status indicates that the generated image is available for download at the given URL. If multiple images were generated, their URLs are all returned in the `result.data` field.

JSON

```
{
  "created": 1685064331,
  "expires": 1685150737,
  "id": "4b755937-3173-4b49-bf3f-da6702a3971a",
  "result": {
    "data": [
      {
        "url": "<URL_TO_IMAGE>"
      },
      {
        "url": "<URL_TO_NEXT_IMAGE>"
      },
      ...
    ]
  },
  "status": "succeeded"
}
```

Delete a generated image from the server (DALL-E 2)

You can use the operation ID returned by the request to delete the corresponding image from the Azure server. Generated images are automatically deleted after 24 hours by default, but you can trigger the deletion earlier if you want to.

HTTP

```
DELETE https://{your-resource-name}.openai.azure.com/openai/operations/images/{operation-id}?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI Resource.
`operation-id`	string	Required	The GUID that identifies the original image generation request.

Supported versions

- `2023-06-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-07-01-preview` (retiring 2024-04-02) Swagger spec

- `2023-08-01-preview` (retiring 2024-04-02) Swagger spec

Example request

Console

```
curl -X DELETE "https://{your-resource-name}.openai.azure.com/openai/operations/images/{operation-id}?api-version=2023-06-01-preview"
-H "Content-Type: application/json"
-H "Api-Key: {api key}"
```

Response

The operation returns a `204` status code if successful. This API only succeeds if the operation is in an end state (not `running`).

Speech to text

Request a speech to text transcription

Transcribes an audio file.

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/audio/transcriptions?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI resource.
`deployment-id`	string	Required	The name of your Whisper model deployment such as *MyWhisperDeployment*. You're required to first deploy a Whisper model before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This value follows the YYYY-MM-DD format.

Supported versions

- `2023-09-01-preview` (retiring 2024-04-02) Swagger spec
- `2023-12-01-preview` Swagger spec

Request body

Expand table

Parameter	Type	Required?	Default	Description
`file`	file	Yes	N/A	The audio file object (not file name) to transcribe, in one of these formats: `flac`, `mp3`, `mp4`, `mpeg`, `mpga`, `m4a`, `ogg`, `wav`, or `webm`. The file size limit for the Azure OpenAI Whisper model is 25 MB. If you need to transcribe a file larger than 25 MB, break it into chunks. Alternatively you can use the Azure AI Speech batch transcription API. You can get sample audio files from the Azure AI Speech SDK repository at GitHub.
`language`	string	No	Null	The language of the input audio such as `fr`. Supplying the input language in ISO-639-1 format improves accuracy and latency. For the list of supported languages, see the OpenAI documentation.
`prompt`	string	No	Null	An optional text to guide the model's style or continue a previous audio segment. The prompt should match the audio language. For more information about prompts including example use cases, see the OpenAI documentation.
`response_format`	string	No	json	The format of the transcript output, in one of these options: json, text, srt, verbose_json, or vtt. The default value is *json*.
`temperature`	number	No	0	The sampling temperature, between 0 and 1. Higher values like 0.8 makes the output more random, while lower values like 0.2 make it more focused and deterministic. If set to 0, the model uses log probability to automatically increase the temperature until certain thresholds are hit.

	The default value is *0*.

Example request

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/audio/transcriptions?api-version=2023-09-01-preview \
  -H "Content-Type: multipart/form-data" \
  -H "api-key: $YOUR_API_KEY" \
  -F file="@./YOUR_AUDIO_FILE_NAME.wav" \
  -F "language=en" \
  -F "prompt=The transcript contains zoology terms and geographical locations." \
  -F "temperature=0" \
  -F "response_format=srt"
```

Example response

srt

```
1
00:00:00,960 --> 00:00:07,680
The ocelot, Lepardus paradalis, is a small wild cat native to the southwestern United States,

2
00:00:07,680 --> 00:00:13,520
Mexico, and Central and South America. This medium-sized cat is characterized by

3
00:00:13,520 --> 00:00:18,960
solid black spots and streaks on its coat, round ears, and white neck and undersides.

4
00:00:19,760 --> 00:00:27,840
It weighs between 8 and 15.5 kilograms, 18 and 34 pounds, and reaches 40 to 50 centimeters

5
00:00:27,840 --> 00:00:34,560
16 to 20 inches at the shoulders. It was first described by Carl Linnaeus in 1758.

6
00:00:35,360 --> 00:00:42,880
Two subspecies are recognized, L. p. paradalis and L. p. mitis. Typically active during twilight

7
00:00:42,880 --> 00:00:48,480
and at night, the ocelot tends to be solitary and territorial. It is efficient at climbing,

8
00:00:48,480 --> 00:00:54,480
leaping, and swimming. It preys on small terrestrial mammals such as armadillo, opossum,

9
00:00:54,480 --> 00:00:56,480
and lagomorphs.
```

Request a speech to text translation

Translates an audio file from another language into English. For the list of supported languages, see the OpenAI documentation.

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/audio/translations?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI resource.
`deployment-id`	string	Required	The name of your Whisper model deployment such as *MyWhisperDeployment*. You're required to first deploy a Whisper model before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This value follows the YYYY-MM-DD format.

Supported versions

- `2023-09-01-preview` (retiring 2024-04-02) Swagger spec

- `2023-12-01-preview` Swagger spec

Request body

Expand table

Parameter	Type	Required?	Default	Description
`file`	file	Yes	N/A	The audio file object (not file name) to transcribe, in one of these formats: flac, mp3, mp4, mpeg, mpga, m4a, ogg, wav, or webm. The file size limit for the Azure OpenAI Whisper model is 25 MB. If you need to transcribe a file larger than 25 MB, break it into chunks. You can download sample audio files from the Azure AI Speech SDK repository at GitHub.
`prompt`	string	No	Null	An optional text to guide the model's style or continue a previous audio segment. The prompt should match the audio language. For more information about prompts including example use cases, see the OpenAI documentation.
`response_format`	string	No	json	The format of the transcript output, in one of these options: json, text, srt, verbose_json, or vtt. The default value is *json*.
`temperature`	number	No	0	The sampling temperature, between 0 and 1. Higher values like 0.8 makes the output more random, while lower values like 0.2 make it more focused and deterministic. If set to 0, the model uses log probability to automatically increase the temperature until certain thresholds are hit. The default value is *0*.

Example request

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/audio/translations?api-version=2023-09-01-preview \
  -H "Content-Type: multipart/form-data" \
  -H "api-key: $YOUR_API_KEY" \
  -F file="@./YOUR_AUDIO_FILE_NAME.wav" \
  -F "temperature=0" \
  -F "response_format=json"
```

Example response

JSON

```
{
  "text": "Hello, my name is Wolfgang and I come from Germany. Where are you heading today?"
}
```

Text to speech

Synthesize text to speech.

HTTP

```
POST https://{your-resource-name}.openai.azure.com/openai/deployments/{deployment-id}/audio/speech?api-version={api-version}
```

Path parameters

Expand table

Parameter	Type	Required?	Description
`your-resource-name`	string	Required	The name of your Azure OpenAI resource.
`deployment-id`	string	Required	The name of your text to speech model deployment such as *MyTextToSpeechDeployment*. You're required to first deploy a text to speech model (such as `tts-1` or `tts-1-hd`) before you can make calls.
`api-version`	string	Required	The API version to use for this operation. This value follows the YYYY-MM-DD format.

Supported versions

- `2024-02-15-preview`

Request body

Expand table

Parameter	Type	Required?	Default	Description
`model`	string	Yes	N/A	One of the available TTS models: `tts-1` or `tts-1-hd`
`input`	string	Yes	N/A	The text to generate audio for. The maximum length is 4096 characters. Specify input text in the language of your choice.[1]
`voice`	string	Yes	N/A	The voice to use when generating the audio. Supported voices are `alloy`, `echo`, `fable`, `onyx`, `nova`, and `shimmer`. Previews of the voices are available in the OpenAI text to speech guide.

[1] The text to speech models generally support the same languages as the Whisper model. For the list of supported languages, see the OpenAI documentation.

Example request

Console

```
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/audio/speech?api-version=2024-02-15-preview \
 -H "api-key: $YOUR_API_KEY" \
 -H "Content-Type: application/json" \
 -d '{
    "model": "tts-hd",
    "input": "I'm excited to try text to speech.",
    "voice": "alloy"
}' --output speech.mp3
```

Example response

The speech is returned as an audio file from the previous request.

Management APIs

Azure OpenAI is deployed as a part of the Azure AI services. All Azure AI services rely on the same set of management APIs for creation, update and delete operations. The management APIs are also used for deploying models within an OpenAI resource.

Management APIs reference documentation

Next steps

Learn about Models, and fine-tuning with the REST API. Learn more about the underlying models that power Azure OpenAI.