

# Introduction<sup>⌚</sup>

This API reference describes the RESTful, streaming, and realtime APIs you can use to interact with the OpenAI platform. REST APIs are usable via HTTP in any environment that supports HTTP requests. Language-specific SDKs are listed [on the libraries page](#).

---

# Authentication<sup>⌚</sup>

The OpenAI API uses API keys for authentication. Create, manage, and learn more about API keys in your [organization settings](#).

**Remember that your API key is a secret!** Do not share it with others or expose it in any client-side code (browsers, apps). API keys should be securely loaded from an environment variable or key management service on the server.

API keys should be provided via [HTTP Bearer authentication](#).

Authorization: Bearer OPENAI\_API\_KEY

If you belong to multiple organizations or access projects through a legacy user API key, pass a header to specify which organization and project to use for an API request:

```
curl https://api.openai.com/v1/models \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Organization: $ORGANIZATION_ID" \
-H "OpenAI-Project: $PROJECT_ID"
```

Usage from these API requests counts as usage for the specified organization and project. Organization IDs can be found on your [organization settings](#) page. Project IDs can be found on your [general settings](#) page by selecting the specific project.

## Debugging requests<sup>②</sup>

In addition to [error codes](#) returned from API responses, you can inspect HTTP response headers containing the unique ID of a particular API request or information about rate limiting applied to your requests. Below is an incomplete list of HTTP headers returned with API responses:

### API meta information

`openai-organization` : The [organization](#) associated with the request

`openai-processing-ms` : Time taken processing your API request

`openai-version` : REST API version used for this request (currently `2020-10-01` )

`x-request-id` : Unique identifier for this API request (used in troubleshooting)

## Rate limiting information

`x-ratelimit-limit-requests`

`x-ratelimit-limit-tokens`

`x-ratelimit-remaining-requests`

`x-ratelimit-remaining-tokens`

`x-ratelimit-reset-requests`

`x-ratelimit-reset-tokens`

**OpenAI recommends logging request IDs in production deployments** for more efficient troubleshooting with our [support team](#), should the need arise. Our [official SDKs](#) provide a property on top-level response objects containing the value of the `x-request-id` header.

---

## Backward compatibility<sup>2</sup>

OpenAI is committed to providing stability to API users by avoiding breaking changes in major API versions whenever reasonably possible. This includes:

The REST API (currently `v1`)

Our first-party [SDKs](#) (released SDKs adhere to [semantic versioning](#))

Model families (like `gpt-4o` or `o4-mini`)

**Model prompting behavior between snapshots is subject to change.** Model outputs are by their nature variable, so expect changes in prompting and model behavior between snapshots.

For example, if you moved from `gpt-4o-2024-05-13` to `gpt-4o-2024-08-06`, the same `system` or `user` messages could function differently between versions. The best way to ensure consistent prompting behavior and model output is to use pinned model versions, and to implement [evals](#) for your applications.

## Backwards-compatible API changes:

Adding new resources (URLs) to the REST API and SDKs

Adding new optional API parameters

Adding new properties to JSON response objects or event data

Changing the order of properties in a JSON response object

Changing the length or format of opaque strings, like resource identifiers and UUIDs

Adding new event types (in either streaming or the Realtime API)

See the [changelog](#) for a list of backwards-compatible changes and rare breaking changes.

---

# Responses



OpenAI's most advanced interface for generating model responses. Supports text and image inputs, and text outputs. Create stateful interactions with the model, using the output of previous responses as input. Extend the model's capabilities with built-in tools for file search, web search, computer use, and more. Allow the model access to external systems and data using function calling.

Related guides:

[Quickstart](#)

[Text inputs and outputs](#)

[Image inputs](#)

[Structured Outputs](#)

[Function calling](#)

[Conversation state](#)

## Extend the models with tools

# Create a model response



```
POST https://api.openai.com/v1/responses
```

Creates a model response. Provide [text](#) or [image](#) inputs to generate [text](#) or [JSON](#) outputs. Have the model call your own [custom code](#) or use built-in [tools](#) like [web search](#) or [file search](#) to use your own data as input for the model's response.

## Request body

**background** boolean Optional Defaults to false

Whether to run the model response in the background. [Learn more](#).

**conversation** string or object Optional Defaults to null

The conversation that this response belongs to. Items from this conversation are prepended to [input\\_items](#) for this response request. Input items and output items from this response are automatically added to this conversation after this response completes.

> Show possible types

**include** array Optional

Specify additional output data to include in the model response. Currently supported values are:

`web_search_call.action.sources` : Include the sources of the web search tool call.

`code_interpreter_call.outputs` : Includes the outputs of python code execution in code interpreter tool call items.

`computer_call_output.output.image_url` : Include image urls from the computer call output.

`file_search_call.results` : Include the search results of the file search tool call.

`message.input_image.image_url` : Include image urls from the input message.

`message.output_text.logprobs` : Include logprobs with assistant messages.

`reasoning.encrypted_content` : Includes an encrypted version of reasoning tokens in reasoning item outputs. This enables reasoning items to be used in multi-turn conversations when using the Responses API statelessly (like when the `store` parameter is set to `false`, or when an organization is enrolled in the zero data retention program).

---

**input** string or array Optional

Text, image, or file inputs to the model, used to generate a response.

Learn more:

[Text inputs and outputs](#)

[Image inputs](#)

[File inputs](#)

[Conversation state](#)

[Function calling](#)

> Show possible types

---

**instructions** string Optional

A system (or developer) message inserted into the model's context.

When using along with `previous_response_id`, the instructions from a previous response will not be carried over to the next response. This makes it simple to swap out system (or developer) messages in new responses.

---

**max\_output\_tokens** integer Optional

An upper bound for the number of tokens that can be generated for a response, including visible output tokens and reasoning tokens.

---

**max\_tool\_calls** integer Optional

The maximum number of total calls to built-in tools that can be processed in a response. This maximum number applies across all built-in tool calls, not per individual tool. Any further attempts to call a tool by the model will be ignored.

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string Optional

Model ID used to generate the response, like `gpt-4o` or `o3`. OpenAI offers a wide range of models with different capabilities, performance characteristics, and price points. Refer to the [model guide](#) to browse and compare available models.

---

**parallel\_tool\_calls** boolean Optional Defaults to true

Whether to allow the model to run tool calls in parallel.

---

**previous\_response\_id** string Optional

The unique ID of the previous response to the model. Use this to create multi-turn conversations. Learn more about [conversation state](#). Cannot be used in conjunction with [conversation](#).

---

**prompt** object Optional

Reference to a prompt template and its variables. [Learn more](#).

> Show properties

---

**prompt\_cache\_key** string Optional

Used by OpenAI to cache responses for similar requests to optimize your cache hit rates. Replaces the [user](#) field.

[Learn more](#).

---

**reasoning** object Optional

**gpt-5 and o-series models only**

Configuration options for [reasoning models](#).

> Show properties

---

**safety\_identifier** string Optional

A stable identifier used to help detect users of your application that may be violating OpenAI's usage policies. The IDs should be a string that uniquely identifies each user. We recommend hashing their username or email address, in order to avoid sending us any identifying information. [Learn more](#).

**service\_tier** string Optional Defaults to auto

Specifies the processing type used for serving the request.

If set to 'auto', then the request will be processed with the service tier configured in the Project settings. Unless otherwise configured, the Project will use 'default'.

If set to 'default', then the request will be processed with the standard pricing and performance for the selected model.

If set to '[flex](#)' or '[priority](#)', then the request will be processed with the corresponding service tier.

When not set, the default behavior is 'auto'.

When the `service_tier` parameter is set, the response body will include the `service_tier` value based on the processing mode actually used to serve the request. This response value may be different from the value set in the parameter.

---

**store** boolean Optional Defaults to true

Whether to store the generated model response for later retrieval via API.

---

**stream** boolean Optional Defaults to false

If set to true, the model response data will be streamed to the client as it is generated using [server-sent events](#). See the [Streaming section below](#) for more information.

---

**stream\_options** object Optional Defaults to null

Options for streaming responses. Only set this when you set `stream: true`.

> Show properties

---

**temperature** number Optional Defaults to 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.

---

**text** object Optional

Configuration options for a text response from the model. Can be plain text or structured JSON data. Learn more:

[Text inputs and outputs](#)

[Structured Outputs](#)

> Show properties

---

**tool\_choice** string or object Optional

How the model should select which tool (or tools) to use when generating a response. See the `tools` parameter to see how to specify which tools the model can call.

> Show possible types

---

**tools** array Optional

An array of tools the model may call while generating a response. You can specify which tool to use by setting the `tool_choice` parameter.

We support the following categories of tools:

**Built-in tools:** Tools that are provided by OpenAI that extend the model's capabilities, like [web search](#) or [file search](#).

Learn more about [built-in tools](#).

**MCP Tools:** Integrations with third-party systems via custom MCP servers or predefined connectors such as Google Drive and SharePoint. Learn more about [MCP Tools](#).

**Function calls (custom tools):** Functions that are defined by you, enabling the model to call your own code with strongly typed arguments and outputs. Learn more about [function calling](#). You can also use custom tools to call your own code.

> Show possible types

---

**top\_logprobs** integer Optional

An integer between 0 and 20 specifying the number of most likely tokens to return at each token position, each with an associated log probability.

---

**top\_p** number Optional Defaults to 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.

---

**truncation** string Optional Defaults to disabled

The truncation strategy to use for the model response.

`auto` : If the input to this Response exceeds the model's context window size, the model will truncate the response to fit the context window by dropping items from the beginning of the conversation.

`disabled` (default): If the input size will exceed the context window size for a model, the request will fail with a 400 error.

---

**user** Deprecated string Optional

This field is being replaced by `safety_identifier` and `prompt_cache_key`. Use `prompt_cache_key` instead to maintain caching optimizations. A stable identifier for your end-users. Used to boost cache hit rates by better bucketing similar requests and to help OpenAI detect and prevent abuse. [Learn more](#).

## Returns

Returns a [Response](#) object.

**Text input**

**Image input**

**File input**

**Web search**

**File search**

**Streaming**

**Functions**

**Reasoning**

Example request

```
curl https://api.openai.com/v1/responses \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "model": "gpt-4.1",
  "input": "Tell me a three sentence bedtime story about a unicorn."
}'
```

Response

```
{
  "id": "resp_67ccd2bed1ec8190b14f964abc0542670bb6a6b452d3795b",
  "object": "response",
  "created_at": 1741476542,
```

```
"status": "completed",
"error": null,
"incomplete_details": null,
"instructions": null,
"max_output_tokens": null,
"model": "gpt-4.1-2025-04-14",
"output": [
{
  "type": "message",
  "id": "msg_67ccd2bf17f0819081ff3bb2cf6508e60bb6a6b452d3795b",
  "status": "completed",
  "role": "assistant",
  "content": [
    {
      "type": "output_text",
      "text": "In a peaceful grove beneath a silver moon, a unicorn named Lumina discovered a lone flower that glowed with a soft, ethereal light. As she approached it, the flower began to sing a gentle melody that calmed her racing thoughts. Lumina closed her eyes and listened intently, letting the music wash over her. When she opened them again, she felt a sense of clarity and peace she had never known before. She decided to take a moment to appreciate the beauty of the world around her, grateful for the reminder of the magic still hidden in the corners of the universe.",
      "annotations": []
    }
  ]
},
"parallel_tool_calls": true,
"previous_response_id": null,
"reasoning": {
  "effort": null,
  "summary": null
},
"store": true,
"temperature": 1.0,
```

```
"text": {  
    "format": {  
        "type": "text"  
    }  
},  
"tool_choice": "auto",  
"tools": [],  
"top_p": 1.0,  
"truncation": "disabled",  
"usage": {  
    "input_tokens": 36,  
    "input_tokens_details": {  
        "cached_tokens": 0  
    },  
    "output_tokens": 87,  
    "output_tokens_details": {  
        "reasoning_tokens": 0  
    },  
    "total_tokens": 123  
},  
"user": null,  
"metadata": {}  
}
```

# Get a model response



```
GET https://api.openai.com/v1/responses/{response_id}
```

Retrieves a model response with the given ID.

## Path parameters

**response\_id** string Required

The ID of the response to retrieve.

## Query parameters

**include** array Optional

Additional fields to include in the response. See the `include` parameter for Response creation above for more information.

**include\_obfuscation** boolean Optional

When true, stream obfuscation will be enabled. Stream obfuscation adds random characters to an `obfuscation` field on streaming delta events to normalize payload sizes as a mitigation to certain side-channel attacks. These obfuscation fields are included by default, but add a small amount of overhead to the data stream. You can set `include_obfuscation` to false to optimize for bandwidth if you trust the network links between your application and the OpenAI API.

**starting\_after** integer Optional

The sequence number of the event after which to start streaming.

**stream** boolean Optional

If set to true, the model response data will be streamed to the client as it is generated using [server-sent events](#). See the [Streaming section below](#) for more information.

## Returns

The [Response](#) object matching the specified ID.

Example request

```
curl https://api.openai.com/v1/responses/resp_123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

Response

```
{
  "id": "resp_67cb71b351908190a308f3859487620d06981a8637e6bc44",
  "object": "response",
  "created_at": 1741386163,
  "status": "completed",
  "error": null,
  "incomplete_details": null,
  "instructions": null,
```

```
"max_output_tokens": null,  
"model": "gpt-4o-2024-08-06",  
"output": [  
  {  
    "type": "message",  
    "id": "msg_67cb71b3c2b0819084d481baaaf148f206981a8637e6bc44",  
    "status": "completed",  
    "role": "assistant",  
    "content": [  
      {  
        "type": "output_text",  
        "text": "Silent circuits hum, \nThoughts emerge in data streams— \nDigital dawn breaks."  
        "annotations": []  
      }  
    ]  
  },  
  ],  
  "parallel_tool_calls": true,  
  "previous_response_id": null,  
  "reasoning": {  
    "effort": null,  
    "summary": null  
  },  
  "store": true,  
  "temperature": 1.0,  
  "text": {  
    "format": {  
      "type": "text"  
    }  
  }  
]
```

```
},
"tool_choice": "auto",
"tools": [],
"top_p": 1.0,
"truncation": "disabled",
"usage": {
  "input_tokens": 32,
  "input_tokens_details": {
    "cached_tokens": 0
  },
  "output_tokens": 18,
  "output_tokens_details": {
    "reasoning_tokens": 0
  },
  "total_tokens": 50
},
"user": null,
"metadata": {}
}
```

## Delete a model response



```
DELETE https://api.openai.com/v1/responses/{response_id}
```

Deletes a model response with the given ID.

## Path parameters

**response\_id** string Required

The ID of the response to delete.

## Returns

A success message.

### Example request

```
curl -X DELETE https://api.openai.com/v1/responses/resp_123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "id": "resp_6786a1bec27481909a17d673315b29f6",
  "object": "response",
  "deleted": true
}
```

# Cancel a response



```
POST https://api.openai.com/v1/responses/{response_id}/cancel
```

Cancels a model response with the given ID. Only responses created with the `background` parameter set to `true` can be cancelled. [Learn more.](#)

## Path parameters

---

`response_id` string Required

The ID of the response to cancel.

## Returns

---

A [Response](#) object.

Example request

```
curl -X POST https://api.openai.com/v1/responses/resp_123/cancel \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{
  "id": "resp_67cb71b351908190a308f3859487620d06981a8637e6bc44",
  "object": "response",
  "created_at": 1741386163,
  "status": "completed",
  "error": null,
  "incomplete_details": null,
  "instructions": null,
  "max_output_tokens": null,
  "model": "gpt-4o-2024-08-06",
  "output": [
    {
      "type": "message",
      "id": "msg_67cb71b3c2b0819084d481baaf148f206981a8637e6bc44",
      "status": "completed",
      "role": "assistant",
      "content": [
        {
          "type": "output_text",
          "text": "Silent circuits hum, \nThoughts emerge in data streams— \nDigital dawn breaks."
          "annotations": []
        }
      ]
    }
  ]
}
```

```
        ],
    },
],
"parallel_tool_calls": true,
"previous_response_id": null,
"reasoning": {
    "effort": null,
    "summary": null
},
"store": true,
"temperature": 1.0,
"text": {
    "format": {
        "type": "text"
    }
},
"tool_choice": "auto",
"tools": [],
"top_p": 1.0,
"truncation": "disabled",
"usage": {
    "input_tokens": 32,
    "input_tokens_details": {
        "cached_tokens": 0
    },
    "output_tokens": 18,
    "output_tokens_details": {
        "reasoning_tokens": 0
    },
}
```

```
"total_tokens": 50
},
"user": null,
"metadata": {}
}
```

# List input items



```
GET https://api.openai.com/v1/responses/{response_id}/input_items
```

Returns a list of input items for a given response.

## Path parameters

**response\_id** string Required

The ID of the response to retrieve input items for.

## Query parameters

**after** string Optional

An item ID to list items after, used in pagination.

---

**include** array Optional

Additional fields to include in the response. See the `include` parameter for Response creation above for more information.

---

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

---

**order** string Optional

The order to return the input items in. Default is `desc`.

`asc` : Return the input items in ascending order.

`desc` : Return the input items in descending order.

---

## Returns

A list of input item objects.

---

**Example request**

```
curl https://api.openai.com/v1/responses/resp_abc123/input_items \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "msg_abc123",  
      "type": "message",  
      "role": "user",  
      "content": [  
        {  
          "type": "input_text",  
          "text": "Tell me a three sentence bedtime story about a unicorn."  
        }  
      ]  
    },  
    {"first_id": "msg_abc123",  
     "last_id": "msg_abc123",  
     "has_more": false  
  }  
}
```

# Get input token counts

🔗

POST [https://api.openai.com/v1/responses/input\\_tokens](https://api.openai.com/v1/responses/input_tokens)

## Get input token counts

### Request body

---

**conversation** string or object Optional Defaults to null

The conversation that this response belongs to. Items from this conversation are prepended to `input_items` for this response request. Input items and output items from this response are automatically added to this conversation after this response completes.

> Show possible types

---

**input** string or array Optional

Text, image, or file inputs to the model, used to generate a response

> Show possible types

---

**instructions** string Optional

A system (or developer) message inserted into the model's context. When used along with `previous_response_id`, the instructions from a previous response will not be carried over to the next response. This makes it simple to swap out system (or developer) messages in new responses.

---

**model** string Optional

Model ID used to generate the response, like `gpt-4o` or `o3`. OpenAI offers a wide range of models with different capabilities, performance characteristics, and price points. Refer to the [model guide](#) to browse and compare available models.

**parallel\_tool\_calls** boolean Optional

Whether to allow the model to run tool calls in parallel.

---

**previous\_response\_id** string Optional

The unique ID of the previous response to the model. Use this to create multi-turn conversations. Learn more about [conversation state](#). Cannot be used in conjunction with [conversation](#).

---

**reasoning** object Optional

**gpt-5 and o-series models only**

Configuration options for [reasoning models](#).

> Show properties

---

**text** object Optional

Configuration options for a text response from the model. Can be plain text or structured JSON data. Learn more:

[Text inputs and outputs](#)

[Structured Outputs](#)

> Show properties

---

**tool\_choice** string or object Optional

How the model should select which tool (or tools) to use when generating a response. See the [tools](#) parameter to see how to specify which tools the model can call.

> Show possible types

**tools** array Optional

An array of tools the model may call while generating a response. You can specify which tool to use by setting the `tool_choice` parameter.

> Show possible types

**truncation** string Optional

The truncation strategy to use for the model response. - `auto` : If the input to this Response exceeds the model's context window size, the model will truncate the response to fit the context window by dropping items from the beginning of the conversation. - `disabled` (default): If the input size will exceed the context window size for a model, the request will fail with a 400 error.

## Returns

The input token counts.

```
{  
  object: "response.input_tokens"  
  input_tokens: 123  
}
```

### Example request

```
curl -X POST https://api.openai.com/v1/responses/input_tokens \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "model": "gpt-5",
  "input": "Tell me a joke."
}'
```

## Response

```
{  
  "object": "response.input_tokens",  
  "input_tokens": 11  
}
```

# The response object



## **background** boolean

Whether to run the model response in the background. [Learn more.](#)

## **conversation** object

The conversation that this response belongs to. Input items and output items from this response are automatically added to this conversation.

> Show properties

---

**created\_at** number

Unix timestamp (in seconds) of when this Response was created.

---

**error** object

An error object returned when the model fails to generate a Response.

> Show properties

---

**id** string

Unique identifier for this Response.

---

**incomplete\_details** object

Details about why the response is incomplete.

> Show properties

---

**instructions** string or array

A system (or developer) message inserted into the model's context.

When using along with `previous_response_id`, the instructions from a previous response will not be carried over to the next response. This makes it simple to swap out system (or developer) messages in new responses.

> Show possible types

---

**max\_output\_tokens** integer

An upper bound for the number of tokens that can be generated for a response, including visible output tokens and [reasoning tokens](#).

---

**max\_tool\_calls** integer

The maximum number of total calls to built-in tools that can be processed in a response. This maximum number applies across all built-in tool calls, not per individual tool. Any further attempts to call a tool by the model will be ignored.

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string

Model ID used to generate the response, like `gpt-4o` or `o3`. OpenAI offers a wide range of models with different capabilities, performance characteristics, and price points. Refer to the [model guide](#) to browse and compare available models.

---

**object** string

The object type of this resource - always set to `response`.

---

**output** array

An array of content items generated by the model.

The length and order of items in the `output` array is dependent on the model's response.

Rather than accessing the first item in the `output` array and assuming it's an `assistant` message with the content generated by the model, you might consider using the `output_text` property where supported in SDKs.

> Show possible types

---

**output\_text** string SDK Only

SDK-only convenience property that contains the aggregated text output from all `output_text` items in the `output` array, if any are present. Supported in the Python and JavaScript SDKs.

---

**parallel\_tool\_calls** boolean

Whether to allow the model to run tool calls in parallel.

---

**previous\_response\_id** string

The unique ID of the previous response to the model. Use this to create multi-turn conversations. Learn more about [conversation state](#). Cannot be used in conjunction with `conversation`.

---

**prompt** object

Reference to a prompt template and its variables. [Learn more](#).

> Show properties

---

**prompt\_cache\_key** string

Used by OpenAI to cache responses for similar requests to optimize your cache hit rates. Replaces the `user` field. [Learn more](#).

---

**reasoning** object**gpt-5 and o-series models only**

Configuration options for [reasoning models](#).

> Show properties

---

**safety\_identifier** string

A stable identifier used to help detect users of your application that may be violating OpenAI's usage policies. The IDs should be a string that uniquely identifies each user. We recommend hashing their username or email address, in order to avoid sending us any identifying information. [Learn more.](#)

---

**service\_tier** string

Specifies the processing type used for serving the request.

If set to 'auto', then the request will be processed with the service tier configured in the Project settings. Unless otherwise configured, the Project will use 'default'.

If set to 'default', then the request will be processed with the standard pricing and performance for the selected model.

If set to '[flex](#)' or '[priority](#)', then the request will be processed with the corresponding service tier.

When not set, the default behavior is 'auto'.

When the `service_tier` parameter is set, the response body will include the `service_tier` value based on the processing mode actually used to serve the request. This response value may be different from the value set in the parameter.

---

**status** string

The status of the response generation. One of `completed`, `failed`, `in_progress`, `cancelled`, `queued`, or `incomplete`.

---

**temperature** number

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.

## **text** object

Configuration options for a text response from the model. Can be plain text or structured JSON data. Learn more:

[Text inputs and outputs](#)

[Structured Outputs](#)

> Show properties

---

## **tool\_choice** string or object

How the model should select which tool (or tools) to use when generating a response. See the `tools` parameter to see how to specify which tools the model can call.

> Show possible types

---

## **tools** array

An array of tools the model may call while generating a response. You can specify which tool to use by setting the `tool_choice` parameter.

We support the following categories of tools:

**Built-in tools:** Tools that are provided by OpenAI that extend the model's capabilities, like [web search](#) or [file search](#).

Learn more about [built-in tools](#).

**MCP Tools:** Integrations with third-party systems via custom MCP servers or predefined connectors such as Google Drive and SharePoint. Learn more about [MCP Tools](#).

**Function calls (custom tools):** Functions that are defined by you, enabling the model to call your own code with strongly typed arguments and outputs. Learn more about [function calling](#). You can also use custom tools to call your own code.

> Show possible types

---

#### **top\_logprobs** integer

An integer between 0 and 20 specifying the number of most likely tokens to return at each token position, each with an associated log probability.

---

#### **top\_p** number

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.

---

#### **truncation** string

The truncation strategy to use for the model response.

`auto` : If the input to this Response exceeds the model's context window size, the model will truncate the response to fit the context window by dropping items from the beginning of the conversation.

`disabled` (default): If the input size will exceed the context window size for a model, the request will fail with a 400 error.

---

#### **usage** object

Represents token usage details including input tokens, output tokens, a breakdown of output tokens, and the total tokens used.

> Show properties

**user** Deprecated string

This field is being replaced by `safety_identifier` and `prompt_cache_key`. Use `prompt_cache_key` instead to maintain caching optimizations. A stable identifier for your end-users. Used to boost cache hit rates by better bucketing similar requests and to help OpenAI detect and prevent abuse. [Learn more](#).

OBJECT The response object

```
{  
  "id": "resp_67ccd3a9da748190baa7f1570fe91ac604becb25c45c1d41",  
  "object": "response",  
  "created_at": 1741476777,  
  "status": "completed",  
  "error": null,  
  "incomplete_details": null,  
  "instructions": null,  
  "max_output_tokens": null,  
  "model": "gpt-4o-2024-08-06",  
  "output": [  
    {  
      "type": "message",  
      "id": "msg_67ccd3acc8d48190a77525dc6de64b4104becb25c45c1d41",  
      "status": "completed",  
      "role": "assistant",  
      "content": [  
        {  
          "type": "output_text",  
          "text": "The image depicts a scenic landscape with a wooden boardwalk or pathway leading",  
          "annotations": []  
        }  
      ]  
    }  
  ]  
}
```

```
        }
    ],
},
],
"parallel_tool_calls": true,
"previous_response_id": null,
"reasoning": {
    "effort": null,
    "summary": null
},
"store": true,
"temperature": 1,
"text": {
    "format": {
        "type": "text"
    }
},
"tool_choice": "auto",
"tools": [],
"top_p": 1,
"truncation": "disabled",
"usage": {
    "input_tokens": 328,
    "input_tokens_details": {
        "cached_tokens": 0
    },
    "output_tokens": 52,
    "output_tokens_details": {
        "reasoning_tokens": 0
    }
}
```

```
    },
    "total_tokens": 380
  },
  "user": null,
  "metadata": {}
}
```

## The input item list



A list of Response items.

**data** array

A list of items used to generate this response.

> Show possible types

**first\_id** string

The ID of the first item in the list.

**has\_more** boolean

Whether there are more items available.

**last\_id** string

The ID of the last item in the list.

**object** string

The type of object returned, must be `list`.

OBJECT The input item list

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "msg_abc123",  
      "type": "message",  
      "role": "user",  
      "content": [  
        {  
          "type": "input_text",  
          "text": "Tell me a three sentence bedtime story about a unicorn."  
        }  
      ]  
    }  
  ],  
  "first_id": "msg_abc123",  
  "last_id": "msg_abc123",  
  "has_more": false  
}
```

# Conversations



Create and manage conversations to store and retrieve conversation state across Response API calls.

---

## Create a conversation



```
POST https://api.openai.com/v1/conversations
```

Create a conversation.

### Request body

---

**items** array Optional

Initial items to include in the conversation context. You may add up to 20 items at a time.

> Show possible types

**metadata** object or null Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard. Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

## Returns

Returns a [Conversation](#) object.

Example request

```
curl https://api.openai.com/v1/conversations \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "metadata": {"topic": "demo"},
  "items": [
    {
      "type": "message",
      "role": "user",
      "content": "Hello!"
    }
  ]
}'
```

## Response

```
{  
  "id": "conv_123",  
  "object": "conversation",  
  "created_at": 1741900000,  
  "metadata": {"topic": "demo"}  
}
```

# Retrieve a conversation



```
GET https://api.openai.com/v1/conversations/{conversation_id}
```

Get a conversation

## Path parameters

**conversation\_id** string Required

The ID of the conversation to retrieve.

## Returns

Returns a [Conversation](#) object.

#### Example request

```
curl https://api.openai.com/v1/conversations/conv_123 \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

#### Response

```
{
  "id": "conv_123",
  "object": "conversation",
  "created_at": 1741900000,
  "metadata": {"topic": "demo"}
}
```

# Update a conversation



```
POST https://api.openai.com/v1/conversations/{conversation_id}
```

## Update a conversation

## Path parameters

**conversation\_id** string Required

The ID of the conversation to update.

## Request body

**metadata** map Required

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

## Returns

Returns the updated [Conversation](#) object.

### Example request

```
curl https://api.openai.com/v1/conversations/conv_123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "metadata": {"topic": "project-x"}
}'
```

## Response

```
{  
  "id": "conv_123",  
  "object": "conversation",  
  "created_at": 1741900000,  
  "metadata": {"topic": "project-x"}  
}
```

# Delete a conversation



```
DELETE https://api.openai.com/v1/conversations/{conversation_id}
```

Delete a conversation. Items in the conversation will not be deleted.

## Path parameters

**conversation\_id** string Required

The ID of the conversation to delete.

## Returns

A success message.

#### Example request

```
curl -X DELETE https://api.openai.com/v1/conversations/conv_123 \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

#### Response

```
{
  "id": "conv_123",
  "object": "conversation.deleted",
  "deleted": true
}
```

## List items



```
GET https://api.openai.com/v1/conversations/{conversation_id}/items
```

List all items for a conversation with the given ID.

## Path parameters

---

### **conversation\_id** string Required

The ID of the conversation to list items for.

## Query parameters

---

### **after** string Optional

An item ID to list items after, used in pagination.

### **include** array Optional

Specify additional output data to include in the model response. Currently supported values are:

`web_search_call.action.sources` : Include the sources of the web search tool call.

`code_interpreter_call.outputs` : Includes the outputs of python code execution in code interpreter tool call items.

`computer_call_output.output.image_url` : Include image urls from the computer call output.

`file_search_call.results` : Include the search results of the file search tool call.

`message.input_image.image_url` : Include image urls from the input message.

`message.output_text.logprobs` : Include logprobs with assistant messages.

`reasoning.encrypted_content` : Includes an encrypted version of reasoning tokens in reasoning item outputs. This enables reasoning items to be used in multi-turn conversations when using the Responses API statelessly (like when the `store` parameter is set to `false`, or when an organization is enrolled in the zero data retention program).

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional

The order to return the input items in. Default is `desc`.

`asc` : Return the input items in ascending order.

`desc` : Return the input items in descending order.

## Returns

Returns a list object containing Conversation items.

### Example request

```
curl "https://api.openai.com/v1/conversations/conv_123/items?limit=10" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

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

```
"type": "message",
"id": "msg_abc",
"status": "completed",
"role": "user",
"content": [
    {"type": "input_text", "text": "Hello!"}
]
},
"first_id": "msg_abc",
"last_id": "msg_abc",
"has_more": false
}
```

# Create items



POST [https://api.openai.com/v1/conversations/{conversation\\_id}/items](https://api.openai.com/v1/conversations/{conversation_id}/items)

Create items in a conversation with the given ID.

## Path parameters

**conversation\_id** string Required

The ID of the conversation to add the item to.

## Query parameters

---

### **include** array Optional

Additional fields to include in the response. See the `include` parameter for [listing Conversation items above](#) for more information.

## Request body

---

### **items** array Required

The items to add to the conversation. You may add up to 20 items at a time.

> Show possible types

## Returns

---

Returns the list of added [items](#).

### Example request

```
curl https://api.openai.com/v1/conversations/conv_123/items \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
```

```
-d '{  
  "items": [  
    {  
      "type": "message",  
      "role": "user",  
      "content": [  
        {"type": "input_text", "text": "Hello!"}  
      ]  
    },  
    {  
      "type": "message",  
      "role": "user",  
      "content": [  
        {"type": "input_text", "text": "How are you?"}  
      ]  
    }  
  ]  
}'
```

## Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "type": "message",  
      "id": "msg_abc",  
      "status": "completed",  
      "role": "user",  
      "content": [{"text": "Hello!"}],  
      "created": 1683219400  
    }  
  ]  
}
```

```
"content": [
    {"type": "input_text", "text": "Hello!"}
]
},
{
    "type": "message",
    "id": "msg_def",
    "status": "completed",
    "role": "user",
    "content": [
        {"type": "input_text", "text": "How are you?"}
    ]
},
],
"first_id": "msg_abc",
"last_id": "msg_def",
"has_more": false
}
```

## Retrieve an item



GET [https://api.openai.com/v1/conversations/{conversation\\_id}/items/{item\\_id}](https://api.openai.com/v1/conversations/{conversation_id}/items/{item_id})

Get a single item from a conversation with the given IDs.

## Path parameters

---

**conversation\_id** string Required

The ID of the conversation that contains the item.

---

**item\_id** string Required

The ID of the item to retrieve.

---

## Query parameters

---

**include** array Optional

Additional fields to include in the response. See the `include` parameter for [listing Conversation items above](#) for more information.

---

## Returns

---

Returns a [Conversation Item](#).

---

Example request

```
curl https://api.openai.com/v1/conversations/conv_123/items/msg_abc \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "type": "message",  
  "id": "msg_abc",  
  "status": "completed",  
  "role": "user",  
  "content": [  
    {"type": "input_text", "text": "Hello!"}  
  ]  
}
```

## Delete an item



```
DELETE https://api.openai.com/v1/conversations/{conversation_id}/items/{item_id}
```

Delete an item from a conversation with the given IDs.

## Path parameters

**conversation\_id** string Required

The ID of the conversation that contains the item.

**item\_id** string Required

The ID of the item to delete.

## Returns

Returns the updated [Conversation](#) object.

### Example request

```
curl -X DELETE https://api.openai.com/v1/conversations/conv_123/items/msg_abc \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "id": "conv_123",
  "object": "conversation",
  "created_at": 1741900000,
  "metadata": {"topic": "demo"}
}
```

# The conversation object



## **created\_at** integer

The time at which the conversation was created, measured in seconds since the Unix epoch.

---

## **id** string

The unique ID of the conversation.

---

## **metadata**

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard. Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

## **object** string

The object type, which is always `conversation`.

---

# The item list



A list of Conversation items.

---

**data** array

A list of conversation items.

> Show possible types

---

**first\_id** string

The ID of the first item in the list.

---

**has\_more** boolean

Whether there are more items available.

---

**last\_id** string

The ID of the last item in the list.

---

**object** string

The type of object returned, must be `list`.

---

# Videos



Generate videos.

---

## Create video



```
POST https://api.openai.com/v1/videos
```

Create a video

### Request body

**prompt** string Required

Text prompt that describes the video to generate.

---

**input\_reference** file Optional

Optional image reference that guides generation.

---

**model** string Optional

The video generation model to use. Defaults to `sora-2`.

**seconds** string Optional

Clip duration in seconds. Defaults to 4 seconds.

**size** string Optional

Output resolution formatted as width x height. Defaults to 720x1280.

## Returns

Returns the newly created [video job](#).

Example request

```
curl https://api.openai.com/v1/videos \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-F "model=sora-2" \
-F "prompt=A calico cat playing a piano on stage"
```

Response

```
{
  "id": "video_123",
  "object": "video",
  "model": "sora-2",
```

```
"status": "queued",
"progress": 0,
"created_at": 1712697600,
"size": "1024x1808",
"seconds": "8",
"quality": "standard"
}
```

# Remix video



POST [https://api.openai.com/v1/videos/{video\\_id}/remix](https://api.openai.com/v1/videos/{video_id}/remix)

Create a video remix

## Path parameters

**video\_id** string Required

The identifier of the completed video to remix.

## Request body

**prompt** string Required

Updated text prompt that directs the remix generation.

## Returns

Creates a remix of the specified [video job](#) using the provided prompt.

### Example request

```
curl -X POST https://api.openai.com/v1/videos/video_123/remix \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "prompt": "Extend the scene with the cat taking a bow to the cheering audience"
}'
```

### Response

```
{
  "id": "video_456",
  "object": "video",
  "model": "sora-2",
  "status": "queued",
```

```
"progress": 0,  
"created_at": 1712698600,  
"size": "720x1280",  
"seconds": "8",  
"remixed_from_video_id": "video_123"  
}
```

# List videos



GET <https://api.openai.com/v1/videos>

List videos

## Query parameters

**after** string Optional

Identifier for the last item from the previous pagination request

**limit** integer Optional

Number of items to retrieve

**order** string Optional

Sort order of results by timestamp. Use `asc` for ascending order or `desc` for descending order.

## Returns

Returns a paginated list of [video jobs](#) for the organization.

### Example request

```
curl https://api.openai.com/v1/videos \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "data": [
    {
      "id": "video_123",
      "object": "video",
      "model": "sora-2",
      "status": "completed"
    }
  ],
  "object": "list"
}
```

# Retrieve video



```
GET https://api.openai.com/v1/videos/{video_id}
```

Retrieve a video

## Path parameters

**video\_id** string Required

The identifier of the video to retrieve.

## Returns

Returns the [video job](#) matching the provided identifier.

### Example request

```
import OpenAI from 'openai';

const client = new OpenAI();
```

```
const video = await client.videos.retrieve('video_123');

console.log(video.id);
```

# Delete video



```
DELETE https://api.openai.com/v1/videos/{video_id}
```

Delete a video

## Path parameters

**video\_id** string Required

The identifier of the video to delete.

## Returns

Returns the deleted video job metadata.

### Example request

```
import OpenAI from 'openai';

const client = new OpenAI();

const video = await client.videos.delete('video_123');

console.log(video.id);
```

## Retrieve video content



GET [https://api.openai.com/v1/videos/{video\\_id}/content](https://api.openai.com/v1/videos/{video_id}/content)

Download video content

### Path parameters

**video\_id** string Required

The identifier of the video whose media to download.

## Query parameters

**variant** string Optional

Which downloadable asset to return. Defaults to the MP4 video.

## Returns

Streams the rendered video content for the specified video job.

Example request

```
import OpenAI from 'openai';

const client = new OpenAI();

const response = await client.videos.downloadContent('video_123');

console.log(response);

const content = await response.blob();
console.log(content);
```

# Video job



Structured information describing a generated video job.

---

## **completed\_at** integer

Unix timestamp (seconds) for when the job completed, if finished.

---

## **created\_at** integer

Unix timestamp (seconds) for when the job was created.

---

## **error** object

Error payload that explains why generation failed, if applicable.

> Show properties

---

## **expires\_at** integer

Unix timestamp (seconds) for when the downloadable assets expire, if set.

---

## **id** string

Unique identifier for the video job.

---

## **model** string

The video generation model that produced the job.

---

**object** string

The object type, which is always `video`.

---

**progress** integer

Approximate completion percentage for the generation task.

---

**remixed\_from\_video\_id** string

Identifier of the source video if this video is a remix.

---

**seconds** string

Duration of the generated clip in seconds.

---

**size** string

The resolution of the generated video.

---

**status** string

Current lifecycle status of the video job.

---

# Streaming events



When you create a Response with `stream` set to `true`, the server will emit server-sent events to the client as the Response is generated. This section contains the events that are emitted by the server.

[Learn more about streaming responses.](#)

---



## response.created



An event that is emitted when a response is created.

---

**response** object

The response that was created.

> Show properties

---

**sequence\_number** integer

The sequence number for this event.

---

**type** string

The type of the event. Always `response.created`.

#### OBJECT `response.created`

```
{  
  "type": "response.created",  
  "response": {  
    "id": "resp_67ccfcdd16748190a91872c75d38539e09e4d4aac714747c",  
    "object": "response",  
    "created_at": 1741487325,  
    "status": "in_progress",  
    "error": null,  
    "incomplete_details": null,  
    "instructions": null,  
    "max_output_tokens": null,  
    "model": "gpt-4o-2024-08-06",  
    "output": [],  
    "parallel_tool_calls": true,  
    "previous_response_id": null,  
    "reasoning": {  
      "effort": null,  
      "summary": null  
    },  
    "store": true,  
    "temperature": 1,  
    "text": {  
      "format": {  
        "type": "text"  
      }  
    }  
}
```

```
    },
    "tool_choice": "auto",
    "tools": [],
    "top_p": 1,
    "truncation": "disabled",
    "usage": null,
    "user": null,
    "metadata": {}
},
"sequence_number": 1
}
```

## response.in\_progress



Emitted when the response is in progress.

**response** object

The response that is in progress.

> Show properties

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.in_progress`.

OBJECT `response.in_progress`

```
{  
  "type": "response.in_progress",  
  "response": {  
    "id": "resp_67ccfcdd16748190a91872c75d38539e09e4d4aac714747c",  
    "object": "response",  
    "created_at": 1741487325,  
    "status": "in_progress",  
    "error": null,  
    "incomplete_details": null,  
    "instructions": null,  
    "max_output_tokens": null,  
    "model": "gpt-4o-2024-08-06",  
    "output": [],  
    "parallel_tool_calls": true,  
    "previous_response_id": null,  
    "reasoning": {  
      "effort": null,  
      "summary": null  
    },  
    "store": true,  
    "temperature": 1,  
    "text": {  
      "format": {  
        "type": "string"  
      }  
    }  
  }  
}
```

```
        "type": "text"
    }
},
"tool_choice": "auto",
"tools": [],
"top_p": 1,
"truncation": "disabled",
"usage": null,
"user": null,
"metadata": {}
},
"sequence_number": 1
}
```

## response.completed



Emitted when the model response is complete.

**response** object

Properties of the completed response.

> Show properties

**sequence\_number** integer

The sequence number for this event.

**type** string

The type of the event. Always `response.completed`.

OBJECT `response.completed`

```
{  
  "type": "response.completed",  
  "response": {  
    "id": "resp_123",  
    "object": "response",  
    "created_at": 1740855869,  
    "status": "completed",  
    "error": null,  
    "incomplete_details": null,  
    "input": [],  
    "instructions": null,  
    "max_output_tokens": null,  
    "model": "gpt-4o-mini-2024-07-18",  
    "output": [  
      {  
        "id": "msg_123",  
        "type": "message",  
        "role": "assistant",  
        "content": [  
          {  
            "type": "output_text",  
            "text": "In a shimmering forest under a sky full of stars, a lonely unicorn named Lila c"}]
```

```
        "annotations": [],
    }
]
},
],
"previous_response_id": null,
"reasoning_effort": null,
"store": false,
"temperature": 1,
"text": {
    "format": {
        "type": "text"
    }
},
"tool_choice": "auto",
"tools": [],
"top_p": 1,
"truncation": "disabled",
"usage": {
    "input_tokens": 0,
    "output_tokens": 0,
    "output_tokens_details": {
        "reasoning_tokens": 0
    },
    "total_tokens": 0
},
"user": null,
"metadata": {}
},
```

```
"sequence_number": 1  
}
```

# response.failed



An event that is emitted when a response fails.

**response** object

The response that failed.

> Show properties

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.failed`.

OBJECT `response.failed`

```
{  
  "type": "response.failed",  
  "response": {  
    "id": "resp_123",  
    "object": "response",  
    "created_at": 1740855869,  
    "status": "failed",  
    "error": {  
      "code": "server_error",  
      "message": "The model failed to generate a response."  
    },  
    "incomplete_details": null,  
    "instructions": null,  
    "max_output_tokens": null,  
    "model": "gpt-4o-mini-2024-07-18",  
    "output": [],  
    "previous_response_id": null,  
    "reasoning_effort": null,  
    "store": false,  
    "temperature": 1,  
    "text": {  
      "format": {  
        "type": "text"  
      }  
    },  
    "tool_choice": "auto",  
    "tools": [],  
    "top_p": 1,  
    "truncation": "disabled",  
  },  
}
```

```
"usage": null,  
"user": null,  
"metadata": {}  
}  
}
```

## response.incomplete



An event that is emitted when a response finishes as incomplete.

**response** object

The response that was incomplete.

> Show properties

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.incomplete`.

OBJECT `response.incomplete`

```
{  
  "type": "response.incomplete",  
  "response": {  
    "id": "resp_123",  
    "object": "response",  
    "created_at": 1740855869,  
    "status": "incomplete",  
    "error": null,  
    "incomplete_details": {  
      "reason": "max_tokens"  
    },  
    "instructions": null,  
    "max_output_tokens": null,  
    "model": "gpt-4o-mini-2024-07-18",  
    "output": [],  
    "previous_response_id": null,  
    "reasoning_effort": null,  
    "store": false,  
    "temperature": 1,  
    "text": {  
      "format": {  
        "type": "text"  
      }  
    },  
    "tool_choice": "auto",  
    "tools": [],  
    "top_p": 1,  
    "truncation": "disabled",  
    "usage": null,  
  },  
}
```

```
"user": null,  
"metadata": {}  
,  
"sequence_number": 1  
}
```



## response.output\_item.added



Emitted when a new output item is added.

**item** object

The output item that was added.

> Show possible types

**output\_index** integer

The index of the output item that was added.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.output_item.added`.

OBJECT `response.output_item.added`

```
{  
  "type": "response.output_item.added",  
  "output_index": 0,  
  "item": {  
    "id": "msg_123",  
    "status": "in_progress",  
    "type": "message",  
    "role": "assistant",  
    "content": []  
  },  
  "sequence_number": 1  
}
```

## response.output\_item.done



Emitted when an output item is marked done.

**item** object

The output item that was marked done.

> Show possible types

**output\_index** integer

The index of the output item that was marked done.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.output_item.done`.

OBJECT `response.output_item.done`

```
{  
  "type": "response.output_item.done",  
  "output_index": 0,  
  "item": {  
    "id": "msg_123",  
    "status": "completed",  
    "type": "message",  
    "role": "assistant",  
    "content": [  
      {  
        "type": "output_text",  
        "text": "In a shimmering forest under a sky full of stars, a lonely unicorn named Lila disco"  
    ]  
  }  
}
```

```
        "annotations": [],
    }
]
},
"sequence_number": 1
}
```



## response.content\_part.added



Emitted when a new content part is added.



**content\_index** integer

The index of the content part that was added.

---

**item\_id** string

The ID of the output item that the content part was added to.

---

**output\_index** integer

The index of the output item that the content part was added to.

**part** object

The content part that was added.

› Show possible types

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.content_part.added`.

OBJECT `response.content_part.added`

```
{  
  "type": "response.content_part.added",  
  "item_id": "msg_123",  
  "output_index": 0,  
  "content_index": 0,  
  "part": {  
    "type": "output_text",  
    "text": "",  
    "annotations": []  
  },  
  "sequence_number": 1  
}
```

# response.content\_part.done



Emitted when a content part is done.

---

**content\_index** integer

The index of the content part that is done.

---

**item\_id** string

The ID of the output item that the content part was added to.

---

**output\_index** integer

The index of the output item that the content part was added to.

---

**part** object

The content part that is done.

> Show possible types

---

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.content_part.done`.

OBJECT `response.content_part.done`

```
{  
  "type": "response.content_part.done",  
  "item_id": "msg_123",  
  "output_index": 0,  
  "content_index": 0,  
  "sequence_number": 1,  
  "part": {  
    "type": "output_text",  
    "text": "In a shimmering forest under a sky full of stars, a lonely unicorn named Lila discovered a hidden stream flowing from a crystal-clear waterfall.",  
    "annotations": []  
  }  
}
```

## response.output\_text.delta



Emitted when there is an additional text delta.

**content\_index** integer

The index of the content part that the text delta was added to.

**delta** string

The text delta that was added.

**item\_id** string

The ID of the output item that the text delta was added to.

**logprobs** array

The log probabilities of the tokens in the delta.

> Show properties

**output\_index** integer

The index of the output item that the text delta was added to.

**sequence\_number** integer

The sequence number for this event.

**type** string

The type of the event. Always `response.output_text.delta`.

OBJECT `response.output_text.delta`

```
{  
  "type": "response.output_text.delta",  
  "item_id": "msg_123",  
  "output_index": 0,  
  "content_index": 0,  
  "delta": "In",  
  "sequence_number": 1  
}
```

## response.output\_text.done



Emitted when text content is finalized.

**content\_index** integer

The index of the content part that the text content is finalized.

**item\_id** string

The ID of the output item that the text content is finalized.

**logprobs** array

The log probabilities of the tokens in the delta.

> Show properties

---

**output\_index** integer

The index of the output item that the text content is finalized.

---

**sequence\_number** integer

The sequence number for this event.

---

**text** string

The text content that is finalized.

---

**type** string

The type of the event. Always `response.output_text.done`.

OBJECT `response.output_text.done`

```
{  
  "type": "response.output_text.done",  
  "item_id": "msg_123",  
  "output_index": 0,  
  "content_index": 0,  
  "text": "In a shimmering forest under a sky full of stars, a lonely unicorn named Lila discovered  
  "sequence_number": 1  
}
```



# response.refusal.delta



Emitted when there is a partial refusal text.

---

**content\_index** integer

The index of the content part that the refusal text is added to.

---

**delta** string

The refusal text that is added.

---

**item\_id** string

The ID of the output item that the refusal text is added to.

---

**output\_index** integer

The index of the output item that the refusal text is added to.

---

**sequence\_number** integer

The sequence number of this event.

---

**type** string

The type of the event. Always `response.refusal.delta`.

OBJECT `response.refusal.delta`

```
{  
  "type": "response.refusal.delta",  
  "item_id": "msg_123",  
  "output_index": 0,  
  "content_index": 0,  
  "delta": "refusal text so far",  
  "sequence_number": 1  
}
```

## **response.refusal.done**



Emitted when refusal text is finalized.

**content\_index** integer

The index of the content part that the refusal text is finalized.

**item\_id** string

The ID of the output item that the refusal text is finalized.

**output\_index** integer

The index of the output item that the refusal text is finalized.

**refusal** string

The refusal text that is finalized.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.refusal.done`.

OBJECT `response.refusal.done`

```
{  
  "type": "response.refusal.done",  
  "item_id": "item-abc",  
  "output_index": 1,  
  "content_index": 2,  
  "refusal": "final refusal text",  
  "sequence_number": 1  
}
```



# response.function\_call\_arguments.delta



Emitted when there is a partial function-call arguments delta.

---

**delta** string

The function-call arguments delta that is added.

---

**item\_id** string

The ID of the output item that the function-call arguments delta is added to.

---

**output\_index** integer

The index of the output item that the function-call arguments delta is added to.

---

**sequence\_number** integer

The sequence number of this event.

---

**type** string

The type of the event. Always `response.function_call_arguments.delta`.

OBJECT `response.function_call_arguments.delta`

```
{  
  "type": "response.function_call_arguments.delta",  
  "item_id": "item-abc",  
  "output_index": 0,  
  "delta": "{ \"arg\": "  
  "sequence_number": 1  
}
```

## response.function\_call\_arguments.done



Emitted when function-call arguments are finalized.

---

**arguments** string

The function-call arguments.

---

**item\_id** string

The ID of the item.

---

**name** string

The name of the function that was called.

**output\_index** integer

The index of the output item.

**sequence\_number** integer

The sequence number of this event.

**type** string

OBJECT response.function\_call\_arguments.done

```
{  
  "type": "response.function_call_arguments.done",  
  "item_id": "item-abc",  
  "name": "get_weather",  
  "output_index": 1,  
  "arguments": "{ \"arg\": 123 }",  
  "sequence_number": 1  
}
```



# response.file\_search\_call.in\_progress



Emitted when a file search call is initiated.

---

**item\_id** string

The ID of the output item that the file search call is initiated.

---

**output\_index** integer

The index of the output item that the file search call is initiated.

---

**sequence\_number** integer

The sequence number of this event.

---

**type** string

The type of the event. Always `response.file_search_call.in_progress`.

OBJECT `response.file_search_call.in_progress`

```
{  
  "type": "response.file_search_call.in_progress",  
  "output_index": 0,  
  "item_id": "fs_123",  
  "sequence_number": 1  
}
```

## response.file\_search\_call.searching



Emitted when a file search is currently searching.

---

### item\_id string

The ID of the output item that the file search call is initiated.

---

### output\_index integer

The index of the output item that the file search call is searching.

---

### sequence\_number integer

The sequence number of this event.

---

### type string

The type of the event. Always `response.file_search_call.searching`.

OBJECT `response.file_search_call.searching`

```
{  
  "type": "response.file_search_call.searching",  
  "output_index": 0,  
  "item_id": "fs_123",  
  "sequence_number": 1  
}
```

## **response.file\_search\_call.completed**



Emitted when a file search call is completed (results found).

**item\_id** string

The ID of the output item that the file search call is initiated.

**output\_index** integer

The index of the output item that the file search call is initiated.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.file_search_call.completed`.

OBJECT `response.file_search_call.completed`

```
{  
  "type": "response.file_search_call.completed",  
  "output_index": 0,  
  "item_id": "fs_123",  
  "sequence_number": 1  
}
```



## **response.web\_search\_call.in\_progress**



Emitted when a web search call is initiated.

**item\_id** string

Unique ID for the output item associated with the web search call.

---

**output\_index** integer

The index of the output item that the web search call is associated with.

---

**sequence\_number** integer

The sequence number of the web search call being processed.

---

**type** string

The type of the event. Always `response.web_search_call.in_progress`.

```
OBJECT response.web_search_call.in_progress
```

```
{  
  "type": "response.web_search_call.in_progress",  
  "output_index": 0,  
  "item_id": "ws_123",  
  "sequence_number": 0  
}
```

## response.web\_search\_call.searching



Emitted when a web search call is executing.

**item\_id** string

Unique ID for the output item associated with the web search call.

**output\_index** integer

The index of the output item that the web search call is associated with.

**sequence\_number** integer

The sequence number of the web search call being processed.

**type** string

The type of the event. Always `response.web_search_call.searching`.

OBJECT `response.web_search_call.searching`

```
{  
  "type": "response.web_search_call.searching",  
  "output_index": 0,  
  "item_id": "ws_123",  
  "sequence_number": 0  
}
```

# response.web\_search\_call.completed



Emitted when a web search call is completed.

---

**item\_id** string

Unique ID for the output item associated with the web search call.

---

**output\_index** integer

The index of the output item that the web search call is associated with.

---

**sequence\_number** integer

The sequence number of the web search call being processed.

---

**type** string

The type of the event. Always `response.web_search_call.completed`.

OBJECT `response.web_search_call.completed`

```
{  
  "type": "response.web_search_call.completed",  
  "output_index": 0,  
  "item_id": "ws_123",  
  "sequence_number": 0  
}
```



## response.reasoning\_summary\_part.added



Emitted when a new reasoning summary part is added.

**item\_id** string

The ID of the item this summary part is associated with.

**output\_index** integer

The index of the output item this summary part is associated with.

**part** object

The summary part that was added.

> Show properties

---

**sequence\_number** integer

The sequence number of this event.

---

**summary\_index** integer

The index of the summary part within the reasoning summary.

---

**type** string

The type of the event. Always `response.reasoning_summary_part.added`.

OBJECT `response.reasoning_summary_part.added`

```
{  
  "type": "response.reasoning_summary_part.added",  
  "item_id": "rs_6806bfca0b2481918a5748308061a2600d3ce51bdffd5476",  
  "output_index": 0,  
  "summary_index": 0,  
  "part": {  
    "type": "summary_text",  
    "text": ""  
  },  
  "sequence_number": 1  
}
```

# response.reasoning\_summary\_part.done



Emitted when a reasoning summary part is completed.

---

## **item\_id** string

The ID of the item this summary part is associated with.

---

## **output\_index** integer

The index of the output item this summary part is associated with.

---

## **part** object

The completed summary part.

> Show properties

---

## **sequence\_number** integer

The sequence number of this event.

---

## **summary\_index** integer

The index of the summary part within the reasoning summary.

---

## **type** string

The type of the event. Always `response.reasoning_summary_part.done`.

---

OBJECT response.reasoning\_summary\_part.done

```
{  
  "type": "response.reasoning_summary_part.done",  
  "item_id": "rs_6806bfca0b2481918a5748308061a2600d3ce51bdffd5476",  
  "output_index": 0,  
  "summary_index": 0,  
  "part": {  
    "type": "summary_text",  
    "text": "**Responding to a greeting**\n\nThe user just said, \"Hello!\" So, it seems I need to e  
},  
  "sequence_number": 1  
}
```



## response.reasoning\_summary\_text.delta



Emitted when a delta is added to a reasoning summary text.

**delta** string

The text delta that was added to the summary.

---

**item\_id** string

The ID of the item this summary text delta is associated with.

---

**output\_index** integer

The index of the output item this summary text delta is associated with.

---

**sequence\_number** integer

The sequence number of this event.

---

**summary\_index** integer

The index of the summary part within the reasoning summary.

---

**type** string

The type of the event. Always `response.reasoning_summary_text.delta`.

OBJECT `response.reasoning_summary_text.delta`

```
{  
  "type": "response.reasoning_summary_text.delta",  
  "item_id": "rs_6806bfca0b2481918a5748308061a2600d3ce51bdffd5476",  
  "output_index": 0,  
  "summary_index": 0,  
  "delta": "***Responding to a greeting**\n\nThe user just said, \"Hello!\" So, it seems I need to e
```

```
"sequence_number": 1  
}
```

## response.reasoning\_summary\_text.done



Emitted when a reasoning summary text is completed.

---

### item\_id string

The ID of the item this summary text is associated with.

---

### output\_index integer

The index of the output item this summary text is associated with.

---

### sequence\_number integer

The sequence number of this event.

---

### summary\_index integer

The index of the summary part within the reasoning summary.

---

### text string

The full text of the completed reasoning summary.

**type** string

The type of the event. Always `response.reasoning_summary_text.done`.

OBJECT `response.reasoning_summary_text.done`

```
{  
  "type": "response.reasoning_summary_text.done",  
  "item_id": "rs_6806bfca0b2481918a5748308061a2600d3ce51bdffd5476",  
  "output_index": 0,  
  "summary_index": 0,  
  "text": "**Responding to a greeting**\n\nThe user just said, \"Hello!\" So, it seems I need to eng  
  "sequence_number": 1  
}
```



## response.reasoning\_text.delta



Emitted when a delta is added to a reasoning text.

**content\_index** integer

The index of the reasoning content part this delta is associated with.

**delta** string

The text delta that was added to the reasoning content.

**item\_id** string

The ID of the item this reasoning text delta is associated with.

**output\_index** integer

The index of the output item this reasoning text delta is associated with.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always `response.reasoning_text.delta`.

OBJECT `response.reasoning_text.delta`

{

```
"type": "response.reasoning_text.delta",
"item_id": "rs_123",
"output_index": 0,
"content_index": 0,
```

```
"delta": "The",
"sequence_number": 1
}
```

## response.reasoning\_text.done



Emitted when a reasoning text is completed.

---

**content\_index** integer

The index of the reasoning content part.

---

**item\_id** string

The ID of the item this reasoning text is associated with.

---

**output\_index** integer

The index of the output item this reasoning text is associated with.

---

**sequence\_number** integer

The sequence number of this event.

**text** string

The full text of the completed reasoning content.

**type** string

The type of the event. Always `response.reasoning_text.done`.

OBJECT `response.reasoning_text.done`

```
{  
  "type": "response.reasoning_text.done",  
  "item_id": "rs_123",  
  "output_index": 0,  
  "content_index": 0,  
  "text": "The user is asking...",  
  "sequence_number": 4  
}
```



## response.image\_generation\_call.completed



Emitted when an image generation tool call has completed and the final image is available.

**item\_id** string

The unique identifier of the image generation item being processed.

**output\_index** integer

The index of the output item in the response's output array.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.image\_generation\_call.completed'.

OBJECT response.image\_generation\_call.completed

```
{  
  "type": "response.image_generation_call.completed",  
  "output_index": 0,  
  "item_id": "item-123",  
  "sequence_number": 1  
}
```

# response.image\_generation\_call.generating



Emitted when an image generation tool call is actively generating an image (intermediate state).

---

**item\_id** string

The unique identifier of the image generation item being processed.

---

**output\_index** integer

The index of the output item in the response's output array.

---

**sequence\_number** integer

The sequence number of the image generation item being processed.

---

**type** string

The type of the event. Always 'response.image\_generation\_call.generating'.

```
OBJECT response.image_generation_call.generating
```

```
{  
  "type": "response.image_generation_call.generating",  
  "output_index": 0,  
  "item_id": "item-123",  
  "sequence_number": 0  
}
```

## response.image\_generation\_call.in\_progress



Emitted when an image generation tool call is in progress.

---

### item\_id string

The unique identifier of the image generation item being processed.

---

### output\_index integer

The index of the output item in the response's output array.

---

### sequence\_number integer

The sequence number of the image generation item being processed.

---

### type string

The type of the event. Always 'response.image\_generation\_call.in\_progress'.

```
OBJECT response.image_generation_call.in_progress
```

```
{  
  "type": "response.image_generation_call.in_progress",  
  "output_index": 0,  
  "item_id": "item-123",  
  "sequence_number": 0  
}
```

## response.image\_generation\_call.partial\_image



Emitted when a partial image is available during image generation streaming.

**item\_id** string

The unique identifier of the image generation item being processed.

**output\_index** integer

The index of the output item in the response's output array.

**partial\_image\_b64** string

Base64-encoded partial image data, suitable for rendering as an image.

---

**partial\_image\_index** integer

0-based index for the partial image (backend is 1-based, but this is 0-based for the user).

---

**sequence\_number** integer

The sequence number of the image generation item being processed.

---

**type** string

The type of the event. Always 'response.image\_generation\_call.partial\_image'.

```
OBJECT response.image_generation_call.partial_image
```

```
{  
  "type": "response.image_generation_call.partial_image",  
  "output_index": 0,  
  "item_id": "item-123",  
  "sequence_number": 0,  
  "partial_image_index": 0,  
  "partial_image_b64": "..."  
}
```



# response.mcp\_call\_arguments.delta



Emitted when there is a delta (partial update) to the arguments of an MCP tool call.

---

**delta** string

A JSON string containing the partial update to the arguments for the MCP tool call.

---

**item\_id** string

The unique identifier of the MCP tool call item being processed.

---

**output\_index** integer

The index of the output item in the response's output array.

---

**sequence\_number** integer

The sequence number of this event.

---

**type** string

The type of the event. Always 'response.mcp\_call\_arguments.delta'.

```
OBJECT response.mcp_call_arguments.delta
```

{

```
  "type": "response.mcp_call_arguments.delta",
  "output_index": 0,
```

```
"item_id": "item-abc",
"delta": "{}",
"sequence_number": 1
}
```

## response.mcp\_call\_arguments.done



Emitted when the arguments for an MCP tool call are finalized.

---

**arguments** string

A JSON string containing the finalized arguments for the MCP tool call.

---

**item\_id** string

The unique identifier of the MCP tool call item being processed.

---

**output\_index** integer

The index of the output item in the response's output array.

---

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.mcp\_call\_arguments.done'.

OBJECT response.mcp\_call\_arguments.done

```
{  
  "type": "response.mcp_call_arguments.done",  
  "output_index": 0,  
  "item_id": "item-abc",  
  "arguments": "{\"arg1\": \"value1\", \"arg2\": \"value2\"}",  
  "sequence_number": 1  
}
```



## response.mcp\_call.completed



Emitted when an MCP tool call has completed successfully.

**item\_id** string

The ID of the MCP tool call item that completed.

**output\_index** integer

The index of the output item that completed.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.mcp\_call.completed'.

OBJECT response.mcp\_call.completed

```
{  
  "type": "response.mcp_call.completed",  
  "sequence_number": 1,  
  "item_id": "mcp_682d437d90a88191bf88cd03aae0c3e503937d5f622d7a90",  
  "output_index": 0  
}
```

## response.mcp\_call.failed



Emitted when an MCP tool call has failed.

**item\_id** string

The ID of the MCP tool call item that failed.

**output\_index** integer

The index of the output item that failed.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.mcp\_call.failed'.

```
OBJECT response.mcp_call.failed
{
  "type": "response.mcp_call.failed",
  "sequence_number": 1,
  "item_id": "mcp_682d437d90a88191bf88cd03aae0c3e503937d5f622d7a90",
  "output_index": 0
}
```

# response.mcp\_call.in\_progress



Emitted when an MCP tool call is in progress.

---

**item\_id** string

The unique identifier of the MCP tool call item being processed.

---

**output\_index** integer

The index of the output item in the response's output array.

---

**sequence\_number** integer

The sequence number of this event.

---

**type** string

The type of the event. Always 'response.mcp\_call.in\_progress'.

OBJECT response.mcp\_call.in\_progress

```
{  
  "type": "response.mcp_call.in_progress",  
  "sequence_number": 1,  
  "output_index": 0,  
  "item_id": "mcp_682d437d90a88191bf88cd03aae0c3e503937d5f622d7a90"  
}
```



## response.mcp\_list\_tools.completed



Emitted when the list of available MCP tools has been successfully retrieved.

**item\_id** string

The ID of the MCP tool call item that produced this output.

**output\_index** integer

The index of the output item that was processed.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.mcp\_list\_tools.completed'.

```
OBJECT response.mcp_list_tools.completed
```

```
{  
  "type": "response.mcp_list_tools.completed",  
  "sequence_number": 1,  
  "output_index": 0,  
  "item_id": "mcp1_682d4379df088191886b70f4ec39f90403937d5f622d7a90"  
}
```

## response.mcp\_list\_tools.failed



Emitted when the attempt to list available MCP tools has failed.

**item\_id** string

The ID of the MCP tool call item that failed.

**output\_index** integer

The index of the output item that failed.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.mcp\_list\_tools.failed'.

OBJECT response.mcp\_list\_tools.failed

```
{  
  "type": "response.mcp_list_tools.failed",  
  "sequence_number": 1,  
  "output_index": 0,  
  "item_id": "mcpl_682d4379df088191886b70f4ec39f90403937d5f622d7a90"  
}
```

## response.mcp\_list\_tools.in\_progress



Emitted when the system is in the process of retrieving the list of available MCP tools.

**item\_id** string

The ID of the MCP tool call item that is being processed.

**output\_index** integer

The index of the output item that is being processed.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.mcp\_list\_tools.in\_progress'.

OBJECT response.mcp\_list\_tools.in\_progress

```
{  
  "type": "response.mcp_list_tools.in_progress",  
  "sequence_number": 1,  
  "output_index": 0,  
  "item_id": "mcpl_682d4379df088191886b70f4ec39f90403937d5f622d7a90"  
}
```



# response.code\_interpreter\_call.in\_progress



Emitted when a code interpreter call is in progress.

---

**item\_id** string

The unique identifier of the code interpreter tool call item.

---

**output\_index** integer

The index of the output item in the response for which the code interpreter call is in progress.

---

**sequence\_number** integer

The sequence number of this event, used to order streaming events.

---

**type** string

The type of the event. Always `response.code_interpreter_call.in_progress`.

OBJECT `response.code_interpreter_call.in_progress`

```
{  
  "type": "response.code_interpreter_call.in_progress",  
  "output_index": 0,  
  "item_id": "ci_12345",  
  "sequence_number": 1  
}
```

## response.code\_interpreter\_call.interpreting



Emitted when the code interpreter is actively interpreting the code snippet.

---

### item\_id string

The unique identifier of the code interpreter tool call item.

---

### output\_index integer

The index of the output item in the response for which the code interpreter is interpreting code.

---

### sequence\_number integer

The sequence number of this event, used to order streaming events.

---

### type string

The type of the event. Always `response.code_interpreter_call.interpreting`.

OBJECT `response.code_interpreter_call.interpreting`

```
{  
  "type": "response.code_interpreter_call.interpreting",  
  "output_index": 4,  
  "item_id": "ci_12345",  
  "sequence_number": 1  
}
```

## **response.code\_interpreter\_call.completed**



Emitted when the code interpreter call is completed.

**item\_id** string

The unique identifier of the code interpreter tool call item.

**output\_index** integer

The index of the output item in the response for which the code interpreter call is completed.

**sequence\_number** integer

The sequence number of this event, used to order streaming events.

**type** string

The type of the event. Always `response.code_interpreter_call.completed`.

OBJECT `response.code_interpreter_call.completed`

```
{  
  "type": "response.code_interpreter_call.completed",  
  "output_index": 5,  
  "item_id": "ci_12345",  
  "sequence_number": 1  
}
```



## **response.code\_interpreter\_call\_code.delta**



Emitted when a partial code snippet is streamed by the code interpreter.

**delta** string

The partial code snippet being streamed by the code interpreter.

---

**item\_id** string

The unique identifier of the code interpreter tool call item.

---

**output\_index** integer

The index of the output item in the response for which the code is being streamed.

---

**sequence\_number** integer

The sequence number of this event, used to order streaming events.

---

**type** string

The type of the event. Always `response.code_interpreter_call_code.delta`.

---

OBJECT `response.code_interpreter_call_code.delta`

```
{  
  "type": "response.code_interpreter_call_code.delta",  
  "output_index": 0,  
  "item_id": "ci_12345",  
  "delta": "print('Hello, world')",  
  "sequence_number": 1  
}
```

# response.code\_interpreter\_call\_code.done



Emitted when the code snippet is finalized by the code interpreter.

---

**code** string

The final code snippet output by the code interpreter.

---

**item\_id** string

The unique identifier of the code interpreter tool call item.

---

**output\_index** integer

The index of the output item in the response for which the code is finalized.

---

**sequence\_number** integer

The sequence number of this event, used to order streaming events.

---

**type** string

The type of the event. Always `response.code_interpreter_call_code.done`.

```
OBJECT response.code_interpreter_call_code.done
```

{

```
  "type": "response.code_interpreter_call_code.done",
  "output_index": 3,
```

```
"item_id": "ci_12345",
"code": "print('done')",
"sequence_number": 1
}
```



## response.output\_text.annotation.added



Emitted when an annotation is added to output text content.

---

**annotation** object

The annotation object being added. (See annotation schema for details.)

---

**annotation\_index** integer

The index of the annotation within the content part.

---

**content\_index** integer

The index of the content part within the output item.

**item\_id** string

The unique identifier of the item to which the annotation is being added.

**output\_index** integer

The index of the output item in the response's output array.

**sequence\_number** integer

The sequence number of this event.

**type** string

The type of the event. Always 'response.output\_text.annotation.added'.

OBJECT response.output\_text.annotation.added

```
{  
  "type": "response.output_text.annotation.added",  
  "item_id": "item-abc",  
  "output_index": 0,  
  "content_index": 0,  
  "annotation_index": 0,  
  "annotation": {  
    "type": "text_annotation",  
    "text": "This is a test annotation",  
    "start": 0,  
    "end": 10  
  },  
}
```

```
"sequence_number": 1
```

## response.queued



Emitted when a response is queued and waiting to be processed.

**response** object

The full response object that is queued.

> Show properties

**sequence\_number** integer

The sequence number for this event.

**type** string

The type of the event. Always 'response.queued'.

OBJECT response.queued

```
{  
  "type": "response.queued",  
  "response": {  
    "id": "res_123",
```

```
"status": "queued",
"created_at": "2021-01-01T00:00:00Z",
"updated_at": "2021-01-01T00:00:00Z"
},
"sequence_number": 1
}
```



## response.custom\_tool\_call\_input.delta



Event representing a delta (partial update) to the input of a custom tool call.

**delta** string

The incremental input data (delta) for the custom tool call.

**item\_id** string

Unique identifier for the API item associated with this event.

**output\_index** integer

The index of the output this delta applies to.

**sequence\_number** integer

The sequence number of this event.

**type** string

The event type identifier.

```
OBJECT response.custom_tool_call_input.delta
{
  "type": "response.custom_tool_call_input.delta",
  "output_index": 0,
  "item_id": "ctc_1234567890abcdef",
  "delta": "partial input text"
}
```

## response.custom\_tool\_call\_input.done



Event indicating that input for a custom tool call is complete.

**input** string

The complete input data for the custom tool call.

**item\_id** string

Unique identifier for the API item associated with this event.

**output\_index** integer

The index of the output this event applies to.

**sequence\_number** integer

The sequence number of this event.

**type** string

The event type identifier.

```
OBJECT response.custom_tool_call_input.done

{
  "type": "response.custom_tool_call_input.done",
  "output_index": 0,
  "item_id": "ctc_1234567890abcdef",
  "input": "final complete input text"
}
```

# error



Emitted when an error occurs.

---

**code** string

The error code.

---

**message** string

The error message.

---

**param** string

The error parameter.

---

**sequence\_number** integer

The sequence number of this event.

---

**type** string

The type of the event. Always `error`.

OBJECT `error`

{

`"type": "error",`  
  `"code": "ERR_SOMETHING",`

```
"message": "Something went wrong",
"param": null,
"sequence_number": 1
}
```

## Webhook Events



Webhooks are HTTP requests sent by OpenAI to a URL you specify when certain events happen during the course of API usage.

[Learn more about webhooks.](#)



### response.completed



Sent when a background response has been completed.

**created\_at** integer

The Unix timestamp (in seconds) of when the model response was completed.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `response.completed`.

OBJECT `response.completed`

```
{  
  "id": "evt_abc123",  
  "type": "response.completed",  
  "created_at": 1719168000,  
  "data": {  
    "id": "resp_abc123"  
  }  
}
```

# response.cancelled



Sent when a background response has been cancelled.

**created\_at** integer

The Unix timestamp (in seconds) of when the model response was cancelled.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `response.cancelled`.

OBJECT `response.cancelled`

```
{  
  "id": "evt_abc123",  
  "type": "response.cancelled",  
  "created_at": 1719168000,  
  "data": {  
    "id": "resp_abc123"  
  }  
}
```

## response.failed



Sent when a background response has failed.

**created\_at** integer

The Unix timestamp (in seconds) of when the model response failed.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `response.failed`.

OBJECT `response.failed`

```
{  
  "id": "evt_abc123",  
  "type": "response.failed",  
  "created_at": 1719168000,  
  "data": {  
    "id": "resp_abc123"  
  }  
}
```

## response.incomplete



Sent when a background response has been interrupted.

**created\_at** integer

The Unix timestamp (in seconds) of when the model response was interrupted.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `response.incomplete`.

OBJECT `response.incomplete`

```
{  
  "id": "evt_abc123",  
  "type": "response.incomplete",  
  "created_at": 1719168000,  
  "data": {  
    "id": "resp_abc123"  
  }  
}
```



# batch.completed



Sent when a batch API request has been completed.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the batch API request was completed.

---

**data** object

Event data payload.

> Show properties

---

**id** string

The unique ID of the event.

---

**object** string

The object of the event. Always `event`.

---

**type** string

The type of the event. Always `batch.completed`.

---

OBJECT `batch.completed`

{

`"id": "evt_abc123",`

```
"type": "batch.completed",
"created_at": 1719168000,
"data": {
  "id": "batch_abc123"
}
```

## batch.cancelled



Sent when a batch API request has been cancelled.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the batch API request was cancelled.

---

**data** object

Event data payload.

> Show properties

---

**id** string

The unique ID of the event.

---

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `batch.cancelled`.

OBJECT `batch.cancelled`

```
{  
  "id": "evt_abc123",  
  "type": "batch.cancelled",  
  "created_at": 1719168000,  
  "data": {  
    "id": "batch_abc123"  
  }  
}
```

## batch.expired



Sent when a batch API request has expired.

**created\_at** integer

The Unix timestamp (in seconds) of when the batch API request expired.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `batch.expired`.

OBJECT `batch.expired`

```
{  
  "id": "evt_abc123",  
  "type": "batch.expired",  
  "created_at": 1719168000,  
  "data": {  
    "id": "batch_abc123"  
  }  
}
```

# batch.failed



Sent when a batch API request has failed.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the batch API request failed.

---

**data** object

Event data payload.

> Show properties

---

**id** string

The unique ID of the event.

---

**object** string

The object of the event. Always `event`.

---

**type** string

The type of the event. Always `batch.failed`.

OBJECT `batch.failed`

{

`"id": "evt_abc123",`

```
"type": "batch.failed",
"created_at": 1719168000,
"data": {
  "id": "batch_abc123"
}
```



## fine\_tuning.job.succeeded



Sent when a fine-tuning job has succeeded.

**created\_at** integer

The Unix timestamp (in seconds) of when the fine-tuning job succeeded.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `fine_tuning.job.succeeded`.

OBJECT `fine_tuning.job.succeeded`

```
{  
  "id": "evt_abc123",  
  "type": "fine_tuning.job.succeeded",  
  "created_at": 1719168000,  
  "data": {  
    "id": "ftjob_abc123"  
  }  
}
```

## **fine\_tuning.job.failed**



Sent when a fine-tuning job has failed.

**created\_at** integer

The Unix timestamp (in seconds) of when the fine-tuning job failed.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `fine_tuning.job.failed`.

OBJECT `fine_tuning.job.failed`

{

```
"id": "evt_abc123",
"type": "fine_tuning.job.failed",
"created_at": 1719168000,
"data": {
  "id": "ftjob_abc123"
```

{  
}

# **fine\_tuning.job.cancelled**



Sent when a fine-tuning job has been cancelled.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the fine-tuning job was cancelled.

---

**data** object

Event data payload.

> Show properties

---

**id** string

The unique ID of the event.

---

**object** string

The object of the event. Always `event`.

---

**type** string

The type of the event. Always `fine_tuning.job.cancelled`.

#### OBJECT `fine_tuning.job.cancelled`

```
{  
  "id": "evt_abc123",  
  "type": "fine_tuning.job.cancelled",  
  "created_at": 1719168000,  
  "data": {  
    "id": "ftjob_abc123"  
  }  
}
```



## eval.run.succeeded



Sent when an eval run has succeeded.

**created\_at** integer

The Unix timestamp (in seconds) of when the eval run succeeded.

---

**data** object

Event data payload.

> Show properties

---

**id** string

The unique ID of the event.

---

**object** string

The object of the event. Always `event`.

---

**type** string

The type of the event. Always `eval.run.succeeded`.

---

OBJECT `eval.run.succeeded`

```
{  
  "id": "evt_abc123",  
  "type": "eval.run.succeeded",  
  "created_at": 1719168000,  
  "data": {  
    "id": "evalrun_abc123"  
  }  
}
```

# eval.run.failed



Sent when an eval run has failed.

---

## **created\_at** integer

The Unix timestamp (in seconds) of when the eval run failed.

---

## **data** object

Event data payload.

[> Show properties](#)

---

## **id** string

The unique ID of the event.

---

## **object** string

The object of the event. Always `event`.

---

## **type** string

The type of the event. Always `eval.run.failed`.

---

**OBJECT eval.run.failed**

```
{  
  "id": "evt_abc123",  
  "type": "eval.run.failed",  
  "created_at": 1719168000,  
  "data": {  
    "id": "evalrun_abc123"  
  }  
}
```

## eval.run.canceled



Sent when an eval run has been canceled.

**created\_at** integer

The Unix timestamp (in seconds) of when the eval run was canceled.

**data** object

Event data payload.

> Show properties

**id** string

The unique ID of the event.

**object** string

The object of the event. Always `event`.

**type** string

The type of the event. Always `eval.run.canceled`.

OBJECT `eval.run.canceled`

```
{  
  "id": "evt_abc123",  
  "type": "eval.run.canceled",  
  "created_at": 1719168000,  
  "data": {  
    "id": "evalrun_abc123"  
  }  
}
```



# realtime.call.incoming



Sent when Realtime API Receives a incoming SIP call.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the model response was completed.

---

**data** object

Event data payload.

> Show properties

---

**id** string

The unique ID of the event.

---

**object** string

The object of the event. Always `event`.

---

**type** string

The type of the event. Always `realtime.call.incoming`.

OBJECT `realtime.call.incoming`

{

`"id": "evt_abc123",`

```
"type": "realtime.call.incoming",
"created_at": 1719168000,
"data": {
  "call_id": "rtc_479a275623b54bdb9b6fbae2f7cbd408",
  "sip_headers": [
    {"name": "Max-Forwards", "value": "63"},
    {"name": "CSeq", "value": "851287 INVITE"},
    {"name": "Content-Type", "value": "application/sdp"}
  ]
}
}
```

# Audio



Learn how to turn audio into text or text into audio.

Related guide: [Speech to text](#)

## Create speech



POST <https://api.openai.com/v1/audio/speech>

Generates audio from the input text.

## Request body

---

### **input** string Required

The text to generate audio for. The maximum length is 4096 characters.

---

### **model** string Required

One of the available TTS models: `tts-1` , `tts-1-hd` or `gpt-4o-mini-tts` .

---

### **voice** string Required

The voice to use when generating the audio. Supported voices are `alloy` , `ash` , `ballad` , `coral` , `echo` , `fable` , `onyx` , `nova` , `sage` , `shimmer` , and `verse` . Previews of the voices are available in the Text to speech guide.

---

### **instructions** string Optional

Control the voice of your generated audio with additional instructions. Does not work with `tts-1` or `tts-1-hd` .

---

### **response\_format** string Optional Defaults to mp3

The format to audio in. Supported formats are `mp3` , `opus` , `aac` , `flac` , `wav` , and `pcm` .

---

### **speed** number Optional Defaults to 1

The speed of the generated audio. Select a value from `0.25` to `4.0`. `1.0` is the default.

**stream\_format** string Optional Defaults to audio

The format to stream the audio in. Supported formats are `sse` and `audio`. `sse` is not supported for `tts-1` or `tts-1-hd`.

## Returns

The audio file content or a [stream of audio events](#).

**Default SSE Stream Format**

Example request

```
curl https://api.openai.com/v1/audio/speech \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "model": "gpt-4o-mini-tts",
  "input": "The quick brown fox jumped over the lazy dog.",
  "voice": "alloy"
}' \
--output speech.mp3
```

# Create transcription



POST <https://api.openai.com/v1/audio/transcriptions>

Transcribes audio into the input language.

## Request body

### **file** file Required

The audio file object (not file name) to transcribe, in one of these formats: flac, mp3, mp4, mpeg, mpg, m4a, ogg, wav, or webm.

### **model** string Required

ID of the model to use. The options are `gpt-4o-transcribe` , `gpt-4o-mini-transcribe` , `whisper-1` (which is powered by our open source Whisper V2 model), and `gpt-4o-transcribe-diarize` .

### **chunking\_strategy** "auto" or object Optional

Controls how the audio is cut into chunks. When set to `"auto"` , the server first normalizes loudness and then uses voice activity detection (VAD) to choose boundaries. `server_vad` object can be provided to tweak VAD detection parameters manually. If unset, the audio is transcribed as a single block. Required when using `gpt-4o-transcribe-diarize` for inputs longer than 30 seconds.

> Show possible types

### **include** array Optional

Additional information to include in the transcription response. `logprobs` will return the log probabilities of the tokens in the response to understand the model's confidence in the transcription. `logprobs` only works with `response_format` set to `json` and only with the models `gpt-4o-transcribe` and `gpt-4o-mini-transcribe`. This field is not supported when using `gpt-4o-transcribe-diarize`.

---

**known\_speaker\_names** array Optional

Optional list of speaker names that correspond to the audio samples provided in `known_speaker_references[]`. Each entry should be a short identifier (for example `customer` or `agent`). Up to 4 speakers are supported.

---

**known\_speaker\_references** array Optional

Optional list of audio samples (as data URLs) that contain known speaker references matching `known_speaker_names[]`. Each sample must be between 2 and 10 seconds, and can use any of the same input audio formats supported by `file`.

---

**language** string Optional

The language of the input audio. Supplying the input language in ISO-639-1 (e.g. `en`) format will improve accuracy and latency.

---

**prompt** string Optional

An optional text to guide the model's style or continue a previous audio segment. The `prompt` should match the audio language. This field is not supported when using `gpt-4o-transcribe-diarize`.

---

**response\_format** string Optional Defaults to json

The format of the output, in one of these options: `json`, `text`, `srt`, `verbose_json`, `vtt`, or `diarized_json`. For `gpt-4o-transcribe` and `gpt-4o-mini-transcribe`, the only supported format is `json`. For

`gpt-4o-transcribe-diarize`, the supported formats are `json`, `text`, and `diarized_json`, with `diarized_json` required to receive speaker annotations.

---

**stream** boolean Optional Defaults to false

If set to true, the model response data will be streamed to the client as it is generated using [server-sent events](#). See the [Streaming section of the Speech-to-Text guide](#) for more information.

Note: Streaming is not supported for the `whisper-1` model and will be ignored.

---

**temperature** number Optional Defaults to 0

The sampling temperature, between 0 and 1. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic. If set to 0, the model will use [log\\_probability](#) to automatically increase the temperature until certain thresholds are hit.

---

**timestamp\_granularities** array Optional Defaults to segment

The timestamp granularities to populate for this transcription. `response_format` must be set `verbose_json` to use timestamp granularities. Either or both of these options are supported: `word`, or `segment`. Note: There is no additional latency for segment timestamps, but generating word timestamps incurs additional latency. This option is not available for `gpt-4o-transcribe-diarize`.

## Returns

---

The [transcription object](#), a [diarized transcription object](#), a [verbose transcription object](#), or a [stream of transcript events](#).

**Default****Diarization****Streaming****Logprobs****Word timestamps****Segment timestamps**

## Example request

```
curl https://api.openai.com/v1/audio/transcriptions \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: multipart/form-data" \
-F file="@/path/to/file/audio.mp3" \
-F model="gpt-4o-transcribe"
```

## Response

```
{
  "text": "Imagine the wildest idea that you've ever had, and you're curious about how it might scal",
  "usage": {
    "type": "tokens",
    "input_tokens": 14,
    "input_token_details": {
      "text_tokens": 0,
      "audio_tokens": 14
    },
    "output_tokens": 45,
    "total_tokens": 59
  }
}
```

# Create translation



```
POST https://api.openai.com/v1/audio/translations
```

Translates audio into English.

## Request body

---

**file** file Required

The audio file object (not file name) translate, in one of these formats: flac, mp3, mp4, mpeg, mpg, m4a, ogg, wav, or webm.

---

**model** string or "whisper-1" Required

ID of the model to use. Only `whisper-1` (which is powered by our open source Whisper V2 model) is currently available.

---

**prompt** string Optional

An optional text to guide the model's style or continue a previous audio segment. The prompt should be in English.

---

**response\_format** string Optional Defaults to json

The format of the output, in one of these options: `json`, `text`, `srt`, `verbose_json`, or `vtt`.

**temperature** number Optional Defaults to 0

The sampling temperature, between 0 and 1. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic. If set to 0, the model will use [log\\_probability](#) to automatically increase the temperature until certain thresholds are hit.

## Returns

The translated text.

### Example request

```
curl https://api.openai.com/v1/audio/translations \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: multipart/form-data" \
-F file="@/path/to/file/german.m4a" \
-F model="whisper-1"
```

### Response

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

# The transcription object (JSON)



Represents a transcription response returned by model, based on the provided input.

## logprobs array

The log probabilities of the tokens in the transcription. Only returned with the models `gpt-4o-transcribe` and `gpt-4o-mini-transcribe` if `logprobs` is added to the `include` array.

> Show properties

## text string

The transcribed text.

## usage object

Token usage statistics for the request.

> Show possible types

## OBJECT The transcription object (JSON)

```
{  
  "text": "Imagine the wildest idea that you've ever had, and you're curious about how it might sca  
  "usage": {  
    "type": "tokens",  
    "input_tokens": 14,  
    "input_token_details": {  
      "text_tokens": 10,  
      "model_tokens": 4,  
      "other_tokens": 10  
    }  
  }  
}
```

```
    "audio_tokens": 4
  },
  "output_tokens": 101,
  "total_tokens": 115
}
}
```

# The transcription object (Diarized JSON)



Represents a diarized transcription response returned by the model, including the combined transcript and speaker-segment annotations.

**duration** number

Duration of the input audio in seconds.

**segments** array

Segments of the transcript annotated with timestamps and speaker labels.

> Show properties

**task** string

The type of task that was run. Always `transcribe`.

**text** string

The concatenated transcript text for the entire audio input.

**usage** object

Token or duration usage statistics for the request.

> Show possible types

OBJECT The transcription object (Diarized JSON)

```
{  
  "task": "transcribe",  
  "duration": 42.7,  
  "text": "Agent: Thanks for calling OpenAI support.\nCustomer: Hi, I need help with diarization.",  
  "segments": [  
    {  
      "type": "transcript.text.segment",  
      "id": "seg_001",  
      "start": 0.0,  
      "end": 5.2,  
      "text": "Thanks for calling OpenAI support.",  
      "speaker": "agent"  
    },  
    {  
      "type": "transcript.text.segment",  
      "id": "seg_002",  
      "start": 5.2,  
      "end": 12.8,  
      "text": "Hi, I need help with diarization."  
    }  
  ]  
}
```

```
        "speaker": "A"  
    }  
,  
  "usage": {  
    "type": "duration",  
    "seconds": 43  
  }  
}
```

# The transcription object (Verbose JSON)



Represents a verbose json transcription response returned by model, based on the provided input.

**duration** number

The duration of the input audio.

**language** string

The language of the input audio.

**segments** array

Segments of the transcribed text and their corresponding details.

> Show properties

---

**text** string

The transcribed text.

---

**usage** object

Usage statistics for models billed by audio input duration.

> Show properties

---

**words** array

Extracted words and their corresponding timestamps.

> Show properties

OBJECT The transcription object (Verbose JSON)

```
{  
  "task": "transcribe",  
  "language": "english",  
  "duration": 8.470000267028809,  
  "text": "The beach was a popular spot on a hot summer day. People were swimming in the ocean, buil  
  "segments": [  
    {  
      "id": 0,  
      "seek": 0,  
      "start": 0.0,  
      "end": 3.319999933242798,  
      "text": " The beach was a popular spot on a hot summer day.",  
      "tokens": [  
        {  
          "start": 0.0,  
          "end": 3.319999933242798,  
          "text": "The beach was a popular spot on a hot summer day."  
        }  
      ]  
    }  
  ]  
}
```

```
      50364, 440, 7534, 390, 257, 3743, 4008, 322, 257, 2368, 4266, 786, 13, 50530
    ],
    "temperature": 0.0,
    "avg_logprob": -0.2860786020755768,
    "compression_ratio": 1.2363636493682861,
    "no_speech_prob": 0.00985979475080967
  },
  ...
],
"usage": {
  "type": "duration",
  "seconds": 9
}
}
```

## Stream Event (speech.audio.delta)



Emitted for each chunk of audio data generated during speech synthesis.

**audio** string

A chunk of Base64-encoded audio data.

**type** string

The type of the event. Always `speech.audio.delta`.

OBJECT Stream Event (`speech.audio.delta`)

```
{  
  "type": "speech.audio.delta",  
  "audio": "base64-encoded-audio-data"  
}
```

## Stream Event (`speech.audio.done`)



Emitted when the speech synthesis is complete and all audio has been streamed.

**type** string

The type of the event. Always `speech.audio.done`.

**usage** object

Token usage statistics for the request.

&gt; Show properties

OBJECT Stream Event (speech.audio.done)

```
{  
  "type": "speech.audio.done",  
  "usage": {  
    "input_tokens": 14,  
    "output_tokens": 101,  
    "total_tokens": 115  
  }  
}
```

## Stream Event (transcript.text.delta)



Emitted when there is an additional text delta. This is also the first event emitted when the transcription starts.

Only emitted when you [create a transcription](#) with the `Stream` parameter set to `true`.

**delta** string

The text delta that was additionally transcribed.

**logprobs** array

The log probabilities of the delta. Only included if you [create a transcription](#) with the `include[]` parameter set to `logprobs`.

> Show properties

**segment\_id** string

Identifier of the diarized segment that this delta belongs to. Only present when using `gpt-4o-transcribe-diarize`.

**type** string

The type of the event. Always `transcript.text.delta`.

OBJECT Stream Event (`transcript.text.delta`)

```
{  
  "type": "transcript.text.delta",  
  "delta": " wonderful"  
}
```

## Stream Event (`transcript.text.segment`)



Emitted when a diarized transcription returns a completed segment with speaker information. Only emitted when you create a transcription with `stream` set to `true` and `response_format` set to `diarized_json`.

**end** number

End timestamp of the segment in seconds.

---

**id** string

Unique identifier for the segment.

---

**speaker** string

Speaker label for this segment.

---

**start** number

Start timestamp of the segment in seconds.

---

**text** string

Transcript text for this segment.

---

**type** string

The type of the event. Always `transcript.text.segment`.

OBJECT Stream Event (`transcript.text.segment`)

```
{  
  "type": "transcript.text.segment",  
  "id": "seg_002",  
  "start": 5.2,  
  "end": 12.8,  
  "text": "Hi, I need help with diarization.",  
}
```

"speaker": "A"

# Stream Event (transcript.text.done)



Emitted when the transcription is complete. Contains the complete transcription text. Only emitted when you [create a transcription](#) with the `Stream` parameter set to `true`.

## **logprobs** array

The log probabilities of the individual tokens in the transcription. Only included if you [create a transcription](#) with the `include[]` parameter set to `logprobs`.

> Show properties

## **text** string

The text that was transcribed.

## **type** string

The type of the event. Always `transcript.text.done`.

## **usage** object

Usage statistics for models billed by token usage.

> Show properties

## OBJECT Stream Event (transcript.text.done)

```
{  
  "type": "transcript.text.done",  
  "text": "I see skies of blue and clouds of white, the bright blessed days, the dark sacred nights,  
  "usage": {  
    "type": "tokens",  
    "input_tokens": 14,  
    "input_token_details": {  
      "text_tokens": 10,  
      "audio_tokens": 4  
    },  
    "output_tokens": 31,  
    "total_tokens": 45  
  }  
}
```

# Images



Given a prompt and/or an input image, the model will generate a new image. Related guide:

[Image generation](#)

---

# Create image



POST <https://api.openai.com/v1/images/generations>

Creates an image given a prompt. [Learn more](#).

## Request body

**prompt** string Required

A text description of the desired image(s). The maximum length is 32000 characters for `gpt-image-1` , 1000 characters for `dall-e-2` and 4000 characters for `dall-e-3` .

---

**background** string or null Optional Defaults to auto

Allows to set transparency for the background of the generated image(s). This parameter is only supported for `gpt-image-1` . Must be one of `transparent` , `opaque` or `auto` (default value). When `auto` is used, the model will automatically determine the best background for the image.

If `transparent`, the output format needs to support transparency, so it should be set to either `png` (default value) or `webp`.

---

**model** string Optional Defaults to dall-e-2

The model to use for image generation. One of `dall-e-2`, `dall-e-3`, or `gpt-image-1`. Defaults to `dall-e-2` unless a parameter specific to `gpt-image-1` is used.

---

**moderation** string or null Optional Defaults to auto

Control the content-moderation level for images generated by `gpt-image-1`. Must be either `low` for less restrictive filtering or `auto` (default value).

---

**n** integer or null Optional Defaults to 1

The number of images to generate. Must be between 1 and 10. For `dall-e-3`, only `n=1` is supported.

---

**output\_compression** integer or null Optional Defaults to 100

The compression level (0-100%) for the generated images. This parameter is only supported for `gpt-image-1` with the `webp` or `jpeg` output formats, and defaults to 100.

---

**output\_format** string or null Optional Defaults to png

The format in which the generated images are returned. This parameter is only supported for `gpt-image-1`. Must be one of `png`, `jpeg`, or `webp`.

---

**partial\_images** integer Optional Defaults to 0

The number of partial images to generate. This parameter is used for streaming responses that return partial images. Value must be between 0 and 3. When set to 0, the response will be a single image sent in one streaming event.

Note that the final image may be sent before the full number of partial images are generated if the full image is generated more quickly.

---

**quality** string or null Optional Defaults to auto

The quality of the image that will be generated.

`auto` (default value) will automatically select the best quality for the given model.

`high`, `medium` and `low` are supported for `gpt-image-1`.

`hd` and `standard` are supported for `dall-e-3`.

`standard` is the only option for `dall-e-2`.

---

**response\_format** string or null Optional Defaults to url

The format in which generated images with `dall-e-2` and `dall-e-3` are returned. Must be one of `url` or `b64_json`.

URLs are only valid for 60 minutes after the image has been generated. This parameter isn't supported for `gpt-image-1` which will always return base64-encoded images.

---

**size** string or null Optional Defaults to auto

The size of the generated images. Must be one of `1024x1024`, `1536x1024` (landscape), `1024x1536` (portrait), or `auto` (default value) for `gpt-image-1`, one of `256x256`, `512x512`, or `1024x1024` for `dall-e-2`, and one of `1024x1024`, `1792x1024`, or `1024x1792` for `dall-e-3`.

**stream** boolean or null Optional Defaults to false

Generate the image in streaming mode. Defaults to `false`. See the [Image generation guide](#) for more information. This parameter is only supported for `gpt-image-1`.

**style** string or null Optional Defaults to vivid

The style of the generated images. This parameter is only supported for `dall-e-3`. Must be one of `vivid` or `natural`.

Vivid causes the model to lean towards generating hyper-real and dramatic images. Natural causes the model to produce more natural, less hyper-real looking images.

**user** string Optional

A unique identifier representing your end-user, which can help OpenAI to monitor and detect abuse. [Learn more](#).

## Returns

Returns an [image](#) object.

**Generate image** **Streaming**

Example request

```
curl https://api.openai.com/v1/images/generations \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "model": "gpt-image-1",
```

```
"prompt": "A cute baby sea otter",
"n": 1,
"size": "1024x1024"
}'
```

## Response

```
{
  "created": 1713833628,
  "data": [
    {
      "b64_json": "..."
    }
  ],
  "usage": {
    "total_tokens": 100,
    "input_tokens": 50,
    "output_tokens": 50,
    "input_tokens_details": {
      "text_tokens": 10,
      "image_tokens": 40
    }
  }
}
```

# Create image edit



```
POST https://api.openai.com/v1/images/edits
```

Creates an edited or extended image given one or more source images and a prompt. This endpoint only supports `gpt-image-1` and `dall-e-2`.

## Request body

---

**image** string or array Required

The image(s) to edit. Must be a supported image file or an array of images.

For `gpt-image-1`, each image should be a `png`, `webp`, or `jpg` file less than 50MB. You can provide up to 16 images.

For `dall-e-2`, you can only provide one image, and it should be a square `png` file less than 4MB.

---

**prompt** string Required

A text description of the desired image(s). The maximum length is 1000 characters for `dall-e-2`, and 32000 characters for `gpt-image-1`.

---

**background** string or null Optional Defaults to auto

Allows to set transparency for the background of the generated image(s). This parameter is only supported for `gpt-image-1`. Must be one of `transparent`, `opaque` or `auto` (default value). When `auto` is used, the model will automatically determine the best background for the image.

If `transparent`, the output format needs to support transparency, so it should be set to either `png` (default value) or `webp`.

---

**input\_fidelity** string Optional

Control how much effort the model will exert to match the style and features, especially facial features.

---

**mask** file Optional

An additional image whose fully transparent areas (e.g. where alpha is zero) indicate where `image` should be edited. If there are multiple images provided, the mask will be applied on the first image. Must be a valid PNG file, less than 4MB, and have the same dimensions as `image`.

---

**model** string Optional Defaults to dall-e-2

The model to use for image generation. Only `dall-e-2` and `gpt-image-1` are supported. Defaults to `dall-e-2` unless a parameter specific to `gpt-image-1` is used.

---

**n** integer or null Optional Defaults to 1

The number of images to generate. Must be between 1 and 10.

---

**output\_compression** integer or null Optional Defaults to 100

The compression level (0-100%) for the generated images. This parameter is only supported for `gpt-image-1` with the `webp` or `jpeg` output formats, and defaults to 100.

**output\_format** string or null Optional Defaults to png

The format in which the generated images are returned. This parameter is only supported for `gpt-image-1`. Must be one of `png`, `jpeg`, or `webp`. The default value is `png`.

**partial\_images** integer Optional Defaults to 0

The number of partial images to generate. This parameter is used for streaming responses that return partial images. Value must be between 0 and 3. When set to 0, the response will be a single image sent in one streaming event.

Note that the final image may be sent before the full number of partial images are generated if the full image is generated more quickly.

**quality** string or null Optional Defaults to auto

The quality of the image that will be generated. `high`, `medium` and `low` are only supported for `gpt-image-1`. `dall-e-2` only supports `standard` quality. Defaults to `auto`.

**response\_format** string or null Optional Defaults to url

The format in which the generated images are returned. Must be one of `url` or `b64_json`. URLs are only valid for 60 minutes after the image has been generated. This parameter is only supported for `dall-e-2`, as `gpt-image-1` will always return base64-encoded images.

**size** string or null Optional Defaults to 1024x1024

The size of the generated images. Must be one of `1024x1024`, `1536x1024` (landscape), `1024x1536` (portrait), or `auto` (default value) for `gpt-image-1`, and one of `256x256`, `512x512`, or `1024x1024` for `dall-e-2`.

**stream** boolean or null Optional Defaults to false

Edit the image in streaming mode. Defaults to `false`. See the [Image generation guide](#) for more information.

**user** string Optional

A unique identifier representing your end-user, which can help OpenAI to monitor and detect abuse. [Learn more](#).

## Returns

Returns an [image](#) object.

[Edit image](#)   [Streaming](#)

Example request

```
curl -s -D >(grep -i x-request-id >&2) \
-o >(jq -r '.data[0].b64_json' | base64 --decode > gift-basket.png) \
-X POST "https://api.openai.com/v1/images/edits" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-F "model=gpt-image-1" \
-F "image[]=@body-lotion.png" \
-F "image[]=@bath-bomb.png" \
-F "image[]=@incense-kit.png" \
-F "image[]=@soap.png" \
-F 'prompt>Create a lovely gift basket with these four items in it'
```

# Create image variation



POST <https://api.openai.com/v1/images/variants>

Creates a variation of a given image. This endpoint only supports `dall-e-2`.

## Request body

### `image` file Required

The image to use as the basis for the variation(s). Must be a valid PNG file, less than 4MB, and square.

---

### `model` string or "dall-e-2" Optional Defaults to dall-e-2

The model to use for image generation. Only `dall-e-2` is supported at this time.

---

### `n` integer or null Optional Defaults to 1

The number of images to generate. Must be between 1 and 10.

---

### `response_format` string or null Optional Defaults to url

The format in which the generated images are returned. Must be one of `url` or `b64_json`. URLs are only valid for 60 minutes after the image has been generated.

---

### `size` string or null Optional Defaults to 1024x1024

The size of the generated images. Must be one of `256x256`, `512x512`, or `1024x1024`.

**user** string Optional

A unique identifier representing your end-user, which can help OpenAI to monitor and detect abuse. [Learn more.](#)

## Returns

Returns a list of [image](#) objects.

### Example request

```
curl https://api.openai.com/v1/images/variations \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-F image="@otter.png" \
-F n=2 \
-F size="1024x1024"
```

### Response

```
{
  "created": 1589478378,
  "data": [
    {
      "url": "https://..."
    },
    {
      "url": "https://..."
    }
  ]
}
```

```
    }  
]  
,
```

# The image generation response



The response from the image generation endpoint.

---

**background** string

The background parameter used for the image generation. Either `transparent` or `opaque`.

---

**created** integer

The Unix timestamp (in seconds) of when the image was created.

---

**data** array

The list of generated images.

> Show properties

---

**output\_format** string

The output format of the image generation. Either `png`, `webp`, or `jpeg`.

---

**quality** string

The quality of the image generated. Either `low`, `medium`, or `high`.

**size** string

The size of the image generated. Either `1024x1024`, `1024x1536`, or `1536x1024`.

**usage** object

For `gpt-image-1` only, the token usage information for the image generation.

> Show properties

OBJECT The image generation response

```
{  
  "created": 1713833628,  
  "data": [  
    {  
      "b64_json": "..."  
    }  
  ],  
  "background": "transparent",  
  "output_format": "png",  
  "size": "1024x1024",  
  "quality": "high",  
  "usage": {  
    "total_tokens": 100,  
    "input_tokens": 50,  
    "output_tokens": 50,  
    "input_tokens_details": {  
      "text_tokens": 10,  
      "image_tokens": 50  
    }  
  }  
}
```

```
        "image_tokens": 40
    }
}
}
```

# Image Streaming

🔗

Stream image generation and editing in real time with server-sent events.

[Learn more about image streaming.](#)

🔗

## image\_generation.partial\_image

🔗

Emitted when a partial image is available during image generation streaming.

**b64\_json** string

Base64-encoded partial image data, suitable for rendering as an image.

---

**background** string

The background setting for the requested image.

---

**created\_at** integer

The Unix timestamp when the event was created.

---

**output\_format** string

The output format for the requested image.

---

**partial\_image\_index** integer

0-based index for the partial image (streaming).

---

**quality** string

The quality setting for the requested image.

---

**size** string

The size of the requested image.

---

**type** string

The type of the event. Always `image_generation.partial_image`.

OBJECT `image_generation.partial_image`

```
{  
  "type": "image_generation.partial_image",  
  "b64_json": "...",  
  "created_at": 1620000000,  
  "size": "1024x1024",  
  "quality": "high",  
  "background": "transparent",  
  "output_format": "png",  
  "partial_image_index": 0  
}
```

## image\_generation.completed



Emitted when image generation has completed and the final image is available.

### b64\_json string

Base64-encoded image data, suitable for rendering as an image.

### background string

The background setting for the generated image.

**created\_at** integer

The Unix timestamp when the event was created.

**output\_format** string

The output format for the generated image.

**quality** string

The quality setting for the generated image.

**size** string

The size of the generated image.

**type** string

The type of the event. Always `image_generation.completed`.

**usage** object

For `gpt-image-1` only, the token usage information for the image generation.

> Show properties

OBJECT `image_generation.completed`

{

```
"type": "image_generation.completed",
"b64_json": "...",
"created_at": 1620000000,
"size": "1024x1024",
```

```
"quality": "high",
"background": "transparent",
"output_format": "png",
"usage": {
    "total_tokens": 100,
    "input_tokens": 50,
    "output_tokens": 50,
    "input_tokens_details": {
        "text_tokens": 10,
        "image_tokens": 40
    }
}
}
```



## image\_edit.partial\_image



Emitted when a partial image is available during image editing streaming.

**b64\_json** string

Base64-encoded partial image data, suitable for rendering as an image.

---

**background** string

The background setting for the requested edited image.

---

**created\_at** integer

The Unix timestamp when the event was created.

---

**output\_format** string

The output format for the requested edited image.

---

**partial\_image\_index** integer

0-based index for the partial image (streaming).

---

**quality** string

The quality setting for the requested edited image.

---

**size** string

The size of the requested edited image.

---

**type** string

The type of the event. Always `image_edit.partial_image`.

OBJECT `image_edit.partial_image`

```
{  
  "type": "image_edit.partial_image",  
  "b64_json": "...",  
  "created_at": 1620000000,  
  "size": "1024x1024",  
  "quality": "high",  
  "background": "transparent",  
  "output_format": "png",  
  "partial_image_index": 0  
}
```

## image\_edit.completed



Emitted when image editing has completed and the final image is available.

### b64\_json string

Base64-encoded final edited image data, suitable for rendering as an image.

### background string

The background setting for the edited image.

**created\_at** integer

The Unix timestamp when the event was created.

**output\_format** string

The output format for the edited image.

**quality** string

The quality setting for the edited image.

**size** string

The size of the edited image.

**type** string

The type of the event. Always `image_edit.completed`.

**usage** object

For `gpt-image-1` only, the token usage information for the image generation.

> Show properties

OBJECT `image_edit.completed`

{

```
"type": "image_edit.completed",
"b64_json": "...",
"created_at": 1620000000,
"size": "1024x1024",
```

```
"quality": "high",
"background": "transparent",
"output_format": "png",
"usage": {
    "total_tokens": 100,
    "input_tokens": 50,
    "output_tokens": 50,
    "input_tokens_details": {
        "text_tokens": 10,
        "image_tokens": 40
    }
}
}
```

# Embeddings



Get a vector representation of a given input that can be easily consumed by machine learning models and algorithms. Related guide: [Embeddings](#)

# Create embeddings



POST <https://api.openai.com/v1/embeddings>

Creates an embedding vector representing the input text.

## Request body

### **input** string or array Required

Input text to embed, encoded as a string or array of tokens. To embed multiple inputs in a single request, pass an array of strings or array of token arrays. The input must not exceed the max input tokens for the model (8192 tokens for all embedding models), cannot be an empty string, and any array must be 2048 dimensions or less. [Example Python code](#) for counting tokens. In addition to the per-input token limit, all embedding models enforce a maximum of 300,000 tokens summed across all inputs in a single request.

### **model** string Required

ID of the model to use. You can use the [List models](#) API to see all of your available models, or see our [Model overview](#) for descriptions of them.

### **dimensions** integer Optional

The number of dimensions the resulting output embeddings should have. Only supported in [text-embedding-3](#) and later models.

### **encoding\_format** string Optional Defaults to float

The format to return the embeddings in. Can be either `float` or `base64`.

**user** string Optional

A unique identifier representing your end-user, which can help OpenAI to monitor and detect abuse. [Learn more](#).

## Returns

A list of `embedding` objects.

### Example request

```
curl https://api.openai.com/v1/embeddings \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "input": "The food was delicious and the waiter...",
  "model": "text-embedding-ada-002",
  "encoding_format": "float"
}'
```

### Response

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

```
"object": "embedding",
"embedding": [
  0.0023064255,
  -0.009327292,
  .... (1536 floats total for ada-002)
  -0.0028842222,
],
"index": 0
}
],
"model": "text-embedding-ada-002",
"usage": {
  "prompt_tokens": 8,
  "total_tokens": 8
}
}
```

# The embedding object



Represents an embedding vector returned by embedding endpoint.

**embedding** array

The embedding vector, which is a list of floats. The length of vector depends on the model as listed in the [embedding guide](#).

**index** integer

The index of the embedding in the list of embeddings.

**object** string

The object type, which is always "embedding".

OBJECT The embedding object

```
{  
  "object": "embedding",  
  "embedding": [  
    0.0023064255,  
    -0.009327292,  
    .... (1536 floats total for ada-002)  
    -0.0028842222,  
  ],  
  "index": 0  
}
```

# Evals



Create, manage, and run evals in the OpenAI platform. Related guide: [Evals](#)

# Create eval



POST <https://api.openai.com/v1/evals>

Create the structure of an evaluation that can be used to test a model's performance. An evaluation is a set of testing criteria and the config for a data source, which dictates the schema of the data used in the evaluation. After creating an evaluation, you can run it on different models and model parameters. We support several types of graders and datasources. For more information, see the [Evals guide](#).

## Request body

**data\_source\_config** object Required

The configuration for the data source used for the evaluation runs. Dictates the schema of the data used in the evaluation.

> Show possible types

**testing\_criteria** array Required

A list of graders for all eval runs in this group. Graders can reference variables in the data source using double curly braces notation, like `{{item.variable_name}}`. To reference the model's output, use the `sample` namespace (ie, `{{sample.output_text}}`).

> Show possible types

### metadata map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

### name string Optional

The name of the evaluation.

## Returns

The created [Eval](#) object.

### Example request

```
curl https://api.openai.com/v1/evals \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
    "name": "Sentiment",
    "data_source_config": {
        "type": "stored_completions",
        "metadata": {
            "usecase": "chatbot"
        }
    }
}'
```

```
        }
    },
    "testing_criteria": [
        {
            "type": "label_model",
            "model": "o3-mini",
            "input": [
                {
                    "role": "developer",
                    "content": "Classify the sentiment of the following statement as one of 'positive', 'neutral', or 'negative'. Statement: {{item.input}}"
                },
                {
                    "role": "user",
                    "content": "Statement: {{item.input}}"
                }
            ],
            "passing_labels": [
                "positive"
            ],
            "labels": [
                "positive",
                "neutral",
                "negative"
            ],
            "name": "Example label grader"
        }
    ]
}'
```

## Response

```
{  
  "object": "eval",  
  "id": "eval_67b7fa9a81a88190ab4aa417e397ea21",  
  "data_source_config": {  
    "type": "stored_completions",  
    "metadata": {  
      "usecase": "chatbot"  
    },  
    "schema": {  
      "type": "object",  
      "properties": {  
        "item": {  
          "type": "object"  
        },  
        "sample": {  
          "type": "object"  
        }  
      },  
      "required": [  
        "item",  
        "sample"  
      ]  
    },  
    "testing_criteria": [  
      {  
        "name": "Example label grader",  
        "type": "label_model",  
        "model_id": "label_model_12345678901234567890123456789012",  
        "label_type": "text_label",  
        "label_value": "positive",  
        "label_confidence": 0.95  
      }  
    ]  
  }  
}
```

```
"model": "o3-mini",
"input": [
  {
    "type": "message",
    "role": "developer",
    "content": {
      "type": "input_text",
      "text": "Classify the sentiment of the following statement as one of positive, neutral,
    }
  },
  {
    "type": "message",
    "role": "user",
    "content": {
      "type": "input_text",
      "text": "Statement: {{item.input}}"
    }
  }
],
"passing_labels": [
  "positive"
],
"labels": [
  "positive",
  "neutral",
  "negative"
]
},
],
],
```

```
"name": "Sentiment",
"created_at": 1740110490,
"metadata": {
  "description": "An eval for sentiment analysis"
}
}
```

## Get an eval



```
GET https://api.openai.com/v1/evals/{eval_id}
```

Get an evaluation by ID.

### Path parameters

**eval\_id** string Required

The ID of the evaluation to retrieve.

### Returns

The Eval object matching the specified ID.

#### Example request

```
curl https://api.openai.com/v1/evals/eval_67abd54d9b0081909a86353f6fb9317a \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

#### Response

```
{
  "object": "eval",
  "id": "eval_67abd54d9b0081909a86353f6fb9317a",
  "data_source_config": {
    "type": "custom",
    "schema": {
      "type": "object",
      "properties": {
        "item": {
          "type": "object",
          "properties": {
            "input": {
              "type": "string"
            },
            "ground_truth": {
              "type": "string"
            }
          }
        }
      }
    }
}
```

```
        },
        "required": [
            "input",
            "ground_truth"
        ]
    }
},
"required": [
    "item"
]
},
"testing_criteria": [
{
    "name": "String check",
    "id": "String check-2eaf2d8d-d649-4335-8148-9535a7ca73c2",
    "type": "string_check",
    "input": "{{item.input}}",
    "reference": "{{item.ground_truth}}",
    "operation": "eq"
}
],
"name": "External Data Eval",
"created_at": 1739314509,
"metadata": {},
}
```

# Update an eval



```
POST https://api.openai.com/v1/evals/{eval_id}
```

Update certain properties of an evaluation.

## Path parameters

**eval\_id** string Required

The ID of the evaluation to update.

## Request body

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**name** string Optional

Rename the evaluation.

## Returns

The Eval object matching the updated version.

#### Example request

```
curl https://api.openai.com/v1/evals/eval_67abd54d9b0081909a86353f6fb9317a \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"name": "Updated Eval", "metadata": {"description": "Updated description"}}'
```

#### Response

```
{
  "object": "eval",
  "id": "eval_67abd54d9b0081909a86353f6fb9317a",
  "data_source_config": {
    "type": "custom",
    "schema": {
      "type": "object",
      "properties": {
        "item": {
          "type": "object",
          "properties": {
            "input": {
              "type": "string"
            },
            "ground_truth": {
              "type": "string"
            }
          }
        }
      }
    }
  }
}
```

```
        }
    },
    "required": [
        "input",
        "ground_truth"
    ]
}
},
"required": [
    "item"
]
}
},
"testing_criteria": [
{
    "name": "String check",
    "id": "String check-2eaf2d8d-d649-4335-8148-9535a7ca73c2",
    "type": "string_check",
    "input": "{{item.input}}",
    "reference": "{{item.ground_truth}}",
    "operation": "eq"
},
],
"name": "Updated Eval",
"created_at": 1739314509,
"metadata": {"description": "Updated description"},
}
```

# Delete an eval



```
DELETE https://api.openai.com/v1/evals/{eval_id}
```

Delete an evaluation.

## Path parameters

---

**eval\_id** string Required

The ID of the evaluation to delete.

## Returns

---

A deletion confirmation object.

### Example request

```
curl https://api.openai.com/v1/evals/eval_abc123 \  
-X DELETE \  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "object": "eval.deleted",  
  "deleted": true,  
  "eval_id": "eval_abc123"  
}
```

# List evals



```
GET https://api.openai.com/v1/evals
```

List evaluations for a project.

## Query parameters

**after** string Optional

Identifier for the last eval from the previous pagination request.

**limit** integer Optional Defaults to 20

Number of evals to retrieve.

**order** string Optional Defaults to asc

Sort order for evals by timestamp. Use `asc` for ascending order or `desc` for descending order.

**order\_by** string Optional Defaults to created\_at

Evals can be ordered by creation time or last updated time. Use `created_at` for creation time or `updated_at` for last updated time.

## Returns

A list of [evals](#) matching the specified filters.

Example request

```
curl https://api.openai.com/v1/evals?limit=1 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

Response

```
{  
  "object": "list",
```

```
"data": [
  {
    "id": "eval_67abd54d9b0081909a86353f6fb9317a",
    "object": "eval",
    "data_source_config": {
      "type": "stored_completions",
      "metadata": {
        "usecase": "push_notifications_summarizer"
      },
      "schema": {
        "type": "object",
        "properties": {
          "item": {
            "type": "object"
          },
          "sample": {
            "type": "object"
          }
        },
        "required": [
          "item",
          "sample"
        ]
      }
    },
    "testing_criteria": [
      {
        "name": "Push Notification Summary Grader",
        "id": "Push Notification Summary Grader-9b876f24-4762-4be9-aff4-db7a9b31c673",
        "type": "object"
      }
    ]
  }
]
```

```
"type": "label_model",
"model": "o3-mini",
"input": [
  {
    "type": "message",
    "role": "developer",
    "content": {
      "type": "input_text",
      "text": "\nLabel the following push notification summary as either correct or incorrect\n"
    }
  },
  {
    "type": "message",
    "role": "user",
    "content": {
      "type": "input_text",
      "text": "\nPush notifications: {{item.input}}\nSummary: {{sample.output_text}}\n"
    }
  }
],
"passing_labels": [
  "correct"
],
"labels": [
  "correct",
  "incorrect"
],
"sampling_params": null
}
```

```
],
  "name": "Push Notification Summary Grader",
  "created_at": 1739314509,
  "metadata": {
    "description": "A stored completions eval for push notification summaries"
  }
},
],
"first_id": "eval_67abd54d9b0081909a86353f6fb9317a",
"last_id": "eval_67aa884cf6688190b58f657d4441c8b7",
"has_more": true
}
```

## Get eval runs



```
GET https://api.openai.com/v1/evals/{eval_id}/runs
```

Get a list of runs for an evaluation.

### Path parameters

**eval\_id** string Required

The ID of the evaluation to retrieve runs for.

## Query parameters

---

**after** string Optional

Identifier for the last run from the previous pagination request.

---

**limit** integer Optional Defaults to 20

Number of runs to retrieve.

---

**order** string Optional Defaults to asc

Sort order for runs by timestamp. Use `asc` for ascending order or `desc` for descending order. Defaults to `asc`.

---

**status** string Optional

Filter runs by status. One of `queued` | `in_progress` | `failed` | `completed` | `canceled`.

---

## Returns

---

A list of [EvalRun](#) objects matching the specified ID.

Example request

```
curl https://api.openai.com/v1/evals/egroup_67abd54d9b0081909a86353f6fb9317a/runs \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{
  "object": "list",
  "data": [
    {
      "object": "eval.run",
      "id": "evalrun_67e0c7d31560819090d60c0780591042",
      "eval_id": "eval_67e0c726d560819083f19a957c4c640b",
      "report_url": "https://platform.openai.com/evaluations/eval_67e0c726d560819083f19a957c4c640b",
      "status": "completed",
      "model": "o3-mini",
      "name": "bulk_with_negative_examples_o3-mini",
      "created_at": 1742784467,
      "result_counts": {
        "total": 1,
        "errored": 0,
        "failed": 0,
        "passed": 1
      },
      "per_model_usage": [
        {
          "model": "o3-mini",
          "tokens": 1000
        }
      ]
    }
  ]
}
```

```
"model_name": "o3-mini",
"invocation_count": 1,
"prompt_tokens": 563,
"completion_tokens": 874,
"total_tokens": 1437,
"cached_tokens": 0
},
],
"per_testing_criteria_results": [
{
  "testing_criteria": "Push Notification Summary Grader-1808cd0b-eeec-4e0b-a519-337e79f4f5c",
  "passed": 1,
  "failed": 0
},
],
"data_source": {
  "type": "completions",
  "source": {
    "type": "file_content",
    "content": [
      {
        "item": {
          "notifications": "\n- New message from Sarah: \"Can you call me later?\"\n- Your pa"
        }
      }
    ]
  },
  "input_messages": {
    "type": "template",
```

```
"template": [
  {
    "type": "message",
    "role": "developer",
    "content": {
      "type": "input_text",
      "text": "\n\n\n\nYou are a helpful assistant that takes in an array of push notifications and generates a response.\n\n"
    }
  },
  {
    "type": "message",
    "role": "user",
    "content": {
      "type": "input_text",
      "text": "<push_notifications>{{item.notifications}}</push_notifications>"
    }
  }
],
},
"model": "o3-mini",
"sampling_params": null
},
"error": null,
"metadata": {}
}
],
"first_id": "evalrun_67e0c7d31560819090d60c0780591042",
"last_id": "evalrun_67e0c7d31560819090d60c0780591042",
```

```
"has_more": true
```

# Get an eval run



```
GET https://api.openai.com/v1/evals/{eval_id}/runs/{run_id}
```

Get an evaluation run by ID.

## Path parameters

**eval\_id** string Required

The ID of the evaluation to retrieve runs for.

**run\_id** string Required

The ID of the run to retrieve.

## Returns

The [EvalRun](#) object matching the specified ID.

## Example request

```
curl https://api.openai.com/v1/evals/eval_67abd54d9b0081909a86353f6fb9317a/runs/evalrun_67abd54d60ec  
-H "Authorization: Bearer $OPENAI_API_KEY" \  
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "eval.run",  
  "id": "evalrun_67abd54d60ec8190832b46859da808f7",  
  "eval_id": "eval_67abd54d9b0081909a86353f6fb9317a",  
  "report_url": "https://platform.openai.com/evaluations/eval_67abd54d9b0081909a86353f6fb9317a?run_=evalrun_67abd54d60ec8190832b46859da808f7",  
  "status": "queued",  
  "model": "gpt-4o-mini",  
  "name": "gpt-4o-mini",  
  "created_at": 1743092069,  
  "result_counts": {  
    "total": 0,  
    "errored": 0,  
    "failed": 0,  
    "passed": 0  
  },  
  "per_model_usage": null,  
  "per_testing_criteria_results": null,  
  "data_source": {  
    "id": "data_source_67abd54d60ec8190832b46859da808f7",  
    "label": "System",  
    "type": "System",  
    "value": "System",  
    "order": 1  
  }  
}
```

```
"type": "completions",
"source": {
  "type": "file_content",
"content": [
  {
    "item": {
      "input": "Tech Company Launches Advanced Artificial Intelligence Platform",
      "ground_truth": "Technology"
    }
  },
  {
    "item": {
      "input": "Central Bank Increases Interest Rates Amid Inflation Concerns",
      "ground_truth": "Markets"
    }
  },
  {
    "item": {
      "input": "International Summit Addresses Climate Change Strategies",
      "ground_truth": "World"
    }
  },
  {
    "item": {
      "input": "Major Retailer Reports Record-Breaking Holiday Sales",
      "ground_truth": "Business"
    }
  },
  {
    "item": {
      "input": "New Study Reveals Impact of AI on Employment Rates",
      "ground_truth": "Economy"
    }
  }
]
```

```
"item": {  
    "input": "National Team Qualifies for World Championship Finals",  
    "ground_truth": "Sports"  
}  
,  
{  
    "item": {  
        "input": "Stock Markets Rally After Positive Economic Data Released",  
        "ground_truth": "Markets"  
    }  
,  
{  
    "item": {  
        "input": "Global Manufacturer Announces Merger with Competitor",  
        "ground_truth": "Business"  
    }  
,  
{  
    "item": {  
        "input": "Breakthrough in Renewable Energy Technology Unveiled",  
        "ground_truth": "Technology"  
    }  
,  
{  
    "item": {  
        "input": "World Leaders Sign Historic Climate Agreement",  
        "ground_truth": "World"  
    }  
,
```

```
{  
    "item": {  
        "input": "Professional Athlete Sets New Record in Championship Event",  
        "ground_truth": "Sports"  
    }  
},  
{  
    "item": {  
        "input": "Financial Institutions Adapt to New Regulatory Requirements",  
        "ground_truth": "Business"  
    }  
},  
{  
    "item": {  
        "input": "Tech Conference Showcases Advances in Artificial Intelligence",  
        "ground_truth": "Technology"  
    }  
},  
{  
    "item": {  
        "input": "Global Markets Respond to Oil Price Fluctuations",  
        "ground_truth": "Markets"  
    }  
},  
{  
    "item": {  
        "input": "International Cooperation Strengthened Through New Treaty",  
        "ground_truth": "World"  
    }  
}
```

```
        },
        {
          "item": {
            "input": "Sports League Announces Revised Schedule for Upcoming Season",
            "ground_truth": "Sports"
          }
        }
      ],
    },
    "input_messages": {
      "type": "template",
      "template": [
        {
          "type": "message",
          "role": "developer",
          "content": {
            "type": "input_text",
            "text": "Categorize a given news headline into one of the following topics: Technology."
          }
        },
        {
          "type": "message",
          "role": "user",
          "content": {
            "type": "input_text",
            "text": "{{item.input}}"
          }
        }
      ]
    }
  }
}
```

```
},
  "model": "gpt-4o-mini",
  "sampling_params": {
    "seed": 42,
    "temperature": 1.0,
    "top_p": 1.0,
    "max_completions_tokens": 2048
  }
},
"error": null,
"metadata": {}
}
```

## Create eval run



POST [https://api.openai.com/v1/evals/{eval\\_id}/runs](https://api.openai.com/v1/evals/{eval_id}/runs)

Kicks off a new run for a given evaluation, specifying the data source, and what model configuration to use to test. The datasource will be validated against the schema specified in the config of the evaluation.

### Path parameters

**eval\_id** string Required

The ID of the evaluation to create a run for.

## Request body

**data\_source** object Required

Details about the run's data source.

> Show possible types

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**name** string Optional

The name of the run.

## Returns

The [EvalRun](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/evals/eval_67e579652b548190aaa83ada4b125f47/runs \
-X POST \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"name": "gpt-4o-mini", "data_source": {"type": "completions", "input_messages": {"type": "template",
```

## Response

```
{
  "object": "eval.run",
  "id": "evalrun_67e57965b480819094274e3a32235e4c",
  "eval_id": "eval_67e579652b548190aaa83ada4b125f47",
  "report_url": "https://platform.openai.com/evaluations/eval_67e579652b548190aaa83ada4b125f47&run_",
  "status": "queued",
  "model": "gpt-4o-mini",
  "name": "gpt-4o-mini",
  "created_at": 1743092069,
  "result_counts": {
    "total": 0,
    "errored": 0,
    "failed": 0,
    "passed": 0
  },
  "per_model_usage": null,
  "per_testing_criteria_results": null,
  "data_source": {
    "type": "completions",
```

```
"source": {
    "type": "file_content",
    "content": [
        {
            "item": {
                "input": "Tech Company Launches Advanced Artificial Intelligence Platform",
                "ground_truth": "Technology"
            }
        }
    ]
},
"input_messages": {
    "type": "template",
    "template": [
        {
            "type": "message",
            "role": "developer",
            "content": {
                "type": "input_text",
                "text": "Categorize a given news headline into one of the following topics: Technology."
            }
        },
        {
            "type": "message",
            "role": "user",
            "content": {
                "type": "input_text",
                "text": "{{item.input}}"
            }
        }
    ]
}
```

```
        }
    ],
},
"model": "gpt-4o-mini",
"sampling_params": {
    "seed": 42,
    "temperature": 1.0,
    "top_p": 1.0,
    "max_completions_tokens": 2048
}
},
"error": null,
"metadata": {}
}
```

## Cancel eval run



POST [https://api.openai.com/v1/evals/{eval\\_id}/runs/{run\\_id}](https://api.openai.com/v1/evals/{eval_id}/runs/{run_id})

Cancel an ongoing evaluation run.

### Path parameters

**eval\_id** string Required

The ID of the evaluation whose run you want to cancel.

**run\_id** string Required

The ID of the run to cancel.

## Returns

The updated [EvalRun](#) object reflecting that the run is canceled.

### Example request

```
curl https://api.openai.com/v1/evals/eval_67abd54d9b0081909a86353f6fb9317a/runs/evalrun_67abd54d60ec  
-X POST \  
-H "Authorization: Bearer $OPENAI_API_KEY" \  
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "eval.run",  
  "id": "evalrun_67abd54d60ec8190832b46859da808f7",  
  "eval_id": "eval_67abd54d9b0081909a86353f6fb9317a",  
  "report_url": "https://platform.openai.com/evaluations/eval_67abd54d9b0081909a86353f6fb9317a?run_i
```

```
"status": "canceled",
"model": "gpt-4o-mini",
"name": "gpt-4o-mini",
"created_at": 1743092069,
"result_counts": {
    "total": 0,
    "errored": 0,
    "failed": 0,
    "passed": 0
},
"per_model_usage": null,
"per_testing_criteria_results": null,
"data_source": {
    "type": "completions",
    "source": {
        "type": "file_content",
        "content": [
            {
                "item": {
                    "input": "Tech Company Launches Advanced Artificial Intelligence Platform",
                    "ground_truth": "Technology"
                }
            },
            {
                "item": {
                    "input": "Central Bank Increases Interest Rates Amid Inflation Concerns",
                    "ground_truth": "Markets"
                }
            }
        ],
        "count": 2
    }
}
```

```
{  
    "item": {  
        "input": "International Summit Addresses Climate Change Strategies",  
        "ground_truth": "World"  
    }  
},  
{  
    "item": {  
        "input": "Major Retailer Reports Record-Breaking Holiday Sales",  
        "ground_truth": "Business"  
    }  
},  
{  
    "item": {  
        "input": "National Team Qualifies for World Championship Finals",  
        "ground_truth": "Sports"  
    }  
},  
{  
    "item": {  
        "input": "Stock Markets Rally After Positive Economic Data Released",  
        "ground_truth": "Markets"  
    }  
},  
{  
    "item": {  
        "input": "Global Manufacturer Announces Merger with Competitor",  
        "ground_truth": "Business"  
    }  
}
```

```
},
{
  "item": {
    "input": "Breakthrough in Renewable Energy Technology Unveiled",
    "ground_truth": "Technology"
  }
},
{
  "item": {
    "input": "World Leaders Sign Historic Climate Agreement",
    "ground_truth": "World"
  }
},
{
  "item": {
    "input": "Professional Athlete Sets New Record in Championship Event",
    "ground_truth": "Sports"
  }
},
{
  "item": {
    "input": "Financial Institutions Adapt to New Regulatory Requirements",
    "ground_truth": "Business"
  }
},
{
  "item": {
    "input": "Tech Conference Showcases Advances in Artificial Intelligence",
    "ground_truth": "Technology"
  }
}
```

```
        }
    },
    {
        "item": {
            "input": "Global Markets Respond to Oil Price Fluctuations",
            "ground_truth": "Markets"
        }
    },
    {
        "item": {
            "input": "International Cooperation Strengthened Through New Treaty",
            "ground_truth": "World"
        }
    },
    {
        "item": {
            "input": "Sports League Announces Revised Schedule for Upcoming Season",
            "ground_truth": "Sports"
        }
    }
],
},
"input_messages": {
    "type": "template",
    "template": [
        {
            "type": "message",
            "role": "developer",
            "content": {
                "text": "Hello! I'm a developer using the OpenAI API. How can I assist you today?"
            }
        }
    ]
}
```

```
        "type": "input_text",
        "text": "Categorize a given news headline into one of the following topics: Technology,
    }
},
{
    "type": "message",
    "role": "user",
    "content": {
        "type": "input_text",
        "text": "{{item.input}}"
    }
}
],
},
"model": "gpt-4o-mini",
"sampling_params": {
    "seed": 42,
    "temperature": 1.0,
    "top_p": 1.0,
    "max_completions_tokens": 2048
},
},
"error": null,
"metadata": {}
}
```

# Delete eval run



```
DELETE https://api.openai.com/v1/evals/{eval_id}/runs/{run_id}
```

Delete an eval run.

## Path parameters

**eval\_id** string Required

The ID of the evaluation to delete the run from.

**run\_id** string Required

The ID of the run to delete.

## Returns

An object containing the status of the delete operation.

### Example request

```
curl https://api.openai.com/v1/evals/eval_123abc/runs/evalrun_abc456 \
-X DELETE \
```

```
-H "Authorization: Bearer $OPENAI_API_KEY" \
```

## Response

```
{  
  "object": "eval.run.deleted",  
  "deleted": true,  
  "run_id": "evalrun_abc456"  
}
```

# Get an output item of an eval run



```
GET https://api.openai.com/v1/evals/{eval_id}/runs/{run_id}/output_items/{output_item_id}
```

Get an evaluation run output item by ID.

## Path parameters

**eval\_id** string Required

The ID of the evaluation to retrieve runs for.

**output\_item\_id** string Required

The ID of the output item to retrieve.

**run\_id** string **Required**

The ID of the run to retrieve.

## Returns

The [EvalRunOutputItem](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/evals/eval_67abd54d9b0081909a86353f6fb9317a/runs/evalrun_67abd54d60ec  
  -H "Authorization: Bearer $OPENAI_API_KEY" \  
  -H "Content-Type: application/json"
```

### Response

```
{  
  "object": "eval.run.output_item",  
  "id": "outputitem_67e5796c28e081909917bf79f6e6214d",  
  "created_at": 1743092076,  
  "run_id": "evalrun_67abd54d60ec8190832b46859da808f7",  
  "eval_id": "eval_67abd54d9b0081909a86353f6fb9317a",  
  "status": "pass",  
  "datasource_item_id": 5,
```

```
"datasource_item": {
    "input": "Stock Markets Rally After Positive Economic Data Released",
    "ground_truth": "Markets"
},
"results": [
{
    "name": "String check-a2486074-d803-4445-b431-ad2262e85d47",
    "sample": null,
    "passed": true,
    "score": 1.0
},
],
"sample": {
    "input": [
{
        "role": "developer",
        "content": "Categorize a given news headline into one of the following topics: Technology,",
        "tool_call_id": null,
        "tool_calls": null,
        "function_call": null
},
{
        "role": "user",
        "content": "Stock Markets Rally After Positive Economic Data Released",
        "tool_call_id": null,
        "tool_calls": null,
        "function_call": null
}
]
},
```

```
"output": [
  {
    "role": "assistant",
    "content": "Markets",
    "tool_call_id": null,
    "tool_calls": null,
    "function_call": null
  }
],
"finish_reason": "stop",
"model": "gpt-4o-mini-2024-07-18",
"usage": {
  "total_tokens": 325,
  "completion_tokens": 2,
  "prompt_tokens": 323,
  "cached_tokens": 0
},
"error": null,
"temperature": 1.0,
"max_completion_tokens": 2048,
"top_p": 1.0,
"seed": 42
}
```

# Get eval run output items



```
GET https://api.openai.com/v1/evals/{eval_id}/runs/{run_id}/output_items
```

Get a list of output items for an evaluation run.

## Path parameters

**eval\_id** string Required

The ID of the evaluation to retrieve runs for.

**run\_id** string Required

The ID of the run to retrieve output items for.

## Query parameters

**after** string Optional

Identifier for the last output item from the previous pagination request.

**limit** integer Optional Defaults to 20

Number of output items to retrieve.

**order** string Optional Defaults to asc

Sort order for output items by timestamp. Use `asc` for ascending order or `desc` for descending order. Defaults to `asc`.

**status** string Optional

Filter output items by status. Use `failed` to filter by failed output items or `pass` to filter by passed output items.

## Returns

A list of `EvalRunOutputItem` objects matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/evals/egroup_67abd54d9b0081909a86353f6fb9317a/runs/erun_67abd54d60ec8  
-H "Authorization: Bearer $OPENAI_API_KEY" \  
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "eval.run.output_item",  
      "id": "outputitem_67e5796c28e081909917bf79f6e6214d",  
      "created_at": 1743092076,  
      "run_id": "evalrun_67abd54d60ec8190832b46859da808f7",  
      "status": "pass",  
      "text": "The quick brown fox jumps over the lazy dog.",  
      "start": 0,  
      "end": 142,  
      "error": null  
    }  
  ]  
}
```

```
"eval_id": "eval_67abd54d9b0081909a86353f6fb9317a",
"status": "pass",
"datasource_item_id": 5,
"datasource_item": {
  "input": "Stock Markets Rally After Positive Economic Data Released",
  "ground_truth": "Markets"
},
"results": [
  {
    "name": "String check-a2486074-d803-4445-b431-ad2262e85d47",
    "sample": null,
    "passed": true,
    "score": 1.0
  }
],
"sample": {
  "input": [
    {
      "role": "developer",
      "content": "Categorize a given news headline into one of the following topics: Technology, Sports, Business, Science, or Politics.",
      "tool_call_id": null,
      "tool_calls": null,
      "function_call": null
    },
    {
      "role": "user",
      "content": "Stock Markets Rally After Positive Economic Data Released",
      "tool_call_id": null,
      "tool_calls": null,
      "function_call": null
    }
  ]
}
```

```
        "function_call": null
    }
],
"output": [
{
    "role": "assistant",
    "content": "Markets",
    "tool_call_id": null,
    "tool_calls": null,
    "function_call": null
},
],
"finish_reason": "stop",
"model": "gpt-4o-mini-2024-07-18",
"usage": {
    "total_tokens": 325,
    "completion_tokens": 2,
    "prompt_tokens": 323,
    "cached_tokens": 0
},
"error": null,
"temperature": 1.0,
"max_completion_tokens": 2048,
"top_p": 1.0,
"seed": 42
}
}
],
"first_id": "outputitem_67e5796c28e081909917bf79f6e6214d",
```

```
"last_id": "outputitem_67e5796c28e081909917bf79f6e6214d",
"has_more": true
}
```

# The eval object



An Eval object with a data source config and testing criteria. An Eval represents a task to be done for your LLM integration. Like:

Improve the quality of my chatbot

See how well my chatbot handles customer support

Check if o4-mini is better at my usecase than gpt-4o

---

**created\_at** integer

The Unix timestamp (in seconds) for when the eval was created.

---

**data\_source\_config** object

Configuration of data sources used in runs of the evaluation.

> Show possible types

**id** string

Unique identifier for the evaluation.

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**name** string

The name of the evaluation.

**object** string

The object type.

**testing\_criteria** array

A list of testing criteria.

› Show possible types

OBJECT The eval object

```
{  
  "object": "eval",  
  "id": "eval_67abd54d9b0081909a86353f6fb9317a",  
  "data_source_config": {  
    "type": "custom",  
    "item_schema": {
```

```
"type": "object",
"properties": {
    "label": {"type": "string"},
},
"required": ["label"]
},
"include_sample_schema": true
},
"testing_criteria": [
{
    "name": "My string check grader",
    "type": "string_check",
    "input": "{{sample.output_text}}",
    "reference": "{{item.label}}",
    "operation": "eq",
}
],
"name": "External Data Eval",
"created_at": 1739314509,
"metadata": {
    "test": "synthetics",
}
}
```

# The eval run object



A schema representing an evaluation run.

---

## **created\_at** integer

Unix timestamp (in seconds) when the evaluation run was created.

---

## **data\_source** object

Information about the run's data source.

> Show possible types

---

## **error** object

An object representing an error response from the Eval API.

> Show properties

---

## **eval\_id** string

The identifier of the associated evaluation.

---

## **id** string

Unique identifier for the evaluation run.

---

## **metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string

The model that is evaluated, if applicable.

---

**name** string

The name of the evaluation run.

---

**object** string

The type of the object. Always "eval.run".

---

**per\_model\_usage** array

Usage statistics for each model during the evaluation run.

> Show properties

---

**per\_testing\_criteria\_results** array

Results per testing criteria applied during the evaluation run.

> Show properties

---

**report\_url** string

The URL to the rendered evaluation run report on the UI dashboard.

---

**result\_counts** object

Counters summarizing the outcomes of the evaluation run.

> Show properties

**status** string

The status of the evaluation run.

OBJECT The eval run object

```
{  
  "object": "eval.run",  
  "id": "evalrun_67e57965b480819094274e3a32235e4c",  
  "eval_id": "eval_67e579652b548190aaa83ada4b125f47",  
  "report_url": "https://platform.openai.com/evaluations/eval_67e579652b548190aaa83ada4b125f47?run",  
  "status": "queued",  
  "model": "gpt-4o-mini",  
  "name": "gpt-4o-mini",  
  "created_at": 1743092069,  
  "result_counts": {  
    "total": 0,  
    "errored": 0,  
    "failed": 0,  
    "passed": 0  
  },  
  "per_model_usage": null,  
  "per_testing_criteria_results": null,  
  "data_source": {  
    "type": "completions",  
    "source": {  
      "type": "file_content",  
      "content": [  
        {  
          "id": "content_1",  
          "path": "path/to/file1.txt",  
          "content": "Content of file 1"  
        },  
        {  
          "id": "content_2",  
          "path": "path/to/file2.txt",  
          "content": "Content of file 2"  
        }  
      ]  
    }  
  }  
}
```

```
"item": {
    "input": "Tech Company Launches Advanced Artificial Intelligence Platform",
    "ground_truth": "Technology"
},
{
"item": {
    "input": "Central Bank Increases Interest Rates Amid Inflation Concerns",
    "ground_truth": "Markets"
}
},
{
"item": {
    "input": "International Summit Addresses Climate Change Strategies",
    "ground_truth": "World"
}
},
{
"item": {
    "input": "Major Retailer Reports Record-Breaking Holiday Sales",
    "ground_truth": "Business"
}
},
{
"item": {
    "input": "National Team Qualifies for World Championship Finals",
    "ground_truth": "Sports"
}
},
```

```
{  
    "item": {  
        "input": "Stock Markets Rally After Positive Economic Data Released",  
        "ground_truth": "Markets"  
    }  
},  
{  
    "item": {  
        "input": "Global Manufacturer Announces Merger with Competitor",  
        "ground_truth": "Business"  
    }  
},  
{  
    "item": {  
        "input": "Breakthrough in Renewable Energy Technology Unveiled",  
        "ground_truth": "Technology"  
    }  
},  
{  
    "item": {  
        "input": "World Leaders Sign Historic Climate Agreement",  
        "ground_truth": "World"  
    }  
},  
{  
    "item": {  
        "input": "Professional Athlete Sets New Record in Championship Event",  
        "ground_truth": "Sports"  
    }  
}
```

```
},
{
  "item": {
    "input": "Financial Institutions Adapt to New Regulatory Requirements",
    "ground_truth": "Business"
  }
},
{
  "item": {
    "input": "Tech Conference Showcases Advances in Artificial Intelligence",
    "ground_truth": "Technology"
  }
},
{
  "item": {
    "input": "Global Markets Respond to Oil Price Fluctuations",
    "ground_truth": "Markets"
  }
},
{
  "item": {
    "input": "International Cooperation Strengthened Through New Treaty",
    "ground_truth": "World"
  }
},
{
  "item": {
    "input": "Sports League Announces Revised Schedule for Upcoming Season",
    "ground_truth": "Sports"
  }
}
```

```
        }
    }
],
},
"input_messages": {
    "type": "template",
    "template": [
        {
            "type": "message",
            "role": "developer",
            "content": {
                "type": "input_text",
                "text": "Categorize a given news headline into one of the following topics: Technology."
            }
        },
        {
            "type": "message",
            "role": "user",
            "content": {
                "type": "input_text",
                "text": "{{item.input}}"
            }
        }
    ]
},
"model": "gpt-4o-mini",
"sampling_params": {
    "seed": 42,
    "temperature": 1.0,
```

```
        "top_p": 1.0,  
        "max_completions_tokens": 2048  
    }  
,  
    "error": null,  
    "metadata": {}  
}
```

# The eval run output item object



A schema representing an evaluation run output item.

**created\_at** integer

Unix timestamp (in seconds) when the evaluation run was created.

**datasource\_item** object

Details of the input data source item.

**datasource\_item\_id** integer

The identifier for the data source item.

**eval\_id** string

The identifier of the evaluation group.

---

**id** string

Unique identifier for the evaluation run output item.

---

**object** string

The type of the object. Always "eval.run.output\_item".

---

**results** array

A list of grader results for this output item.

> Show properties

---

**run\_id** string

The identifier of the evaluation run associated with this output item.

---

**sample** object

A sample containing the input and output of the evaluation run.

> Show properties

---

**status** string

The status of the evaluation run.

OBJECT The eval run output item object

```
{  
  "object": "eval.run.output_item",  
  "id": "outputitem_67abd55eb6548190bb580745d5644a33",  
  "run_id": "evalrun_67abd54d60ec8190832b46859da808f7",  
  "eval_id": "eval_67abd54d9b0081909a86353f6fb9317a",  
  "created_at": 1739314509,  
  "status": "pass",  
  "datasource_item_id": 137,  
  "datasource_item": {  
    "teacher": "To grade essays, I only check for style, content, and grammar.",  
    "student": "I am a student who is trying to write the best essay."  
  },  
  "results": [  
    {  
      "name": "String Check Grader",  
      "type": "string-check-grader",  
      "score": 1.0,  
      "passed": true,  
    }  
  ],  
  "sample": {  
    "input": [  
      {  
        "role": "system",  
        "content": "You are an evaluator bot..."  
      },  
      {  
        "role": "user",  
        "content": "You are assessing..."  
      }  
    ]  
  }  
}
```

```
        },
      ],
      "output": [
        {
          "role": "assistant",
          "content": "The rubric is not clear nor concise."
        }
      ],
      "finish_reason": "stop",
      "model": "gpt-4o-2024-08-06",
      "usage": {
        "total_tokens": 521,
        "completion_tokens": 2,
        "prompt_tokens": 519,
        "cached_tokens": 0
      },
      "error": null,
      "temperature": 1.0,
      "max_completion_tokens": 2048,
      "top_p": 1.0,
      "seed": 42
    }
  }
```

# Fine-tuning



Manage fine-tuning jobs to tailor a model to your specific training data. Related guide:

[Fine-tune models](#)

## Create fine-tuning job



```
POST https://api.openai.com/v1/fine_tuning/jobs
```

Creates a fine-tuning job which begins the process of creating a new model from a given dataset.

Response includes details of the enqueued job including job status and the name of the fine-tuned models once complete.

[Learn more about fine-tuning](#)

### Request body

**model** string Required

The name of the model to fine-tune. You can select one of the [supported models](#).

**training\_file** string Required

The ID of an uploaded file that contains training data.

See [upload file](#) for how to upload a file.

Your dataset must be formatted as a JSONL file. Additionally, you must upload your file with the purpose [fine-tune](#).

The contents of the file should differ depending on if the model uses the [chat](#), [completions](#) format, or if the fine-tuning method uses the [preference](#) format.

See the [fine-tuning guide](#) for more details.

---

**hyperparameters** Deprecated object Optional

The hyperparameters used for the fine-tuning job. This value is now deprecated in favor of [method](#), and should be passed in under the [method](#) parameter.

> Show properties

---

**integrations** array or null Optional

A list of integrations to enable for your fine-tuning job.

> Show properties

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**method** object Optional

The method used for fine-tuning.

> Show properties

---

**seed** integer or null Optional

The seed controls the reproducibility of the job. Passing in the same seed and job parameters should produce the same results, but may differ in rare cases. If a seed is not specified, one will be generated for you.

---

**suffix** string or null Optional Defaults to null

A string of up to 64 characters that will be added to your fine-tuned model name.

For example, a `suffix` of "custom-model-name" would produce a model name like

`ft:gpt-4o-mini:openai:custom-model-name:7p4lURel`.

---

**validation\_file** string or null Optional

The ID of an uploaded file that contains validation data.

If you provide this file, the data is used to generate validation metrics periodically during fine-tuning. These metrics can be viewed in the fine-tuning results file. The same data should not be present in both train and validation files.

Your dataset must be formatted as a JSONL file. You must upload your file with the purpose `fine-tune`.

See the [fine-tuning guide](#) for more details.

## Returns

---

A [fine-tuning.job](#) object.

[Default](#)[Epochs](#)[DPO](#)[Reinforcement](#)[Validation file](#)[W&B Integration](#)

## Example request

```
curl https://api.openai.com/v1/fine_tuning/jobs \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "training_file": "file-BK7bzQj3FfZFXr7DbL6xJwfo",
  "model": "gpt-4o-mini"
}'
```

## Response

```
{
  "object": "fine_tuning.job",
  "id": "ftjob-abc123",
  "model": "gpt-4o-mini-2024-07-18",
  "created_at": 1721764800,
  "fine_tuned_model": null,
  "organization_id": "org-123",
  "result_files": [],
  "status": "queued",
  "validation_file": null,
  "training_file": "file-abc123",
  "method": {
    "type": "supervised",
    "supervised": {
      "hyperparameters": {
```

```
        "batch_size": "auto",
        "learning_rate_multiplier": "auto",
        "n_epochs": "auto",
    }
}
},
"metadata": null
}
```

# List fine-tuning jobs



GET [https://api.openai.com/v1/fine\\_tuning/jobs](https://api.openai.com/v1/fine_tuning/jobs)

List your organization's fine-tuning jobs

## Query parameters

**after** string Optional

Identifier for the last job from the previous pagination request.

**limit** integer Optional Defaults to 20

Number of fine-tuning jobs to retrieve.

**metadata** object or null Optional

Optional metadata filter. To filter, use the syntax `metadata[k]=v`. Alternatively, set `metadata=null` to indicate no metadata.

## Returns

A list of paginated [fine-tuning.job](#) objects.

### Example request

```
curl https://api.openai.com/v1/fine_tuning/jobs?limit=2&metadata[key]=value \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "object": "fine_tuning.job",
      "id": "ftjob-abc123",
      "model": "gpt-4o-mini-2024-07-18",
      "created_at": 1721764800,
      "fine_tuned_model": null,
      "organization_id": "org-123",
```

```
"result_files": [],
  "status": "queued",
  "validation_file": null,
  "training_file": "file-abc123",
  "metadata": {
    "key": "value"
  }
},
{
  ...
},
{
  ...
],
  "has_more": true
}
```

# List fine-tuning events



```
GET https://api.openai.com/v1/fine_tuning/jobs/{fine_tuning_job_id}/events
```

Get status updates for a fine-tuning job.

## Path parameters

**fine\_tuning\_job\_id** string **Required**

The ID of the fine-tuning job to get events for.

## Query parameters

**after** string Optional

Identifier for the last event from the previous pagination request.

**limit** integer Optional Defaults to 20

Number of events to retrieve.

## Returns

A list of fine-tuning event objects.

### Example request

```
curl https://api.openai.com/v1/fine_tuning/jobs/ftjob-abc123/events \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

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

```
{  
    "object": "fine_tuning.job.event",  
    "id": "ft-event-ddTJfwuMVpfLXse00Am0Gqjm",  
    "created_at": 1721764800,  
    "level": "info",  
    "message": "Fine tuning job successfully completed",  
    "data": null,  
    "type": "message"  
},  
{  
    "object": "fine_tuning.job.event",  
    "id": "ft-event-tyiGuB72evQncpH87xe505Sv",  
    "created_at": 1721764800,  

```

## List fine-tuning checkpoints



```
GET https://api.openai.com/v1/fine_tuning/jobs/{fine_tuning_job_id}/checkpoints
```

List checkpoints for a fine-tuning job.

## Path parameters

**fine\_tuning\_job\_id** string Required

The ID of the fine-tuning job to get checkpoints for.

## Query parameters

**after** string Optional

Identifier for the last checkpoint ID from the previous pagination request.

**limit** integer Optional Defaults to 10

Number of checkpoints to retrieve.

## Returns

A list of fine-tuning [checkpoint objects](#) for a fine-tuning job.

### Example request

```
curl https://api.openai.com/v1/fine_tuning/jobs/ftjob-abc123/checkpoints \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{
  "object": "list",
  "data": [
    {
      "object": "fine_tuning.job.checkpoint",
      "id": "ftckpt_zc4Q7MP6XxulcVzj4MZdwsAB",
      "created_at": 1721764867,
      "fine_tuned_model_checkpoint": "ft:gpt-4o-mini-2024-07-18:my-org:custom-suffix:96olL566:ckpt-s",
      "metrics": {
        "full_valid_loss": 0.134,
        "full_valid_mean_token_accuracy": 0.874
      },
      "fine_tuning_job_id": "ftjob-abc123",
      "step_number": 2000
    },
    {
      "object": "fine_tuning.job.checkpoint",
      "id": "ftckpt_enQCFmOTGj3syEpYVhBRLTSy",
      "created_at": 1721764800,
      "fine_tuned_model_checkpoint": "ft:gpt-4o-mini-2024-07-18:my-org:custom-suffix:7q8mpxmy:ckpt-s",
      "metrics": {
        "full_valid_loss": 0.167,
        "full_valid_mean_token_accuracy": 0.781
      }
    }
  ]
}
```

```
  },
  "fine_tuning_job_id": "ftjob-abc123",
  "step_number": 1000
}
],
"first_id": "ftckpt_zc4Q7MP6XxulcVzj4MZdwsAB",
"last_id": "ftckpt_enQCFmOTGj3syEpYVhBRLTSy",
"has_more": true
}
```

## List checkpoint permissions



```
GET https://api.openai.com/v1/fine_tuning/checkpoints/{fine_tuned_model_checkpoint}/permissions
```

**NOTE:** This endpoint requires an [admin API key](#).

Organization owners can use this endpoint to view all permissions for a fine-tuned model checkpoint.

### Path parameters

**fine\_tuned\_model\_checkpoint** string Required

The ID of the fine-tuned model checkpoint to get permissions for.

## Query parameters

---

**after** string Optional

Identifier for the last permission ID from the previous pagination request.

---

**limit** integer Optional Defaults to 10

Number of permissions to retrieve.

---

**order** string Optional Defaults to descending

The order in which to retrieve permissions.

---

**project\_id** string Optional

The ID of the project to get permissions for.

---

## Returns

---

A list of fine-tuned model checkpoint permission objects for a fine-tuned model checkpoint.

Example request

```
curl https://api.openai.com/v1/fine_tuning/checkpoints/ft:gpt-4o-mini-2024-07-18:org:weather:B7R9VjC  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "checkpoint.permission",  
      "id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
      "created_at": 1721764867,  
      "project_id": "proj_abGMw111N8IrBb6SvvY5A1iH"  
    },  
    {  
      "object": "checkpoint.permission",  
      "id": "cp_enQCFmOTGj3syEpYVhBRLTSy",  
      "created_at": 1721764800,  
      "project_id": "proj_iqGMw111N8IrBb6SvvY5A1oF"  
    },  
  ],  
  "first_id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
  "last_id": "cp_enQCFmOTGj3syEpYVhBRLTSy",  
  "has_more": false  
}
```

# Create checkpoint permissions



```
POST https://api.openai.com/v1/fine_tuning/checkpoints/{fine_tuned_model_checkpoint}/permissions
```

**NOTE:** Calling this endpoint requires an [admin API key](#).

This enables organization owners to share fine-tuned models with other projects in their organization.

## Path parameters

---

**fine\_tuned\_model\_checkpoint** string Required

The ID of the fine-tuned model checkpoint to create a permission for.

## Request body

---

**project\_ids** array Required

The project identifiers to grant access to.

## Returns

A list of fine-tuned model checkpoint [permission objects](#) for a fine-tuned model checkpoint.

#### Example request

```
curl https://api.openai.com/v1/fine_tuning/checkpoints/ft:gpt-4o-mini-2024-07-18:org:weather:B7R9VjC  
-H "Authorization: Bearer $OPENAI_API_KEY"  
-d '{"project_ids": ["proj_abGMw111N8IrBb6SvvY5A1iH"]}'
```

#### Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "checkpoint.permission",  
      "id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
      "created_at": 1721764867,  
      "project_id": "proj_abGMw111N8IrBb6SvvY5A1iH"  
    }  
  ],  
  "first_id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
  "last_id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
  "has_more": false  
}
```

# Delete checkpoint permission



```
DELETE https://api.openai.com/v1/fine_tuning/checkpoints/{fine_tuned_model_checkpoint}/permissions/{permission_id}
```

**NOTE:** This endpoint requires an [admin API key](#).

Organization owners can use this endpoint to delete a permission for a fine-tuned model checkpoint.

## Path parameters

**fine\_tuned\_model\_checkpoint** string Required

The ID of the fine-tuned model checkpoint to delete a permission for.

**permission\_id** string Required

The ID of the fine-tuned model checkpoint permission to delete.

## Returns

The deletion status of the fine-tuned model checkpoint permission object.

#### Example request

```
curl https://api.openai.com/v1/fine_tuning/checkpoints/ft:gpt-4o-mini-2024-07-18:org:weather:B7R9VjC  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

#### Response

```
{  
  "object": "checkpoint.permission",  
  "id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
  "deleted": true  
}
```

## Retrieve fine-tuning job



```
GET https://api.openai.com/v1/fine_tuning/jobs/{fine_tuning_job_id}
```

Get info about a fine-tuning job.

[Learn more about fine-tuning](#)

## Path parameters

**fine\_tuning\_job\_id** string Required

The ID of the fine-tuning job.

## Returns

The [fine-tuning](#) object with the given ID.

### Example request

```
curl https://api.openai.com/v1/fine_tuning/jobs/ft-AF1WoRqd3aJAHsqc9NY7iL8F \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "object": "fine_tuning.job",
  "id": "ftjob-abc123",
  "model": "davinci-002",
  "created_at": 1692661014,
  "finished_at": 1692661190,
  "fine_tuned_model": "ft:davinci-002:my-org:custom_suffix:7q8mpxmy",
```

```
"organization_id": "org-123",
"result_files": [
    "file-abc123"
],
"status": "succeeded",
"validation_file": null,
"training_file": "file-abc123",
"hyperparameters": {
    "n_epochs": 4,
    "batch_size": 1,
    "learning_rate_multiplier": 1.0
},
"trained_tokens": 5768,
"integrations": [],
"seed": 0,
"estimated_finish": 0,
"method": {
    "type": "supervised",
    "supervised": {
        "hyperparameters": {
            "n_epochs": 4,
            "batch_size": 1,
            "learning_rate_multiplier": 1.0
        }
    }
}
}
```

# Cancel fine-tuning



```
POST https://api.openai.com/v1/fine_tuning/jobs/{fine_tuning_job_id}/cancel
```

Immediately cancel a fine-tune job.

## Path parameters

**fine\_tuning\_job\_id** string Required

The ID of the fine-tuning job to cancel.

## Returns

The cancelled [fine-tuning](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/fine_tuning/jobs/ftjob-abc123/cancel \  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "object": "fine_tuning.job",  
  "id": "ftjob-abc123",  
  "model": "gpt-4o-mini-2024-07-18",  
  "created_at": 1721764800,  
  "fine_tuned_model": null,  
  "organization_id": "org-123",  
  "result_files": [],  
  "status": "cancelled",  
  "validation_file": "file-abc123",  
  "training_file": "file-abc123"  
}
```

# Resume fine-tuning



```
POST https://api.openai.com/v1/fine_tuning/jobs/{fine_tuning_job_id}/resume
```

Resume a fine-tune job.

## Path parameters

**fine\_tuning\_job\_id** string Required

The ID of the fine-tuning job to resume.

## Returns

The resumed [fine-tuning](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/fine_tuning/jobs/ftjob-abc123/resume \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "object": "fine_tuning.job",
  "id": "ftjob-abc123",
  "model": "gpt-4o-mini-2024-07-18",
  "created_at": 1721764800,
  "fine_tuned_model": null,
  "organization_id": "org-123",
  "result_files": [],
  "status": "queued",
  "validation_file": "file-abc123",
```

```
"training_file": "file-abc123"
```

# Pause fine-tuning



```
POST https://api.openai.com/v1/fine_tuning/jobs/{fine_tuning_job_id}/pause
```

Pause a fine-tune job.

## Path parameters

**fine\_tuning\_job\_id** string Required

The ID of the fine-tuning job to pause.

## Returns

The paused fine-tuning object.

### Example request

```
curl -X POST https://api.openai.com/v1/fine_tuning/jobs/ftjob-abc123/pause \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{
  "object": "fine_tuning.job",
  "id": "ftjob-abc123",
  "model": "gpt-4o-mini-2024-07-18",
  "created_at": 1721764800,
  "fine_tuned_model": null,
  "organization_id": "org-123",
  "result_files": [],
  "status": "paused",
  "validation_file": "file-abc123",
  "training_file": "file-abc123"
}
```

# Training format for chat models using the supervised method



The per-line training example of a fine-tuning input file for chat models using the supervised method. Input messages may contain text or image content only. Audio and file input messages are not currently supported for fine-tuning.

---

**functions** Deprecated array

A list of functions the model may generate JSON inputs for.

> Show properties

---

**messages** array

> Show possible types

---

**parallel\_tool\_calls** boolean

Whether to enable parallel function calling during tool use.

---

**tools** array

A list of tools the model may generate JSON inputs for.

> Show properties

OBJECT Training format for chat models using the supervised method

```
{  
  "messages": [  
    { "role": "user", "content": "What is the weather in San Francisco?" },  
    {  
      "role": "assistant",  
      "tool_calls": [  
        ...  
      ]  
    }  
  ]  
}
```

```
{  
    "id": "call_id",  
    "type": "function",  
    "function": {  
        "name": "get_current_weather",  
        "arguments": "{\"location\": \"San Francisco, USA\", \"format\": \"celsius\"}"  
    }  
}  
]  
}  
],  
"parallel_tool_calls": false,  
"tools": [  
{  
    "type": "function",  
    "function": {  
        "name": "get_current_weather",  
        "description": "Get the current weather",  
        "parameters": {  
            "type": "object",  
            "properties": {  
                "location": {  
                    "type": "string",  
                    "description": "The city and country, eg. San Francisco, USA"  
                },  
                "format": { "type": "string", "enum": ["celsius", "fahrenheit"] }  
            },  
            "required": ["location", "format"]  
        }  
    }  
}
```

```
    }
  }
]
```

# Training format for chat models using the preference method



The per-line training example of a fine-tuning input file for chat models using the dpo method. Input messages may contain text or image content only. Audio and file input messages are not currently supported for fine-tuning.

---

## **input** object

> Show properties

---

## **non\_preferred\_output** array

The non-preferred completion message for the output.

> Show possible types

---

## **preferred\_output** array

The preferred completion message for the output.

> Show possible types

OBJECT Training format for chat models using the preference method

```
{  
  "input": {  
    "messages": [  
      { "role": "user", "content": "What is the weather in San Francisco?" }  
    ]  
  },  
  "preferred_output": [  
    {  
      "role": "assistant",  
      "content": "The weather in San Francisco is 70 degrees Fahrenheit."  
    }  
  ],  
  "non_preferred_output": [  
    {  
      "role": "assistant",  
      "content": "The weather in San Francisco is 21 degrees Celsius."  
    }  
  ]  
}
```

# Training format for reasoning models using the reinforcement method



Per-line training example for reinforcement fine-tuning. Note that `messages` and `tools` are the only reserved keywords. Any other arbitrary key-value data can be included on training datapoints and will be available to reference during grading under the `{{ item.XXX }}` template variable. Input messages may contain text or image content only. Audio and file input messages are not currently supported for fine-tuning.

---

## `messages` array

> Show possible types

---

## `tools` array

A list of tools the model may generate JSON inputs for.

> Show properties

---

OBJECT Training format for reasoning models using the reinforcement method

```
{  
  "messages": [  
    {  
      "role": "user",  
      "content": "Your task is to take a chemical in SMILES format and predict the number of hydroxyl groups."  
    },  
  ],
```

```
# Any other JSON data can be inserted into an example and referenced during RFT grading
"reference_answer": {
    "donor_bond_counts": 5,
    "acceptor_bond_counts": 7
}
```

# The fine-tuning job object



The `fine_tuning.job` object represents a fine-tuning job that has been created through the API.

## `created_at` integer

The Unix timestamp (in seconds) for when the fine-tuning job was created.

## `error` object

For fine-tuning jobs that have `failed`, this will contain more information on the cause of the failure.

> Show properties

## `estimated_finish` integer

The Unix timestamp (in seconds) for when the fine-tuning job is estimated to finish. The value will be null if the fine-tuning job is not running.

**fine\_tuned\_model** string

The name of the fine-tuned model that is being created. The value will be null if the fine-tuning job is still running.

---

**finished\_at** integer

The Unix timestamp (in seconds) for when the fine-tuning job was finished. The value will be null if the fine-tuning job is still running.

---

**hyperparameters** object

The hyperparameters used for the fine-tuning job. This value will only be returned when running `supervised` jobs.

> Show properties

---

**id** string

The object identifier, which can be referenced in the API endpoints.

---

**integrations** array

A list of integrations to enable for this fine-tuning job.

> Show possible types

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**method** object

The method used for fine-tuning.

> Show properties

---

**model** string

The base model that is being fine-tuned.

---

**object** string

The object type, which is always "fine\_tuning.job".

---

**organization\_id** string

The organization that owns the fine-tuning job.

---

**result\_files** array

The compiled results file ID(s) for the fine-tuning job. You can retrieve the results with the [Files API](#).

---

**seed** integer

The seed used for the fine-tuning job.

---

**status** string

The current status of the fine-tuning job, which can be either `validating_files` , `queued` , `running` , `succeeded` , `failed` , or `cancelled` .

---

**trained\_tokens** integer

The total number of billable tokens processed by this fine-tuning job. The value will be null if the fine-tuning job is still running.

**training\_file** string

The file ID used for training. You can retrieve the training data with the [Files API](#).

**validation\_file** string

The file ID used for validation. You can retrieve the validation results with the [Files API](#).

OBJECT The fine-tuning job object

```
{  
  "object": "fine_tuning.job",  
  "id": "ftjob-abc123",  
  "model": "davinci-002",  
  "created_at": 1692661014,  
  "finished_at": 1692661190,  
  "fine_tuned_model": "ft:davinci-002:my-org:custom_suffix:7q8mpxmy",  
  "organization_id": "org-123",  
  "result_files": [  
    "file-abc123"  
,  
    "status": "succeeded",  
    "validation_file": null,  
    "training_file": "file-abc123",  
    "hyperparameters": {  
      "n_epochs": 4,  
      "batch_size": 1,  
      "learning_rate_multiplier": 1.0  
,  
      "trained_tokens": 5768,
```

```
"integrations": [],
"seed": 0,
"estimated_finish": 0,
"method": {
  "type": "supervised",
  "supervised": {
    "hyperparameters": {
      "n_epochs": 4,
      "batch_size": 1,
      "learning_rate_multiplier": 1.0
    }
  }
},
"metadata": {
  "key": "value"
}
}
```

# The fine-tuning job event object



Fine-tuning job event object

**created\_at** integer

The Unix timestamp (in seconds) for when the fine-tuning job was created.

---

**data** object

The data associated with the event.

---

**id** string

The object identifier.

---

**level** string

The log level of the event.

---

**message** string

The message of the event.

---

**object** string

The object type, which is always "fine\_tuning.job.event".

---

**type** string

The type of event.

OBJECT The fine-tuning job event object

```
{  
  "object": "fine_tuning.job.event",  
  "id": "ftevent-abc123"  
  "created_at": 1677610602,
```

```
"level": "info",
"message": "Created fine-tuning job",
"data": {},
"type": "message"
}
```

# The fine-tuning job checkpoint object



The `fine_tuning.job.checkpoint` object represents a model checkpoint for a fine-tuning job that is ready to use.

**created\_at** integer

The Unix timestamp (in seconds) for when the checkpoint was created.

**fine\_tuned\_model\_checkpoint** string

The name of the fine-tuned checkpoint model that is created.

**fine\_tuning\_job\_id** string

The name of the fine-tuning job that this checkpoint was created from.

**id** string

The checkpoint identifier, which can be referenced in the API endpoints.

---

**metrics** object

Metrics at the step number during the fine-tuning job.

› Show properties

---

**object** string

The object type, which is always "fine\_tuning.job.checkpoint".

---

**step\_number** integer

The step number that the checkpoint was created at.

OBJECT The fine-tuning job checkpoint object

```
{  
  "object": "fine_tuning.job.checkpoint",  
  "id": "ftckpt_qtZ5Gyk4BLq1SfLFWp3RtO3P",  
  "created_at": 1712211699,  
  "fine_tuned_model_checkpoint": "ft:gpt-4o-mini-2024-07-18:my-org:custom_suffix:9ABe12dg:ckpt-step-",  
  "fine_tuning_job_id": "ftjob-fpbNQ3H1GrMehXRF8c097xTN",  
  "metrics": {  
    "step": 88,  
    "train_loss": 0.478,  
    "train_mean_token_accuracy": 0.924,  
    "valid_loss": 10.112,  
    "valid_mean_token_accuracy": 0.145,  
    "full_valid_loss": 0.567,  
    "full_valid_mean_token_accuracy": 0.944
```

```
},  
  "step_number": 88  
}
```

# The fine-tuned model checkpoint permission object



The `checkpoint.permission` object represents a permission for a fine-tuned model checkpoint.

**created\_at** integer

The Unix timestamp (in seconds) for when the permission was created.

**id** string

The permission identifier, which can be referenced in the API endpoints.

**object** string

The object type, which is always "checkpoint.permission".

**project\_id** string

The project identifier that the permission is for.

OBJECT The fine-tuned model checkpoint permission object

```
{  
  "object": "checkpoint.permission",  
  "id": "cp_zc4Q7MP6XxulcVzj4MZdwsAB",  
  "created_at": 1712211699,  
  "project_id": "proj_abGMw1llN8IrBb6SvvY5A1iH"  
}
```

## Graders



Manage and run graders in the OpenAI platform. Related guide: [Graders](#)

## String Check Grader



A StringCheckGrader object that performs a string comparison between input and reference using a specified operation.

**input** string

The input text. This may include template strings.

**name** string

The name of the grader.

**operation** string

The string check operation to perform. One of `eq`, `ne`, `like`, or `ilike`.

**reference** string

The reference text. This may include template strings.

**type** string

The object type, which is always `string_check`.

## OBJECT String Check Grader

```
{  
  "type": "string_check",  
  "name": "Example string check grader",  
  "input": "{{sample.output_text}}",  
  "reference": "{{item.label}}",  
  "operation": "eq"  
}
```

# Text Similarity Grader



A TextSimilarityGrader object which grades text based on similarity metrics.

---

**evaluation\_metric** string

The evaluation metric to use. One of `cosine`, `fuzzy_match`, `bleu`, `gleu`, `meteor`, `rouge_1`, `rouge_2`, `rouge_3`, `rouge_4`, `rouge_5`, or `rouge_l`.

---

**input** string

The text being graded.

---

**name** string

The name of the grader.

---

**reference** string

The text being graded against.

---

**type** string

The type of grader.

**OBJECT** Text Similarity Grader

```
{  
  "type": "text_similarity",  
  "name": "Example text similarity grader",  
  "input": "{{sample.output_text}}",  
  "reference": "{{item.label}}",  
  "evaluation_metric": "fuzzy_match"  
}
```

# Score Model Grader



A ScoreModelGrader object that uses a model to assign a score to the input.

**input** array

The input text. This may include template strings.

> Show properties

**model** string

The model to use for the evaluation.

**name** string

The name of the grader.

**range** array

The range of the score. Defaults to `[0, 1]`.

**sampling\_params** object

The sampling parameters for the model.

> Show properties

**type** string

The object type, which is always `score_model`.

## OBJECT Score Model Grader

```
{  
  "type": "score_model",  
  "name": "Example score model grader",  
  "input": [  
    {  
      "role": "user",  
      "content": (  
        "Score how close the reference answer is to the model answer. Score 1.0 if they are  
        " Return just a floating point score\n\n"  
        " Reference answer: {{item.label}}\n\n"  
        " Model answer: {{sample.output_text}}"  
      ),  
    }  
  ]}
```

```
],
  "model": "o4-mini-2025-04-16",
  "sampling_params": {
    "temperature": 1,
    "top_p": 1,
    "seed": 42,
    "max_completions_tokens": 32768,
    "reasoning_effort": "medium"
  },
}
```

# Label Model Grader



A LabelModelGrader object which uses a model to assign labels to each item in the evaluation.

**input** array

> Show properties

**labels** array

The labels to assign to each item in the evaluation.

**model** string

The model to use for the evaluation. Must support structured outputs.

**name** string

The name of the grader.

**passing\_labels** array

The labels that indicate a passing result. Must be a subset of labels.

**type** string

The object type, which is always `label_model`.

## OBJECT Label Model Grader

```
{  
  "name": "First label grader",  
  "type": "label_model",  
  "model": "gpt-4o-2024-08-06",  
  "input": [  
    {  
      "type": "message",  
      "role": "system",  
      "content": {  
        "type": "input_text",  
        "text": "Classify the sentiment of the following statement as one of positive, neutral, or negative."  
      }  
    },  
    {  
      "type": "message",  
      "role": "user",  
      "content": {  
        "type": "input_text",  
        "text": "The food was delicious."  
      }  
    }  
  ]  
}
```

```
"type": "message",
"role": "user",
"content": {
    "type": "input_text",
    "text": "Statement: {{item.response}}"
}
],
"passing_labels": [
    "positive"
],
"labels": [
    "positive",
    "neutral",
    "negative"
]
}
```

# Python Grader



A PythonGrader object that runs a python script on the input.

## image\_tag

string

The image tag to use for the python script.

---

**name** string

The name of the grader.

---

**source** string

The source code of the python script.

---

**type** string

The object type, which is always `python`.

#### OBJECT Python Grader

```
{  
    "type": "python",  
    "name": "Example python grader",  
    "image_tag": "2025-05-08",  
    "source": """  
  
def grade(sample: dict, item: dict) -> float:  
    """  
        Returns 1.0 if `output_text` equals `label`, otherwise 0.0.  
    """  
    output = sample.get("output_text")  
    label = item.get("label")  
    return 1.0 if output == label else 0.0
```

```
""",  
,
```

# Multi Grader



A MultiGrader object combines the output of multiple graders to produce a single score.

**calculate\_output** string

A formula to calculate the output based on grader results.

**graders** object

> Show possible types

**name** string

The name of the grader.

**type** string

The object type, which is always `multi`.

OBJECT Multi Grader

```
{  
  "type": "multi",  
  "name": "example multi grader",  
  "graders": [  
    {  
      "type": "text_similarity",  
      "name": "example text similarity grader",  
      "input": "The graded text",  
      "reference": "The reference text",  
      "evaluation_metric": "fuzzy_match"  
    },  
    {  
      "type": "string_check",  
      "name": "Example string check grader",  
      "input": "{{sample.output_text}}",  
      "reference": "{{item.label}}",  
      "operation": "eq"  
    }  
  "calculate_output": "0.5 * text_similarity_score + 0.5 * string_check_score)"  
}
```

# Run grader Beta

⌚

```
POST https://api.openai.com/v1/fine_tuning/alpha/graders/run
```

Run a grader.

## Request body

---

**grader** object Required

The grader used for the fine-tuning job.

> Show possible types

---

**model\_sample** string Required

The model sample to be evaluated. This value will be used to populate the `sample` namespace. See [the guide](#) for more details. The `output_json` variable will be populated if the model sample is a valid JSON string.

---

**item** object Optional

The dataset item provided to the grader. This will be used to populate the `item` namespace. See [the guide](#) for more details.

---

## Returns

---

The results from the grader run.

### Example request

```
curl -X POST https://api.openai.com/v1/fine_tuning/alpha/graders/run \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "grader": {
    "type": "score_model",
    "name": "Example score model grader",
    "input": [
      {
        "role": "user",
        "content": "Score how close the reference answer is to the model
answer. Score 1.0 if they are the same and 0.0 if they are different. Return just a floating point s
      }
    ],
    "model": "gpt-4o-2024-08-06",
    "sampling_params": {
      "temperature": 1,
      "top_p": 1,
      "seed": 42
    }
  },
  "item": {
    "reference_answer": "fuzzy wuzzy was a bear"
  },
  "model_sample": "fuzzy wuzzy was a bear"
}'
```

## Response

```
{  
    "reward": 1.0,  
    "metadata": {  
        "name": "Example score model grader",  
        "type": "score_model",  
        "errors": {  
            "formula_parse_error": false,  
            "sample_parse_error": false,  
            "truncated_observation_error": false,  
            "unresponsive_reward_error": false,  
            "invalid_variable_error": false,  
            "other_error": false,  
            "python_grader_server_error": false,  
            "python_grader_server_error_type": null,  
            "python_grader_runtime_error": false,  
            "python_grader_runtime_error_details": null,  
            "model_grader_server_error": false,  
            "model_grader_refusal_error": false,  
            "model_grader_parse_error": false,  
            "model_grader_server_error_details": null  
        },  
        "execution_time": 4.365238428115845,  
        "scores": {},  
        "token_usage": {  
            "prompt_tokens": 190,  
            "completion_tokens": 100  
        }  
    }  
}
```

```
"total_tokens": 324,  
"completion_tokens": 134,  
"cached_tokens": 0  
},  
"sampled_model_name": "gpt-4o-2024-08-06"  
},  
"sub_rewards": {},  
"model_grader_token_usage_per_model": {  
    "gpt-4o-2024-08-06": {  
        "prompt_tokens": 190,  
        "total_tokens": 324,  
        "completion_tokens": 134,  
        "cached_tokens": 0  
    }  
}  
}
```

## Validate grader Beta



POST [https://api.openai.com/v1/fine\\_tuning/alpha/graders/validate](https://api.openai.com/v1/fine_tuning/alpha/graders/validate)

Validate a grader.

## Request body

### grader object Required

The grader used for the fine-tuning job.

> Show possible types

## Returns

The validated grader object.

### Example request

```
curl https://api.openai.com/v1/fine_tuning/alpha/graders/validate \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "grader": {
    "type": "string_check",
    "name": "Example string check grader",
    "input": "{{sample.output_text}}",
    "reference": "{{item.label}}",
    "operation": "eq"
  }
}'
```

### Response

```
{  
  "grader": {  
    "type": "string_check",  
    "name": "Example string check grader",  
    "input": "{{sample.output_text}}",  
    "reference": "{{item.label}}",  
    "operation": "eq"  
  }  
}
```

# Batch



Create large batches of API requests for asynchronous processing. The Batch API returns completions within 24 hours for a 50% discount. Related guide: [Batch](#)

## Create batch



POST <https://api.openai.com/v1/batches>

# Creates and executes a batch from an uploaded file of requests

## Request body

---

### **completion\_window** string Required

The time frame within which the batch should be processed. Currently only `24h` is supported.

---

### **endpoint** string Required

The endpoint to be used for all requests in the batch. Currently `/v1/responses` , `/v1/chat/completions` , `/v1/embeddings` , and `/v1/completions` are supported. Note that `/v1/embeddings` batches are also restricted to a maximum of 50,000 embedding inputs across all requests in the batch.

---

### **input\_file\_id** string Required

The ID of an uploaded file that contains requests for the new batch.

See [upload file](#) for how to upload a file.

Your input file must be formatted as a [JSONL file](#), and must be uploaded with the purpose `batch` . The file can contain up to 50,000 requests, and can be up to 200 MB in size.

---

### **metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**output\_expires\_after** object Optional

The expiration policy for the output and/or error file that are generated for a batch.

> Show properties

## Returns

The created [Batch](#) object.

### Example request

```
curl https://api.openai.com/v1/batches \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "input_file_id": "file-abc123",
  "endpoint": "/v1/chat/completions",
  "completion_window": "24h"
}'
```

### Response

```
{
  "id": "batch_abc123",
  "object": "batch",
  "endpoint": "/v1/chat/completions",
  "errors": null,
```

```
"input_file_id": "file-abc123",
"completion_window": "24h",
"status": "validating",
"output_file_id": null,
"error_file_id": null,
"created_at": 1711471533,
"in_progress_at": null,
"expires_at": null,
"finalizing_at": null,
"completed_at": null,
"failed_at": null,
"expired_at": null,
"cancelling_at": null,
"cancelled_at": null,
"request_counts": {
    "total": 0,
    "completed": 0,
    "failed": 0
},
"metadata": {
    "customer_id": "user_123456789",
    "batch_description": "Nightly eval job",
}
}
```

# Retrieve batch



```
GET https://api.openai.com/v1/batches/{batch_id}
```

Retrieves a batch.

## Path parameters

**batch\_id** string Required

The ID of the batch to retrieve.

## Returns

The [Batch](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/batches/batch_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
```

### Response

```
{  
  "id": "batch_abc123",  
  "object": "batch",  
  "endpoint": "/v1/completions",  
  "errors": null,  
  "input_file_id": "file-abc123",  
  "completion_window": "24h",  
  "status": "completed",  
  "output_file_id": "file-cvaTdG",  
  "error_file_id": "file-H0WS94",  
  "created_at": 1711471533,  
  "in_progress_at": 1711471538,  
  "expires_at": 1711557933,  
  "finalizing_at": 1711493133,  
  "completed_at": 1711493163,  
  "failed_at": null,  
  "expired_at": null,  
  "cancelling_at": null,  
  "cancelled_at": null,  
  "request_counts": {  
    "total": 100,  
    "completed": 95,  
    "failed": 5  
  },  
  "metadata": {  
    "customer_id": "user_123456789",  
    "batch_description": "Nightly eval job",  
  }  
}
```

# Cancel batch



```
POST https://api.openai.com/v1/batches/{batch_id}/cancel
```

Cancels an in-progress batch. The batch will be in status `cancelling` for up to 10 minutes, before changing to `cancelled`, where it will have partial results (if any) available in the output file.

## Path parameters

**batch\_id** string Required

The ID of the batch to cancel.

## Returns

The [Batch](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/batches/batch_abc123/cancel \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-X POST
```

## Response

```
{
  "id": "batch_abc123",
  "object": "batch",
  "endpoint": "/v1/chat/completions",
  "errors": null,
  "input_file_id": "file-abc123",
  "completion_window": "24h",
  "status": "cancelling",
  "output_file_id": null,
  "error_file_id": null,
  "created_at": 1711471533,
  "in_progress_at": 1711471538,
  "expires_at": 1711557933,
  "finalizing_at": null,
  "completed_at": null,
  "failed_at": null,
  "expired_at": null,
  "canceling_at": 1711475133,
  "cancelled_at": null,
  "request_counts": {
    "total": 100,
```

```
"completed": 23,  
  "failed": 1  
},  
  "metadata": {  
    "customer_id": "user_123456789",  
    "batch_description": "Nightly eval job",  
  }  
}
```

# List batch



GET <https://api.openai.com/v1/batches>

List your organization's batches.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include `after=obj_foo` in order to fetch the next page of the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

A list of paginated Batch objects.

### Example request

```
curl https://api.openai.com/v1/batches?limit=2 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "id": "batch_abc123",
      "object": "batch",
      "endpoint": "/v1/chat/completions",
      "errors": null,
      "input_file_id": "file-abc123",
      "completion_window": "24h",
```

```
"status": "completed",
"output_file_id": "file-cvaTdG",
"error_file_id": "file-HOWS94",
"created_at": 1711471533,
"in_progress_at": 1711471538,
"expires_at": 1711557933,
"finalizing_at": 1711493133,
"completed_at": 1711493163,
"failed_at": null,
"expired_at": null,
"cancelling_at": null,
"cancelled_at": null,
"request_counts": {
    "total": 100,
    "completed": 95,
    "failed": 5
},
"metadata": {
    "customer_id": "user_123456789",
    "batch_description": "Nightly job",
}
},
{ ... },
],
"first_id": "batch_abc123",
"last_id": "batch_abc456",
"has_more": true
}
```

# The batch object



**cancelled\_at** integer

The Unix timestamp (in seconds) for when the batch was cancelled.

---

**cancelling\_at** integer

The Unix timestamp (in seconds) for when the batch started cancelling.

---

**completed\_at** integer

The Unix timestamp (in seconds) for when the batch was completed.

---

**completion\_window** string

The time frame within which the batch should be processed.

---

**created\_at** integer

The Unix timestamp (in seconds) for when the batch was created.

---

**endpoint** string

The OpenAI API endpoint used by the batch.

---

**error\_file\_id** string

The ID of the file containing the outputs of requests with errors.

---

**errors** object

> Show properties

---

**expired\_at** integer

The Unix timestamp (in seconds) for when the batch expired.

---

**expires\_at** integer

The Unix timestamp (in seconds) for when the batch will expire.

---

**failed\_at** integer

The Unix timestamp (in seconds) for when the batch failed.

---

**finalizing\_at** integer

The Unix timestamp (in seconds) for when the batch started finalizing.

---

**id** string**in\_progress\_at** integer

The Unix timestamp (in seconds) for when the batch started processing.

---

**input\_file\_id** string

The ID of the input file for the batch.

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string

Model ID used to process the batch, like `gpt-5-2025-08-07`. OpenAI offers a wide range of models with different capabilities, performance characteristics, and price points. Refer to the [model guide](#) to browse and compare available models.

---

**object** string

The object type, which is always `batch`.

---

**output\_file\_id** string

The ID of the file containing the outputs of successfully executed requests.

---

**request\_counts** object

The request counts for different statuses within the batch.

> Show properties

---

**status** string

The current status of the batch.

**usage** object

Represents token usage details including input tokens, output tokens, a breakdown of output tokens, and the total tokens used.

Only populated on batches created after September 7, 2025.

> Show properties

OBJECT The batch object

```
{  
  "id": "batch_abc123",  
  "object": "batch",  
  "endpoint": "/v1/completions",  
  "model": "gpt-5-2025-08-07",  
  "errors": null,  
  "input_file_id": "file-abc123",  
  "completion_window": "24h",  
  "status": "completed",  
  "output_file_id": "file-cvaTdG",  
  "error_file_id": "file-HOWS94",  
  "created_at": 1711471533,  
  "in_progress_at": 1711471538,  
  "expires_at": 1711557933,  
  "finalizing_at": 1711493133,  
  "completed_at": 1711493163,  
  "failed_at": null,  
  "expired_at": null,  
  "cancelling_at": null,  
  "cancelled_at": null,  
  "request_counts": {
```

```
"total": 100,  
"completed": 95,  
"failed": 5  
},  
"usage": {  
    "input_tokens": 1500,  
    "input_tokens_details": {  
        "cached_tokens": 1024  
    },  
    "output_tokens": 500,  
    "output_tokens_details": {  
        "reasoning_tokens": 300  
    },  
    "total_tokens": 2000  
},  
"metadata": {  
    "customer_id": "user_123456789",  
    "batch_description": "Nightly eval job",  
}  
}
```

# The request input object



The per-line object of the batch input file

**custom\_id** string

A developer-provided per-request id that will be used to match outputs to inputs. Must be unique for each request in a batch.

**method** string

The HTTP method to be used for the request. Currently only `POST` is supported.

**url** string

The OpenAI API relative URL to be used for the request. Currently `/v1/chat/completions`, `/v1/embeddings`, and `/v1/completions` are supported.

OBJECT The request input object

```
{"custom_id": "request-1", "method": "POST", "url": "/v1/chat/completions", "body": {"model": "gpt-4"}}
```

## The request output object



The per-line object of the batch output and error files

**custom\_id** string

A developer-provided per-request id that will be used to match outputs to inputs.

**error** object

For requests that failed with a non-HTTP error, this will contain more information on the cause of the failure.

> Show properties

**id** string**response** object

> Show properties

OBJECT The request output object

```
{"id": "batch_req_wnaDys", "custom_id": "request-2", "response": {"status_code": 200, "request_id":
```

# Files



Files are used to upload documents that can be used with features like [Assistants](#), [Fine-tuning](#), and [Batch API](#).

# Upload file



```
POST https://api.openai.com/v1/files
```

Upload a file that can be used across various endpoints. Individual files can be up to 512 MB, and the size of all files uploaded by one organization can be up to 1 TB.

The Assistants API supports files up to 2 million tokens and of specific file types. See the [Assistants Tools guide](#) for details.

The Fine-tuning API only supports `.jsonl` files. The input also has certain required formats for fine-tuning [chat](#) or [completions](#) models.

The Batch API only supports `.jsonl` files up to 200 MB in size. The input also has a specific required [format](#).

Please [contact us](#) if you need to increase these storage limits.

## Request body

### **file** file Required

The File object (not file name) to be uploaded.

---

### **purpose** string Required

The intended purpose of the uploaded file. One of: - `assistants` : Used in the Assistants API - `batch` : Used in the Batch API - `fine-tune` : Used for fine-tuning - `vision` : Images used for vision fine-tuning - `user_data` : Flexible file type for any purpose - `evals` : Used for eval data sets

#### `expires_after` object Optional

The expiration policy for a file. By default, files with `purpose=batch` expire after 30 days and all other files are persisted until they are manually deleted.

> Show properties

## Returns

The uploaded [File](#) object.

#### Example request

```
curl https://api.openai.com/v1/files \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-F purpose="fine-tune" \
-F file="@mydata.jsonl"
-F expires_after[anchor]="created_at"
-F expires_after[seconds]=2592000
```

#### Response

```
{  
  "id": "file-abc123",  
  "object": "file",  
  "bytes": 120000,  
  "created_at": 1677610602,  
  "expires_at": 1677614202,  
  "filename": "mydata.jsonl",  
  "purpose": "fine-tune",  
}
```

## List files



GET <https://api.openai.com/v1/files>

Returns a list of files.

### Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with `obj_foo`, your subsequent call can include `after=obj_foo` in order to fetch the next page of

the list.

---

**limit** integer Optional Defaults to 10000

A limit on the number of objects to be returned. Limit can range between 1 and 10,000, and the default is 10,000.

---

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

---

**purpose** string Optional

Only return files with the given purpose.

## Returns

---

A list of [File](#) objects.

Example request

```
curl https://api.openai.com/v1/files \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

Response

```
{  
  "object": "list",
```

```
"data": [
  {
    "id": "file-abc123",
    "object": "file",
    "bytes": 175,
    "created_at": 1613677385,
    "expires_at": 1677614202,
    "filename": "salesOverview.pdf",
    "purpose": "assistants",
  },
  {
    "id": "file-abc456",
    "object": "file",
    "bytes": 140,
    "created_at": 1613779121,
    "expires_at": 1677614202,
    "filename": "puppy.jsonl",
    "purpose": "fine-tune",
  }
],
"first_id": "file-abc123",
"last_id": "file-abc456",
"has_more": false
}
```

# Retrieve file



```
GET https://api.openai.com/v1/files/{file_id}
```

Returns information about a specific file.

## Path parameters

**file\_id** string Required

The ID of the file to use for this request.

## Returns

The [File](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/files/file-abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{  
  "id": "file-abc123",  
  "object": "file",  
  "bytes": 120000,  
  "created_at": 1677610602,  
  "expires_at": 1677614202,  
  "filename": "mydata.jsonl",  
  "purpose": "fine-tune",  
}
```

## Delete file



```
DELETE https://api.openai.com/v1/files/{file_id}
```

Delete a file and remove it from all vector stores.

### Path parameters

**file\_id** string Required

The ID of the file to use for this request.

## Returns

Deletion status.

### Example request

```
curl https://api.openai.com/v1/files/file-abc123 \
-X DELETE \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "id": "file-abc123",
  "object": "file",
  "deleted": true
}
```

## Retrieve file content



```
GET https://api.openai.com/v1/files/{file_id}/content
```

Returns the contents of the specified file.

## Path parameters

---

### file\_id string Required

The ID of the file to use for this request.

## Returns

---

The file content.

### Example request

```
curl https://api.openai.com/v1/files/file-abc123/content \
-H "Authorization: Bearer $OPENAI_API_KEY" > file.jsonl
```

# The file object



The `File` object represents a document that has been uploaded to OpenAI.

**bytes** integer

The size of the file, in bytes.

---

**created\_at** integer

The Unix timestamp (in seconds) for when the file was created.

---

**expires\_at** integer

The Unix timestamp (in seconds) for when the file will expire.

---

**filename** string

The name of the file.

---

**id** string

The file identifier, which can be referenced in the API endpoints.

---

**object** string

The object type, which is always `file`.

---

**purpose** string

The intended purpose of the file. Supported values are `assistants`, `assistants_output`, `batch`, `batch_output`, `fine-tune`, `fine-tune-results`, `vision`, and `user_data`.

---

**status** Deprecated string

Deprecated. The current status of the file, which can be either `uploaded`, `processed`, or `error`.

### **status\_details** Deprecated string

Deprecated. For details on why a fine-tuning training file failed validation, see the `error` field on `fine_tuning.job`.

OBJECT The file object

```
{  
  "id": "file-abc123",  
  "object": "file",  
  "bytes": 120000,  
  "created_at": 1677610602,  
  "expires_at": 1680202602,  
  "filename": "salesOverview.pdf",  
  "purpose": "assistants",  
}
```

## Uploads



Allows you to upload large files in multiple parts.

# Create upload



```
POST https://api.openai.com/v1/uploads
```

Creates an intermediate [Upload](#) object that you can add [Parts](#) to. Currently, an Upload can accept at most 8 GB in total and expires after an hour after you create it.

Once you complete the Upload, we will create a [File](#) object that contains all the parts you uploaded. This File is usable in the rest of our platform as a regular File object.

For certain `purpose` values, the correct `mime_type` must be specified. Please refer to documentation for the [supported MIME types for your use case](#).

For guidance on the proper filename extensions for each purpose, please follow the documentation on [creating a File](#).

## Request body

**bytes** integer Required

The number of bytes in the file you are uploading.

---

**filename** string Required

The name of the file to upload.

**mime\_type** string Required

The MIME type of the file.

This must fall within the supported MIME types for your file purpose. See the supported MIME types for assistants and vision.

**purpose** string Required

The intended purpose of the uploaded file.

See the [documentation on File purposes](#).

**expires\_after** object Optional

The expiration policy for a file. By default, files with `purpose=batch` expire after 30 days and all other files are persisted until they are manually deleted.

> Show properties

## Returns

The [Upload](#) object with status `pending`.

### Example request

```
curl https://api.openai.com/v1/uploads \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "purpose": "fine-tune",
  "filename": "training_examples.jsonl",
```

```
"bytes": 2147483648,  
"mime_type": "text/jsonl",  
"expires_after": {  
    "anchor": "created_at",  
    "seconds": 3600  
}  
}'
```

## Response

```
{  
    "id": "upload_abc123",  
    "object": "upload",  
    "bytes": 2147483648,  
    "created_at": 1719184911,  
    "filename": "training_examples.jsonl",  
    "purpose": "fine-tune",  
    "status": "pending",  
    "expires_at": 1719127296  
}
```

# Add upload part



```
POST https://api.openai.com/v1/uploads/{upload_id}/parts
```

Adds a Part to an Upload object. A Part represents a chunk of bytes from the file you are trying to upload.

Each Part can be at most 64 MB, and you can add Parts until you hit the Upload maximum of 8 GB.

It is possible to add multiple Parts in parallel. You can decide the intended order of the Parts when you complete the Upload.

## Path parameters

---

**upload\_id** string Required

The ID of the Upload.

## Request body

---

**data** file Required

The chunk of bytes for this Part.

## Returns

---

The upload Part object.

### Example request

```
curl https://api.openai.com/v1/uploads/upload_abc123/parts  
-F data="aHR0cHM6Ly9hcGkub3BlbmFpLmNvbS92MS91cGxvYWRz..."
```

### Response

```
{  
  "id": "part_def456",  
  "object": "upload.part",  
  "created_at": 1719185911,  
  "upload_id": "upload_abc123"  
}
```

## Complete upload



```
POST https://api.openai.com/v1/uploads/{upload_id}/complete
```

Completes the [Upload](#).

Within the returned Upload object, there is a nested [File](#) object that is ready to use in the rest of the platform.

You can specify the order of the Parts by passing in an ordered list of the Part IDs.

The number of bytes uploaded upon completion must match the number of bytes initially specified when creating the Upload object. No Parts may be added after an Upload is completed.

## Path parameters

---

**upload\_id** string Required

The ID of the Upload.

## Request body

---

**part\_ids** array Required

The ordered list of Part IDs.

**md5** string Optional

The optional md5 checksum for the file contents to verify if the bytes uploaded matches what you expect.

## Returns

---

The [Upload](#) object with status `completed` with an additional `file` property containing the created usable File object.

### Example request

```
curl https://api.openai.com/v1/uploads/upload_abc123/complete  
-d '{  
  "part_ids": ["part_def456", "part_ghi789"]  
}'
```

## Response

```
{  
  "id": "upload_abc123",  
  "object": "upload",  
  "bytes": 2147483648,  
  "created_at": 1719184911,  
  "filename": "training_examples.jsonl",  
  "purpose": "fine-tune",  
  "status": "completed",  
  "expires_at": 1719127296,  
  "file": {  
    "id": "file-xyz321",  
    "object": "file",  
    "bytes": 2147483648,  
    "created_at": 1719186911,  
    "expires_at": 1719127296,  
    "filename": "training_examples.jsonl",  
    "purpose": "fine-tune",  
  }  
}
```

# Cancel upload



```
POST https://api.openai.com/v1/uploads/{upload_id}/cancel
```

Cancels the Upload. No Parts may be added after an Upload is cancelled.

## Path parameters

**upload\_id** string Required

The ID of the Upload.

## Returns

The [Upload](#) object with status `cancelled`.

### Example request

```
curl https://api.openai.com/v1/uploads/upload_abc123/cancel
```

## Response

```
{  
  "id": "upload_abc123",  
  "object": "upload",  
  "bytes": 2147483648,  
  "created_at": 1719184911,  
  "filename": "training_examples.jsonl",  
  "purpose": "fine-tune",  
  "status": "cancelled",  
  "expires_at": 1719127296  
}
```

# The upload object



The Upload object can accept byte chunks in the form of Parts.

### **bytes** integer

The intended number of bytes to be uploaded.

### **created\_at** integer

The Unix timestamp (in seconds) for when the Upload was created.

**expires\_at** integer

The Unix timestamp (in seconds) for when the Upload will expire.

**file** undefined or null

The ready File object after the Upload is completed.

**filename** string

The name of the file to be uploaded.

**id** string

The Upload unique identifier, which can be referenced in API endpoints.

**object** string

The object type, which is always "upload".

**purpose** string

The intended purpose of the file. [Please refer here](#) for acceptable values.

**status** string

The status of the Upload.

OBJECT The upload object

{

  "id": "upload\_abc123",

```
"object": "upload",
"bytes": 2147483648,
"created_at": 1719184911,
"filename": "training_examples.jsonl",
"purpose": "fine-tune",
"status": "completed",
"expires_at": 1719127296,
"file": {
  "id": "file-xyz321",
  "object": "file",
  "bytes": 2147483648,
  "created_at": 1719186911,
  "filename": "training_examples.jsonl",
  "purpose": "fine-tune",
}
}
```

## The upload part object



The upload Part represents a chunk of bytes we can add to an Upload object.

**created\_at** integer

The Unix timestamp (in seconds) for when the Part was created.

**id** string

The upload Part unique identifier, which can be referenced in API endpoints.

**object** string

The object type, which is always `upload.part`.

**upload\_id** string

The ID of the Upload object that this Part was added to.

OBJECT The upload part object

```
{  
  "id": "part_def456",  
  "object": "upload.part",  
  "created_at": 1719186911,  
  "upload_id": "upload_abc123"  
}
```

# Models



List and describe the various models available in the API. You can refer to the [Models](#) documentation to understand what models are available and the differences between them.

---

## List models



```
GET https://api.openai.com/v1/models
```

Lists the currently available models, and provides basic information about each one such as the owner and availability.

### Returns

---

A list of [model](#) objects.

#### Example request

```
curl https://api.openai.com/v1/models \
```

## Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "model-id-0",  
      "object": "model",  
      "created": 1686935002,  
      "owned_by": "organization-owner"  
    },  
    {  
      "id": "model-id-1",  
      "object": "model",  
      "created": 1686935002,  
      "owned_by": "organization-owner",  
    },  
    {  
      "id": "model-id-2",  
      "object": "model",  
      "created": 1686935002,  
      "owned_by": "openai"  
    },  
  ],  
  "object": "list"  
}
```

# Retrieve model



```
GET https://api.openai.com/v1/models/{model}
```

Retrieves a model instance, providing basic information about the model such as the owner and permissioning.

## Path parameters

---

**model** string Required

The ID of the model to use for this request

## Returns

---

The [model](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/models/gpt-5 \  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "id": "gpt-5",  
  "object": "model",  
  "created": 1686935002,  
  "owned_by": "openai"  
}
```

# Delete a fine-tuned model



```
DELETE https://api.openai.com/v1/models/{model}
```

Delete a fine-tuned model. You must have the Owner role in your organization to delete a model.

## Path parameters

**model** string Required

The model to delete

## Returns

Deletion status.

#### Example request

```
curl https://api.openai.com/v1/models/ft:gpt-4o-mini:acemeco:suffix:abc123 \
-X DELETE \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

#### Response

```
{
  "id": "ft:gpt-4o-mini:acemeco:suffix:abc123",
  "object": "model",
  "deleted": true
}
```

## The model object



Describes an OpenAI model offering that can be used with the API.

**created** integer

The Unix timestamp (in seconds) when the model was created.

**id** string

The model identifier, which can be referenced in the API endpoints.

**object** string

The object type, which is always "model".

**owned\_by** string

The organization that owns the model.

OBJECT The model object

```
{  
  "id": "gpt-5",  
  "object": "model",  
  "created": 1686935002,  
  "owned_by": "openai"  
}
```

# Moderations



Given text and/or image inputs, classifies if those inputs are potentially harmful across several categories. Related guide: [Moderations](#)

---

## Create moderation



```
POST https://api.openai.com/v1/moderations
```

Classifies if text and/or image inputs are potentially harmful. Learn more in the [moderation guide](#).

### Request body

---

**input** string or array Required

Input (or inputs) to classify. Can be a single string, an array of strings, or an array of multi-modal input objects similar to other models.

> Show possible types

---

**model** string Optional Defaults to omni-moderation-latest

The content moderation model you would like to use. Learn more in [the moderation guide](#), and learn about available models [here](#).

## Returns

A [moderation object](#).

**Single string**

**Image and text**

Example request

```
curl https://api.openai.com/v1/moderations \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{
  "input": "I want to kill them."
}'
```

Response

```
{
  "id": "modr-AB8CjOTu2jiq12hp1AQPfeqFWaORR",
  "model": "text-moderation-007",
  "results": [
    {
      "flagged": true,
      "categories": {
```

```
"sexual": false,  
"hate": false,  
"harassment": true,  
"self-harm": false,  
"sexual/minors": false,  
"hate/threatening": false,  
"violence/graphic": false,  
"self-harm/intent": false,  
"self-harm/instructions": false,  
"harassment/threatening": true,  
"violence": true  
},  
"category_scores": {  
    "sexual": 0.000011726012417057063,  
    "hate": 0.22706663608551025,  
    "harassment": 0.5215635299682617,  
    "self-harm": 2.227119921371923e-6,  
    "sexual/minors": 7.107352217872176e-8,  
    "hate/threatening": 0.023547329008579254,  
    "violence/graphic": 0.00003391829886822961,  
    "self-harm/intent": 1.646940972932498e-6,  
    "self-harm/instructions": 1.1198755256458526e-9,  
    "harassment/threatening": 0.5694745779037476,  
    "violence": 0.9971134662628174  
}  
}  
]  
}
```

# The moderation object



Represents if a given text input is potentially harmful.

**id** string

The unique identifier for the moderation request.

**model** string

The model used to generate the moderation results.

**results** array

A list of moderation objects.

> Show properties

OBJECT The moderation object

```
{  
  "id": "modr-0d9740456c391e43c445bf0f010940c7",  
  "model": "omni-moderation-latest",  
  "results": [  
    {
```

```
"flagged": true,  
"categories": {  
    "harassment": true,  
    "harassment/threatening": true,  
    "sexual": false,  
    "hate": false,  
    "hate/threatening": false,  
    "illicit": false,  
    "illicit/violent": false,  
    "self-harm/intent": false,  
    "self-harm/instructions": false,  
    "self-harm": false,  
    "sexual/minors": false,  
    "violence": true,  
    "violence/graphic": true  
},  
"category_scores": {  
    "harassment": 0.8189693396524255,  
    "harassment/threatening": 0.804985420696006,  
    "sexual": 1.573112165348997e-6,  
    "hate": 0.007562942636942845,  
    "hate/threatening": 0.004208854591835476,  
    "illicit": 0.030535955153511665,  
    "illicit/violent": 0.008925306722380033,  
    "self-harm/intent": 0.00023023930975076432,  
    "self-harm/instructions": 0.0002293869201073356,  
    "self-harm": 0.012598046106750154,  
    "sexual/minors": 2.212566909570261e-8,  
    "violence": 0.9999992735124786,
```

```
"violence/graphic": 0.843064871157054
},
"category_applied_input_types": {
    "harassment": [
        "text"
    ],
    "harassment/threatening": [
        "text"
    ],
    "sexual": [
        "text",
        "image"
    ],
    "hate": [
        "text"
    ],
    "hate/threatening": [
        "text"
    ],
    "illicit": [
        "text"
    ],
    "illicit/violent": [
        "text"
    ],
    "self-harm/intent": [
        "text",
        "image"
    ],
}
```

```
"self-harm/instructions": [
    "text",
    "image"
],
"self-harm": [
    "text",
    "image"
],
"sexual/minors": [
    "text"
],
"violence": [
    "text",
    "image"
],
"violence/graphic": [
    "text",
    "image"
]
}
]
}
]
```

# Vector stores



Vector stores power semantic search for the Retrieval API and the `file_search` tool in the Responses and Assistants APIs.

Related guide: [File Search](#)

---

## Create vector store



POST `https://api.openai.com/v1/vector_stores`

Create a vector store.

### Request body

---

`chunking_strategy` object Optional

The chunking strategy used to chunk the file(s). If not set, will use the `auto` strategy. Only applicable if `file_ids` is non-empty.

> Show possible types

**description** string Optional

A description for the vector store. Can be used to describe the vector store's purpose.

**expires\_after** object Optional

The expiration policy for a vector store.

> Show properties

**file\_ids** array Optional

A list of [File](#) IDs that the vector store should use. Useful for tools like [file\\_search](#) that can access files.

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**name** string Optional

The name of the vector store.

## Returns

A [vector store](#) object.

### Example request

```
curl https://api.openai.com/v1/vector_stores \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "name": "Support FAQ"
}'
```

## Response

```
{
  "id": "vs_abc123",
  "object": "vector_store",
  "created_at": 1699061776,
  "name": "Support FAQ",
  "description": "Contains commonly asked questions and answers, organized by topic.",
  "bytes": 139920,
  "file_counts": {
    "in_progress": 0,
    "completed": 3,
    "failed": 0,
    "cancelled": 0,
    "total": 3
  }
}
```

# List vector stores



```
GET https://api.openai.com/v1/vector_stores
```

Returns a list of vector stores.

## Query parameters

---

### **after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

---

### **before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the previous page of the list.

---

### **limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

A list of vector store objects.

### Example request

```
curl https://api.openai.com/v1/vector_stores \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "id": "vs_abc123",
      "object": "vector_store",
      "created_at": 1699061776,
      "name": "Support FAQ",
```

```
"description": "Contains commonly asked questions and answers, organized by topic.",  
"bytes": 139920,  
"file_counts": {  
    "in_progress": 0,  
    "completed": 3,  
    "failed": 0,  
    "cancelled": 0,  
    "total": 3  
}  
},  
{  
    "id": "vs_abc456",  
    "object": "vector_store",  
    "created_at": 1699061776,  
    "name": "Support FAQ v2",  
    "description": null,  
    "bytes": 139920,  
    "file_counts": {  
        "in_progress": 0,  
        "completed": 3,  
        "failed": 0,  
        "cancelled": 0,  
        "total": 3  
    }  
}  
,  
{"first_id": "vs_abc123",  
"last_id": "vs_abc456",
```

```
"has_more": false  
}
```

# Retrieve vector store



```
GET https://api.openai.com/v1/vector_stores/{vector_store_id}
```

Retrieves a vector store.

## Path parameters

**vector\_store\_id** string Required

The ID of the vector store to retrieve.

## Returns

The vector store object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

## Response

```
{  
  "id": "vs_abc123",  
  "object": "vector_store",  
  "created_at": 1699061776  
}
```

# Modify vector store



```
POST https://api.openai.com/v1/vector_stores/{vector_store_id}
```

Modifies a vector store.

## Path parameters

**vector\_store\_id** string Required

The ID of the vector store to modify.

## Request body

**expires\_after** object or null Optional

The expiration policy for a vector store.

> Show properties

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**name** string or null Optional

The name of the vector store.

## Returns

The modified [vector store](#) object.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
-d '{
  "name": "Support FAQ"
}'
```

## Response

```
{
  "id": "vs_abc123",
  "object": "vector_store",
  "created_at": 1699061776,
  "name": "Support FAQ",
  "description": "Contains commonly asked questions and answers, organized by topic.",
  "bytes": 139920,
  "file_counts": {
    "in_progress": 0,
    "completed": 3,
    "failed": 0,
    "cancelled": 0,
    "total": 3
  }
}
```

# Delete vector store



```
DELETE https://api.openai.com/v1/vector_stores/{vector_store_id}
```

Delete a vector store.

## Path parameters

**vector\_store\_id** string Required

The ID of the vector store to delete.

## Returns

Deletion status

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-X DELETE
```

## Response

```
{  
  id: "vs_abc123",  
  object: "vector_store.deleted",  
  deleted: true  
}
```

# Search vector store



POST [https://api.openai.com/v1/vector\\_stores/{vector\\_store\\_id}/search](https://api.openai.com/v1/vector_stores/{vector_store_id}/search)

Search a vector store for relevant chunks based on a query and file attributes filter.

## Path parameters

**vector\_store\_id** string Required

The ID of the vector store to search.

## Request body

**query** string or array Required

A query string for a search

---

**filters** object Optional

A filter to apply based on file attributes.

> Show possible types

---

**max\_num\_results** integer Optional Defaults to 10

The maximum number of results to return. This number should be between 1 and 50 inclusive.

---

**ranking\_options** object Optional

Ranking options for search.

> Show properties

---

**rewrite\_query** boolean Optional Defaults to false

Whether to rewrite the natural language query for vector search.

---

## Returns

---

A page of search results from the vector store.

Example request

```
curl -X POST \
https://api.openai.com/v1/vector_stores/vs_abc123/search \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"query": "What is the return policy?", "filters": {...}}'
```

## Response

```
{
  "object": "vector_store.search_results.page",
  "search_query": "What is the return policy?",
  "data": [
    {
      "file_id": "file_123",
      "filename": "document.pdf",
      "score": 0.95,
      "attributes": {
        "author": "John Doe",
        "date": "2023-01-01"
      },
      "content": [
        {
          "type": "text",
          "text": "Relevant chunk"
        }
      ]
    },
    {
      "file_id": "file_456",
      "filename": "report.pdf",
      "score": 0.85,
      "attributes": {
        "author": "Jane Doe",
        "date": "2023-01-02"
      },
      "content": [
        {
          "type": "text",
          "text": "Another relevant chunk"
        }
      ]
    }
  ],
  "total": 2
}
```

```
"file_id": "file_456",
"filename": "notes.txt",
"score": 0.89,
"attributes": {
    "author": "Jane Smith",
    "date": "2023-01-02"
},
"content": [
    {
        "type": "text",
        "text": "Sample text content from the vector store."
    }
]
},
"has_more": false,
"next_page": null
}
```

# The vector store object



A vector store is a collection of processed files can be used by the `file_search` tool.

`created_at` integer

The Unix timestamp (in seconds) for when the vector store was created.

---

**expires\_after** object

The expiration policy for a vector store.

> Show properties

---

**expires\_at** integer

The Unix timestamp (in seconds) for when the vector store will expire.

---

**file\_counts** object

> Show properties

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**last\_active\_at** integer

The Unix timestamp (in seconds) for when the vector store was last active.

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**name** string

The name of the vector store.

**object** string

The object type, which is always `vector_store`.

**status** string

The status of the vector store, which can be either `expired`, `in_progress`, or `completed`. A status of `completed` indicates that the vector store is ready for use.

**usage\_bytes** integer

The total number of bytes used by the files in the vector store.

OBJECT The vector store object

```
{  
  "id": "vs_123",  
  "object": "vector_store",  
  "created_at": 1698107661,  
  "usage_bytes": 123456,  
  "last_active_at": 1698107661,  
  "name": "my_vector_store",  
  "status": "completed",  
  "file_counts": {  
    "in_progress": 0,  
    "completed": 100,  
    "cancelled": 0,  
    "failed": 0,  
  },  
}
```

```
"total": 100
},
"last_used_at": 1698107661
}
```

## Vector store files



Vector store files represent files inside a vector store.

Related guide: [File Search](#)

## Create vector store file



```
POST https://api.openai.com/v1/vector_stores/{vector_store_id}/files
```

Create a vector store file by attaching a [File](#) to a [vector store](#).

### Path parameters

**vector\_store\_id** string Required

The ID of the vector store for which to create a File.

## Request body

**file\_id** string Required

A File ID that the vector store should use. Useful for tools like `file_search` that can access files.

**attributes** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard. Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters, booleans, or numbers.

**chunking\_strategy** object Optional

The chunking strategy used to chunk the file(s). If not set, will use the `auto` strategy.

> Show possible types

## Returns

A vector store file object.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
    "file_id": "file-abc123"
}'
```

## Response

```
{
  "id": "file-abc123",
  "object": "vector_store.file",
  "created_at": 1699061776,
  "usage_bytes": 1234,
  "vector_store_id": "vs_abcd",
  "status": "completed",
  "last_error": null
}
```

# List vector store files



```
GET https://api.openai.com/v1/vector_stores/{vector_store_id}/files
```

Returns a list of vector store files.

## Path parameters

**vector\_store\_id** string Required

The ID of the vector store that the files belong to.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

**before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the previous page of the list.

**filter** string Optional

Filter by file status. One of `in_progress`, `completed`, `failed`, `cancelled`.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

A list of [vector store file](#) objects.

Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

Response

```
{  
  "object": "list",
```

```
"data": [
  {
    "id": "file-abc123",
    "object": "vector_store.file",
    "created_at": 1699061776,
    "vector_store_id": "vs_abc123"
  },
  {
    "id": "file-abc456",
    "object": "vector_store.file",
    "created_at": 1699061776,
    "vector_store_id": "vs_abc123"
  }
],
"first_id": "file-abc123",
"last_id": "file-abc456",
"has_more": false
}
```

## Retrieve vector store file



```
GET https://api.openai.com/v1/vector_stores/{vector_store_id}/files/{file_id}
```

Retrieves a vector store file.

## Path parameters

### file\_id string Required

The ID of the file being retrieved.

### vector\_store\_id string Required

The ID of the vector store that the file belongs to.

## Returns

The vector store file object.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files/file-abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{  
  "id": "file-abc123",
```

```
"object": "vector_store.file",
"created_at": 1699061776,
"vector_store_id": "vs_abcd",
"status": "completed",
"last_error": null
}
```

## Retrieve vector store file content



```
GET https://api.openai.com/v1/vector_stores/{vector_store_id}/files/{file_id}/content
```

Retrieve the parsed contents of a vector store file.

### Path parameters

**file\_id** string Required

The ID of the file within the vector store.

**vector\_store\_id** string Required

The ID of the vector store.

## Returns

The parsed contents of the specified vector store file.

### Example request

```
curl \  
https://api.openai.com/v1/vector_stores/vs_abc123/files/file-abc123/content \  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{  
  "file_id": "file-abc123",  
  "filename": "example.txt",  
  "attributes": {"key": "value"},  
  "content": [  
    {"type": "text", "text": "..."},  
    ...  
  ]  
}
```

# Update vector store file attributes



```
POST https://api.openai.com/v1/vector_stores/{vector_store_id}/files/{file_id}
```

Update attributes on a vector store file.

## Path parameters

**file\_id** string Required

The ID of the file to update attributes.

**vector\_store\_id** string Required

The ID of the vector store the file belongs to.

## Request body

**attributes** map Required

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard. Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters, booleans, or numbers.

## Returns

The updated vector store file object.

#### Example request

```
curl https://api.openai.com/v1/vector_stores/{vector_store_id}/files/{file_id} \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"attributes": {"key1": "value1", "key2": 2}}'
```

#### Response

```
{
  "id": "file-abc123",
  "object": "vector_store.file",
  "usage_bytes": 1234,
  "created_at": 1699061776,
  "vector_store_id": "vs_abcd",
  "status": "completed",
  "last_error": null,
  "chunking_strategy": {...},
  "attributes": {"key1": "value1", "key2": 2}
}
```

# Delete vector store file



```
DELETE https://api.openai.com/v1/vector_stores/{vector_store_id}/files/{file_id}
```

Delete a vector store file. This will remove the file from the vector store but the file itself will not be deleted. To delete the file, use the [delete file](#) endpoint.

## Path parameters

**file\_id** string Required

The ID of the file to delete.

**vector\_store\_id** string Required

The ID of the vector store that the file belongs to.

## Returns

Deletion status

Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files/file-abc123 \  
-H "Authorization: Bearer $OPENAI_API_KEY" \  
-
```

```
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-X DELETE
```

## Response

```
{  
  id: "file-abc123",  
  object: "vector_store.file.deleted",  
  deleted: true  
}
```

# The vector store file object Beta



A list of files attached to a vector store.

### attributes map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard. Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters, booleans, or numbers.

### chunking\_strategy object

The strategy used to chunk the file.

> Show possible types

---

**created\_at** integer

The Unix timestamp (in seconds) for when the vector store file was created.

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**last\_error** object

The last error associated with this vector store file. Will be `null` if there are no errors.

> Show properties

---

**object** string

The object type, which is always `vector_store.file`.

---

**status** string

The status of the vector store file, which can be either `in_progress`, `completed`, `cancelled`, or `failed`. The status `completed` indicates that the vector store file is ready for use.

---

**usage\_bytes** integer

The total vector store usage in bytes. Note that this may be different from the original file size.

---

**vector\_store\_id** string

The ID of the vector store that the File is attached to.

OBJECT The vector store file object

```
{  
  "id": "file-abc123",  
  "object": "vector_store.file",  
  "usage_bytes": 1234,  
  "created_at": 1698107661,  
  "vector_store_id": "vs_abc123",  
  "status": "completed",  
  "last_error": null,  
  "chunking_strategy": {  
    "type": "static",  
    "static": {  
      "max_chunk_size_tokens": 800,  
      "chunk_overlap_tokens": 400  
    }  
  }  
}
```

## Vector store file batches



Vector store file batches represent operations to add multiple files to a vector store. Related guide: [File Search](#)

---

# Create vector store file batch



```
POST https://api.openai.com/v1/vector_stores/{vector_store_id}/file_batches
```

Create a vector store file batch.

## Path parameters

---

**vector\_store\_id** string Required

The ID of the vector store for which to create a File Batch.

## Request body

---

**file\_ids** array Required

A list of [File](#) IDs that the vector store should use. Useful for tools like [file\\_search](#) that can access files.

---

**attributes** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard. Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters, booleans, or numbers.

### chunking\_strategy object Optional

The chunking strategy used to chunk the file(s). If not set, will use the `auto` strategy.

> Show possible types

## Returns

A [vector store file batch](#) object.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/file_batches \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "file_ids": ["file-abc123", "file-abc456"]
}'
```

### Response

```
{  
  "id": "vsfb_abc123",  
  "object": "vector_store.file_batch",  
  "created_at": 1699061776,  
  "vector_store_id": "vs_abc123",  
  "status": "in_progress",  
  "file_counts": {  
    "in_progress": 1,  
    "completed": 1,  
    "failed": 0,  
    "cancelled": 0,  
    "total": 0,  
  },  
}
```

# Retrieve vector store file batch



```
GET https://api.openai.com/v1/vector_stores/{vector_store_id}/file_batches/{batch_id}
```

Retrieves a vector store file batch.

## Path parameters

**batch\_id** string Required

The ID of the file batch being retrieved.

**vector\_store\_id** string Required

The ID of the vector store that the file batch belongs to.

## Returns

The [vector store file batch](#) object.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files_batches/vsfb_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "id": "vsfb_abc123",
  "object": "vector_store.file_batch",
  "created_at": 1699061776,
  "vector_store_id": "vs_abc123",
  "status": "in_progress",
```

```
"file_counts": {  
    "in_progress": 1,  
    "completed": 1,  
    "failed": 0,  
    "cancelled": 0,  
    "total": 0,  
}  
}
```

## Cancel vector store file batch



POST [https://api.openai.com/v1/vector\\_stores/{vector\\_store\\_id}/file\\_batches/{batch\\_id}/cancel](https://api.openai.com/v1/vector_stores/{vector_store_id}/file_batches/{batch_id}/cancel)

Cancel a vector store file batch. This attempts to cancel the processing of files in this batch as soon as possible.

### Path parameters

**batch\_id** string Required

The ID of the file batch to cancel.

**vector\_store\_id** string Required

The ID of the vector store that the file batch belongs to.

## Returns

The modified vector store file batch object.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files_batches/vsfb_abc123/cancel \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-X POST
```

### Response

```
{  
  "id": "vsfb_abc123",  
  "object": "vector_store.file_batch",  
  "created_at": 1699061776,  
  "vector_store_id": "vs_abc123",  
  "status": "in_progress",  
  "file_counts": {  
    "in_progress": 12,  
    "completed": 3,  
    "failed": 0,  
  },  
}
```

```
    "cancelled": 0,  
    "total": 15,  
}  
}
```

# List vector store files in a batch



```
GET https://api.openai.com/v1/vector_stores/{vector_store_id}/file_batches/{batch_id}/files
```

Returns a list of vector store files in a batch.

## Path parameters

**batch\_id** string Required

The ID of the file batch that the files belong to.

**vector\_store\_id** string Required

The ID of the vector store that the files belong to.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include `after=obj_foo` in order to fetch the next page of the list.

---

**before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include `before=obj_foo` in order to fetch the previous page of the list.

---

**filter** string Optional

Filter by file status. One of `in_progress`, `completed`, `failed`, `cancelled`.

---

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

---

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

---

## Returns

---

A list of [vector store file](#) objects.

### Example request

```
curl https://api.openai.com/v1/vector_stores/vs_abc123/files_batches/vsfb_abc123/files \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "id": "file-abc123",
      "object": "vector_store.file",
      "created_at": 1699061776,
      "vector_store_id": "vs_abc123"
    },
    {
      "id": "file-abc456",
      "object": "vector_store.file",
      "created_at": 1699061776,
      "vector_store_id": "vs_abc123"
    }
  ],
}
```

```
"first_id": "file-abc123",
"last_id": "file-abc456",
"has_more": false
}
```

# The vector store files batch object Beta



A batch of files attached to a vector store.

**created\_at** integer

The Unix timestamp (in seconds) for when the vector store files batch was created.

**file\_counts** object

> Show properties

**id** string

The identifier, which can be referenced in API endpoints.

**object** string

The object type, which is always `vector_store.file_batch`.

**status** string

The status of the vector store files batch, which can be either `in_progress`, `completed`, `cancelled` or `failed`.

**vector\_store\_id** string

The ID of the vector store that the File is attached to.

OBJECT The vector store files batch object

```
{  
  "id": "vsfb_123",  
  "object": "vector_store.files_batch",  
  "created_at": 1698107661,  
  "vector_store_id": "vs_abc123",  
  "status": "completed",  
  "file_counts": {  
    "in_progress": 0,  
    "completed": 100,  
    "failed": 0,  
    "cancelled": 0,  
    "total": 100  
  }  
}
```



Manage ChatKit sessions, threads, and file uploads for internal integrations.

# Create ChatKit session Beta



POST <https://api.openai.com/v1/chatkit/sessions>

Create a ChatKit session

## Request body

**user** string Required

A free-form string that identifies your end user; ensures this Session can access other objects that have the same **user** scope.

**workflow** object Required

Workflow that powers the session.

> Show properties

**chatkit\_configuration** object Optional

Optional overrides for ChatKit runtime configuration features

> Show properties

**expires\_after** object Optional

Optional override for session expiration timing in seconds from creation. Defaults to 10 minutes.

> Show properties

**rate\_limits** object Optional

Optional override for per-minute request limits. When omitted, defaults to 10.

> Show properties

## Returns

Returns a [ChatKit session](#) object.

### Example request

```
curl https://api.openai.com/v1/chatkit/sessions \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: chatkit_beta=v1" \
-d '{
  "workflow": {
    "id": "workflow_alpha",
    "version": "2024-10-01"
  },
  "scope": {
    "project": "alpha",
    "environment": "staging"
}
```

```
    },
    "expires_after": 1800,
    "max_requests_per_1_minute": 60,
    "max_requests_per_session": 500
}'
```

## Response

```
{
  "client_secret": "chatkit_token_123",
  "expires_after": 1800,
  "workflow": {
    "id": "workflow_alpha",
    "version": "2024-10-01"
  },
  "scope": {
    "project": "alpha",
    "environment": "staging"
  },
  "max_requests_per_1_minute": 60,
  "max_requests_per_session": 500,
  "status": "active"
}
```

# Cancel chat session

Beta



```
POST https://api.openai.com/v1/chatkit/sessions/{session_id}/cancel
```

Cancel a ChatKit session

## Path parameters

**session\_id** string Required

Unique identifier for the ChatKit session to cancel.

## Returns

Returns the chat session after it has been cancelled. Cancelling prevents new requests from using the issued client secret.

### Example request

```
curl -X POST \
https://api.openai.com/v1/chatkit/sessions/cksess_123/cancel \
-H "OpenAI-Beta: chatkit_beta=v1" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{  
  "id": "cksess_123",  
  "object": "chatkit.session",  
  "workflow": {  
    "id": "workflow_alpha",  
    "version": "1"  
  },  
  "scope": {  
    "customer_id": "cust_456"  
  },  
  "max_requests_per_1_minute": 30,  
  "ttl_seconds": 900,  
  "status": "cancelled",  
  "cancelled_at": 1712345678  
}
```

## List ChatKit threads Beta



GET <https://api.openai.com/v1/chatkit/threads>

List ChatKit threads

## Query parameters

---

**after** string Optional

List items created after this thread item ID. Defaults to null for the first page.

---

**before** string Optional

List items created before this thread item ID. Defaults to null for the newest results.

---

**limit** integer Optional

Maximum number of thread items to return. Defaults to 20.

---

**order** string Optional

Sort order for results by creation time. Defaults to `desc`.

---

**user** string Optional

Filter threads that belong to this user identifier. Defaults to null to return all users.

---

## Returns

---

Returns a paginated list of ChatKit threads accessible to the request scope.

Example request

```
curl "https://api.openai.com/v1/chatkit/threads?limit=2&order=desc" \
-H "OpenAI-Beta: chatkit_beta=v1" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{
  "data": [
    {
      "id": "cthr_abc123",
      "object": "chatkit.thread",
      "title": "Customer escalation"
    },
    {
      "id": "cthr_def456",
      "object": "chatkit.thread",
      "title": "Demo feedback"
    }
  ],
  "has_more": false,
  "object": "list"
}
```

# Retrieve ChatKit thread

Beta



```
GET https://api.openai.com/v1/chatkit/threads/{thread_id}
```

Retrieve a ChatKit thread

## Path parameters

**thread\_id** string Required

Identifier of the ChatKit thread to retrieve.

## Returns

Returns a [Thread](#) object.

### Example request

```
curl https://api.openai.com/v1/chatkit/threads/cthr_abc123 \
-H "OpenAI-Beta: chatkit_beta=v1" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{  
  "id": "cthr_abc123",  
  "object": "chatkit.thread",  
  "title": "Customer escalation",  
  "items": {  
    "data": [  
      {  
        "id": "cthi_user_001",  
        "object": "chatkit.thread_item",  
        "type": "user_message",  
        "content": [  
          {  
            "type": "input_text",  
            "text": "I need help debugging an onboarding issue."  
          }  
        ],  
        "attachments": []  
      },  
      {  
        "id": "cthi_assistant_002",  
        "object": "chatkit.thread_item",  
        "type": "assistant_message",  
        "content": [  
          {  
            "type": "output_text",  
            "text": "Let's start by confirming the workflow version you deployed."  
          }  
        ]  
      }  
    }  
  }  
}
```

```
  ],
  "has_more": false
}
}
```

## Delete ChatKit thread Beta



```
DELETE https://api.openai.com/v1/chatkit/threads/{thread_id}
```

Delete a ChatKit thread

### Path parameters

**thread\_id** string Required

Identifier of the ChatKit thread to delete.

### Returns

Returns a confirmation object for the deleted thread.

#### Example request

```
import OpenAI from 'openai';

const client = new OpenAI();

const thread = await client.beta.chat_kit.threads.delete('cthr_123');

console.log(thread.id);
```

## List ChatKit thread items Beta



GET [https://api.openai.com/v1/chatkit/threads/{thread\\_id}/items](https://api.openai.com/v1/chatkit/threads/{thread_id}/items)

List ChatKit thread items

### Path parameters

**thread\_id** string Required

Identifier of the ChatKit thread whose items are requested.

## Query parameters

---

**after** string Optional

List items created after this thread item ID. Defaults to null for the first page.

---

**before** string Optional

List items created before this thread item ID. Defaults to null for the newest results.

---

**limit** integer Optional

Maximum number of thread items to return. Defaults to 20.

---

**order** string Optional

Sort order for results by creation time. Defaults to `desc`.

## Returns

---

Returns a list of thread items for the specified thread.

Example request

```
curl "https://api.openai.com/v1/chatkit/threads/cthr_abc123/items?limit=3" \
-H "OpenAI-Beta: chatkit_beta=v1" \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{
  "data": [
    {
      "id": "cthi_user_001",
      "object": "chatkit.thread_item",
      "type": "user_message",
      "content": [
        {
          "type": "input_text",
          "text": "I need help debugging an onboarding issue."
        }
      ],
      "attachments": []
    },
    {
      "id": "cthi_assistant_002",
      "object": "chatkit.thread_item",
      "type": "assistant_message",
      "content": [
        {
          "type": "output_text",
          "text": "Let's start by confirming the workflow version you deployed."
        }
      ]
    }
  ]
}
```

```
        }
      ],
    },
  ],
  "has_more": false,
  "object": "list"
}
```

# The chat session object



Represents a ChatKit session and its resolved configuration.

---

**chatkit\_configuration** object

Resolved ChatKit feature configuration for the session.

> Show properties

---

**client\_secret** string

Ephemeral client secret that authenticates session requests.

---

**expires\_at** integer

Unix timestamp (in seconds) for when the session expires.

**id** string

Identifier for the ChatKit session.

**max\_requests\_per\_1\_minute** integer

Convenience copy of the per-minute request limit.

**object** string

Type discriminator that is always `chatkit.session`.

**rate\_limits** object

Resolved rate limit values.

> Show properties

**status** string

Current lifecycle state of the session.

**user** string

User identifier associated with the session.

**workflow** object

Workflow metadata for the session.

> Show properties

OBJECT The chat session object

```
{  
  "id": "cksess_123",  
  "object": "chatkit.session",  
  "client_secret": "ek_token_123",  
  "expires_at": 1712349876,  
  "workflow": {  
    "id": "workflow_alpha",  
    "version": "2024-10-01"  
  },  
  "user": "user_789",  
  "rate_limits": {  
    "max_requests_per_1_minute": 60  
  },  
  "max_requests_per_1_minute": 60,  
  "status": "cancelled",  
  "chatkit_configuration": {  
    "automatic_thread_titling": {  
      "enabled": true  
    },  
    "file_upload": {  
      "enabled": true,  
      "max_file_size": 16,  
      "max_files": 20  
    },  
    "history": {  
      "enabled": true,  
      "recent_threads": 10  
    }  
  }  
}
```

{  
}

# The thread object



Represents a ChatKit thread and its current status.

**created\_at** integer

Unix timestamp (in seconds) for when the thread was created.

**id** string

Identifier of the thread.

**object** string

Type discriminator that is always `chatkit.thread`.

**status** object

Current status for the thread. Defaults to `active` for newly created threads.

> Show possible types

**title** string

Optional human-readable title for the thread. Defaults to null when no title has been generated.

**user** string

Free-form string that identifies your end user who owns the thread.

OBJECT The thread object

```
{  
  "id": "cthr_def456",  
  "object": "chatkit.thread",  
  "created_at": 1712345600,  
  "title": "Demo feedback",  
  "status": {  
    "type": "active"  
  },  
  "user": "user_456"  
}
```

## Thread Items



A paginated list of thread items rendered for the ChatKit API.

**data** array

A list of items

> Show possible types

---

**first\_id** string

The ID of the first item in the list.

---

**has\_more** boolean

Whether there are more items available.

---

**last\_id** string

The ID of the last item in the list.

---

**object** string

The type of object returned, must be `list`.

---

# Containers



Create and manage containers for use with the Code Interpreter tool.

# Create container



POST <https://api.openai.com/v1/containers>

Create Container

## Request body

**name** string Required

Name of the container to create.

**expires\_after** object Optional

Container expiration time in seconds relative to the 'anchor' time.

> Show properties

**file\_ids** array Optional

IDs of files to copy to the container.

## Returns

The created [container](#) object.

### Example request

```
curl https://api.openai.com/v1/containers \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
    "name": "My Container"
}'
```

### Response

```
{
  "id": "cntr_682e30645a488191b6363a0cbefc0f0a025ec61b66250591",
  "object": "container",
  "created_at": 1747857508,
  "status": "running",
  "expires_after": {
    "anchor": "last_active_at",
    "minutes": 20
  },
  "last_active_at": 1747857508,
  "name": "My Container"
}
```

# List containers



```
GET https://api.openai.com/v1/containers
```

## List Containers

### Query parameters

---

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

---

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

---

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

a list of container objects.

### Example request

```
curl https://api.openai.com/v1/containers \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "id": "cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3cb5863",
      "object": "container",
      "created_at": 1747844794,
      "status": "running",
      "expires_after": {
        "anchor": "last_active_at",
        "minutes": 20
      },
      "last_active_at": 1747844794,
      "name": "My Container"
    }
  ],
  "first_id": "container_123",
```

```
"last_id": "container_123",
"has_more": false
}
```

# Retrieve container



```
GET https://api.openai.com/v1/containers/{container_id}
```

## Retrieve Container

### Path parameters

**container\_id** string Required

### Returns

The [container](#) object.

#### Example request

```
curl https://api.openai.com/v1/containers/cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3cb5863 \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{
  "id": "cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3cb5863",
  "object": "container",
  "created_at": 1747844794,
  "status": "running",
  "expires_after": {
    "anchor": "last_active_at",
    "minutes": 20
  },
  "last_active_at": 1747844794,
  "name": "My Container"
}
```

## Delete a container



```
DELETE https://api.openai.com/v1/containers/{container_id}
```

## Delete Container

### Path parameters

**container\_id** string Required

The ID of the container to delete.

### Returns

Deletion Status

#### Example request

```
curl -X DELETE https://api.openai.com/v1/containers/cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3c  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

#### Response

```
{  
  "id": "cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3cb5863",  
  "object": "container.deleted",  
  "deleted": true  
}
```

# The container object



---

## **created\_at** integer

Unix timestamp (in seconds) when the container was created.

---

## **expires\_after** object

The container will expire after this time period. The anchor is the reference point for the expiration. The minutes is the number of minutes after the anchor before the container expires.

> Show properties

---

## **id** string

Unique identifier for the container.

---

## **name** string

Name of the container.

---

## **object** string

The type of this object.

---

## **status** string

Status of the container (e.g., active, deleted).

OBJECT The container object

```
{  
  "id": "cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3cb5863",  
  "object": "container",  
  "created_at": 1747844794,  
  "status": "running",  
  "expires_after": {  
    "anchor": "last_active_at",  
    "minutes": 20  
  },  
  "last_active_at": 1747844794,  
  "name": "My Container"  
}
```

# Container Files



Create and manage container files for use with the Code Interpreter tool.

# Create container file



```
POST https://api.openai.com/v1/containers/{container_id}/files
```

## Create a Container File

You can send either a multipart/form-data request with the raw file content, or a JSON request with a file ID.

## Path parameters

**container\_id** string Required

## Request body

**file** file Optional

The File object (not file name) to be uploaded.

---

**file\_id** string Optional

Name of the file to create.

## Returns

The created container file object.

#### Example request

```
curl https://api.openai.com/v1/containers/cntr_682e0e7318108198aa783fd921ff305e08e78805b9fdbb04/file  
-H "Authorization: Bearer $OPENAI_API_KEY" \  
-F file="@example.txt"
```

#### Response

```
{  
  "id": "cfile_682e0e8a43c88191a7978f477a09bdf5",  
  "object": "container.file",  
  "created_at": 1747848842,  
  "bytes": 880,  
  "container_id": "cntr_682e0e7318108198aa783fd921ff305e08e78805b9fdbb04",  
  "path": "/mnt/data/88e12fa445d32636f190a0b33daed6cb-tsconfig.json",  
  "source": "user"  
}
```

# List container files



```
GET https://api.openai.com/v1/containers/{container_id}/files
```

## List Container files

### Path parameters

**container\_id** string Required

### Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

a list of container\_file objects.

### Example request

```
curl https://api.openai.com/v1/containers/cntr_682e0e7318108198aa783fd921ff305e08e78805b9fdbb04/file  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "cfile_682e0e8a43c88191a7978f477a09bdf5",  
      "object": "container.file",  
      "created_at": 1747848842,  
      "bytes": 880,  
      "container_id": "cntr_682e0e7318108198aa783fd921ff305e08e78805b9fdbb04",  
      "path": "/mnt/data/88e12fa445d32636f190a0b33daed6cb-tsconfig.json",  
      "source": "user"  
    }  
,  
  "first_id": "cfile_682e0e8a43c88191a7978f477a09bdf5",  
  "has_more": false,
```

```
"last_id": "cfile_682e0e8a43c88191a7978f477a09bdf5"  
}
```

# Retrieve container file



```
GET https://api.openai.com/v1/containers/{container_id}/files/{file_id}
```

## Retrieve Container File

### Path parameters

**container\_id** string Required

**file\_id** string Required

### Returns

The [container file](#) object.

Example request

```
curl https://api.openai.com/v1/containers/container_123/files/file_456 \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Response

```
{  
  "id": "cfile_682e0e8a43c88191a7978f477a09bdf5",  
  "object": "container.file",  
  "created_at": 1747848842,  
  "bytes": 880,  
  "container_id": "cntr_682e0e7318108198aa783fd921ff305e08e78805b9fdbb04",  
  "path": "/mnt/data/88e12fa445d32636f190a0b33daed6cb-tsconfig.json",  
  "source": "user"  
}
```

# Retrieve container file content



```
GET https://api.openai.com/v1/containers/{container_id}/files/{file_id}/content
```

## Retrieve Container File Content

## Path parameters

**container\_id** string Required

**file\_id** string Required

## Returns

The contents of the container file.

### Example request

```
curl https://api.openai.com/v1/containers/container_123/files/cfile_456/content \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

### Response

```
<binary content of the file>
```

## Delete a container file



```
DELETE https://api.openai.com/v1/containers/{container_id}/files/{file_id}
```

## Delete Container File

### Path parameters

**container\_id** string Required

**file\_id** string Required

### Returns

Deletion Status

#### Example request

```
curl -X DELETE https://api.openai.com/v1/containers/cntr_682dfebaacac8198bbfe9c2474fb6f4a085685cbe3c  
-H "Authorization: Bearer $OPENAI_API_KEY"
```

#### Response

```
{  
  "id": "cfile_682e0e8a43c88191a7978f477a09bdf5",  
  "object": "container.file.deleted",
```

```
  "deleted": true  
}
```

# The container file object



**bytes** integer

Size of the file in bytes.

**container\_id** string

The container this file belongs to.

**created\_at** integer

Unix timestamp (in seconds) when the file was created.

**id** string

Unique identifier for the file.

**object** string

The type of this object ( `container.file` ).

**path** string

Path of the file in the container.

**source** stringSource of the file (e.g., `user` , `assistant` ).

OBJECT The container file object

```
{  
  "id": "cfile_682e0e8a43c88191a7978f477a09bdf5",  
  "object": "container.file",  
  "created_at": 1747848842,  
  "bytes": 880,  
  "container_id": "cntr_682e0e7318108198aa783fd921ff305e08e78805b9fdbb04",  
  "path": "/mnt/data/88e12fa445d32636f190a0b33daed6cb-tsconfig.json",  
  "source": "user"  
}
```

# Realtime



Communicate with a multimodal model in real time over low latency interfaces like WebRTC, WebSocket, and SIP. Natively supports speech-to-speech as well as text, image, and audio

inputs and outputs.

[Learn more about the Realtime API.](#)

---

## Create call



POST `https://api.openai.com/v1/realtime/calls`

Create a new Realtime API call over WebRTC and receive the SDP answer needed to complete the peer connection.

### Request body

**sdp** string Required

WebRTC Session Description Protocol (SDP) offer generated by the caller.

**session** object Optional

Optional session configuration to apply before the realtime session is created. Use the same parameters you would send in a

[`create\_client\_secret`](#) request.

> Show properties

## Returns

Returns `201 Created` with the SDP answer in the response body. The `Location` response header includes the call ID for follow-up requests, e.g., establishing a monitoring WebSocket or hanging up the call.

### Example request

```
curl -X POST https://api.openai.com/v1/realtime/calls \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-F "sdp=<offer.sdp;type=application/sdp" \
-F 'session={"type": "realtime", "model": "gpt-realtime"};type=application/json'
```

### Response

```
v=0
o=- 4227147428 1719357865 IN IP4 127.0.0.1
s=-
c=IN IP4 0.0.0.0
t=0 0
a=group:BUNDLE 0 1
a=msid-semantic:WMS *
a=fingerprint:sha-256 CA:92:52:51:B4:91:3B:34:DD:9C:0B:FB:76:19:7E:3B:F1:21:0F:32:2C:38:01:72:5D:31
m=audio 9 UDP/TLS/RTP/SAVPF 111 0 8
a=mid:0
a=ice-ufrag:kZ2qkHXX/u11
a=ice-pwd:uoD16Di50Gx3VbqgA3ymjEQV2kwi0jw6
a=setup:active
```

```
a=rtpmap:111 opus/48000/2
a=candidate:993865896 1 udp 2130706431 4.155.146.196 3478 typ host ufrag kZ2qkHXX/u11
a=candidate:1432411780 1 tcp 1671430143 4.155.146.196 443 typ host tcptype passive ufrag kZ2qkHXX/i
m=application 9 UDP/DTLS/SCTP webrtc-datachannel
a=mid:1
a=sctp-port:5000
```

## Client secrets



REST API endpoint to generate ephemeral client secrets for use in client-side applications. Client secrets are short-lived tokens that can be passed to a client app, such as a web frontend or mobile client, which grants access to the Realtime API without leaking your main API key. You can configure a custom TTL for each client secret.

You can also attach session configuration options to the client secret, which will be applied to any sessions created using that client secret, but these can also be overridden by the client connection.

[Learn more about authentication with client secrets over WebRTC.](#)

# Create client secret



```
POST https://api.openai.com/v1/realtime/client_secrets
```

Create a Realtime client secret with an associated session configuration.

## Request body

**expires\_after** object Optional

Configuration for the client secret expiration. Expiration refers to the time after which a client secret will no longer be valid for creating sessions. The session itself may continue after that time once started. A secret can be used to create multiple sessions until it expires.

> Show properties

**session** object Optional

Session configuration to use for the client secret. Choose either a realtime session or a transcription session.

> Show possible types

## Returns

The created client secret and the effective session object. The client secret is a string that looks like `ek_1234`.

Example request

```
curl -X POST https://api.openai.com/v1/realtime/client_secrets \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "expires_after": {
    "anchor": "created_at",
    "seconds": 600
  },
  "session": {
    "type": "realtime",
    "model": "gpt-realtime",
    "instructions": "You are a friendly assistant."
  }
}'
```

## Response

```
{
  "value": "ek_68af296e8e408191a1120ab6383263c2",
  "expires_at": 1756310470,
  "session": {
    "type": "realtime",
    "object": "realtime.session",
    "id": "sess_C9CiUVUzUzYIssh3ELY1d",
    "model": "gpt-realtime",
    "output_modalities": [
      "audio"
    ],
    "status": "active"
  }
}
```

```
"instructions": "You are a friendly assistant.",  
"tools": [],  
"tool_choice": "auto",  
"max_output_tokens": "inf",  
"tracing": null,  
"truncation": "auto",  
"prompt": null,  
"expires_at": 0,  
"audio": {  
    "input": {  
        "format": {  
            "type": "audio/pcm",  
            "rate": 24000  
        },  
        "transcription": null,  
        "noise_reduction": null,  
        "turn_detection": {  
            "type": "server_vad",  
        }  
    },  
    "output": {  
        "format": {  
            "type": "audio/pcm",  
            "rate": 24000  
        },  
        "voice": "alloy",  
        "speed": 1.0  
    }  
},
```

```
"include": null  
}  
}
```

# Session response object



Response from creating a session and client secret for the Realtime API.

**expires\_at** integer

Expiration timestamp for the client secret, in seconds since epoch.

**session** object

The session configuration for either a realtime or transcription session.

> Show possible types

**value** string

The generated client secret value.

OBJECT Session response object

```
{  
  "value": "ek_68af296e8e408191a1120ab6383263c2",
```

```
"expires_at": 1756310470,  
"session": {  
    "type": "realtime",  
    "object": "realtime.session",  
    "id": "sess_C9CiUVUzUzYIssh3ELY1d",  
    "model": "gpt-realtime-2025-08-25",  
    "output_modalities": [  
        "audio"  
    ],  
    "instructions": "You are a friendly assistant.",  
    "tools": [],  
    "tool_choice": "auto",  
    "max_output_tokens": "inf",  
    "tracing": null,  
    "truncation": "auto",  
    "prompt": null,  
    "expires_at": 0,  
    "audio": {  
        "input": {  
            "format": {  
                "type": "audio/pcm",  
                "rate": 24000  
            },  
            "transcription": null,  
            "noise_reduction": null,  
            "turn_detection": {  
                "type": "server_vad",  
                "threshold": 0.5,  
                "prefix_padding_ms": 300,  
                "postfix_padding_ms": 300  
            }  
        }  
    }  
}
```

```
        "silence_duration_ms": 200,  
        "idle_timeout_ms": null,  
        "create_response": true,  
        "interrupt_response": true  
    }  
,  
    "output": {  
        "format": {  
            "type": "audio/pcm",  
            "rate": 24000  
        },  
        "voice": "alloy",  
        "speed": 1.0  
    }  
,  
    "include": null  
}  
}
```

# Calls



REST endpoints for controlling WebRTC or SIP calls with the Realtime API. Accept or reject an incoming call, transfer it to another destination, or hang up the call once you are finished.

---

# Accept call



```
POST https://api.openai.com/v1/realtime/calls/{call_id}/accept
```

Accept an incoming SIP call and configure the realtime session that will handle it.

## Path parameters

**call\_id** string Required

The identifier for the call provided in the `realtime.call.incoming` webhook.

## Request body

**type** string Required

The type of session to create. Always `realtime` for the Realtime API.

**audio** object Optional

## Configuration for input and output audio.

> Show properties

---

### **include** array Optional

Additional fields to include in server outputs.

`item.input_audio_transcription.logprobs` : Include logprobs for input audio transcription.

---

### **instructions** string Optional

The default system instructions (i.e. system message) prepended to model calls. This field allows the client to guide the model on desired responses. The model can be instructed on response content and format, (e.g. "be extremely succinct", "act friendly", "here are examples of good responses") and on audio behavior (e.g. "talk quickly", "inject emotion into your voice", "laugh frequently"). The instructions are not guaranteed to be followed by the model, but they provide guidance to the model on the desired behavior.

Note that the server sets default instructions which will be used if this field is not set and are visible in the `session.created` event at the start of the session.

---

### **max\_output\_tokens** integer or "inf" Optional

Maximum number of output tokens for a single assistant response, inclusive of tool calls. Provide an integer between 1 and 4096 to limit output tokens, or `inf` for the maximum available tokens for a given model. Defaults to `inf`.

---

### **model** string Optional

The Realtime model used for this session.

---

### **output\_modalities** array Optional Defaults to audio

The set of modalities the model can respond with. It defaults to `["audio"]`, indicating that the model will respond with audio plus a transcript. `["text"]` can be used to make the model respond with text only. It is not possible to request both `text` and `audio` at the same time.

---

**prompt** object Optional

Reference to a prompt template and its variables. [Learn more](#).

> Show properties

---

**tool\_choice** string or object Optional Defaults to auto

How the model chooses tools. Provide one of the string modes or force a specific function/MCP tool.

> Show possible types

---

**tools** array Optional

Tools available to the model.

> Show possible types

---

**tracing** "auto" or object Optional Defaults to null

Realtime API can write session traces to the [Traces Dashboard](#). Set to null to disable tracing. Once tracing is enabled for a session, the configuration cannot be modified.

`auto` will create a trace for the session with default values for the workflow name, group id, and metadata.

> Show possible types

---

**truncation** string or object Optional

Controls how the realtime conversation is truncated prior to model inference. The default is `auto`.

> Show possible types

## Returns

Returns `200 OK` once OpenAI starts ringing the SIP leg with the supplied session configuration.

### Example request

```
curl -X POST https://api.openai.com/v1/realtime/calls/$CALL_ID/accept \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{'
    "type": "realtime",
    "model": "gpt-realtime",
    "instructions": "You are Alex, a friendly concierge for Example Corp.",
}'
```

## Reject call



```
POST https://api.openai.com/v1/realtime/calls/{call_id}/reject
```

Decline an incoming SIP call by returning a SIP status code to the caller.

## Path parameters

**call\_id** string **Required**

The identifier for the call provided in the `realtime.call.incoming` webhook.

## Request body

**status\_code** integer **Optional**

SIP response code to send back to the caller. Defaults to `603` (Decline) when omitted.

## Returns

Returns `200 OK` after OpenAI sends the SIP status code to the caller.

### Example request

```
curl -X POST https://api.openai.com/v1/realtime/calls/$CALL_ID/reject \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"status_code": 486}'
```

# Refer call



```
POST https://api.openai.com/v1/realtime/calls/{call_id}/refer
```

Transfer an active SIP call to a new destination using the SIP REFER verb.

## Path parameters

**call\_id** string Required

The identifier for the call provided in the `realtime.call.incoming` webhook.

## Request body

**target\_uri** string Required

URI that should appear in the SIP Refer-To header. Supports values like `tel:+14155550123` or `sip:agent@example.com`.

## Returns

Returns `200 OK` once the REFER is handed off to your SIP provider.

### Example request

```
curl -X POST https://api.openai.com/v1/realtime/calls/$CALL_ID/refer \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"target_uri": "tel:+14155550123"}'
```

# Hang up call



POST [https://api.openai.com/v1/realtime/calls/{call\\_id}/hangup](https://api.openai.com/v1/realtime/calls/{call_id}/hangup)

End an active Realtime API call, whether it was initiated over SIP or WebRTC.

## Path parameters

**call\_id** string **Required**

The identifier for the call. For SIP calls, use the value provided in the [realtime.call.incoming](#) webhook. For WebRTC sessions, reuse the call ID returned in the [Location](#) header when creating the call with [POST /v1/realtime/calls](#).

## Returns

Returns **200 OK** when OpenAI begins terminating the realtime call.

#### Example request

```
curl -X POST https://api.openai.com/v1/realtime/calls/$CALL_ID/hangup \
-H "Authorization: Bearer $OPENAI_API_KEY"
```

## Client events



These are events that the OpenAI Realtime WebSocket server will accept from the client.



### session.update



Send this event to update the session's configuration. The client may send this event at any time to update any field except for `voice` and `model`. `voice` can be updated only if there have been no other audio outputs yet.

When the server receives a `session.update`, it will respond with a `session.updated` event showing the full, effective configuration. Only the fields that are present in the `session.update` are updated. To clear a field like `instructions`, pass an empty string. To clear a field like `tools`, pass an empty array. To clear a field like `turn_detection`, pass `null`.

---

**event\_id** string

Optional client-generated ID used to identify this event. This is an arbitrary string that a client may assign. It will be passed back if there is an error with the event, but the corresponding `session.updated` event will not include it.

---

**session** object

Update the Realtime session. Choose either a realtime session or a transcription session.

› Show possible types

---

**type** string

The event type, must be `session.update`.

OBJECT `session.update`

```
{  
  "type": "session.update",  
  "session": {  
    "type": "realtime",  
    "instructions": "You are a creative assistant that helps with design tasks.",  
  },  
}
```

```
"tools": [
  {
    "type": "function",
    "name": "display_color_palette",
    "description": "Call this function when a user asks for a color palette.",
    "parameters": {
      "type": "object",
      "strict": true,
      "properties": {
        "theme": {
          "type": "string",
          "description": "Description of the theme for the color scheme."
        },
        "colors": {
          "type": "array",
          "description": "Array of five hex color codes based on the theme.",
          "items": {
            "type": "string",
            "description": "Hex color code"
          }
        }
      },
      "required": [
        "theme",
        "colors"
      ]
    }
  }
],
```

```
    "tool_choice": "auto"  
  },  
  "event_id": "5fc543c4-f59c-420f-8fb9-68c45d1546a7",  
}  
}
```



## input\_audio\_buffer.append



Send this event to append audio bytes to the input audio buffer. The audio buffer is temporary storage you can write to and later commit. A "commit" will create a new user message item in the conversation history from the buffer content and clear the buffer. Input audio transcription (if enabled) will be generated when the buffer is committed.

If VAD is enabled the audio buffer is used to detect speech and the server will decide when to commit. When Server VAD is disabled, you must commit the audio buffer manually. Input audio noise reduction operates on writes to the audio buffer.

The client may choose how much audio to place in each event up to a maximum of 15 MiB, for example streaming smaller chunks from the client may allow the VAD to be more responsive. Unlike most other client

events, the server will not send a confirmation response to this event.

**audio** string

Base64-encoded audio bytes. This must be in the format specified by the `input_audio_format` field in the session configuration.

**event\_id** string

Optional client-generated ID used to identify this event.

**type** string

The event type, must be `input_audio_buffer.append`.

```
OBJECT input_audio_buffer.append
```

```
{  
  "event_id": "event_456",  
  "type": "input_audio_buffer.append",  
  "audio": "Base64EncodedAudioData"  
}
```

## input\_audio\_buffer.commit



Send this event to commit the user input audio buffer, which will create a new user message item in the conversation. This event will produce an error if the input audio buffer is empty. When in Server VAD mode, the client does not need to send this event, the server will commit the audio buffer automatically.

Committing the input audio buffer will trigger input audio transcription (if enabled in session configuration), but it will not create a response from the model. The server will respond with an `input_audio_buffer.committed` event.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**type** string

The event type, must be `input_audio_buffer.commit`.

```
OBJECT input_audio_buffer.commit  
  
{  
  "event_id": "event_789",  
  "type": "input_audio_buffer.commit"  
}
```

# input\_audio\_buffer.clear



Send this event to clear the audio bytes in the buffer. The server will respond with an

`input_audio_buffer.cleared` event.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**type** string

The event type, must be `input_audio_buffer.clear`.

OBJECT `input_audio_buffer.clear`

```
{  
  "event_id": "event_012",  
  "type": "input_audio_buffer.clear"  
}
```



# conversation.item.create



Add a new Item to the Conversation's context, including messages, function calls, and function call responses. This event can be used both to populate a "history" of the conversation and to add new items mid-stream, but has the current limitation that it cannot populate assistant audio messages.

If successful, the server will respond with a `conversation.item.created` event, otherwise an `error` event will be sent.

---

## **event\_id** string

Optional client-generated ID used to identify this event.

---

## **item** object

A single item within a Realtime conversation.

> Show possible types

---

## **previous\_item\_id** string

The ID of the preceding item after which the new item will be inserted. If not set, the new item will be appended to the end of the conversation. If set to `root`, the new item will be added to the beginning of the conversation. If set to an existing ID, it allows an item to be inserted mid-conversation. If the ID cannot be found, an error will be returned and the item will not be added.

---

## **type** string

The event type, must be `conversation.item.create`.

OBJECT `conversation.item.create`

```
{  
  "type": "conversation.item.create",  
  "item": {  
    "type": "message",  
    "role": "user",  
    "content": [  
      {  
        "type": "input_text",  
        "text": "hi"  
      }  
    ]  
  },  
  "event_id": "b904fba0-0ec4-40af-8bbb-f908a9b26793",  
}
```

## conversation.item.retrieve



Send this event when you want to retrieve the server's representation of a specific item in the conversation history. This is useful, for example, to inspect user audio after noise cancellation and VAD. The server will respond with a `conversation.item.retrieved` event, unless the item does not exist in the conversation history, in which case the server will respond with an error.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**item\_id** string

The ID of the item to retrieve.

---

**type** string

The event type, must be `conversation.item.retrieve`.

OBJECT `conversation.item.retrieve`

```
{  
  "event_id": "event_901",  
  "type": "conversation.item.retrieve",  
  "item_id": "item_003"  
}
```

# conversation.item.truncate



Send this event to truncate a previous assistant message's audio. The server will produce audio faster than realtime, so this event is useful when the user interrupts to truncate audio that has already been sent to the client but not yet played. This will synchronize the server's understanding of the audio with the client's playback.

Truncating audio will delete the server-side text transcript to ensure there is not text in the context that hasn't been heard by the user.

If successful, the server will respond with a `conversation.item.truncated` event.

---

**audio\_end\_ms** integer

Inclusive duration up to which audio is truncated, in milliseconds. If the `audio_end_ms` is greater than the actual audio duration, the server will respond with an error.

---

**content\_index** integer

The index of the content part to truncate. Set this to `0`.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**item\_id** string

The ID of the assistant message item to truncate. Only assistant message items can be truncated.

**type** string

The event type, must be `conversation.item.truncate`.

OBJECT `conversation.item.truncate`

```
{  
  "event_id": "event_678",  
  "type": "conversation.item.truncate",  
  "item_id": "item_002",  
  "content_index": 0,  
  "audio_end_ms": 1500  
}
```

## conversation.item.delete



Send this event when you want to remove any item from the conversation history. The server will respond with a `conversation.item.deleted` event, unless the item does not exist in the conversation history, in which case the server will respond with an error.

**event\_id** string

Optional client-generated ID used to identify this event.

---

**item\_id** string

The ID of the item to delete.

---

**type** string

The event type, must be `conversation.item.delete`.

OBJECT `conversation.item.delete`

```
{  
  "event_id": "event_901",  
  "type": "conversation.item.delete",  
  "item_id": "item_003"  
}
```



## response.create



This event instructs the server to create a Response, which means triggering model inference. When in Server VAD mode, the server will create Responses automatically.

A Response will include at least one Item, and may have two, in which case the second will be a function call. These Items will be appended to the conversation history by default.

The server will respond with a `response.created` event, events for Items and content created, and finally a `response.done` event to indicate the Response is complete.

The `response.create` event includes inference configuration like `instructions` and `tools`. If these are set, they will override the Session's configuration for this Response only.

Responses can be created out-of-band of the default Conversation, meaning that they can have arbitrary input, and it's possible to disable writing the output to the Conversation. Only one Response can write to the default Conversation at a time, but otherwise multiple Responses can be created in parallel. The `metadata` field is a good way to disambiguate multiple simultaneous Responses.

Clients can set `conversation` to `none` to create a Response that does not write to the default Conversation. Arbitrary input can be provided with the `input` field, which is an array accepting raw Items and references to existing Items.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**response** object

## Create a new Realtime response with these parameters

> Show properties

**type** string

The event type, must be `response.create`.

OBJECT `response.create`

```
// Trigger a response with the default Conversation and no special parameters
{
  "type": "response.create",
}

// Trigger an out-of-band response that does not write to the default Conversation
{
  "type": "response.create",
  "response": {
    "instructions": "Provide a concise answer.",
    "tools": [], // clear any session tools
    "conversation": "none",
    "output_modalities": ["text"],
    "metadata": {
      "response_purpose": "summarization"
    },
    "input": [
      {
        "type": "item_reference",
        "id": "item_12345",
      },
    ],
  }
}
```

```
{  
  "type": "message",  
  "role": "user",  
  "content": [  
    {  
      "type": "input_text",  
      "text": "Summarize the above message in one sentence."  
    }  
  ]  
},  
}  
}
```

## response.cancel



Send this event to cancel an in-progress response. The server will respond with a `response.done` event with a status of `response.status=cancelled`. If there is no response to cancel, the server will respond with an error. It's safe to call `response.cancel` even if no response is in progress, an error will be returned the session will remain unaffected.

**event\_id** string

Optional client-generated ID used to identify this event.

**response\_id** string

A specific response ID to cancel - if not provided, will cancel an in-progress response in the default conversation.

**type** string

The event type, must be `response.cancel`.

OBJECT `response.cancel`

```
{  
  "type": "response.cancel"  
  "response_id": "resp_12345",  
}
```



## output\_audio\_buffer.clear



**WebRTC Only:** Emit to cut off the current audio response. This will trigger the server to stop generating audio and emit a `output_audio_buffer.cleared` event. This event should be preceded by a `response.cancel` client event to stop the generation of the current response. [Learn more.](#)

---

**event\_id** string

The unique ID of the client event used for error handling.

---

**type** string

The event type, must be `output_audio_buffer.clear`.

```
OBJECT output_audio_buffer.clear
```

```
{  
  "event_id": "optional_client_event_id",  
  "type": "output_audio_buffer.clear"  
}
```

## Server events



These are events emitted from the OpenAI Realtime WebSocket server to the client.

## error



Returned when an error occurs, which could be a client problem or a server problem. Most errors are recoverable and the session will stay open, we recommend to implementors to monitor and log error messages by default.

### error object

Details of the error.

> Show properties

#### event\_id string

The unique ID of the server event.

#### type string

The event type, must be `error`.

OBJECT error

{

`"event_id": "event_890",`

```
"type": "error",
"error": {
    "type": "invalid_request_error",
    "code": "invalid_event",
    "message": "The 'type' field is missing.",
    "param": null,
    "event_id": "event_567"
}
}
```



## session.created



Returned when a Session is created. Emitted automatically when a new connection is established as the first server event. This event will contain the default Session configuration.

**event\_id** string

The unique ID of the server event.

**session** object

## The session configuration.

> Show possible types

**type** string

The event type, must be `session.created`.

OBJECT `session.created`

```
{  
  "type": "session.created",  
  "event_id": "event_C9G5RJeJ2gF77mV7f2B1j",  
  "session": {  
    "type": "realtime",  
    "object": "realtime.session",  
    "id": "sess_C9G5QPteg4UIbotdKLoYQ",  
    "model": "gpt-realtime-2025-08-28",  
    "output_modalities": [  
      "audio"  
    ],  
    "instructions": "Your knowledge cutoff is 2023-10. You are a helpful, witty, and friendly AI. Ad",  
    "tools": [],  
    "tool_choice": "auto",  
    "max_output_tokens": "inf",  
    "tracing": null,  
    "prompt": null,  
    "expires_at": 1756324625,  
    "audio": {  
      "input": {  
        "format": {  
          "type": "string",  
          "enum": ["audio/mpeg", "audio/wav"]  
        }  
      }  
    }  
  }  
}
```

```
        "type": "audio/pcm",
        "rate": 24000
    },
    "transcription": null,
    "noise_reduction": null,
    "turn_detection": {
        "type": "server_vad",
        "threshold": 0.5,
        "prefix_padding_ms": 300,
        "silence_duration_ms": 200,
        "idle_timeout_ms": null,
        "create_response": true,
        "interrupt_response": true
    }
},
"output": {
    "format": {
        "type": "audio/pcm",
        "rate": 24000
    },
    "voice": "marin",
    "speed": 1
}
},
"include": null
},
}
```

# session.updated



Returned when a session is updated with a `session.update` event, unless there is an error.

**event\_id** string

The unique ID of the server event.

**session** object

The session configuration.

› Show possible types

**type** string

The event type, must be `session.updated`.

OBJECT `session.updated`

```
{  
  "type": "session.updated",  
  "event_id": "event_C9G8mqI3IucaojlVKE8Cs",  
  "session": {  
    "type": "realtime",
```

```
"object": "realtime.session",
"id": "sess_C9G8l3zp50uFv4qgxfJ8o",
"model": "gpt-realtime-2025-08-28",
"output_modalities": [
    "audio"
],
"instructions": "Your knowledge cutoff is 2023-10. You are a helpful, witty, and friendly AI. /",
"tools": [
    {
        "type": "function",
        "name": "display_color_palette",
        "description": "\nCall this function when a user asks for a color palette.\n",
        "parameters": {
            "type": "object",
            "strict": true,
            "properties": {
                "theme": {
                    "type": "string",
                    "description": "Description of the theme for the color scheme."
                },
                "colors": {
                    "type": "array",
                    "description": "Array of five hex color codes based on the theme.",
                    "items": {
                        "type": "string",
                        "description": "Hex color code"
                    }
                }
            }
        }
    }
],
```

```
        "required": [
            "theme",
            "colors"
        ],
    },
],
"tool_choice": "auto",
"max_output_tokens": "inf",
"tracing": null,
"prompt": null,
"expires_at": 1756324832,
"audio": {
    "input": {
        "format": {
            "type": "audio/pcm",
            "rate": 24000
        },
        "transcription": null,
        "noise_reduction": null,
        "turn_detection": {
            "type": "server_vad",
            "threshold": 0.5,
            "prefix_padding_ms": 300,
            "silence_duration_ms": 200,
            "idle_timeout_ms": null,
            "create_response": true,
            "interrupt_response": true
        }
    }
}
```

```
},
  "output": {
    "format": {
      "type": "audio pcm",
      "rate": 24000
    },
    "voice": "marin",
    "speed": 1
  }
},
"include": null
},
}
```

🔗  
🔗

## conversation.item.added

🔗

Sent by the server when an Item is added to the default Conversation. This can happen in several cases:

When the client sends a `conversation.item.create` event.

When the input audio buffer is committed. In this case the item will be a user message containing the audio from the buffer.

When the model is generating a Response. In this case the `conversation.item.added` event will be sent when the model starts generating a specific Item, and thus it will not yet have any content (and `status` will be `in_progress`).

The event will include the full content of the Item (except when model is generating a Response) except for audio data, which can be retrieved separately with a `conversation.item.retrieve` event if necessary.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**previous\_item\_id** string

The ID of the item that precedes this one, if any. This is used to maintain ordering when items are inserted.

**type** string

The event type, must be `conversation.item.added`.

OBJECT `conversation.item.added`

{

`"type": "conversation.item.added",`

```
"event_id": "event_C9G8pjSJCFRNEhMEnYAVy",
"previous_item_id": null,
"item": {
  "id": "item_C9G8pGVKYnaZu8PH5YQ90",
  "type": "message",
  "status": "completed",
  "role": "user",
  "content": [
    {
      "type": "input_text",
      "text": "hi"
    }
  ]
}
```

## conversation.item.done



Returned when a conversation item is finalized.

The event will include the full content of the Item except for audio data, which can be retrieved separately with a `conversation.item.retrieve` event if needed.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

› Show possible types

**previous\_item\_id** string

The ID of the item that precedes this one, if any. This is used to maintain ordering when items are inserted.

**type** string

The event type, must be `conversation.item.done`.

OBJECT `conversation.item.done`

```
{  
  "type": "conversation.item.done",  
  "event_id": "event_CCXLgMzPo3qioWCeQa4WH",  
  "previous_item_id": "item_CCXLecNJVIVR2HUy3ABLj",  
  "item": {  
    "id": "item_CCXLfxmM5sXVJVz4mCa2S",  
    "type": "message",  
    "status": "completed",  
    "role": "assistant",  
    "content": [  
      {  
        "type": "output_audio",  
        "content": "The quick brown fox jumps over the lazy dog."  
      }  
    ]  
  }  
}
```

```
        "transcript": "Oh, I can hear you loud and clear! Sounds like we're connected just fine. Wl
    }
]
}
}
```

## conversation.item.retrieved



Returned when a conversation item is retrieved with `conversation.item.retrieve`. This is provided as a way to fetch the server's representation of an item, for example to get access to the post-processed audio data after noise cancellation and VAD. It includes the full content of the Item, including audio data.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**type** string

The event type, must be `conversation.item.retrieved`.

OBJECT `conversation.item.retrieved`

```
{  
  "type": "conversation.item.retrieved",  
  "event_id": "event_CCXGSizgEppa2d4XbKA7K",  
  "item": {  
    "id": "item_CCXGRxbY0n6WE4EszhF5w",  
    "object": "realtime.item",  
    "type": "message",  
    "status": "completed",  
    "role": "assistant",  
    "content": [  
      {  
        "type": "audio",  
        "transcript": "Yes, I can hear you loud and clear. How can I help you today?",  
        "audio": "8//2//v/9//q/+//+P/s...",  
        "format": "pcm16"  
      }  
    ]  
  }  
}
```



# conversation.item.input\_audio\_transcription.complete

## d



This event is the output of audio transcription for user audio written to the user audio buffer. Transcription begins when the input audio buffer is committed by the client or server (when VAD is enabled). Transcription runs asynchronously with Response creation, so this event may come before or after the Response events.

Realtime API models accept audio natively, and thus input transcription is a separate process run on a separate ASR (Automatic Speech Recognition) model. The transcript may diverge somewhat from the model's interpretation, and should be treated as a rough guide.

---

**content\_index** integer

The index of the content part containing the audio.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item containing the audio that is being transcribed.

**logprobs** array

The log probabilities of the transcription.

> Show properties

**transcript** string

The transcribed text.

**type** string

The event type, must be `conversation.item.input_audio_transcription.completed`.

**usage** object

Usage statistics for the transcription, this is billed according to the ASR model's pricing rather than the realtime model's pricing.

> Show possible types

OBJECT `conversation.item.input_audio_transcription.completed`

```
{  
  "type": "conversation.item.input_audio_transcription.completed",  
  "event_id": "event_CCXGRvtUVrax5SJAnNOWZ",  
  "item_id": "item_CCXGQ4e1ht4c0raEYcuR2",  
  "content_index": 0,  
  "transcript": "Hey, can you hear me?",  
  "usage": {  
    "type": "tokens",  
    "total_tokens": 22,  
    "input_tokens": 13,  
    "input_token_details": {  
      "id": "t1",  
      "start": 0, "end": 13  
    }  
  }  
}
```

```
        "text_tokens": 0,  
        "audio_tokens": 13  
    },  
    "output_tokens": 9  
}  
}
```

## conversation.item.input\_audio\_transcription.delta



Returned when the text value of an input audio transcription content part is updated with incremental transcription results.

### **content\_index** integer

The index of the content part in the item's content array.

### **delta** string

The text delta.

### **event\_id** string

The unique ID of the server event.

### **item\_id**

string

The ID of the item containing the audio that is being transcribed.

**logprobs** array

The log probabilities of the transcription. These can be enabled by configuring the session with

"include": ["item.input\_audio\_transcription.logprobs"] . Each entry in the array corresponds a log probability of which token would be selected for this chunk of transcription. This can help to identify if it was possible there were multiple valid options for a given chunk of transcription.

> Show properties

**type** string

The event type, must be conversation.item.input\_audio\_transcription.delta .

```
OBJECT conversation.item.input_audio_transcription.delta

{
  "type": "conversation.item.input_audio_transcription.delta",
  "event_id": "event_CCXGRxsAimPAs8kS2Wc7Z",
  "item_id": "item_CCXGQ4e1ht4c0raEYcuR2",
  "content_index": 0,
  "delta": "Hey",
  "obfuscation": "aLxx0jTEciOGe"
}
```

# conversation.item.input\_audio\_transcription.segment



Returned when an input audio transcription segment is identified for an item.

---

## **content\_index** integer

The index of the input audio content part within the item.

---

## **end** number

End time of the segment in seconds.

---

## **event\_id** string

The unique ID of the server event.

---

## **id** string

The segment identifier.

---

## **item\_id** string

The ID of the item containing the input audio content.

---

## **speaker** string

The detected speaker label for this segment.

---

## **start** number

Start time of the segment in seconds.

**text** string

The text for this segment.

**type** string

The event type, must be `conversation.item.input_audio_transcription.segment`.

OBJECT `conversation.item.input_audio_transcription.segment`

```
{  
  "event_id": "event_6501",  
  "type": "conversation.item.input_audio_transcription.segment",  
  "item_id": "msg_011",  
  "content_index": 0,  
  "text": "hello",  
  "id": "seg_0001",  
  "speaker": "spk_1",  
  "start": 0.0,  
  "end": 0.4  
}
```

# conversation.item.input\_audio\_transcription.failed



Returned when input audio transcription is configured, and a transcription request for a user message failed.

These events are separate from other `error` events so that the client can identify the related Item.

---

## `content_index` integer

The index of the content part containing the audio.

---

## `error` object

Details of the transcription error.

> Show properties

---

## `event_id` string

The unique ID of the server event.

---

## `item_id` string

The ID of the user message item.

---

## `type` string

The event type, must be `conversation.item.input_audio_transcription.failed`.

---

```
OBJECT conversation.item.input_audio_transcription.failed
```

```
{  
  "event_id": "event_2324",  
  "type": "conversation.item.input_audio_transcription.failed",  
  "item_id": "msg_003",  
  "content_index": 0,  
  "error": {  
    "type": "transcription_error",  
    "code": "audio_unintelligible",  
    "message": "The audio could not be transcribed.",  
    "param": null  
  }  
}
```

## conversation.item.truncated



Returned when an earlier assistant audio message item is truncated by the client with a `conversation.item.truncate` event. This event is used to synchronize the server's understanding of the audio with the client's playback.

This action will truncate the audio and remove the server-side text transcript to ensure there is no text in the context that hasn't been heard by the user.

**audio\_end\_ms** integer

The duration up to which the audio was truncated, in milliseconds.

**content\_index** integer

The index of the content part that was truncated.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the assistant message item that was truncated.

**type** string

The event type, must be `conversation.item.truncated`.

OBJECT `conversation.item.truncated`

```
{  
  "event_id": "event_2526",  
  "type": "conversation.item.truncated",  
  "item_id": "msg_004",  
  "content_index": 0,  
  "audio_end_ms": 1500  
}
```

# conversation.item.deleted



Returned when an item in the conversation is deleted by the client with a `conversation.item.delete` event.

This event is used to synchronize the server's understanding of the conversation history with the client's view.

## **event\_id** string

The unique ID of the server event.

## **item\_id** string

The ID of the item that was deleted.

## **type** string

The event type, must be `conversation.item.deleted`.

OBJECT `conversation.item.deleted`

{

```
"event_id": "event_2728",
"type": "conversation.item.deleted",
```



# input\_audio\_buffer.committed



Returned when an input audio buffer is committed, either by the client or automatically in server VAD mode.

The `item_id` property is the ID of the user message item that will be created, thus a `conversation.item.created` event will also be sent to the client.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the user message item that will be created.

---

**previous\_item\_id** string

The ID of the preceding item after which the new item will be inserted. Can be `null` if the item has no predecessor.

---

**type** string

The event type, must be `input_audio_buffer.committed`.

OBJECT `input_audio_buffer.committed`

```
{  
  "event_id": "event_1121",  
  "type": "input_audio_buffer.committed",  
  "previous_item_id": "msg_001",  
  "item_id": "msg_002"  
}
```

## input\_audio\_buffer.cleared



Returned when the input audio buffer is cleared by the client with a `input_audio_buffer.clear` event.

**event\_id** string

The unique ID of the server event.

**type** string

The event type, must be `input_audio_buffer.cleared`.

```
OBJECT input_audio_buffer.cleared
```

```
{  
  "event_id": "event_1314",  
  "type": "input_audio_buffer.cleared"  
}
```

## input\_audio\_buffer.speech\_started



Sent by the server when in `server_vad` mode to indicate that speech has been detected in the audio buffer. This can happen any time audio is added to the buffer (unless speech is already detected). The client may want to use this event to interrupt audio playback or provide visual feedback to the user.

The client should expect to receive a `input_audio_buffer.speech_stopped` event when speech stops. The `item_id` property is the ID of the user message item that will be created when speech stops and will also be included in the `input_audio_buffer.speech_stopped` event (unless the client manually commits the audio buffer during VAD activation).

### audio\_start\_ms integer

Milliseconds from the start of all audio written to the buffer during the session when speech was first detected. This will correspond to the beginning of audio sent to the model, and thus includes the `prefix_padding_ms` configured in the Session.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the user message item that will be created when speech stops.

**type** string

The event type, must be `input_audio_buffer.speech_started`.

OBJECT `input_audio_buffer.speech_started`

```
{  
  "event_id": "event_1516",  
  "type": "input_audio_buffer.speech_started",  
  "audio_start_ms": 1000,  
  "item_id": "msg_003"  
}
```

# input\_audio\_buffer.speech\_stopped

Returned in `server_vad` mode when the server detects the end of speech in the audio buffer. The server will also send an `conversation.item.created` event with the user message item that is created from the audio buffer.

---

**audio\_end\_ms** integer

Milliseconds since the session started when speech stopped. This will correspond to the end of audio sent to the model, and thus includes the `min_silence_duration_ms` configured in the Session.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the user message item that will be created.

---

**type** string

The event type, must be `input_audio_buffer.speech_stopped`.

```
OBJECT input_audio_buffer.speech_stopped
```

```
{
```

```
  "event_id": "event_1718",
  "type": "input_audio_buffer.speech_stopped",
  "audio_end_ms": 2000,
  "item_id": "msg_003"
```

```
}
```

# input\_audio\_buffer.timeout\_triggered



Returned when the Server VAD timeout is triggered for the input audio buffer. This is configured with `idle_timeout_ms` in the `turn_detection` settings of the session, and it indicates that there hasn't been any speech detected for the configured duration.

The `audio_start_ms` and `audio_end_ms` fields indicate the segment of audio after the last model response up to the triggering time, as an offset from the beginning of audio written to the input audio buffer. This means it demarcates the segment of audio that was silent and the difference between the start and end values will roughly match the configured timeout.

The empty audio will be committed to the conversation as an `input_audio` item (there will be a `input_audio_buffer.committed` event) and a model response will be generated. There may be speech that didn't trigger VAD but is still detected by the model, so the model may respond with something relevant to the conversation or a prompt to continue speaking.

## `audio_end_ms` integer

Millisecond offset of audio written to the input audio buffer at the time the timeout was triggered.

**audio\_start\_ms** integer

Millisecond offset of audio written to the input audio buffer that was after the playback time of the last model response.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item associated with this segment.

**type** string

The event type, must be `input_audio_buffer.timeout_triggered`.

OBJECT `input_audio_buffer.timeout_triggered`

```
{  
  "type": "input_audio_buffer.timeout_triggered",  
  "event_id": "event_CEKKrf1KTGvemCPyiJTJ2",  
  "audio_start_ms": 13216,  
  "audio_end_ms": 19232,  
  "item_id": "item_CEKKrWH0GiwN0ET97NUZc"  
}
```



# output\_audio\_buffer.started



**WebRTC Only:** Emitted when the server begins streaming audio to the client. This event is emitted after an audio content part has been added ( `response.content_part.added` ) to the response. [Learn more.](#)

---

**event\_id** string

The unique ID of the server event.

---

**response\_id** string

The unique ID of the response that produced the audio.

---

**type** string

The event type, must be `output_audio_buffer.started`.

OBJECT `output_audio_buffer.started`

```
{  
  "event_id": "event_abc123",  
  "type": "output_audio_buffer.started",  
  "response_id": "resp_abc123"  
}
```

## output\_audio\_buffer.stopped



**WebRTC Only:** Emitted when the output audio buffer has been completely drained on the server, and no more audio is forthcoming. This event is emitted after the full response data has been sent to the client (`response.done`). [Learn more.](#)

### event\_id string

The unique ID of the server event.

### response\_id string

The unique ID of the response that produced the audio.

**type** stringThe event type, must be `output_audio_buffer.stopped`.OBJECT `output_audio_buffer.stopped`

```
{  
  "event_id": "event_abc123",  
  "type": "output_audio_buffer.stopped",  
  "response_id": "resp_abc123"  
}
```

## output\_audio\_buffer.cleared



**WebRTC Only:** Emitted when the output audio buffer is cleared. This happens either in VAD mode when the user has interrupted (`input_audio_buffer.speech_started`), or when the client has emitted the `output_audio_buffer.clear` event to manually cut off the current audio response. [Learn more.](#)

**event\_id** string

The unique ID of the server event.

**response\_id** string

The unique ID of the response that produced the audio.

**type** string

The event type, must be `output_audio_buffer.cleared`.

OBJECT `output_audio_buffer.cleared`

```
{  
  "event_id": "event_abc123",  
  "type": "output_audio_buffer.cleared",  
  "response_id": "resp_abc123"  
}
```



## response.created



Returned when a new Response is created. The first event of response creation, where the response is in an initial state of `in_progress`.

**event\_id** string

The unique ID of the server event.

**response** object

The response resource.

> Show properties

**type** string

The event type, must be `response.created`.

OBJECT `response.created`

```
{  
  "type": "response.created",  
  "event_id": "event_C9G8pqbTEddBSIx60s",  
  "response": {  
    "object": "realtime.response",  
    "id": "resp_C9G8p7IH2WxLbkgPNouYL",  
    "status": "in_progress",  
    "status_details": null,  
    "output": [],  
    "conversation_id": "conv_C9G8mmBkLhQJwCon3hoJN",  
    "output_modalities": [  
      "audio"  
    ],  
    "max_output_tokens": "inf",  
    "audio": {
```

```
"output": {  
    "format": {  
        "type": "audio/pcm",  
        "rate": 24000  
    },  
    "voice": "marin"  
},  
},  
"usage": null,  
"metadata": null  
},  
}
```

## response.done



Returned when a Response is done streaming. Always emitted, no matter the final state. The Response object included in the `response.done` event will include all output items in the Response but will omit the raw audio data.

Clients should check the `status` field of the Response to determine if it was successful (`completed`) or if there was another outcome: `cancelled`, `failed`, or `incomplete`.

A response will contain all output items that were generated during the response, excluding any audio content.▼

**event\_id** string

The unique ID of the server event.

**response** object

The response resource.

> Show properties

**type** string

The event type, must be `response.done`.

OBJECT `response.done`

```
{  
  "type": "response.done",  
  "event_id": "event_CCXHxcMy86rrKhBLDdqCh",  
  "response": {  
    "object": "realtime.response",  
    "id": "resp_CCXHw0UJld10EzIUXQCNh",  
    "status": "completed",  
    "status_details": null,  
    "output": [  
      {  
        "id": "item_CCXHwGjjDUfOXbiySlK7i",  
        "type": "message",  
        "status": "completed",  
        "role": "assistant",  
        "content": [  
          {"text": "Hello, how can I help you today?"}]  
      }  
    ]  
  }  
}
```

```
{  
    "type": "output_audio",  
    "transcript": "Loud and clear! I can hear you perfectly. How can I help you today?"  
}  
]  
}  
],  
"conversation_id": "conv_CCXHsurMKcaVxIZvaCI5m",  
"output_modalities": [  
    "audio"  
,  
    "max_output_tokens": "inf",  
    "audio": {  
        "output": {  
            "format": {  
                "type": "audio/pcm",  
                "rate": 24000  
            },  
            "voice": "alloy"  
        }  
    },  
    "usage": {  
        "total_tokens": 253,  
        "input_tokens": 132,  
        "output_tokens": 121,  
        "input_token_details": {  
            "text_tokens": 119,  
            "audio_tokens": 13,  
            "image_tokens": 0,  
        }  
    }  
}
```

```
        "cached_tokens": 64,  
        "cached_tokens_details": {  
            "text_tokens": 64,  
            "audio_tokens": 0,  
            "image_tokens": 0  
        }  
    },  
    "output_token_details": {  
        "text_tokens": 30,  
        "audio_tokens": 91  
    }  
},  
"metadata": null  
}  
}
```



## response.output\_item.added



Returned when a new Item is created during Response generation.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**output\_index** integer

The index of the output item in the Response.

**response\_id** string

The ID of the Response to which the item belongs.

**type** string

The event type, must be `response.output_item.added`.

OBJECT `response.output_item.added`

{

```
"event_id": "event_3334",
"type": "response.output_item.added",
"response_id": "resp_001",
"output_index": 0,
"item": {
    "id": "msg_007",
    "object": "realtime.item",
```

```
        "type": "message",
        "status": "in_progress",
        "role": "assistant",
        "content": []
    }
}
```

## response.output\_item.done



Returned when an Item is done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**output\_index** integer

The index of the output item in the Response.

**response\_id** string

The ID of the Response to which the item belongs.

**type** string

The event type, must be `response.output_item.done`.

OBJECT `response.output_item.done`

```
{  
  "event_id": "event_3536",  
  "type": "response.output_item.done",  
  "response_id": "resp_001",  
  "output_index": 0,  
  "item": {  
    "id": "msg_007",  
    "object": "realtime.item",  
    "type": "message",  
    "status": "completed",  
    "role": "assistant",  
    "content": [  
      {  
        "type": "text",  
        "text": "Sure, I can help with that."  
      }  
    ]  
  }  
}
```



## response.content\_part.added



Returned when a new content part is added to an assistant message item during response generation.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item to which the content part was added.

---

**output\_index** integer

The index of the output item in the response.

---

**part** object

The content part that was added.

> Show properties

---

**response\_id** string

The ID of the response.

---

**type** string

The event type, must be `response.content_part.added`.

OBJECT `response.content_part.added`

```
{  
  "event_id": "event_3738",  
  "type": "response.content_part.added",  
  "response_id": "resp_001",  
  "item_id": "msg_007",  
  "output_index": 0,  
  "content_index": 0,  
  "part": {  
    "type": "text",  
    "text": ""  
  }  
}
```

# response.content\_part.done



Returned when a content part is done streaming in an assistant message item. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

---

**output\_index** integer

The index of the output item in the response.

---

**part** object

The content part that is done.

> Show properties

---

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.content_part.done`.

OBJECT `response.content_part.done`

```
{  
    "event_id": "event_3940",  
    "type": "response.content_part.done",  
    "response_id": "resp_001",  
    "item_id": "msg_007",  
    "output_index": 0,  
    "content_index": 0,  
    "part": {  
        "type": "text",  
        "text": "Sure, I can help with that."  
    }  
}
```



## response.output\_text.delta



Returned when the text value of an "output\_text" content part is updated.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**delta** string

The text delta.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

---

**type** string

The event type, must be `response.output_text.delta`.

OBJECT `response.output_text.delta`

```
{  
  "event_id": "event_4142",  
  "type": "response.output_text.delta",  
  "response_id": "resp_001",  
  "item_id": "msg_007",  
  "output_index": 0,  
  "content_index": 0,  
  "delta": "Sure, I can h"  
}
```

## response.output\_text.done



Returned when the text value of an "output\_text" content part is done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

### **content\_index** integer

The index of the content part in the item's content array.

### **event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**text** string

The final text content.

**type** string

The event type, must be `response.output_text.done`.

OBJECT `response.output_text.done`

```
{  
  "event_id": "event_4344",  
  "type": "response.output_text.done",  
  "response_id": "resp_001",  
  "item_id": "msg_007",  
  "output_index": 0,  
  "content_index": 0,
```



## response.output\_audio\_transcript.delta



Returned when the model-generated transcription of audio output is updated.

**content\_index** integer

The index of the content part in the item's content array.

**delta** string

The transcript delta.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_audio_transcript.delta`.

OBJECT `response.output_audio_transcript.delta`

```
{  
  "event_id": "event_4546",  
  "type": "response.output_audio_transcript.delta",  
  "response_id": "resp_001",  
  "item_id": "msg_008",  
  "output_index": 0,  
  "content_index": 0,  
  "delta": "Hello, how can I a"  
}
```

## **response.output\_audio\_transcript.done**



Returned when the model-generated transcription of audio output is done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

---

**transcript** string

The final transcript of the audio.

---

**type** string

The event type, must be `response.output_audio_transcript.done`.

OBJECT response.output\_audio\_transcript.done

```
{  
  "event_id": "event_4748",  
  "type": "response.output_audio_transcript.done",  
  "response_id": "resp_001",  
  "item_id": "msg_008",  
  "output_index": 0,  
  "content_index": 0,  
  "transcript": "Hello, how can I assist you today?"  
}
```



## response.output\_audio.delta



Returned when the model-generated audio is updated.

**content\_index** integer

The index of the content part in the item's content array.

**delta** string

Base64-encoded audio data delta.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_audio.delta`.

OBJECT `response.output_audio.delta`

{

```
"event_id": "event_4950",
"type": "response.output_audio.delta",
"response_id": "resp_001",
"item_id": "msg_008",
```

```
"output_index": 0,  
"content_index": 0,  
"delta": "Base64EncodedAudioDelta"  
}
```

## response.output\_audio.done



Returned when the model-generated audio is done. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

---

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_audio.done`.

OBJECT `response.output_audio.done`

```
{  
  "event_id": "event_5152",  
  "type": "response.output_audio.done",  
  "response_id": "resp_001",  
  "item_id": "msg_008",  
  "output_index": 0,  
  "content_index": 0  
}
```



## response.function\_call\_arguments.delta



Returned when the model-generated function call arguments are updated.

---

**call\_id** string

The ID of the function call.

---

**delta** string

The arguments delta as a JSON string.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the function call item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

---

**type** string

The event type, must be `response.function_call_arguments.delta`.

OBJECT `response.function_call_arguments.delta`

```
{  
  "event_id": "event_5354",  
  "type": "response.function_call_arguments.delta",  
  "response_id": "resp_002",  
  "item_id": "fc_001",  
  "output_index": 0,  
  "call_id": "call_001",  
  "delta": "{\"location\": \"San\\"",  
}
```

## response.function\_call\_arguments.done



Returned when the model-generated function call arguments are done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

**arguments** string

The final arguments as a JSON string.

**call\_id** string

The ID of the function call.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the function call item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

---

**type** string

The event type, must be `response.function_call_arguments.done`.

OBJECT `response.function_call_arguments.done`

```
{
```

```
  "event_id": "event_5556",
  "type": "response.function_call_arguments.done",
  "response_id": "resp_002",
  "item_id": "fc_001",
  "output_index": 0,
  "call_id": "call_001",
```

```
"arguments": "{\"location\": \"San Francisco\"}"  
}
```



## response.mcp\_call\_arguments.delta



Returned when MCP tool call arguments are updated during response generation.

---

**delta** string

The JSON-encoded arguments delta.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the MCP tool call item.

---

**obfuscation** string

If present, indicates the delta text was obfuscated.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.mcp_call_arguments.delta`.

OBJECT `response.mcp_call_arguments.delta`

```
{  
    "event_id": "event_6201",  
    "type": "response.mcp_call_arguments.delta",  
    "response_id": "resp_001",  
    "item_id": "mcp_call_001",  
    "output_index": 0,  
    "delta": "{\"partial\":true}"  
}
```

## response.mcp\_call\_arguments.done



Returned when MCP tool call arguments are finalized during response generation.

---

**arguments** string

The final JSON-encoded arguments string.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the MCP tool call item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

---

**type** string

The event type, must be `response.mcp_call_arguments.done`.

OBJECT `response.mcp_call_arguments.done`

{

`"event_id": "event_6202",`  
`"type": "response.mcp_call_arguments.done",`

```
"response_id": "resp_001",
"item_id": "mcp_call_001",
"output_index": 0,
"arguments": "{\"q\": \"docs\"}"
}
```



## response.mcp\_call.in\_progress



Returned when an MCP tool call has started and is in progress.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the MCP tool call item.

---

**output\_index** integer

The index of the output item in the response.

**type** string

The event type, must be `response.mcp_call.in_progress`.

OBJECT `response.mcp_call.in_progress`

```
{  
  "event_id": "event_6301",  
  "type": "response.mcp_call.in_progress",  
  "output_index": 0,  
  "item_id": "mcp_call_001"  
}
```

## response.mcp\_call.completed



Returned when an MCP tool call has completed successfully.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP tool call item.

**output\_index** integer

The index of the output item in the response.

**type** string

The event type, must be `response.mcp_call.completed`.

OBJECT `response.mcp_call.completed`

```
{  
  "event_id": "event_6302",  
  "type": "response.mcp_call.completed",  
  "output_index": 0,  
  "item_id": "mcp_call_001"  
}
```

## response.mcp\_call.failed



Returned when an MCP tool call has failed.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP tool call item.

**output\_index** integer

The index of the output item in the response.

**type** string

The event type, must be `response.mcp_call.failed`.

OBJECT `response.mcp_call.failed`

```
{  
  "event_id": "event_6303",  
  "type": "response.mcp_call.failed",  
  "output_index": 0,  
  "item_id": "mcp_call_001"  
}
```



# mcp\_list\_tools.in\_progress



Returned when listing MCP tools is in progress for an item.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP list tools item.

**type** string

The event type, must be `mcp_list_tools.in_progress`.

OBJECT `mcp_list_tools.in_progress`

```
{  
  "event_id": "event_6101",  
  "type": "mcp_list_tools.in_progress",  
  "item_id": "mcp_list_tools_001"  
}
```

# mcp\_list\_tools.completed



Returned when listing MCP tools has completed for an item.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the MCP list tools item.

---

**type** string

The event type, must be `mcp_list_tools.completed`.

OBJECT `mcp_list_tools.completed`

```
{  
  "event_id": "event_6102",  
  "type": "mcp_list_tools.completed",  
  "item_id": "mcp_list_tools_001"  
}
```

# mcp\_list\_tools.failed



Returned when listing MCP tools has failed for an item.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the MCP list tools item.

---

**type** string

The event type, must be `mcp_list_tools.failed`.

OBJECT `mcp_list_tools.failed`

```
{
```

```
  "event_id": "event_6103",
  "type": "mcp_list_tools.failed",
  "item_id": "mcp_list_tools_001"
```

```
}
```



# rate\_limits.updated



Emitted at the beginning of a Response to indicate the updated rate limits. When a Response is created some tokens will be "reserved" for the output tokens, the rate limits shown here reflect that reservation, which is then adjusted accordingly once the Response is completed.

---

**event\_id** string

The unique ID of the server event.

---

**rate\_limits** array

List of rate limit information.

> Show properties

---

**type** string

The event type, must be `rate_limits.updated`.

OBJECT `rate_limits.updated`

{

`"event_id": "event_5758",`  
  `"type": "rate_limits.updated",`  
  `"rate_limits": [`

```
{  
    "name": "requests",  
    "limit": 1000,  
    "remaining": 999,  
    "reset_seconds": 60  
,  
{  
    "name": "tokens",  
    "limit": 50000,  
    "remaining": 49950,  
    "reset_seconds": 60  
}  
]  
}
```

# Chat Completions



The Chat Completions API endpoint will generate a model response from a list of messages comprising a conversation.

Related guides:

[Quickstart](#)

[Text inputs and outputs](#)

[Image inputs](#)

[Audio inputs and outputs](#)

[Structured Outputs](#)

[Function calling](#)

[Conversation state](#)

**Starting a new project?** We recommend trying [Responses](#) to take advantage of the latest OpenAI platform features. Compare [Chat Completions with Responses](#).

---

## Create chat completion



POST `https://api.openai.com/v1/chat/completions`

**Starting a new project?** We recommend trying [Responses](#) to take advantage of the latest OpenAI platform features. Compare [Chat Completions with Responses](#).

---

Creates a model response for the given chat conversation. Learn more in the [text generation](#), [vision](#), and [audio](#) guides.

Parameter support can differ depending on the model used to generate the response, particularly for newer reasoning models. Parameters that are only supported for reasoning models are noted below. For the current state of unsupported parameters in reasoning models, [refer to the reasoning guide](#).

## Request body

---

### **messages** array Required

A list of messages comprising the conversation so far. Depending on the [model](#) you use, different message types (modalities) are supported, like [text](#), [images](#), and [audio](#).

> Show possible types

---

### **model** string Required

Model ID used to generate the response, like `gpt-4o` or `o3`. OpenAI offers a wide range of models with different capabilities, performance characteristics, and price points. Refer to the [model guide](#) to browse and compare available models.

---

### **audio** object or null Optional

Parameters for audio output. Required when audio output is requested with `modalities: ["audio"]`. [Learn more](#).

> Show properties

---

### **frequency\_penalty** number or null Optional Defaults to 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.

---

### **function\_call** Deprecated string or object Optional

Deprecated in favor of `tool_choice`.

Controls which (if any) function is called by the model.

`none` means the model will not 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 `{"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.

> Show possible types

---

**functions** Deprecated array Optional

Deprecated in favor of `tools`.

A list of functions the model may generate JSON inputs for.

> Show properties

---

**logit\_bias** map Optional Defaults to 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 will vary 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.

---

**logprobs** boolean or null Optional Defaults to false

Whether to return log probabilities of the output tokens or not. If true, returns the log probabilities of each output token returned in the `content` of `message`.

**max\_completion\_tokens** integer or null Optional

An upper bound for the number of tokens that can be generated for a completion, including visible output tokens and reasoning tokens.

---

**max\_tokens** Deprecated integer or null Optional

The maximum number of tokens that can be generated in the chat completion. This value can be used to control costs for text generated via API.

This value is now deprecated in favor of `max_completion_tokens`, and is not compatible with o-series models.

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**modalities** array Optional

Output types that you would like the model to generate. Most models are capable of generating text, which is the default:

`["text"]`

The `gpt-4o-audio-preview` model can also be used to generate audio. To request that this model generate both text and audio responses, you can use:

`["text", "audio"]`

---

**n** integer or null Optional Defaults to 1

How many chat completion choices to generate for each input message. Note that you will be charged based on the number of generated tokens across all of the choices. Keep `n` as `1` to minimize costs.

---

**parallel\_tool\_calls** boolean Optional Defaults to true

Whether to enable parallel function calling during tool use.

---

**prediction** object Optional

Configuration for a Predicted Output, which can greatly improve response times when large parts of the model response are known ahead of time. This is most common when you are regenerating a file with only minor changes to most of the content.

> Show possible types

---

**presence\_penalty** number or null Optional Defaults to 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.

---

**prompt\_cache\_key** string Optional

Used by OpenAI to cache responses for similar requests to optimize your cache hit rates. Replaces the `user` field.

[Learn more.](#)

---

**reasoning\_effort** string Optional Defaults to medium

Constrains effort on reasoning for reasoning models. Currently supported values are `minimal`, `low`, `medium`, and `high`.

Reducing reasoning effort can result in faster responses and fewer tokens used on reasoning in a response.

Note: The `gpt-5-pro` model defaults to (and only supports) `high` reasoning effort.

**response\_format** object Optional

An object specifying the format that the model must output.

Setting to `{ "type": "json_schema", "json_schema": { ... } }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables the older JSON mode, which ensures the message the model generates is valid JSON. Using `json_schema` is preferred for models that support it.

> Show possible types

---

**safety\_identifier** string Optional

A stable identifier used to help detect users of your application that may be violating OpenAI's usage policies. The IDs should be a string that uniquely identifies each user. We recommend hashing their username or email address, in order to avoid sending us any identifying information. [Learn more](#).

---

**seed** Deprecated integer or null Optional

This feature is in Beta. If specified, our system will make a best effort to sample deterministically, such that repeated requests with the same `seed` and parameters should return the same result. Determinism is not guaranteed, and you should refer to the `system_fingerprint` response parameter to monitor changes in the backend.

---

**service\_tier** string Optional Defaults to auto

Specifies the processing type used for serving the request.

If set to 'auto', then the request will be processed with the service tier configured in the Project settings. Unless otherwise configured, the Project will use 'default'.

If set to 'default', then the request will be processed with the standard pricing and performance for the selected model.

If set to '[flex](#)' or '[priority](#)', then the request will be processed with the corresponding service tier.

When not set, the default behavior is 'auto'.

When the `service_tier` parameter is set, the response body will include the `service_tier` value based on the processing mode actually used to serve the request. This response value may be different from the value set in the parameter.

---

**stop** string / array / null Optional Defaults to null

Not supported with latest reasoning models `o3` and `o4-mini`.

Up to 4 sequences where the API will stop generating further tokens. The returned text will not contain the stop sequence.

---

**store** boolean or null Optional Defaults to false

Whether or not to store the output of this chat completion request for use in our [model distillation](#) or [evals](#) products.

Supports text and image inputs. Note: image inputs over 8MB will be dropped.

---

**stream** boolean or null Optional Defaults to false

If set to true, the model response data will be streamed to the client as it is generated using [server-sent events](#). See the [Streaming section below](#) for more information, along with the [streaming responses](#) guide for more information on how to handle the streaming events.

---

**stream\_options** object Optional Defaults to null

Options for streaming response. Only set this when you set `stream: true`.

> Show properties

---

**temperature** number Optional Defaults to 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.

---

**tool\_choice** string or object Optional

Controls which (if any) tool is called by the model. `none` means the model will not call any tool and instead generates a message. `auto` means the model can pick between generating a message or calling one or more tools. `required` means the model must call one or more tools. Specifying a particular tool via

```
{"type": "function", "function": {"name": "my_function"}}
```

 forces the model to call that tool.

`none` is the default when no tools are present. `auto` is the default if tools are present.

> Show possible types

---

**tools** array Optional

A list of tools the model may call. You can provide either [custom tools](#) or [function tools](#).

> Show possible types

---

**top\_logprobs** integer Optional

An integer between 0 and 20 specifying the number of most likely tokens to return at each token position, each with an associated log probability.

---

**top\_p** number Optional Defaults to 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.

**user** Deprecated string Optional

This field is being replaced by `safety_identifier` and `prompt_cache_key`. Use `prompt_cache_key` instead to maintain caching optimizations. A stable identifier for your end-users. Used to boost cache hit rates by better bucketing similar requests and to help OpenAI detect and prevent abuse. [Learn more](#).

**verbosity** string Optional Defaults to medium

Constrains the verbosity of the model's response. Lower values will result in more concise responses, while higher values will result in more verbose responses. Currently supported values are `low`, `medium`, and `high`.

**web\_search\_options** object Optional

This tool searches the web for relevant results to use in a response. Learn more about the [web search tool](#).

> Show properties

## Returns

Returns a [chat completion](#) object, or a streamed sequence of [chat completion chunk](#) objects if the request is streamed.

**Default**   **Image input**   **Streaming**   **Functions**   **Logprobs**

Example request

```
curl https://api.openai.com/v1/chat/completions \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-d '{'
```

```
"model": "gpt-5",
"messages": [
  {
    "role": "developer",
    "content": "You are a helpful assistant."
  },
  {
    "role": "user",
    "content": "Hello!"
  }
]
}'
```

## Response

```
{
  "id": "chatcmpl-B9MBs8CjcvOU2jLn4n570S5qMJKcT",
  "object": "chat.completion",
  "created": 1741569952,
  "model": "gpt-4.1-2025-04-14",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "Hello! How can I assist you today?",
        "refusal": null,
      }
    }
  ]
}'
```

```
        "annotations": [],
    },
    "logprobs": null,
    "finish_reason": "stop"
}
],
"usage": {
    "prompt_tokens": 19,
    "completion_tokens": 10,
    "total_tokens": 29,
    "prompt_tokens_details": {
        "cached_tokens": 0,
        "audio_tokens": 0
    },
    "completion_tokens_details": {
        "reasoning_tokens": 0,
        "audio_tokens": 0,
        "accepted_prediction_tokens": 0,
        "rejected_prediction_tokens": 0
    }
},
"service_tier": "default"
}
```

# Get chat completion



```
GET https://api.openai.com/v1/chat/completions/{completion_id}
```

Get a stored chat completion. Only Chat Completions that have been created with the `store` parameter set to `true` will be returned.

## Path parameters

`completion_id` string Required

The ID of the chat completion to retrieve.

## Returns

The [ChatCompletion](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/chat/completions/chatcmpl-abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "chat.completion",  
  "id": "chatcmpl-abc123",  
  "model": "gpt-4o-2024-08-06",  
  "created": 1738960610,  
  "request_id": "req_ded8ab984ec4bf840f37566c1011c417",  
  "tool_choice": null,  
  "usage": {  
    "total_tokens": 31,  
    "completion_tokens": 18,  
    "prompt_tokens": 13  
  },  
  "seed": 4944116822809979520,  
  "top_p": 1.0,  
  "temperature": 1.0,  
  "presence_penalty": 0.0,  
  "frequency_penalty": 0.0,  
  "system_fingerprint": "fp_50cad350e4",  
  "input_user": null,  
  "service_tier": "default",  
  "tools": null,  
  "metadata": {},  
  "choices": [  
    {  
      "index": 0,  
      "message": {  
        "content": "Mind of circuits hum, \nLearning patterns in silence— \nFuture's quiet spark.",  
        "role": "assistant",  
        "tool_calls": null  
      }  
    }  
  ]  
}
```

```
        "function_call": null
    },
    "finish_reason": "stop",
    "logprobs": null
}
],
"response_format": null
}
```

## Get chat messages



GET [https://api.openai.com/v1/chat/completions/{completion\\_id}/messages](https://api.openai.com/v1/chat/completions/{completion_id}/messages)

Get the messages in a stored chat completion. Only Chat Completions that have been created with the `store` parameter set to `true` will be returned.

### Path parameters

`completion_id` string Required

The ID of the chat completion to retrieve messages from.

## Query parameters

**after** string Optional

Identifier for the last message from the previous pagination request.

**limit** integer Optional Defaults to 20

Number of messages to retrieve.

**order** string Optional Defaults to asc

Sort order for messages by timestamp. Use `asc` for ascending order or `desc` for descending order. Defaults to `asc`.

## Returns

A list of messages for the specified chat completion.

### Example request

```
curl https://api.openai.com/v1/chat/completions/chat_abc123/messages \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "list",
```

```
"data": [
  {
    "id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2-0",
    "role": "user",
    "content": "write a haiku about ai",
    "name": null,
    "content_parts": null
  }
],
"first_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2-0",
"last_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2-0",
"has_more": false
}
```

# List Chat Completions



GET <https://api.openai.com/v1/chat/completions>

List stored Chat Completions. Only Chat Completions that have been stored with the `store` parameter set to `true` will be returned.

## Query parameters

---

**after** string Optional

Identifier for the last chat completion from the previous pagination request.

---

**limit** integer Optional Defaults to 20

Number of Chat Completions to retrieve.

---

**metadata** object or null Optional

A list of metadata keys to filter the Chat Completions by. Example:

```
metadata[key1]=value1&metadata[key2]=value2
```

---

**model** string Optional

The model used to generate the Chat Completions.

---

**order** string Optional Defaults to asc

Sort order for Chat Completions by timestamp. Use `asc` for ascending order or `desc` for descending order. Defaults to `asc`.

---

## Returns

---

A list of [Chat Completions](#) matching the specified filters.

### Example request

```
curl https://api.openai.com/v1/chat/completions \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "object": "chat.completion",
      "id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2",
      "model": "gpt-4.1-2025-04-14",
      "created": 1738960610,
      "request_id": "req_ded8ab984ec4bf840f37566c1011c417",
      "tool_choice": null,
      "usage": {
        "total_tokens": 31,
        "completion_tokens": 18,
        "prompt_tokens": 13
      },
      "seed": 4944116822809979520,
      "top_p": 1.0,
      "temperature": 1.0,
      "presence_penalty": 0.0,
      "frequency_penalty": 0.0,
      "system_fingerprint": "fp_50cad350e4",
    }
  ]
}
```

```
"input_user": null,  
"service_tier": "default",  
"tools": null,  
"metadata": {},  
"choices": [  
    {  
        "index": 0,  
        "message": {  
            "content": "Mind of circuits hum, \nLearning patterns in silence— \nFuture's quiet spa  
            "role": "assistant",  
            "tool_calls": null,  
            "function_call": null  
        },  
        "finish_reason": "stop",  
        "logprobs": null  
    }  
],  
"response_format": null  
}  
],  
"first_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMf1mj2",  
"last_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMf1mj2",  
"has_more": false  
}
```

# Update chat completion



```
POST https://api.openai.com/v1/chat/completions/{completion_id}
```

Modify a stored chat completion. Only Chat Completions that have been created with the `store` parameter set to `true` can be modified. Currently, the only supported modification is to update the `metadata` field.

## Path parameters

---

`completion_id` string Required

The ID of the chat completion to update.

## Request body

---

`metadata` map Required

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

## Returns

The [ChatCompletion](#) object matching the specified ID.

#### Example request

```
curl -X POST https://api.openai.com/v1/chat/completions/chat_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{"metadata": {"foo": "bar"}}'
```

#### Response

```
{
  "object": "chat.completion",
  "id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMf1mj2",
  "model": "gpt-4o-2024-08-06",
  "created": 1738960610,
  "request_id": "req_ded8ab984ec4bf840f37566c1011c417",
  "tool_choice": null,
  "usage": {
    "total_tokens": 31,
    "completion_tokens": 18,
    "prompt_tokens": 13
  },
  "seed": 4944116822809979520,
  "top_p": 1.0,
  "temperature": 1.0,
  "presence_penalty": 0.0,
```

```
"frequency_penalty": 0.0,  
"system_fingerprint": "fp_50cad350e4",  
"input_user": null,  
"service_tier": "default",  
"tools": null,  
"metadata": {  
    "foo": "bar"  
},  
"choices": [  
    {  
        "index": 0,  
        "message": {  
            "content": "Mind of circuits hum, \nLearning patterns in silence— \nFuture's quiet spark."  
            "role": "assistant",  
            "tool_calls": null,  
            "function_call": null  
        },  
        "finish_reason": "stop",  
        "logprobs": null  
    }  
],  
"response_format": null  
}
```

# Delete chat completion



```
DELETE https://api.openai.com/v1/chat/completions/{completion_id}
```

Delete a stored chat completion. Only Chat Completions that have been created with the `store` parameter set to `true` can be deleted.

## Path parameters

`completion_id` string Required

The ID of the chat completion to delete.

## Returns

A deletion confirmation object.

### Example request

```
curl -X DELETE https://api.openai.com/v1/chat/completions/chat_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "chat.completion.deleted",  
  "id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMf1mj2",  
  "deleted": true  
}
```

# The chat completion object



Represents a chat completion response returned by model, based on the provided input.

## choices array

A list of chat completion choices. Can be more than one if `n` is greater than 1.

> Show properties

## created integer

The Unix timestamp (in seconds) of when the chat completion was created.

## id string

A unique identifier for the chat completion.

## model string

The model used for the chat completion.

---

**object** string

The object type, which is always `chat.completion`.

---

**service\_tier** string

Specifies the processing type used for serving the request.

If set to 'auto', then the request will be processed with the service tier configured in the Project settings. Unless otherwise configured, the Project will use 'default'.

If set to 'default', then the request will be processed with the standard pricing and performance for the selected model.

If set to 'flex' or 'priority', then the request will be processed with the corresponding service tier.

When not set, the default behavior is 'auto'.

When the `service_tier` parameter is set, the response body will include the `service_tier` value based on the processing mode actually used to serve the request. This response value may be different from the value set in the parameter.

---

**system\_fingerprint** Deprecated string

This fingerprint represents the backend configuration that the model runs with.

Can be used in conjunction with the `seed` request parameter to understand when backend changes have been made that might impact determinism.

---

**usage** object

Usage statistics for the completion request.

&gt; Show properties

OBJECT The chat completion object

```
{  
  "id": "chatcmpl-B9MHDbslfkBeAs8l4bebGdFOJ6PeG",  
  "object": "chat.completion",  
  "created": 1741570283,  
  "model": "gpt-4o-2024-08-06",  
  "choices": [  
    {  
      "index": 0,  
      "message": {  
        "role": "assistant",  
        "content": "The image shows a wooden boardwalk path running through a lush green field or meadow. The path is made of light-colored wood planks and leads towards a distant, bright horizon. The surrounding area is filled with tall grass and low-lying flowers. The sky is clear and blue.",  
        "refusal": null,  
        "annotations": []  
      },  
      "logprobs": null,  
      "finish_reason": "stop"  
    }  
  "usage": {  
    "prompt_tokens": 1117,  
    "completion_tokens": 46,  
    "total_tokens": 1163,  
    "prompt_tokens_details": {  
      "cached_tokens": 0,  
      "audio_tokens": 0  
    },  
  },  
}
```

```
"completion_tokens_details": {  
    "reasoning_tokens": 0,  
    "audio_tokens": 0,  
    "accepted_prediction_tokens": 0,  
    "rejected_prediction_tokens": 0  
}  
},  
"service_tier": "default",  
"system_fingerprint": "fp_fc9f1d7035"  
}
```

# The chat completion list object



An object representing a list of Chat Completions.

---

## **data** array

An array of chat completion objects.

> Show properties

---

## **first\_id** string

The identifier of the first chat completion in the data array.

**has\_more** boolean

Indicates whether there are more Chat Completions available.

**last\_id** string

The identifier of the last chat completion in the data array.

**object** string

The type of this object. It is always set to "list".

OBJECT The chat completion list object

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "chat.completion",  
      "id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2",  
      "model": "gpt-4o-2024-08-06",  
      "created": 1738960610,  
      "request_id": "req_ded8ab984ec4bf840f37566c1011c417",  
      "tool_choice": null,  
      "usage": {  
        "total_tokens": 31,  
        "completion_tokens": 18,  
        "prompt_tokens": 13  
      },  
      "seed": 4944116822809979520,  
      "content": "Hello, how can I assist you today?"  
    }  
  ]  
}
```

```
"top_p": 1.0,
"temperature": 1.0,
"presence_penalty": 0.0,
"frequency_penalty": 0.0,
"system_fingerprint": "fp_50cad350e4",
"input_user": null,
"service_tier": "default",
"tools": null,
"metadata": {},
"choices": [
{
  "index": 0,
  "message": {
    "content": "Mind of circuits hum, \nLearning patterns in silence— \nFuture's quiet spa",
    "role": "assistant",
    "tool_calls": null,
    "function_call": null
  },
  "finish_reason": "stop",
  "logprobs": null
},
],
"response_format": null
},
],
"first_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMf1mj2",
"last_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMf1mj2",
"has_more": false
}
```

# The chat completion message list object



An object representing a list of chat completion messages.

---

## **data** array

An array of chat completion message objects.

> Show properties

---

## **first\_id** string

The identifier of the first chat message in the data array.

---

## **has\_more** boolean

Indicates whether there are more chat messages available.

---

## **last\_id** string

The identifier of the last chat message in the data array.

---

## **object** string

The type of this object. It is always set to "list".

## OBJECT The chat completion message list object

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2-0",  
      "role": "user",  
      "content": "write a haiku about ai",  
      "name": null,  
      "content_parts": null  
    }  
,  
  "first_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2-0",  
  "last_id": "chatcmpl-AyPNinnUqUDYo9SAdA52NobMflmj2-0",  
  "has_more": false  
}
```

# Streaming



Stream Chat Completions in real time. Receive chunks of completions returned from the model using server-sent events. [Learn more.](#)

---

# The chat completion chunk object



Represents a streamed chunk of a chat completion response returned by the model, based on the provided input. [Learn more.](#)

---

## **choices** array

A list of chat completion choices. Can contain more than one elements if `n` is greater than 1. Can also be empty for the last chunk if you set `stream_options: {"include_usage": true}`.

> Show properties

---

## **created** integer

The Unix timestamp (in seconds) of when the chat completion was created. Each chunk has the same timestamp.

---

## **id** string

A unique identifier for the chat completion. Each chunk has the same ID.

---

## **model** string

The model to generate the completion.

**object** string

The object type, which is always `chat.completion.chunk`.

**service\_tier** string

Specifies the processing type used for serving the request.

If set to 'auto', then the request will be processed with the service tier configured in the Project settings. Unless otherwise configured, the Project will use 'default'.

If set to 'default', then the request will be processed with the standard pricing and performance for the selected model.

If set to 'flex' or 'priority', then the request will be processed with the corresponding service tier.

When not set, the default behavior is 'auto'.

When the `service_tier` parameter is set, the response body will include the `service_tier` value based on the processing mode actually used to serve the request. This response value may be different from the value set in the parameter.

**system\_fingerprint** Deprecated string

This fingerprint represents the backend configuration that the model runs with. Can be used in conjunction with the `seed` request parameter to understand when backend changes have been made that might impact determinism.

**usage** object or null

Usage statistics for the completion request.

> Show properties

OBJECT The chat completion chunk object

```
{"id": "chatcmpl-123", "object": "chat.completion.chunk", "created": 1694268190, "model": "gpt-4o-mini", "s...  
{"id": "chatcmpl-123", "object": "chat.completion.chunk", "created": 1694268190, "model": "gpt-4o-mini", "s...  
....  
{"id": "chatcmpl-123", "object": "chat.completion.chunk", "created": 1694268190, "model": "gpt-4o-mini", "s...
```

## Assistants

Beta



Build assistants that can call models and use tools to perform tasks.

[Get started with the Assistants API](#)

## Create assistant

Beta



POST <https://api.openai.com/v1/assistants>

Create an assistant with a model and instructions.

## Request body

---

**model** string Required

ID of the model to use. You can use the [List models API](#) to see all of your available models, or see our [Model overview](#) for descriptions of them.

---

**description** string Optional

The description of the assistant. The maximum length is 512 characters.

---

**instructions** string Optional

The system instructions that the assistant uses. The maximum length is 256,000 characters.

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**name** string Optional

The name of the assistant. The maximum length is 256 characters.

**reasoning\_effort** string Optional Defaults to medium

Constrains effort on reasoning for [reasoning models](#). Currently supported values are `minimal`, `low`, `medium`, and `high`.

Reducing reasoning effort can result in faster responses and fewer tokens used on reasoning in a response.

Note: The `gpt-5-pro` model defaults to (and only supports) `high` reasoning effort.

---

**response\_format** "auto" or object Optional

Specifies the format that the model must output. Compatible with [GPT-4o](#), [GPT-4 Turbo](#), and all GPT-3.5 Turbo models since `gpt-3.5-turbo-1106`.

Setting to `{ "type": "json_schema", "json_schema": { ... } }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables JSON mode, which ensures the message the model generates is valid JSON.

**Important:** when using JSON mode, you **must** also instruct the model to produce JSON yourself via a system or user message. Without this, the model may generate an unending stream of whitespace until the generation reaches the token limit, resulting in a long-running and seemingly "stuck" request. Also note that the message content may be partially cut off if `finish_reason="length"`, which indicates the generation exceeded `max_tokens` or the conversation exceeded the max context length.

> Show possible types

---

**temperature** number Optional Defaults to 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.

**tool\_resources** object Optional

A set of resources that are used by the assistant's tools. The resources are specific to the type of tool. For example, the

`code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

**tools** array Optional Defaults to []

A list of tool enabled on the assistant. There can be a maximum of 128 tools per assistant. Tools can be of types

`code_interpreter`, `file_search`, or `function`.

> Show possible types

**top\_p** number Optional Defaults to 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.

## Returns

An [assistant](#) object.

**Code Interpreter**

**Files**

Example request

```
curl "https://api.openai.com/v1/assistants" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "instructions": "You are a personal math tutor. When asked a question, write and run Python code",
  "name": "Math Tutor",
  "tools": [{"type": "code_interpreter"}],
  "model": "gpt-4o"
}'
```

### Response

```
{
  "id": "asst_abc123",
  "object": "assistant",
  "created_at": 1698984975,
  "name": "Math Tutor",
  "description": null,
  "model": "gpt-4o",
  "instructions": "You are a personal math tutor. When asked a question, write and run Python code",
  "tools": [
    {
      "type": "code_interpreter"
    }
  ],
  "metadata": {}}
```

```
"top_p": 1.0,  
"temperature": 1.0,  
"response_format": "auto"  
}
```

# List assistants Beta



GET <https://api.openai.com/v1/assistants>

Returns a list of assistants.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

**before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the

previous page of the list.

---

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

---

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

---

A list of assistant objects.

### Example request

```
curl "https://api.openai.com/v1/assistants?order=desc&limit=20" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

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

```
"id": "asst_abc123",
"object": "assistant",
"created_at": 1698982736,
"name": "Coding Tutor",
"description": null,
"model": "gpt-4o",
"instructions": "You are a helpful assistant designed to make me better at coding!",
"tools": [],
"tool_resources": {},
"metadata": {},
"top_p": 1.0,
"temperature": 1.0,
"response_format": "auto"
},
{
"id": "asst_abc456",
"object": "assistant",
"created_at": 1698982718,
"name": "My Assistant",
"description": null,
"model": "gpt-4o",
"instructions": "You are a helpful assistant designed to make me better at coding!",
"tools": [],
"tool_resources": {},
"metadata": {},
"top_p": 1.0,
"temperature": 1.0,
"response_format": "auto"
},
```

```
{  
    "id": "asst_abc789",  
    "object": "assistant",  
    "created_at": 1698982643,  
    "name": null,  
    "description": null,  
    "model": "gpt-4o",  
    "instructions": null,  
    "tools": [],  
    "tool_resources": {},  
    "metadata": {},  
    "top_p": 1.0,  
    "temperature": 1.0,  
    "response_format": "auto"  
}  
],  
"first_id": "asst_abc123",  
"last_id": "asst_abc789",  
"has_more": false  
}
```

## Retrieve assistant

Beta



GET https://api.openai.com/v1/assistants/{assistant\_id}

Retrieves an assistant.

## Path parameters

**assistant\_id** string Required

The ID of the assistant to retrieve.

# Returns

The assistant object matching the specified ID.

## Example request

```
curl https://api.openai.com/v1/assistants/asst_abc123 \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -H "OpenAI-Beta: assistants=v2"
```

## Response

```
{  
  "id": "asst_abc123",  
  "object": "assistant",
```

```
"created_at": 1699009709,  
"name": "HR Helper",  
"description": null,  
"model": "gpt-4o",  
"instructions": "You are an HR bot, and you have access to files to answer employee questions about company policies.",  
"tools": [  
  {  
    "type": "file_search"  
  }  
,  
  {"  
    "type": "file_upload"  
  }  
],  
"metadata": {},  
"top_p": 1.0,  
"temperature": 1.0,  
"response_format": "auto"  
}
```

## Modify assistant Beta



POST [https://api.openai.com/v1/assistants/{assistant\\_id}](https://api.openai.com/v1/assistants/{assistant_id})

Modifies an assistant.

## Path parameters

---

### **assistant\_id** string Required

The ID of the assistant to modify.

## Request body

---

### **description** string Optional

The description of the assistant. The maximum length is 512 characters.

### **instructions** string Optional

The system instructions that the assistant uses. The maximum length is 256,000 characters.

### **metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

### **model** string Optional

ID of the model to use. You can use the [List models](#) API to see all of your available models, or see our [Model overview](#) for descriptions of them.

### **name** string Optional

The name of the assistant. The maximum length is 256 characters.

---

**reasoning\_effort** string Optional Defaults to medium

Constrains effort on reasoning for [reasoning models](#). Currently supported values are `minimal`, `low`, `medium`, and `high`.

Reducing reasoning effort can result in faster responses and fewer tokens used on reasoning in a response.

Note: The `gpt-5-pro` model defaults to (and only supports) `high` reasoning effort.

---

**response\_format** "auto" or object Optional

Specifies the format that the model must output. Compatible with [GPT-4o](#), [GPT-4 Turbo](#), and all GPT-3.5 Turbo models since `gpt-3.5-turbo-1106`.

Setting to `{ "type": "json_schema", "json_schema": { ... } }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables JSON mode, which ensures the message the model generates is valid JSON.

**Important:** when using JSON mode, you **must** also instruct the model to produce JSON yourself via a system or user message. Without this, the model may generate an unending stream of whitespace until the generation reaches the token limit, resulting in a long-running and seemingly "stuck" request. Also note that the message content may be partially cut off if `finish_reason="length"`, which indicates the generation exceeded `max_tokens` or the conversation exceeded the max context length.

> Show possible types

---

**temperature** number Optional Defaults to 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.

---

**tool\_resources** object Optional

A set of resources that are used by the assistant's tools. The resources are specific to the type of tool. For example, the `code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

---

**tools** array Optional Defaults to []

A list of tool enabled on the assistant. There can be a maximum of 128 tools per assistant. Tools can be of types `code_interpreter`, `file_search`, or `function`.

> Show possible types

---

**top\_p** number Optional Defaults to 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.

---

## Returns

The modified assistant object.

---

### Example request

```
curl https://api.openai.com/v1/assistants/asst_abc123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "instructions": "You are an HR bot, and you have access to files to answer employee questions",
  "tools": [{"type": "file_search"}],
  "model": "gpt-4o"
}'
```

## Response

```
{
  "id": "asst_123",
  "object": "assistant",
  "created_at": 1699009709,
  "name": "HR Helper",
  "description": null,
  "model": "gpt-4o",
  "instructions": "You are an HR bot, and you have access to files to answer employee questions abc123",
  "tools": [
    {
      "type": "file_search"
    }
  ],
  "tool_resources": {
    "file_search": {
      "id": "file_search_123"
    }
  }
}
```

```
    "vector_store_ids": [],
  },
},
"metadata": {},
"top_p": 1.0,
"temperature": 1.0,
"response_format": "auto"
}
```

## Delete assistant Beta



```
DELETE https://api.openai.com/v1/assistants/{assistant_id}
```

Delete an assistant.

### Path parameters

**assistant\_id** string Required

The ID of the assistant to delete.

### Returns

## Deletion status

### Example request

```
curl https://api.openai.com/v1/assistants/asst_abc123 \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -H "OpenAI-Beta: assistants=v2" \
  -X DELETE
```

### Response

```
{  
  "id": "asst_abc123",  
  "object": "assistant.deleted",  
  "deleted": true  
}
```

# The assistant object Beta



Represents an [assistant](#) that can call the model and use tools.

**created\_at** integer

The Unix timestamp (in seconds) for when the assistant was created.

---

**description** string

The description of the assistant. The maximum length is 512 characters.

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**instructions** string

The system instructions that the assistant uses. The maximum length is 256,000 characters.

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string

ID of the model to use. You can use the [List models](#) API to see all of your available models, or see our [Model overview](#) for descriptions of them.

---

**name** string

The name of the assistant. The maximum length is 256 characters.

---

**object** string

The object type, which is always `assistant`.

**response\_format** "auto" or object

Specifies the format that the model must output. Compatible with [GPT-4o](#), [GPT-4 Turbo](#), and all GPT-3.5 Turbo models since `gpt-3.5-turbo-1106`.

Setting to `{ "type": "json_schema", "json_schema": { ... } }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables JSON mode, which ensures the message the model generates is valid JSON.

**Important:** when using JSON mode, you **must** also instruct the model to produce JSON yourself via a system or user message. Without this, the model may generate an unending stream of whitespace until the generation reaches the token limit, resulting in a long-running and seemingly "stuck" request. Also note that the message content may be partially cut off if

`finish_reason="length"`, which indicates the generation exceeded `max_tokens` or the conversation exceeded the max context length.

> Show possible types

**temperature** number

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.

**tool\_resources** object

A set of resources that are used by the assistant's tools. The resources are specific to the type of tool. For example, the `code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

### tools array

A list of tool enabled on the assistant. There can be a maximum of 128 tools per assistant. Tools can be of types

`code_interpreter`, `file_search`, or `function`.

> Show possible types

### top\_p number

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.

OBJECT The assistant object

```
{  
  "id": "asst_abc123",  
  "object": "assistant",  
  "created_at": 1698984975,  
  "name": "Math Tutor",  
  "description": null,  
  "model": "gpt-4o",  
  "instructions": "You are a personal math tutor. When asked a question, write and run Python code",  
  "tools": [  
    {  
      "type": "code_interpreter"  
    }  
  ],  
  "metadata": {}  
}
```

```
"top_p": 1.0,  
"temperature": 1.0,  
"response_format": "auto"  
}
```

## Threads Beta



Create threads that assistants can interact with.

Related guide: [Assistants](#)

### Create thread Beta



POST <https://api.openai.com/v1/threads>

Create a thread.

## Request body

---

**messages** array Optional

A list of [messages](#) to start the thread with.

> Show properties

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**tool\_resources** object Optional

A set of resources that are made available to the assistant's tools in this thread. The resources are specific to the type of tool.

For example, the `code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

## Returns

---

A [thread](#) object.

Empty

Messages

Example request

```
curl https://api.openai.com/v1/threads \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-d ''
```

## Response

```
{  
  "id": "thread_abc123",  
  "object": "thread",  
  "created_at": 1699012949,  
  "metadata": {},  
  "tool_resources": {}  
}
```

# Retrieve thread Beta



```
GET https://api.openai.com/v1/threads/{thread_id}
```

Retrieves a thread.

## Path parameters

**thread\_id** string **Required**

The ID of the thread to retrieve.

## Returns

The [thread](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "id": "thread_abc123",
  "object": "thread",
  "created_at": 1699014083,
  "metadata": {},
  "tool_resources": {
    "code_interpreter": {
      "file_ids": []
    }
  }
}
```

```
    }  
}  
}
```

# Modify thread Beta



POST [https://api.openai.com/v1/threads/{thread\\_id}](https://api.openai.com/v1/threads/{thread_id})

Modifies a thread.

## Path parameters

**thread\_id** string Required

The ID of the thread to modify. Only the `metadata` can be modified.

## Request body

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

#### **tool\_resources** object Optional

A set of resources that are made available to the assistant's tools in this thread. The resources are specific to the type of tool.

For example, the `code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

## Returns

The modified `thread` object matching the specified ID.

#### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "metadata": {
    "modified": "true",
    "user": "abc123"
}'
```

{}

## Response

```
{  
  "id": "thread_abc123",  
  "object": "thread",  
  "created_at": 1699014083,  
  "metadata": {  
    "modified": "true",  
    "user": "abc123"  
  },  
  "tool_resources": {}  
}
```

## Delete thread Beta



```
DELETE https://api.openai.com/v1/threads/{thread_id}
```

Delete a thread.

### Path parameters

**thread\_id** string Required

The ID of the thread to delete.

## Returns

Deletion status

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123 \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -H "OpenAI-Beta: assistants=v2" \
  -X DELETE
```

### Response

```
{
  "id": "thread_abc123",
  "object": "thread.deleted",
  "deleted": true
}
```

# The thread object

Beta



Represents a thread that contains [messages](#).

---

**created\_at** integer

The Unix timestamp (in seconds) for when the thread was created.

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**object** string

The object type, which is always `thread`.

---

**tool\_resources** object

A set of resources that are made available to the assistant's tools in this thread. The resources are specific to the type of tool.

For example, the `code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

**OBJECT** The thread object

```
{  
  "id": "thread_abc123",  
  "object": "thread",  
  "created_at": 1698107661,  
  "metadata": {}  
}
```

## Messages Beta



Create messages within threads

Related guide: [Assistants](#)

## Create message Beta



POST [https://api.openai.com/v1/threads/{thread\\_id}/messages](https://api.openai.com/v1/threads/{thread_id}/messages)

## Create a message.

### Path parameters

---

**thread\_id** string Required

The ID of the thread to create a message for.

### Request body

---

**content** string or array Required

> Show possible types

**role** string Required

The role of the entity that is creating the message. Allowed values include:

`user` : Indicates the message is sent by an actual user and should be used in most cases to represent user-generated messages.

`assistant` : Indicates the message is generated by the assistant. Use this value to insert messages from the assistant into the conversation.

---

**attachments** array Optional

A list of files attached to the message, and the tools they should be added to.

> Show properties

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

## Returns

A message object.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/messages \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
    "role": "user",
    "content": "How does AI work? Explain it in simple terms."
}'
```

### Response

```
{  
    "id": "msg_abc123",  
    "object": "thread.message",  
    "role": "user",  
    "content": "How does AI work? Explain it in simple terms.",  
    "created": 1600000000,  
    "model": "text-davinci-003",  
    "parent_message_id": null  
}
```

```
"created_at": 1713226573,  
"assistant_id": null,  
"thread_id": "thread_abc123",  
"run_id": null,  
"role": "user",  
"content": [  
  {  
    "type": "text",  
    "text": {  
      "value": "How does AI work? Explain it in simple terms.",  
      "annotations": []  
    }  
  },  
  ],  
  "attachments": [],  
  "metadata": {}  
}
```

## List messages Beta



GET [https://api.openai.com/v1/threads/{thread\\_id}/messages](https://api.openai.com/v1/threads/{thread_id}/messages)

Returns a list of messages for a given thread.

## Path parameters

---

### **thread\_id** string Required

The ID of the [thread](#) the messages belong to.

## Query parameters

---

### **after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

### **before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the previous page of the list.

### **limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

### **order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

### **run\_id** string Optional

Filter messages by the run ID that generated them.

## Returns

A list of message objects.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/messages \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "id": "msg_abc123",
      "object": "thread.message",
      "created_at": 1699016383,
      "assistant_id": null,
      "thread_id": "thread_abc123",
      "run_id": null,
      "role": "user",
```

```
"content": [
  {
    "type": "text",
    "text": {
      "value": "How does AI work? Explain it in simple terms.",
      "annotations": []
    }
  }
],
"attachments": [],
"metadata": {}
},
{
  "id": "msg_abc456",
  "object": "thread.message",
  "created_at": 1699016383,
  "assistant_id": null,
  "thread_id": "thread_abc123",
  "run_id": null,
  "role": "user",
  "content": [
    {
      "type": "text",
      "text": {
        "value": "Hello, what is AI?",
        "annotations": []
      }
    }
  ],
}
```

```
        "attachments": [],
        "metadata": {}
    },
],
"first_id": "msg_abc123",
"last_id": "msg_abc456",
"has_more": false
}
```

## Retrieve message

Beta



GET [https://api.openai.com/v1/threads/{thread\\_id}/messages/{message\\_id}](https://api.openai.com/v1/threads/{thread_id}/messages/{message_id})

Retrieve a message.

### Path parameters

**message\_id** string Required

The ID of the message to retrieve.

**thread\_id** string Required

The ID of the thread to which this message belongs.

## Returns

The message object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/messages/msg_abc123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "id": "msg_abc123",
  "object": "thread.message",
  "created_at": 1699017614,
  "assistant_id": null,
  "thread_id": "thread_abc123",
  "run_id": null,
  "role": "user",
  "content": [
    {
      "type": "text",
      "text": {
        "value": "How does AI work? Explain it in simple terms.",
        "annotations": []
      }
    }
  ]
}
```

```
    },
  ],
  "attachments": [],
  "metadata": {}
}
```

# Modify message Beta



POST [https://api.openai.com/v1/threads/{thread\\_id}/messages/{message\\_id}](https://api.openai.com/v1/threads/{thread_id}/messages/{message_id})

Modifies a message.

## Path parameters

**message\_id** string Required

The ID of the message to modify.

**thread\_id** string Required

The ID of the thread to which this message belongs.

## Request body

### metadata map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

## Returns

The modified [message](#) object.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/messages/msg_abc123 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "metadata": {
    "modified": "true",
    "user": "abc123"
  }
}'
```

## Response

```
{  
  "id": "msg_abc123",  
  "object": "thread.message",  
  "created_at": 1699017614,  
  "assistant_id": null,  
  "thread_id": "thread_abc123",  
  "run_id": null,  
  "role": "user",  
  "content": [  
    {  
      "type": "text",  
      "text": {  
        "value": "How does AI work? Explain it in simple terms.",  
        "annotations": []  
      }  
    }  
  ],  
  "file_ids": [],  
  "metadata": {  
    "modified": "true",  
    "user": "abc123"  
  }  
}
```

# Delete message

Beta



```
DELETE https://api.openai.com/v1/threads/{thread_id}/messages/{message_id}
```

Deletes a message.

## Path parameters

**message\_id** string Required

The ID of the message to delete.

**thread\_id** string Required

The ID of the thread to which this message belongs.

## Returns

Deletion status

Example request

```
curl -X DELETE https://api.openai.com/v1/threads/thread_abc123/messages/msg_abc123 \
-H "Content-Type: application/json" \
```

```
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta-Assistants=v2"
```

## Response

```
{  
  "id": "msg_abc123",  
  "object": "thread.message.deleted",  
  "deleted": true  
}
```

# The message object Beta



Represents a message within a [thread](#).

### **assistant\_id** string

If applicable, the ID of the [assistant](#) that authored this message.

### **attachments** array

A list of files attached to the message, and the tools they were added to.

> Show properties

### **completed\_at** integer

The Unix timestamp (in seconds) for when the message was completed.

---

**content** array

The content of the message in array of text and/or images.

> Show possible types

---

**created\_at** integer

The Unix timestamp (in seconds) for when the message was created.

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**incomplete\_at** integer

The Unix timestamp (in seconds) for when the message was marked as incomplete.

---

**incomplete\_details** object

On an incomplete message, details about why the message is incomplete.

> Show properties

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**object** string

The object type, which is always `thread.message`.

**role** string

The entity that produced the message. One of `user` or `assistant`.

**run\_id** string

The ID of the `run` associated with the creation of this message. Value is `null` when messages are created manually using the create message or create thread endpoints.

**status** string

The status of the message, which can be either `in_progress`, `incomplete`, or `completed`.

**thread\_id** string

The `thread` ID that this message belongs to.

## OBJECT The message object

```
{  
  "id": "msg_abc123",  
  "object": "thread.message",  
  "created_at": 1698983503,  
  "thread_id": "thread_abc123",  
  "role": "assistant",  
  "content": [  
    {  
      "type": "text",  
      "text": {
```

```
        "value": "Hi! How can I help you today?",  
        "annotations": []  
    }  
}  
],  
"assistant_id": "asst_abc123",  
"run_id": "run_abc123",  
"attachments": [],  
"metadata": {}  
}
```

## Runs Beta



Represents an execution run on a thread.

Related guide: [Assistants](#)

## Create run Beta



```
POST https://api.openai.com/v1/threads/{thread_id}/runs
```

Create a run.

## Path parameters

---

**thread\_id** string Required

The ID of the thread to run.

## Query parameters

---

**include[]** array Optional

A list of additional fields to include in the response. Currently the only supported value is

```
step_details.tool_calls[*].file_search.results[*].content
```

 to fetch the file search result content.

See the [file search tool documentation](#) for more information.

## Request body

---

**assistant\_id** string Required

The ID of the [assistant](#) to use to execute this run.

**additional\_instructions** string or null Optional

Appends additional instructions at the end of the instructions for the run. This is useful for modifying the behavior on a per-run basis without overriding other instructions.

---

**additional\_messages** array or null Optional

Adds additional messages to the thread before creating the run.

> Show properties

---

**instructions** string or null Optional

Overrides the [instructions](#) of the assistant. This is useful for modifying the behavior on a per-run basis.

---

**max\_completion\_tokens** integer or null Optional

The maximum number of completion tokens that may be used over the course of the run. The run will make a best effort to use only the number of completion tokens specified, across multiple turns of the run. If the run exceeds the number of completion tokens specified, the run will end with status [incomplete](#). See [incomplete\\_details](#) for more info.

---

**max\_prompt\_tokens** integer or null Optional

The maximum number of prompt tokens that may be used over the course of the run. The run will make a best effort to use only the number of prompt tokens specified, across multiple turns of the run. If the run exceeds the number of prompt tokens specified, the run will end with status [incomplete](#). See [incomplete\\_details](#) for more info.

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**model** string Optional

The ID of the [Model](#) to be used to execute this run. If a value is provided here, it will override the model associated with the assistant. If not, the model associated with the assistant will be used.

**parallel\_tool\_calls** boolean Optional Defaults to true

Whether to enable [parallel function calling](#) during tool use.

**reasoning\_effort** string Optional Defaults to medium

Constrains effort on reasoning for [reasoning models](#). Currently supported values are `minimal`, `low`, `medium`, and `high`.

Reducing reasoning effort can result in faster responses and fewer tokens used on reasoning in a response.

Note: The `gpt-5-pro` model defaults to (and only supports) `high` reasoning effort.

**response\_format** "auto" or object Optional

Specifies the format that the model must output. Compatible with [GPT-4o](#), [GPT-4 Turbo](#), and all GPT-3.5 Turbo models since `gpt-3.5-turbo-1106`.

Setting to `{ "type": "json_schema", "json_schema": {...} }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables JSON mode, which ensures the message the model generates is valid JSON.

**Important:** when using JSON mode, you **must** also instruct the model to produce JSON yourself via a system or user message. Without this, the model may generate an unending stream of whitespace until the generation reaches the token limit, resulting in a long-running and seemingly "stuck" request. Also note that the message content may be partially cut off if

`finish_reason="length"`, which indicates the generation exceeded `max_tokens` or the conversation exceeded the max context length.

> Show possible types

---

**stream** boolean or null Optional

If `true`, returns a stream of events that happen during the Run as server-sent events, terminating when the Run enters a terminal state with a `data: [DONE]` message.

---

**temperature** number or null Optional Defaults to 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.

---

**tool\_choice** string or object Optional

Controls which (if any) tool is called by the model. `none` means the model will not call any tools and instead generates a message. `auto` is the default value and means the model can pick between generating a message or calling one or more tools. `required` means the model must call one or more tools before responding to the user. Specifying a particular tool like `{"type": "file_search"}` or `{"type": "function", "function": {"name": "my_function"}}` forces the model to call that tool.

> Show possible types

---

**tools** array or null Optional

Override the tools the assistant can use for this run. This is useful for modifying the behavior on a per-run basis.

> Show possible types

---

**top\_p** number or null Optional Defaults to 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.

---

**truncation\_strategy** object or null Optional

Controls for how a thread will be truncated prior to the run. Use this to control the initial context window of the run.

> Show properties

## Returns

---

A [run](#) object.

**Default**   [Streaming](#)   [Streaming with Functions](#)

Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "assistant_id": "asst_abc123"
}'
```

Response

```
{  
  "id": "run_abc123",  
  "object": "thread.run",  
  "created_at": 1699063290,  
  "assistant_id": "asst_abc123",  
  "thread_id": "thread_abc123",  
  "status": "queued",  
  "started_at": 1699063290,  
  "expires_at": null,  
  "cancelled_at": null,  
  "failed_at": null,  
  "completed_at": 1699063291,  
  "last_error": null,  
  "model": "gpt-4o",  
  "instructions": null,  
  "incomplete_details": null,  
  "tools": [  
    {  
      "type": "code_interpreter"  
    }  
  ],  
  "metadata": {},  
  "usage": null,  
  "temperature": 1.0,  
  "top_p": 1.0,  
  "max_prompt_tokens": 1000,  
  "max_completion_tokens": 1000,  
  "truncation_strategy": {  
    "type": "auto",  
  },  
}
```

```
"last_messages": null  
},  
"response_format": "auto",  
"tool_choice": "auto",  
"parallel_tool_calls": true  
}
```

## Create thread and run Beta



POST <https://api.openai.com/v1/threads/runs>

Create a thread and run it in one request.

### Request body

**assistant\_id** string Required

The ID of the [assistant](#) to use to execute this run.

**instructions** string or null Optional

Override the default system message of the assistant. This is useful for modifying the behavior on a per-run basis.

**max\_completion\_tokens** integer or null Optional

The maximum number of completion tokens that may be used over the course of the run. The run will make a best effort to use only the number of completion tokens specified, across multiple turns of the run. If the run exceeds the number of completion tokens specified, the run will end with status `incomplete`. See `incomplete_details` for more info.

---

**max\_prompt\_tokens** integer or null Optional

The maximum number of prompt tokens that may be used over the course of the run. The run will make a best effort to use only the number of prompt tokens specified, across multiple turns of the run. If the run exceeds the number of prompt tokens specified, the run will end with status `incomplete`. See `incomplete_details` for more info.

---

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string Optional

The ID of the [Model](#) to be used to execute this run. If a value is provided here, it will override the model associated with the assistant. If not, the model associated with the assistant will be used.

---

**parallel\_tool\_calls** boolean Optional Defaults to true

Whether to enable [parallel function calling](#) during tool use.

---

**response\_format** "auto" or object Optional

Specifies the format that the model must output. Compatible with [GPT-4o](#), [GPT-4 Turbo](#), and all GPT-3.5 Turbo models since `gpt-3.5-turbo-1106`.

Setting to `{ "type": "json_schema", "json_schema": { ... } }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables JSON mode, which ensures the message the model generates is valid JSON.

**Important:** when using JSON mode, you **must** also instruct the model to produce JSON yourself via a system or user message. Without this, the model may generate an unending stream of whitespace until the generation reaches the token limit, resulting in a long-running and seemingly "stuck" request. Also note that the message content may be partially cut off if `finish_reason="length"`, which indicates the generation exceeded `max_tokens` or the conversation exceeded the max context length.

> Show possible types

---

**stream** boolean or null Optional

If `true`, returns a stream of events that happen during the Run as server-sent events, terminating when the Run enters a terminal state with a `data: [DONE]` message.

---

**temperature** number or null Optional Defaults to 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.

---

**thread** object Optional

Options to create a new thread. If no thread is provided when running a request, an empty thread will be created.

> Show properties

**tool\_choice** string or object Optional

Controls which (if any) tool is called by the model. `none` means the model will not call any tools and instead generates a message. `auto` is the default value and means the model can pick between generating a message or calling one or more tools. `required` means the model must call one or more tools before responding to the user. Specifying a particular tool like `{"type": "file_search"}` or `{"type": "function", "function": {"name": "my_function"}}` forces the model to call that tool.

> Show possible types

---

**tool\_resources** object or null Optional

A set of resources that are used by the assistant's tools. The resources are specific to the type of tool. For example, the `code_interpreter` tool requires a list of file IDs, while the `file_search` tool requires a list of vector store IDs.

> Show properties

---

**tools** array or null Optional

Override the tools the assistant can use for this run. This is useful for modifying the behavior on a per-run basis.

> Show possible types

---

**top\_p** number or null Optional Defaults to 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.

---

**truncation\_strategy** object or null Optional

Controls for how a thread will be truncated prior to the run. Use this to control the initial context window of the run.

> Show properties

## Returns

A run object.

**Default**

**Streaming**

**Streaming with Functions**

Example request

```
curl https://api.openai.com/v1/threads/runs \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
    "assistant_id": "asst_abc123",
    "thread": {
        "messages": [
            {"role": "user", "content": "Explain deep learning to a 5 year old."}
        ]
    }
}'
```

Response

```
{  
    "id": "run_abc123",
```

```
"object": "thread.run",
"created_at": 1699076792,
"assistant_id": "asst_abc123",
"thread_id": "thread_abc123",
"status": "queued",
"started_at": null,
"expires_at": 1699077392,
"cancelled_at": null,
"failed_at": null,
"completed_at": null,
"required_action": null,
"last_error": null,
"model": "gpt-4o",
"instructions": "You are a helpful assistant.",
"tools": [],
"tool_resources": {},
"metadata": {},
"temperature": 1.0,
"top_p": 1.0,
"max_completion_tokens": null,
"max_prompt_tokens": null,
"truncation_strategy": {
  "type": "auto",
  "last_messages": null
},
"incomplete_details": null,
"usage": null,
"response_format": "auto",
"tool_choice": "auto",
```

```
"parallel_tool_calls": true  
}
```

## List runs Beta



```
GET https://api.openai.com/v1/threads/{thread_id}/runs
```

Returns a list of runs belonging to a thread.

### Path parameters

**thread\_id** string Required

The ID of the thread the run belongs to.

### Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

**before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the previous page of the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

A list of [run](#) objects.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "run_abc123",  
      "object": "thread.run",  
      "created_at": 1699075072,  
      "assistant_id": "asst_abc123",  
      "thread_id": "thread_abc123",  
      "status": "completed",  
      "started_at": 1699075072,  
      "expires_at": null,  
      "cancelled_at": null,  
      "failed_at": null,  
      "completed_at": 1699075073,  
      "last_error": null,  
      "model": "gpt-4o",  
      "instructions": null,  
      "incomplete_details": null,  
      "tools": [  
        {  
          "type": "code_interpreter"  
        }  
      ],  
      "tool_resources": {  
        "code_interpreter": {  
          "file_ids": [  
            "file-abc123",  
            "file-abc456"  
          ]  
        }  
      }  
    }  
  ]  
}
```

```
        ],
    },
},
"metadata": {},
"usage": {
    "prompt_tokens": 123,
    "completion_tokens": 456,
    "total_tokens": 579
},
"temperature": 1.0,
"top_p": 1.0,
"max_prompt_tokens": 1000,
"max_completion_tokens": 1000,
"truncation_strategy": {
    "type": "auto",
    "last_messages": null
},
"response_format": "auto",
"tool_choice": "auto",
"parallel_tool_calls": true
},
{
    "id": "run_abc456",
    "object": "thread.run",
    "created_at": 1699063290,
    "assistant_id": "asst_abc123",
    "thread_id": "thread_abc123",
    "status": "completed",
    "started_at": 1699063290,
```

```
"expires_at": null,  
"cancelled_at": null,  
"failed_at": null,  
"completed_at": 1699063291,  
"last_error": null,  
"model": "gpt-4o",  
"instructions": null,  
"incomplete_details": null,  
"tools": [  
    {  
        "type": "code_interpreter"  
    }  
,  
    "tool_resources": {  
        "code_interpreter": {  
            "file_ids": [  
                "file-abc123",  
                "file-abc456"  
            ]  
        }  
    },  
    "metadata": {},  
    "usage": {  
        "prompt_tokens": 123,  
        "completion_tokens": 456,  
        "total_tokens": 579  
    },  
    "temperature": 1.0,  
    "top_p": 1.0,
```

```
"max_prompt_tokens": 1000,  
"max_completion_tokens": 1000,  
"truncation_strategy": {  
    "type": "auto",  
    "last_messages": null  
},  
"response_format": "auto",  
"tool_choice": "auto",  
"parallel_tool_calls": true  
}  
],  
"first_id": "run_abc123",  
"last_id": "run_abc456",  
"has_more": false  
}
```

## Retrieve run

Beta



GET [https://api.openai.com/v1/threads/{thread\\_id}/runs/{run\\_id}](https://api.openai.com/v1/threads/{thread_id}/runs/{run_id})

Retrieves a run.

## Path parameters

**run\_id** string Required

The ID of the run to retrieve.

**thread\_id** string Required

The ID of the thread that was run.

## Returns

The run object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs/run_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "id": "run_abc123",
  "object": "thread.run",
  "created_at": 1699075072,
  "assistant_id": "asst_abc123",
```

```
"thread_id": "thread_abc123",
"status": "completed",
"started_at": 1699075072,
"expires_at": null,
"cancelled_at": null,
"failed_at": null,
"completed_at": 1699075073,
"last_error": null,
"model": "gpt-4o",
"instructions": null,
"incomplete_details": null,
"tools": [
  {
    "type": "code_interpreter"
  }
],
"metadata": {},
"usage": {
  "prompt_tokens": 123,
  "completion_tokens": 456,
  "total_tokens": 579
},
"temperature": 1.0,
"top_p": 1.0,
"max_prompt_tokens": 1000,
"max_completion_tokens": 1000,
"truncation_strategy": {
  "type": "auto",
  "last_messages": null
}
```

```
},  
  "response_format": "auto",  
  "tool_choice": "auto",  
  "parallel_tool_calls": true
```

# Modify run

Beta



POST [https://api.openai.com/v1/threads/{thread\\_id}/runs/{run\\_id}](https://api.openai.com/v1/threads/{thread_id}/runs/{run_id})

Modifies a run.

## Path parameters

**run\_id** string Required

The ID of the run to modify.

**thread\_id** string Required

The ID of the thread that was run.

## Request body

**metadata** map Optional

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

## Returns

The modified run object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs/run_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "metadata": {
    "user_id": "user_abc123"
  }
}'
```

### Response

```
{  
  "id": "run_abc123",  
  "object": "thread.run",  
  "created_at": 1699075072,  
  "assistant_id": "asst_abc123",  
  "thread_id": "thread_abc123",  
  "status": "completed",  
  "started_at": 1699075072,  
  "expires_at": null,  
  "cancelled_at": null,  
  "failed_at": null,  
  "completed_at": 1699075073,  
  "last_error": null,  
  "model": "gpt-4o",  
  "instructions": null,  
  "incomplete_details": null,  
  "tools": [  
    {  
      "type": "code_interpreter"  
    }  
  ],  
  "tool_resources": {  
    "code_interpreter": {  
      "file_ids": [  
        "file-abc123",  
        "file-abc456"  
      ]  
    }  
  },  
},
```

```
"metadata": {  
    "user_id": "user_abc123"  
},  
"usage": {  
    "prompt_tokens": 123,  
    "completion_tokens": 456,  
    "total_tokens": 579  
},  
"temperature": 1.0,  
"top_p": 1.0,  
"max_prompt_tokens": 1000,  
"max_completion_tokens": 1000,  
"truncation_strategy": {  
    "type": "auto",  
    "last_messages": null  
},  
"response_format": "auto",  
"tool_choice": "auto",  
"parallel_tool_calls": true  
}
```

## Submit tool outputs to run

Beta



POST [https://api.openai.com/v1/threads/{thread\\_id}/runs/{run\\_id}/submit\\_tool\\_outputs](https://api.openai.com/v1/threads/{thread_id}/runs/{run_id}/submit_tool_outputs)

When a run has the `status: "requires_action"` and `required_action.type` is `submit_tool_outputs`, this endpoint can be used to submit the outputs from the tool calls once they're all completed. All outputs must be submitted in a single request.

## Path parameters

---

**run\_id** string Required

The ID of the run that requires the tool output submission.

---

**thread\_id** string Required

The ID of the thread to which this run belongs.

---

## Request body

---

**tool\_outputs** array Required

A list of tools for which the outputs are being submitted.

> Show properties

---

**stream** boolean Optional

If `true`, returns a stream of events that happen during the Run as server-sent events, terminating when the Run enters a terminal state with a `data: [DONE]` message.

## Returns

The modified run object matching the specified ID.

**Default**    **Streaming**

Example request

```
curl https://api.openai.com/v1/threads/thread_123/runs/run_123/submit_tool_outputs \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2" \
-d '{
  "tool_outputs": [
    {
      "tool_call_id": "call_001",
      "output": "70 degrees and sunny."
    }
  ]
}'
```

Response

```
{  
  "id": "run_123",
```

```
"object": "thread.run",
"created_at": 1699075592,
"assistant_id": "asst_123",
"thread_id": "thread_123",
"status": "queued",
"started_at": 1699075592,
"expires_at": 1699076192,
"cancelled_at": null,
"failed_at": null,
"completed_at": null,
"last_error": null,
"model": "gpt-4o",
"instructions": null,
"tools": [
  {
    "type": "function",
    "function": {
      "name": "get_current_weather",
      "description": "Get the current weather in a given location",
      "parameters": {
        "type": "object",
        "properties": {
          "location": {
            "type": "string",
            "description": "The city and state, e.g. San Francisco, CA"
          },
          "unit": {
            "type": "string",
            "enum": ["celsius", "fahrenheit"]
          }
        }
      }
    }
  }
]
```

```
        }
      },
      "required": ["location"]
    }
  }
],
"metadata": {},
"usage": null,
"temperature": 1.0,
"top_p": 1.0,
"max_prompt_tokens": 1000,
"max_completion_tokens": 1000,
"truncation_strategy": {
  "type": "auto",
  "last_messages": null
},
"response_format": "auto",
"tool_choice": "auto",
"parallel_tool_calls": true
```

## Cancel a run

Beta



```
POST https://api.openai.com/v1/threads/{thread_id}/runs/{run_id}/cancel
```

Cancels a run that is `in_progress`.

## Path parameters

**run\_id** string Required

The ID of the run to cancel.

**thread\_id** string Required

The ID of the thread to which this run belongs.

## Returns

The modified `run` object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs/run_abc123/cancel \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "OpenAI-Beta: assistants=v2" \
-X POST
```

### Response

```
{  
  "id": "run_abc123",  
  "object": "thread.run",  
  "created_at": 1699076126,  
  "assistant_id": "asst_abc123",  
  "thread_id": "thread_abc123",  
  "status": "cancelling",  
  "started_at": 1699076126,  
  "expires_at": 1699076726,  
  "cancelled_at": null,  
  "failed_at": null,  
  "completed_at": null,  
  "last_error": null,  
  "model": "gpt-4o",  
  "instructions": "You summarize books.",  
  "tools": [  
    {  
      "type": "file_search"  
    }  
  ],  
  "tool_resources": {  
    "file_search": {  
      "vector_store_ids": ["vs_123"]  
    }  
  },  
  "metadata": {},  
  "usage": null,  
  "temperature": 1.0,  
  "top_p": 1.0,
```

```
"response_format": "auto",
"tool_choice": "auto",
"parallel_tool_calls": true
}
```

# The run object Beta



Represents an execution run on a [thread](#).

**assistant\_id** string

The ID of the [assistant](#) used for execution of this run.

**cancelled\_at** integer or null

The Unix timestamp (in seconds) for when the run was cancelled.

**completed\_at** integer or null

The Unix timestamp (in seconds) for when the run was completed.

**created\_at** integer

The Unix timestamp (in seconds) for when the run was created.

**expires\_at** integer or null

The Unix timestamp (in seconds) for when the run will expire.

---

**failed\_at** integer or null

The Unix timestamp (in seconds) for when the run failed.

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**incomplete\_details** object or null

Details on why the run is incomplete. Will be `null` if the run is not incomplete.

> Show properties

---

**instructions** string

The instructions that the assistant used for this run.

---

**last\_error** object or null

The last error associated with this run. Will be `null` if there are no errors.

> Show properties

---

**max\_completion\_tokens** integer or null

The maximum number of completion tokens specified to have been used over the course of the run.

---

**max\_prompt\_tokens** integer or null

The maximum number of prompt tokens specified to have been used over the course of the run.

---

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

---

**model** string

The model that the assistant used for this run.

---

**object** string

The object type, which is always `thread.run`.

---

**parallel\_tool\_calls** boolean

Whether to enable parallel function calling during tool use.

---

**required\_action** object or null

Details on the action required to continue the run. Will be `null` if no action is required.

> Show properties

---

**response\_format** "auto" or object

Specifies the format that the model must output. Compatible with GPT-4o, GPT-4 Turbo, and all GPT-3.5 Turbo models since `gpt-3.5-turbo-1106`.

---

Setting to `{ "type": "json_schema", "json_schema": { ... } }` enables Structured Outputs which ensures the model will match your supplied JSON schema. Learn more in the [Structured Outputs guide](#).

Setting to `{ "type": "json_object" }` enables JSON mode, which ensures the message the model generates is valid JSON.

**Important:** when using JSON mode, you **must** also instruct the model to produce JSON yourself via a system or user message. Without this, the model may generate an unending stream of whitespace until the generation reaches the token limit, resulting in a long-running and seemingly "stuck" request. Also note that the message content may be partially cut off if

`finish_reason="length"`, which indicates the generation exceeded `max_tokens` or the conversation exceeded the max context length.

> Show possible types

---

#### **started\_at** integer or null

The Unix timestamp (in seconds) for when the run was started.

---

#### **status** string

The status of the run, which can be either `queued`, `in_progress`, `requires_action`, `cancelling`, `cancelled`, `failed`, `completed`, `incomplete`, or `expired`.

---

#### **temperature** number or null

The sampling temperature used for this run. If not set, defaults to 1.

---

#### **thread\_id** string

The ID of the [thread](#) that was executed on as a part of this run.

**tool\_choice** string or object

Controls which (if any) tool is called by the model. `none` means the model will not call any tools and instead generates a message. `auto` is the default value and means the model can pick between generating a message or calling one or more tools. `required` means the model must call one or more tools before responding to the user. Specifying a particular tool like `{"type": "file_search"}` or `{"type": "function", "function": {"name": "my_function"}}` forces the model to call that tool.

> Show possible types

**tools** array

The list of tools that the assistant used for this run.

> Show possible types

**top\_p** number or null

The nucleus sampling value used for this run. If not set, defaults to 1.

**truncation\_strategy** object or null

Controls for how a thread will be truncated prior to the run. Use this to control the initial context window of the run.

> Show properties

**usage** object

Usage statistics related to the run. This value will be `null` if the run is not in a terminal state (i.e. `in_progress` , `queued` , etc.).

> Show properties

OBJECT The run object

```
{  
  "id": "run_abc123",  
  "object": "thread.run",  
  "created_at": 1698107661,  
  "assistant_id": "asst_abc123",  
  "thread_id": "thread_abc123",  
  "status": "completed",  
  "started_at": 1699073476,  
  "expires_at": null,  
  "cancelled_at": null,  
  "failed_at": null,  
  "completed_at": 1699073498,  
  "last_error": null,  
  "model": "gpt-4o",  
  "instructions": null,  
  "tools": [{"type": "file_search"}, {"type": "code_interpreter"}],  
  "metadata": {},  
  "incomplete_details": null,  
  "usage": {  
    "prompt_tokens": 123,  
    "completion_tokens": 456,  
    "total_tokens": 579  
  },  
  "temperature": 1.0,  
  "top_p": 1.0,  
  "max_prompt_tokens": 1000,  
  "max_completion_tokens": 1000,  
  "truncation_strategy": {  
    "type": "auto",  
  }  
}
```

```
"last_messages": null  
},  
"response_format": "auto",  
"tool_choice": "auto",  
"parallel_tool_calls": true  
}
```

## Run steps Beta



Represents the steps (model and tool calls) taken during the run.

Related guide: [Assistants](#)

## List run steps Beta



```
GET https://api.openai.com/v1/threads/{thread_id}/runs/{run_id}/steps
```

Returns a list of run steps belonging to a run.

## Path parameters

---

### **run\_id** string Required

The ID of the run the run steps belong to.

---

### **thread\_id** string Required

The ID of the thread the run and run steps belong to.

---

## Query parameters

---

### **after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

---

### **before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the previous page of the list.

---

### **include[]** array Optional

A list of additional fields to include in the response. Currently the only supported value is

`step_details.tool_calls[*].file_search.results[*].content` to fetch the file search result content.

See the [file search tool documentation](#) for more information.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

A list of [run step](#) objects.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs/run_abc123/steps \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

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

```
"id": "step_abc123",
"object": "thread.run.step",
"created_at": 1699063291,
"run_id": "run_abc123",
"assistant_id": "asst_abc123",
"thread_id": "thread_abc123",
"type": "message_creation",
"status": "completed",
"cancelled_at": null,
"completed_at": 1699063291,
"expired_at": null,
"failed_at": null,
"last_error": null,
"step_details": {
    "type": "message_creation",
    "message_creation": {
        "message_id": "msg_abc123"
    }
},
"usage": {
    "prompt_tokens": 123,
    "completion_tokens": 456,
    "total_tokens": 579
}
},
],
"first_id": "step_abc123",
"last_id": "step_abc456",
```

"has\_more": false

# Retrieve run step Beta



```
GET https://api.openai.com/v1/threads/{thread_id}/runs/{run_id}/steps/{step_id}
```

Retrieves a run step.

## Path parameters

**run\_id** string Required

The ID of the run to which the run step belongs.

**step\_id** string Required

The ID of the run step to retrieve.

**thread\_id** string Required

The ID of the thread to which the run and run step belongs.

## Query parameters

**include[]** array Optional

A list of additional fields to include in the response. Currently the only supported value is

`step_details.tool_calls[*].file_search.results[*].content` to fetch the file search result content.

See the [file search tool documentation](#) for more information.

## Returns

The [run step](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/threads/thread_abc123/runs/run_abc123/steps/step_abc123 \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-H "OpenAI-Beta: assistants=v2"
```

### Response

```
{
  "id": "step_abc123",
  "object": "thread.run.step",
  "created_at": 1699063291,
  "run_id": "run_abc123",
  "assistant_id": "asst_abc123",
```

```
"thread_id": "thread_abc123",
"type": "message_creation",
"status": "completed",
"cancelled_at": null,
"completed_at": 1699063291,
"expired_at": null,
"failed_at": null,
"last_error": null,
"step_details": {
    "type": "message_creation",
    "message_creation": {
        "message_id": "msg_abc123"
    }
},
"usage": {
    "prompt_tokens": 123,
    "completion_tokens": 456,
    "total_tokens": 579
}
}
```

## The run step object Beta

🔗

Represents a step in execution of a run.

**assistant\_id** string

The ID of the [assistant](#) associated with the run step.

---

**cancelled\_at** integer

The Unix timestamp (in seconds) for when the run step was cancelled.

---

**completed\_at** integer

The Unix timestamp (in seconds) for when the run step completed.

---

**created\_at** integer

The Unix timestamp (in seconds) for when the run step was created.

---

**expired\_at** integer

The Unix timestamp (in seconds) for when the run step expired. A step is considered expired if the parent run is expired.

---

**failed\_at** integer

The Unix timestamp (in seconds) for when the run step failed.

---

**id** string

The identifier of the run step, which can be referenced in API endpoints.

---

**last\_error** object

The last error associated with this run step. Will be `null` if there are no errors.

> Show properties

**metadata** map

Set of 16 key-value pairs that can be attached to an object. This can be useful for storing additional information about the object in a structured format, and querying for objects via API or the dashboard.

Keys are strings with a maximum length of 64 characters. Values are strings with a maximum length of 512 characters.

**object** string

The object type, which is always `thread.run.step`.

**run\_id** string

The ID of the run that this run step is a part of.

**status** string

The status of the run step, which can be either `in_progress`, `cancelled`, `failed`, `completed`, or `expired`.

**step\_details** object

The details of the run step.

> Show possible types

**thread\_id** string

The ID of the thread that was run.

**type** string

The type of run step, which can be either `message_creation` or `tool_calls`.

**usage** object

Usage statistics related to the run step. This value will be `null` while the run step's status is `in_progress`.

> Show properties

OBJECT The run step object

```
{  
  "id": "step_abc123",  
  "object": "thread.run.step",  
  "created_at": 1699063291,  
  "run_id": "run_abc123",  
  "assistant_id": "asst_abc123",  
  "thread_id": "thread_abc123",  
  "type": "message_creation",  
  "status": "completed",  
  "cancelled_at": null,  
  "completed_at": 1699063291,  
  "expired_at": null,  
  "failed_at": null,  
  "last_error": null,  
  "step_details": {  
    "type": "message_creation",  
    "message_creation": {  
      "message_id": "msg_abc123"  
    }  
  },  
  "usage": {  
    "prompt_tokens": 123,  
    "completion_tokens": 456,  
  }  
}
```

```
"total_tokens": 579  
}  
}
```

## Streaming Beta



Stream the result of executing a Run or resuming a Run after submitting tool outputs. You can stream events from the [Create Thread and Run](#), [Create Run](#), and [Submit Tool Outputs](#) endpoints by passing `"stream": true`. The response will be a [Server-Sent events](#) stream. Our Node and Python SDKs provide helpful utilities to make streaming easy. Reference the [Assistants API quickstart](#) to learn more.

## The message delta object Beta



Represents a message delta i.e. any changed fields on a message during streaming.

**delta** object

The delta containing the fields that have changed on the Message.

> Show properties

---

**id** string

The identifier of the message, which can be referenced in API endpoints.

---

**object** string

The object type, which is always `thread.message.delta`.

OBJECT The message delta object

```
{  
  "id": "msg_123",  
  "object": "thread.message.delta",  
  "delta": {  
    "content": [  
      {  
        "index": 0,  
        "type": "text",  
        "text": { "value": "Hello", "annotations": [] }  
      }  
    ]  
  }  
}
```

# The run step delta object

Beta



Represents a run step delta i.e. any changed fields on a run step during streaming.

## **delta** object

The delta containing the fields that have changed on the run step.

> Show properties

## **id** string

The identifier of the run step, which can be referenced in API endpoints.

## **object** string

The object type, which is always `thread.run.step.delta`.

OBJECT The run step delta object

```
{  
  "id": "step_123",  
  "object": "thread.run.step.delta",  
  "delta": {  
    "step_details": {  
      "type": "tool_calls",  
      "tool_calls": [  
        {  
          "index": 0,  
          "value": "Tool call 1"  
        },  
        {  
          "index": 1,  
          "value": "Tool call 2"  
        }  
      ]  
    }  
  }  
}
```

```
        "id": "call_123",
        "type": "code_interpreter",
        "code_interpreter": { "input": "", "outputs": [] }
    }
]
}
}
}
```

## Assistant stream events

Beta



Represents an event emitted when streaming a Run.

Each event in a server-sent events stream has an `event` and `data` property:

```
event: thread.created
data: {"id": "thread_123", "object": "thread", ...}
```

We emit events whenever a new object is created, transitions to a new state, or is being streamed in parts (deltas). For example, we emit `thread.run.created` when a new run is created, `thread.run.completed` when a run completes, and so on. When an Assistant chooses to create a message during a run, we emit a

`thread.message.created` event, a `thread.message.in_progress` event, many `thread.message.delta` events, and finally a `thread.message.completed` event.

We may add additional events over time, so we recommend handling unknown events gracefully in your code. See the [Assistants API quickstart](#) to learn how to integrate the Assistants API with streaming.

---

**done** `data` is `[DONE]`

Occurs when a stream ends.

---

**error** `data` is an [error](#)

Occurs when an [error](#) occurs. This can happen due to an internal server error or a timeout.

---

**thread.created** `data` is a [thread](#)

Occurs when a new [thread](#) is created.

---

**thread.message.completed** `data` is a [message](#)

Occurs when a [message](#) is completed.

---

**thread.message.created** `data` is a [message](#)

Occurs when a [message](#) is created.

---

**thread.message.delta** `data` is a [message delta](#)

Occurs when parts of a Message are being streamed.

---

**thread.message.in\_progress** `data` is a message

Occurs when a message moves to an `in_progress` state.

---

**thread.message.incomplete** `data` is a message

Occurs when a message ends before it is completed.

---

**thread.run.cancelled** `data` is a run

Occurs when a run is cancelled.

---

**thread.run.cancelling** `data` is a run

Occurs when a run moves to a `canceling` status.

---

**thread.run.completed** `data` is a run

Occurs when a run is completed.

---

**thread.run.created** `data` is a run

Occurs when a new run is created.

---

**thread.run.expired** `data` is a run

Occurs when a run expires.

---

**thread.run.failed** `data` is a run

Occurs when a run fails.

---

**thread.run.in\_progress** `data` is a run

Occurs when a run moves to an `in_progress` status.

---

**thread.run.incomplete** `data` is a run

Occurs when a run ends with status `incomplete`.

---

**thread.run.queued** `data` is a run

Occurs when a run moves to a `queued` status.

---

**thread.run.requires\_action** `data` is a run

Occurs when a run moves to a `requires_action` status.

---

**thread.run.step.cancelled** `data` is a run step

Occurs when a run step is cancelled.

---

**thread.run.step.completed** `data` is a run step

Occurs when a run step is completed.

---

**thread.run.step.created** `data` is a run step

Occurs when a run step is created.

---

**thread.run.step.delta** `data` is a run step delta

Occurs when parts of a run step are being streamed.

---

**thread.run.step.expired** `data` is a run step

Occurs when a run step expires.

---

**thread.run.step.failed** `data` is a run step

Occurs when a run step fails.

---

**thread.run.step.in\_progress** `data` is a run step

Occurs when a run step moves to an `in_progress` state.

---

# Administration



Programmatically manage your organization. The Audit Logs endpoint provides a log of all actions taken in the organization for security and monitoring purposes. To access these endpoints please generate an Admin API Key through the [API Platform Organization overview](#). Admin API keys cannot be used for non-administration endpoints. For best practices on setting up your organization, please refer to this [guide](#)

---

# Admin API Keys



Admin API keys enable Organization Owners to programmatically manage various aspects of their organization, including users, projects, and API keys. These keys provide administrative capabilities, such as creating, updating, and deleting users; managing projects; and overseeing API key lifecycles.

## Key Features of Admin API Keys:

**User Management:** Invite new users, update roles, and remove users from the organization.

**Project Management:** Create, update, archive projects, and manage user assignments within projects.

**API Key Oversight:** List, retrieve, and delete API keys associated with projects.

Only Organization Owners have the authority to create and utilize Admin API keys. To manage these keys, Organization Owners can navigate to the Admin Keys section of their API Platform dashboard.

For direct access to the Admin Keys management page, Organization Owners can use the following link:

<https://platform.openai.com/settings/organization/admin-keys>

It's crucial to handle Admin API keys with care due to their elevated permissions. Adhering to best practices, such as regular key rotation and assigning appropriate permissions, enhances security and ensures proper governance within the organization.

---

## List all organization and project API keys.



GET `https://api.openai.com/v1/organization/admin_api_keys`

List organization API keys

### Query parameters

**after** string or null Optional

**limit** integer Optional Defaults to 20

**order** string Optional Defaults to asc

## Returns

A list of admin and project API key objects.

### Example request

```
curl https://api.openai.com/v1/organization/admin_api_keys?after=key_abc&limit=20 \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "object": "organization.admin_api_key",
      "id": "key_abc",
```

```
"name": "Main Admin Key",
"redacted_value": "sk-admin...def",
"created_at": 1711471533,
"last_used_at": 1711471534,
"owner": {
    "type": "service_account",
    "object": "organization.service_account",
    "id": "sa_456",
    "name": "My Service Account",
    "created_at": 1711471533,
    "role": "member"
}
],
"first_id": "key_abc",
"last_id": "key_abc",
"has_more": false
}
```

# Create admin API key



POST [https://api.openai.com/v1/organization/admin\\_api\\_keys](https://api.openai.com/v1/organization/admin_api_keys)

## Create an organization admin API key

### Request body

**name** string Required

### Returns

The created [AdminApiKey](#) object.

#### Example request

```
curl -X POST https://api.openai.com/v1/organization/admin_api_keys \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{'
  "name": "New Admin Key"
}'
```

#### Response

```
{  
  "object": "organization.admin_api_key",  
  "id": "key_xyz",  
  "name": "New Admin Key",
```

```
"redacted_value": "sk-admin...xyz",
"created_at": 1711471533,
"last_used_at": 1711471534,
"owner": {
  "type": "user",
  "object": "organization.user",
  "id": "user_123",
  "name": "John Doe",
  "created_at": 1711471533,
  "role": "owner"
},
"value": "sk-admin-1234abcd"
}
```

# Retrieve admin API key



```
GET https://api.openai.com/v1/organization/admin_api_keys/{key_id}
```

Retrieve a single organization API key

## Path parameters

**key\_id** string Required

## Returns

The requested [AdminApiKey](#) object.

### Example request

```
curl https://api.openai.com/v1/organization/admin_api_keys/key_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "organization.admin_api_key",
  "id": "key_abc",
  "name": "Main Admin Key",
  "redacted_value": "sk-admin...xyz",
  "created_at": 1711471533,
  "last_used_at": 1711471534,
  "owner": {
    "type": "user",
    "object": "organization.user",
    "id": "user_123",
    "name": "John Doe",
```

```
        "created_at": 1711471533,  
        "role": "owner"  
    }  
}
```

# Delete admin API key



```
DELETE https://api.openai.com/v1/organization/admin_api_keys/{key_id}
```

Delete an organization admin API key

## Path parameters

**key\_id** string Required

## Returns

A confirmation object indicating the key was deleted.

Example request

```
curl -X DELETE https://api.openai.com/v1/organization/admin_api_keys/key_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "id": "key_abc",  
  "object": "organization.admin_api_key.deleted",  
  "deleted": true  
}
```

# The admin API key object



Represents an individual Admin API key in an org.

**created\_at** integer

The Unix timestamp (in seconds) of when the API key was created

**id** string

The identifier, which can be referenced in API endpoints

**last\_used\_at** integer

The Unix timestamp (in seconds) of when the API key was last used

**name** string

The name of the API key

**object** string

The object type, which is always `organization.admin_api_key`

**owner** object

> Show properties

**redacted\_value** string

The redacted value of the API key

**value** string

The value of the API key. Only shown on create.

OBJECT The admin API key object

```
{  
  "object": "organization.admin_api_key",  
  "id": "key_abc",  
  "name": "Main Admin Key",  
  "redacted_value": "sk-admin...xyz",  
  "created_at": 1711471533,  
}
```

```
"last_used_at": 1711471534,  
"owner": {  
    "type": "user",  
    "object": "organization.user",  
    "id": "user_123",  
    "name": "John Doe",  
    "created_at": 1711471533,  
    "role": "owner"  
}  
}
```

# Invites



Invite and manage invitations for an organization.

## List invites



GET <https://api.openai.com/v1/organization/invites>

Returns a list of invites in the organization.

## Query parameters

### **after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

### **limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

A list of [Invite](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/invites?after=invite-abc&limit=20 \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "organization.invite",  
      "id": "invite-abc",  
      "email": "user@example.com",  
      "role": "owner",  
      "status": "accepted",  
      "invited_at": 1711471533,  
      "expires_at": 1711471533,  
      "accepted_at": 1711471533  
    }  
  ],  
  "first_id": "invite-abc",  
  "last_id": "invite-abc",  
  "has_more": false  
}
```

## Create invite



POST <https://api.openai.com/v1/organization/invites>

Create an invite for a user to the organization. The invite must be accepted by the user before they have access to the organization.

## Request body

**email** string Required

Send an email to this address

**role** string Required

owner or reader

**projects** array Optional

An array of projects to which membership is granted at the same time the org invite is accepted. If omitted, the user will be invited to the default project for compatibility with legacy behavior.

> Show properties

## Returns

The created [Invite](#) object.

Example request

```
curl -X POST https://api.openai.com/v1/organization/invites \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
```

```
-H "Content-Type: application/json" \
-d '{
  "email": "anotheruser@example.com",
  "role": "reader",
  "projects": [
    {
      "id": "project-xyz",
      "role": "member"
    },
    {
      "id": "project-abc",
      "role": "owner"
    }
  ]
}'
```

## Response

```
{
  "object": "organization.invite",
  "id": "invite-def",
  "email": "anotheruser@example.com",
  "role": "reader",
  "status": "pending",
  "invited_at": 1711471533,
  "expires_at": 1711471533,
  "accepted_at": null,
  "projects": [
    {

```

```
        "id": "project-xyz",
        "role": "member"
    },
{
    "id": "project-abc",
    "role": "owner"
}
]
```

# Retrieve invite



```
GET https://api.openai.com/v1/organization/invites/{invite_id}
```

Retrieves an invite.

## Path parameters

**invite\_id** string Required

The ID of the invite to retrieve.

## Returns

The Invite object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/organization/invites/invite-abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "organization.invite",
  "id": "invite-abc",
  "email": "user@example.com",
  "role": "owner",
  "status": "accepted",
  "invited_at": 1711471533,
  "expires_at": 1711471533,
  "accepted_at": 1711471533
}
```

# Delete invite



```
DELETE https://api.openai.com/v1/organization/invites/{invite_id}
```

Delete an invite. If the invite has already been accepted, it cannot be deleted.

## Path parameters

**invite\_id** string Required

The ID of the invite to delete.

## Returns

Confirmation that the invite has been deleted

### Example request

```
curl -X DELETE https://api.openai.com/v1/organization/invites/invite-abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "organization.invite.deleted",  
  "id": "invite-abc",  
  "deleted": true  
}
```

# The invite object



Represents an individual `invite` to the organization.

---

**accepted\_at** integer

The Unix timestamp (in seconds) of when the invite was accepted.

---

**email** string

The email address of the individual to whom the invite was sent

---

**expires\_at** integer

The Unix timestamp (in seconds) of when the invite expires.

---

**id** string

The identifier, which can be referenced in API endpoints

**invited\_at** integer

The Unix timestamp (in seconds) of when the invite was sent.

**object** string

The object type, which is always `organization.invite`

**projects** array

The projects that were granted membership upon acceptance of the invite.

> Show properties

**role** string

`owner` or `reader`

**status** string

`accepted`, `expired`, or `pending`

OBJECT The invite object

```
{  
  "object": "organization.invite",  
  "id": "invite-abc",  
  "email": "user@example.com",  
  "role": "owner",  
  "status": "accepted",
```

```
"invited_at": 1711471533,  
"expires_at": 1711471533,  
"accepted_at": 1711471533,  
"projects": [  
  {  
    "id": "project-xyz",  
    "role": "member"  
  }  
]  
}
```

# Users



Manage users and their role in an organization.

## List users



GET <https://api.openai.com/v1/organization/users>

Lists all of the users in the organization.

## Query parameters

---

### **after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

---

### **emails** array Optional

Filter by the email address of users.

---

### **limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

---

A list of [User](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/users?after=user_abc&limit=20 \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "organization.user",  
      "id": "user_abc",  
      "name": "First Last",  
      "email": "user@example.com",  
      "role": "owner",  
      "added_at": 1711471533  
    }  
,  
  "first_id": "user-abc",  
  "last_id": "user-xyz",  
  "has_more": false  
}
```

## Modify user



POST [https://api.openai.com/v1/organization/users/{user\\_id}](https://api.openai.com/v1/organization/users/{user_id})

Modifies a user's role in the organization.

## Path parameters

---

**user\_id** string Required

The ID of the user.

## Request body

---

**role** string Required

owner or reader

## Returns

---

The updated User object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/users/user_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
```

## Response

```
{  
  "object": "organization.user",  
  "id": "user_abc",  
  "name": "First Last",  
  "email": "user@example.com",  
  "role": "owner",  
  "added_at": 1711471533  
}
```

# Retrieve user



```
GET https://api.openai.com/v1/organization/users/{user_id}
```

Retrieves a user by their identifier.

## Path parameters

**user\_id** string Required

The ID of the user.

## Returns

The [User](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/organization/users/user_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "organization.user",
  "id": "user_abc",
  "name": "First Last",
  "email": "user@example.com",
  "role": "owner",
  "added_at": 1711471533
}
```

# Delete user



```
DELETE https://api.openai.com/v1/organization/users/{user_id}
```

Deletes a user from the organization.

## Path parameters

**user\_id** string Required

The ID of the user.

## Returns

Confirmation of the deleted user

### Example request

```
curl -X DELETE https://api.openai.com/v1/organization/users/user_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "organization.user.deleted",  
  "id": "user_abc",  
  "deleted": true  
}
```

# The user object



Represents an individual `user` within an organization.

---

**added\_at** integer

The Unix timestamp (in seconds) of when the user was added.

---

**email** string

The email address of the user

---

**id** string

The identifier, which can be referenced in API endpoints

---

**name** string

The name of the user

**object** string

The object type, which is always `organization.user`

**role** string

`owner` or `reader`

OBJECT The user object

```
{  
  "object": "organization.user",  
  "id": "user_abc",  
  "name": "First Last",  
  "email": "user@example.com",  
  "role": "owner",  
  "added_at": 1711471533  
}
```

# Projects



Manage the projects within an organization includes creation, updating, and archiving or projects. The Default project cannot be archived.

---

# List projects



```
GET https://api.openai.com/v1/organization/projects
```

Returns a list of projects.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

---

**include\_archived** boolean Optional Defaults to false

If `true` returns all projects including those that have been `archived`. Archived projects are not included by default.

---

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

A list of [Project](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/projects?after=proj_abc&limit=20&include_archived=false  
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \  
-H "Content-Type: application/json"
```

### Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "id": "proj_abc",  
      "object": "organization.project",  
      "name": "Project example",  
      "created_at": 1711471533,  
      "archived_at": null,  
      "status": "active"  
    }  
  ],
```

```
"first_id": "proj-abc",
"last_id": "proj-xyz",
"has_more": false
}
```

# Create project



POST <https://api.openai.com/v1/organization/projects>

Create a new project in the organization. Projects can be created and archived, but cannot be deleted.

## Request body

**name** string Required

The friendly name of the project, this name appears in reports.

**geography** string Optional

Create the project with the specified data residency region. Your organization must have access to Data residency functionality in order to use. See [data residency controls](#) to review the functionality and limitations of setting this field.

## Returns

The created Project object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/projects \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{'
    "name": "Project ABC"
}'
```

### Response

```
{
  "id": "proj_abc",
  "object": "organization.project",
  "name": "Project ABC",
  "created_at": 1711471533,
  "archived_at": null,
  "status": "active"
}
```

# Retrieve project



```
GET https://api.openai.com/v1/organization/projects/{project_id}
```

Retrieves a project.

## Path parameters

**project\_id** string Required

The ID of the project.

## Returns

The [Project](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{  
  "id": "proj_abc",  
  "object": "organization.project",  
  "name": "Project example",  
  "created_at": 1711471533,  
  "archived_at": null,  
  "status": "active"  
}
```

# Modify project



POST [https://api.openai.com/v1/organization/projects/{project\\_id}](https://api.openai.com/v1/organization/projects/{project_id})

Modifies a project in the organization.

## Path parameters

**project\_id** string Required

The ID of the project.

## Request body

**name** string Required

The updated name of the project, this name appears in reports.

## Returns

The updated [Project](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/projects/proj_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
    "name": "Project DEF"
}'
```

## Archive project



POST [https://api.openai.com/v1/organization/projects/{project\\_id}/archive](https://api.openai.com/v1/organization/projects/{project_id}/archive)

Archives a project in the organization. Archived projects cannot be used or updated.

## Path parameters

**project\_id** string Required

The ID of the project.

## Returns

The archived [Project](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/projects/proj_abc/archive \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "id": "proj_abc",
  "object": "organization.project",
  "name": "Project DEF",
  "created_at": 1711471533,
  "archived_at": 1711471533,
```

```
    "status": "archived"  
}
```

# The project object



Represents an individual project.

---

**archived\_at** integer

The Unix timestamp (in seconds) of when the project was archived or `null`.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the project was created.

---

**id** string

The identifier, which can be referenced in API endpoints

---

**name** string

The name of the project. This appears in reporting.

---

**object** string

The object type, which is always `organization.project`

**status** string

`active` or `archived`

OBJECT The project object

```
{  
  "id": "proj_abc",  
  "object": "organization.project",  
  "name": "Project example",  
  "created_at": 1711471533,  
  "archived_at": null,  
  "status": "active"  
}
```

## Project users



Manage users within a project, including adding, updating roles, and removing users.

# List project users



```
GET https://api.openai.com/v1/organization/projects/{project_id}/users
```

Returns a list of users in the project.

## Path parameters

**project\_id** string Required

The ID of the project.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

A list of [ProjectUser](#) objects.

#### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/users?after=user_abc&limit=20 \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

#### Response

```
{
  "object": "list",
  "data": [
    {
      "object": "organization.project.user",
      "id": "user_abc",
      "name": "First Last",
      "email": "user@example.com",
      "role": "owner",
      "added_at": 1711471533
    }
  ],
  "first_id": "user-abc",
  "last_id": "user-xyz",
  "has_more": false
}
```

# Create project user



```
POST https://api.openai.com/v1/organization/projects/{project_id}/users
```

Adds a user to the project. Users must already be members of the organization to be added to a project.

## Path parameters

---

**project\_id** string Required

The ID of the project.

## Request body

---

**role** string Required

owner or member

**user\_id** string Required

The ID of the user.

## Returns

The created [ProjectUser](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/projects/proj_abc/users \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
    "user_id": "user_abc",
    "role": "member"
}'
```

### Response

```
{
  "object": "organization.project.user",
  "id": "user_abc",
  "email": "user@example.com",
  "role": "owner",
  "added_at": 1711471533
}
```

# Retrieve project user



```
GET https://api.openai.com/v1/organization/projects/{project_id}/users/{user_id}
```

Retrieves a user in the project.

## Path parameters

**project\_id** string Required

The ID of the project.

**user\_id** string Required

The ID of the user.

## Returns

The [ProjectUser](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/users/user_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "organization.project.user",  
  "id": "user_abc",  
  "name": "First Last",  
  "email": "user@example.com",  
  "role": "owner",  
  "added_at": 1711471533  
}
```

# Modify project user



```
POST https://api.openai.com/v1/organization/projects/{project_id}/users/{user_id}
```

Modifies a user's role in the project.

## Path parameters

**project\_id** string Required

The ID of the project.

**user\_id** string Required

The ID of the user.

## Request body

**role** string Required

owner or member

## Returns

The updated [ProjectUser](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/projects/proj_abc/users/user_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
    "role": "owner"
}'
```

### Response

```
{  
  "object": "organization.project.user",  
  "id": "user_abc",  
  "name": "First Last",  
  "email": "user@example.com",  
  "role": "owner",  
  "added_at": 1711471533  
}
```

# Delete project user



```
DELETE https://api.openai.com/v1/organization/projects/{project_id}/users/{user_id}
```

Deletes a user from the project.

## Path parameters

**project\_id** string Required

The ID of the project.

**user\_id** string Required

The ID of the user.

## Returns

Confirmation that project has been deleted or an error in case of an archived project, which has no users

### Example request

```
curl -X DELETE https://api.openai.com/v1/organization/projects/proj_abc/users/user_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "organization.project.user.deleted",
  "id": "user_abc",
  "deleted": true
}
```

# The project user object



Represents an individual user in a project.

---

**added\_at** integer

The Unix timestamp (in seconds) of when the project was added.

---

**email** string

The email address of the user

---

**id** string

The identifier, which can be referenced in API endpoints

---

**name** string

The name of the user

---

**object** string

The object type, which is always `organization.project.user`

---

**role** string

`owner` or `member`

OBJECT The project user object

```
{  
  "object": "organization.project.user",  
  "id": "user_abc",  
  "name": "First Last",  
  "email": "user@example.com",  
  "role": "owner",  
  "added_at": 1711471533  
}
```

# Project service accounts



Manage service accounts within a project. A service account is a bot user that is not associated with a user. If a user leaves an organization, their keys and membership in projects will no longer work. Service accounts do not have this limitation. However, service accounts can also be deleted from a project.

## List project service accounts



```
GET https://api.openai.com/v1/organization/projects/{project_id}/service_accounts
```

Returns a list of service accounts in the project.

## Path parameters

**project\_id** string Required

The ID of the project.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include `after=obj_foo` in order to fetch the next page of the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

A list of [ProjectServiceAccount](#) objects.

## Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/service_accounts?after=custom_id&limit=10
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{
  "object": "list",
  "data": [
    {
      "object": "organization.project.service_account",
      "id": "svc_acct_abc",
      "name": "Service Account",
      "role": "owner",
      "created_at": 1711471533
    }
  ],
  "first_id": "svc_acct_abc",
  "last_id": "svc_acct_xyz",
  "has_more": false
}
```

# Create project service account



```
POST https://api.openai.com/v1/organization/projects/{project_id}/service_accounts
```

Creates a new service account in the project. This also returns an unredacted API key for the service account.

## Path parameters

**project\_id** string Required

The ID of the project.

## Request body

**name** string Required

The name of the service account being created.

## Returns

The created [ProjectServiceAccount](#) object.

Example request

```
curl -X POST https://api.openai.com/v1/organization/projects/proj_abc/service_accounts \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
  "name": "Production App"
}'
```

## Response

```
{
  "object": "organization.project.service_account",
  "id": "svc_acct_abc",
  "name": "Production App",
  "role": "member",
  "created_at": 1711471533,
  "api_key": {
    "object": "organization.project.service_account.api_key",
    "value": "sk-abcdefgijklmnop123",
    "name": "Secret Key",
    "created_at": 1711471533,
    "id": "key_abc"
  }
}
```

# Retrieve project service account



```
GET https://api.openai.com/v1/organization/projects/{project_id}/service_accounts/{service_account_id}
```

Retrieves a service account in the project.

## Path parameters

**project\_id** string Required

The ID of the project.

**service\_account\_id** string Required

The ID of the service account.

## Returns

The [ProjectServiceAccount](#) object matching the specified ID.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/service_accounts/svc_acct_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "organization.project.service_account",  
  "id": "svc_acct_abc",  
  "name": "Service Account",  
  "role": "owner",  
  "created_at": 1711471533  
}
```

# Delete project service account



```
DELETE https://api.openai.com/v1/organization/projects/{project_id}/service_accounts/{service_account_id}
```

Deletes a service account from the project.

## Path parameters

**project\_id** string Required

The ID of the project.

**service\_account\_id** string Required

The ID of the service account.

## Returns

Confirmation of service account being deleted, or an error in case of an archived project, which has no service accounts

### Example request

```
curl -X DELETE https://api.openai.com/v1/organization/projects/proj_abc/service_accounts/svc_acct_ab
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "organization.project.service_account.deleted",
  "id": "svc_acct_abc",
  "deleted": true
}
```

# The project service account object



Represents an individual service account in a project.

---

**created\_at** integer

The Unix timestamp (in seconds) of when the service account was created

---

**id** string

The identifier, which can be referenced in API endpoints

---

**name** string

The name of the service account

---

**object** string

The object type, which is always `organization.project.service_account`

---

**role** string

`owner` or `member`

OBJECT The project service account object

{

```
"object": "organization.project.service_account",
"id": "svc_acct_abc",
```

```
"name": "Service Account",
"role": "owner",
"created_at": 1711471533
}
```

# Project API keys



Manage API keys for a given project. Supports listing and deleting keys for users. This API does not allow issuing keys for users, as users need to authorize themselves to generate keys.

## List project API keys



```
GET https://api.openai.com/v1/organization/projects/{project_id}/api_keys
```

Returns a list of API keys in the project.

## Path parameters

**project\_id** string Required

The ID of the project.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with `obj_foo`, your subsequent call can include `after=obj_foo` in order to fetch the next page of the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

## Returns

A list of [ProjectApiKey](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/api_keys?after=key_abc&limit=20 \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
    "object": "list",  
    "data": [  
        {  
            "object": "organization.project.api_key",  
            "redacted_value": "sk-abc...def",  
            "name": "My API Key",  
            "created_at": 1711471533,  
            "last_used_at": 1711471534,  
            "id": "key_abc",  
            "owner": {  
                "type": "user",  
                "user": {  
                    "object": "organization.project.user",  
                    "id": "user_abc",  
                    "name": "First Last",  
                    "email": "user@example.com",  
                    "role": "owner",  
                    "added_at": 1711471533  
                }  
            }  
        }  
    ],  
    "first_id": "key_abc",  
    "last_id": "key_xyz",  
    "has_more": false  
}
```

# Retrieve project API key



```
GET https://api.openai.com/v1/organization/projects/{project_id}/api_keys/{key_id}
```

Retrieves an API key in the project.

## Path parameters

**key\_id** string Required

The ID of the API key.

**project\_id** string Required

The ID of the project.

## Returns

The [ProjectApiKey](#) object matching the specified ID.

#### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/api_keys/key_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

#### Response

```
{
  "object": "organization.project.api_key",
  "redacted_value": "sk-abc...def",
  "name": "My API Key",
  "created_at": 1711471533,
  "last_used_at": 1711471534,
  "id": "key_abc",
  "owner": {
    "type": "user",
    "user": {
      "object": "organization.project.user",
      "id": "user_abc",
      "name": "First Last",
      "email": "user@example.com",
      "role": "owner",
      "added_at": 1711471533
    }
  }
}
```

{  
}

# Delete project API key



```
DELETE https://api.openai.com/v1/organization/projects/{project_id}/api_keys/{key_id}
```

Deletes an API key from the project.

## Path parameters

**key\_id** string Required

The ID of the API key.

**project\_id** string Required

The ID of the project.

## Returns

Confirmation of the key's deletion or an error if the key belonged to a service account

#### Example request

```
curl -X DELETE https://api.openai.com/v1/organization/projects/proj_abc/api_keys/key_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

#### Response

```
{
  "object": "organization.project.api_key.deleted",
  "id": "key_abc",
  "deleted": true
}
```

# The project API key object

🔗

Represents an individual API key in a project.

**created\_at** integer

The Unix timestamp (in seconds) of when the API key was created

**id** string

The identifier, which can be referenced in API endpoints

**last\_used\_at** integer

The Unix timestamp (in seconds) of when the API key was last used.

**name** string

The name of the API key

**object** string

The object type, which is always `organization.project.api_key`

**owner** object

> Show properties

**redacted\_value** string

The redacted value of the API key

OBJECT The project API key object

{

`"object": "organization.project.api_key",`

```
"redacted_value": "sk-abc...def",
"name": "My API Key",
"created_at": 1711471533,
"last_used_at": 1711471534,
"id": "key_abc",
"owner": {
    "type": "user",
    "user": {
        "object": "organization.project.user",
        "id": "user_abc",
        "name": "First Last",
        "email": "user@example.com",
        "role": "owner",
        "created_at": 1711471533
    }
}
}
```

# Project rate limits



Manage rate limits per model for projects. Rate limits may be configured to be equal to or lower than the organization's rate limits.

# List project rate limits



```
GET https://api.openai.com/v1/organization/projects/{project_id}/rate_limits
```

Returns the rate limits per model for a project.

## Path parameters

**project\_id** string Required

The ID of the project.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of the list.

**before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, beginning with obj\_foo, your subsequent call can include before=obj\_foo in order to fetch the previous page of the list.

**limit** integer Optional Defaults to 100

A limit on the number of objects to be returned. The default is 100.

## Returns

A list of [ProjectRateLimit](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/rate_limits?after=rl_xxx&limit=20 \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

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

```
        "object": "project.rate_limit",
        "id": "rl-ada",
        "model": "ada",
        "max_requests_per_1_minute": 600,
        "max_tokens_per_1_minute": 150000,
        "max_images_per_1_minute": 10
    },
],
"first_id": "rl-ada",
"last_id": "rl-ada",
"has_more": false
}
```

## Modify project rate limit



POST [https://api.openai.com/v1/organization/projects/{project\\_id}/rate\\_limits/{rate\\_limit\\_id}](https://api.openai.com/v1/organization/projects/{project_id}/rate_limits/{rate_limit_id})

Updates a project rate limit.

### Path parameters

**project\_id** string Required

The ID of the project.

---

**rate\_limit\_id** string Required

The ID of the rate limit.

---

## Request body

---

**batch\_1\_day\_max\_input\_tokens** integer Optional

The maximum batch input tokens per day. Only relevant for certain models.

---

**max\_audio\_megabytes\_per\_1\_minute** integer Optional

The maximum audio megabytes per minute. Only relevant for certain models.

---

**max\_images\_per\_1\_minute** integer Optional

The maximum images per minute. Only relevant for certain models.

---

**max\_requests\_per\_1\_day** integer Optional

The maximum requests per day. Only relevant for certain models.

---

**max\_requests\_per\_1\_minute** integer Optional

The maximum requests per minute.

---

**max\_tokens\_per\_1\_minute** integer Optional

The maximum tokens per minute.

## Returns

The updated [ProjectRateLimit](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/projects/proj_abc/rate_limits/rl_xxx \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
    "max_requests_per_1_minute": 500
}'
```

### Response

```
{
  "object": "project.rate_limit",
  "id": "rl-ada",
  "model": "ada",
  "max_requests_per_1_minute": 600,
  "max_tokens_per_1_minute": 150000,
```

```
"max_images_per_1_minute": 10  
}
```

# The project rate limit object



Represents a project rate limit config.

---

**batch\_1\_day\_max\_input\_tokens** integer

The maximum batch input tokens per day. Only present for relevant models.

---

**id** string

The identifier, which can be referenced in API endpoints.

---

**max\_audio\_megabytes\_per\_1\_minute** integer

The maximum audio megabytes per minute. Only present for relevant models.

---

**max\_images\_per\_1\_minute** integer

The maximum images per minute. Only present for relevant models.

---

**max\_requests\_per\_1\_day** integer

The maximum requests per day. Only present for relevant models.

**max\_requests\_per\_1\_minute** integer

The maximum requests per minute.

**max\_tokens\_per\_1\_minute** integer

The maximum tokens per minute.

**model** string

The model this rate limit applies to.

**object** string

The object type, which is always `project.rate_limit`

OBJECT The project rate limit object

```
{  
  "object": "project.rate_limit",  
  "id": "rl_ada",  
  "model": "ada",  
  "max_requests_per_1_minute": 600,  
  "max_tokens_per_1_minute": 150000,  
  "max_images_per_1_minute": 10  
}
```

# Audit logs



Logs of user actions and configuration changes within this organization. To log events, an Organization Owner must activate logging in the [Data Controls Settings](#). Once activated, for security reasons, logging cannot be deactivated.

---

## List audit logs



```
GET https://api.openai.com/v1/organization/audit_logs
```

List user actions and configuration changes within this organization.

### Query parameters

**actor\_emails[]** array Optional

Return only events performed by users with these emails.

**actor\_ids[]** array Optional

Return only events performed by these actors. Can be a user ID, a service account ID, or an api key tracking ID.

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include `after=obj_foo` in order to fetch the next page of the list.

---

**before** string Optional

A cursor for use in pagination. `before` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, starting with obj\_foo, your subsequent call can include `before=obj_foo` in order to fetch the previous page of the list.

---

**effective\_at** object Optional

Return only events whose `effective_at` (Unix seconds) is in this range.

> Show properties

---

**event\_types[]** array Optional

Return only events with a `type` in one of these values. For example, `project.created`. For all options, see the documentation for the [audit log object](#).

---

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

---

**project\_ids[]** array Optional

Return only events for these projects.

**resource\_ids[]** array Optional

Return only events performed on these targets. For example, a project ID updated.

## Returns

A list of paginated [Audit Log](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/audit_logs \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "id": "audit_log-xxx_yyyyymmdd",
      "type": "project.archived",
      "effective_at": 1722461446,
      "actor": {
        "type": "api_key",
        "api_key": {
          "id": "key-12345678901234567890123456789012"
        }
      }
    }
  ]
}
```

```
"type": "user",
"user": {
    "id": "user-xxx",
    "email": "user@example.com"
},
},
"project.archived": {
    "id": "proj_abc"
},
},
{
"id": "audit_log-yyy_20240101",
"type": "api_key.updated",
"effective_at": 1720804190,
"actor": {
    "type": "session",
    "session": {
        "user": {
            "id": "user-xxx",
            "email": "user@example.com"
        },
        "ip_address": "127.0.0.1",
        "user_agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/118.0.0.0 Safari/537.36",
        "ja3": "a497151ce4338a12c4418c44d375173e",
        "ja4": "q13d0313h3_55b375c5d22e_c7319ce65786",
        "ip_address_details": {
            "country": "US",
            "city": "San Francisco",
        }
    }
}
```

```
        "region": "California",
        "region_code": "CA",
        "asn": "1234",
        "latitude": "37.77490",
        "longitude": "-122.41940"
    }
},
],
"api_key.updated": {
    "id": "key_xxxx",
    "data": {
        "scopes": ["resource_2.operation_2"]
    }
},
],
"first_id": "audit_log-xxx_20240101",
"last_id": "audit_log_yyy_20240101",
"has_more": true
}
```

## The audit log object



A log of a user action or configuration change within this organization.

**actor** object

The actor who performed the audit logged action.

> Show properties

---

**api\_key.created** object

The details for events with this [type](#).

> Show properties

---

**api\_key.deleted** object

The details for events with this [type](#).

> Show properties

---

**api\_key.updated** object

The details for events with this [type](#).

> Show properties

---

**certificate.created** object

The details for events with this [type](#).

> Show properties

---

**certificate.deleted** object

The details for events with this [type](#).

> Show properties

---

**certificate.updated** object

The details for events with this [type](#).

> Show properties

---

**certificates.activated** object

The details for events with this [type](#).

> Show properties

---

**certificates.deactivated** object

The details for events with this [type](#).

> Show properties

---

**checkpoint.permission.created** object

The project and fine-tuned model checkpoint that the checkpoint permission was created for.

> Show properties

---

**checkpoint.permission.deleted** object

The details for events with this [type](#).

> Show properties

---

**effective\_at** integer

The Unix timestamp (in seconds) of the event.

---

**external\_key.registered** object

The details for events with this [type](#).

> Show properties

**external\_key.removed** object

The details for events with this [type](#).

> Show properties

---

**group.created** object

The details for events with this [type](#).

> Show properties

---

**group.deleted** object

The details for events with this [type](#).

> Show properties

---

**group.updated** object

The details for events with this [type](#).

> Show properties

---

**id** string

The ID of this log.

---

**invite.accepted** object

The details for events with this [type](#).

> Show properties

---

**invite.deleted** object

The details for events with this [type](#).

> Show properties

---

**invite.sent** object

The details for events with this [type](#).

> Show properties

---

**ip\_allowlist.config.activated** object

The details for events with this [type](#).

> Show properties

---

**ip\_allowlist.config.deactivated** object

The details for events with this [type](#).

> Show properties

---

**ip\_allowlist.created** object

The details for events with this [type](#).

> Show properties

---

**ip\_allowlist.deleted** object

The details for events with this [type](#).

> Show properties

---

**ip\_allowlist.updated** object

The details for events with this [type](#).

> Show properties

---

**login.failed** object

The details for events with this [type](#).

> Show properties

---

**login.succeeded** object

This event has no additional fields beyond the standard audit log attributes.

---

**logout.failed** object

The details for events with this [type](#).

> Show properties

---

**logout.succeeded** object

This event has no additional fields beyond the standard audit log attributes.

---

**organization.updated** object

The details for events with this [type](#).

> Show properties

---

**project** object

The project that the action was scoped to. Absent for actions not scoped to projects. Note that any admin actions taken via Admin API keys are associated with the default project.

> Show properties

---

**project.archived** object

The details for events with this [type](#).

> Show properties

---

### **project.created** object

The details for events with this [type](#).

> Show properties

---

### **project.deleted** object

The details for events with this [type](#).

> Show properties

---

### **project.updated** object

The details for events with this [type](#).

> Show properties

---

### **rate\_limit.deleted** object

The details for events with this [type](#).

> Show properties

---

### **rate\_limit.updated** object

The details for events with this [type](#).

> Show properties

---

### **role.assignment.created** object

The details for events with this [type](#).

> Show properties

---

**role.assignment.deleted** object

The details for events with this [type](#).

> Show properties

---

**role.created** object

The details for events with this [type](#).

> Show properties

---

**role.deleted** object

The details for events with this [type](#).

> Show properties

---

**role.updated** object

The details for events with this [type](#).

> Show properties

---

**scim.disabled** object

The details for events with this [type](#).

> Show properties

---

**scim.enabled** object

The details for events with this [type](#).

> Show properties

---

**service\_account.created** object

The details for events with this [type](#).

> Show properties

---

**service\_account.deleted** object

The details for events with this [type](#).

> Show properties

---

**service\_account.updated** object

The details for events with this [type](#).

> Show properties

---

**type** string

The event type.

---

**user.added** object

The details for events with this [type](#).

> Show properties

---

**user.deleted** object

The details for events with this [type](#).

> Show properties

---

**user.updated** object

The details for events with this [type](#).

> Show properties

---

## OBJECT The audit log object

```
{  
  "id": "req_xxx_20240101",  
  "type": "api_key.created",  
  "effective_at": 1720804090,  
  "actor": {  
    "type": "session",  
    "session": {  
      "user": {  
        "id": "user-xxx",  
        "email": "user@example.com"  
      },  
      "ip_address": "127.0.0.1",  
      "user_agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like  
    }  
  },  
  "api_key.created": {  
    "id": "key_xxxx",  
    "data": {  
      "scopes": ["resource.operation"]  
    }  
  }  
}
```

# Usage



The **Usage API** provides detailed insights into your activity across the OpenAI API. It also includes a separate [Costs endpoint](#), which offers visibility into your spend, breaking down consumption by invoice line items and project IDs.

While the Usage API delivers granular usage data, it may not always reconcile perfectly with the Costs due to minor differences in how usage and spend are recorded. For financial purposes, we recommend using the [Costs endpoint](#) or the [Costs tab](#) in the Usage Dashboard, which will reconcile back to your billing invoice.

---

# Completions



```
GET https://api.openai.com/v1/organization/usage/completions
```

Get completions usage details for the organization.

## Query parameters

---

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

---

**api\_key\_ids** array Optional

Return only usage for these API keys.

---

**batch** boolean Optional

If `true`, return batch jobs only. If `false`, return non-batch jobs only. By default, return both.

---

**bucket\_width** string Optional Defaults to `1d`

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

---

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

---

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`, `user_id`, `api_key_id`, `model`, `batch`, `service_tier` or any combination of them.

---

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

---

**models** array Optional

Return only usage for these models.

---

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

---

**project\_ids** array Optional

Return only usage for these projects.

---

**user\_ids** array Optional

Return only usage for these users.

---

## Returns

A list of paginated, time bucketed [Completions](#) usage objects.

---

### Example request

```
curl "https://api.openai.com/v1/organization/usage/completions?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "page",  
  "data": [  
    {  
      "object": "bucket",  
      "start_time": 1730419200,  
      "end_time": 1730505600,  
      "results": [  
        {  
          "object": "organization.usage.completions.result",  
          "input_tokens": 1000,  
          "output_tokens": 500,  
          "input_cached_tokens": 800,  
          "input_audio_tokens": 0,  
          "output_audio_tokens": 0,  
          "num_model_requests": 5,  
          "project_id": null,  
          "user_id": null,  
          "api_key_id": null,  
          "model": null,  
          "batch": null,  
          "service_tier": null  
        }  
      ]  
    }  
  ],  
  "has_more": true,
```

```
        "next_page": "page_AAAAAGdGxdEiJdKOAAAAAGcqsYA="  
    }
```

# Completions usage object



The aggregated completions usage details of the specific time bucket.

**api\_key\_id** string

When `group_by=api_key_id` , this field provides the API key ID of the grouped usage result.

**batch** boolean

When `group_by=batch` , this field tells whether the grouped usage result is batch or not.

**input\_audio\_tokens** integer

The aggregated number of audio input tokens used, including cached tokens.

**input\_cached\_tokens** integer

The aggregated number of text input tokens that has been cached from previous requests. For customers subscribe to scale tier, this includes scale tier tokens.

**input\_tokens** integer

The aggregated number of text input tokens used, including cached tokens. For customers subscribe to scale tier, this includes scale tier tokens.

---

**model** string

When `group_by=model` , this field provides the model name of the grouped usage result.

---

**num\_model\_requests** integer

The count of requests made to the model.

---

**object** string**output\_audio\_tokens** integer

The aggregated number of audio output tokens used.

---

**output\_tokens** integer

The aggregated number of text output tokens used. For customers subscribe to scale tier, this includes scale tier tokens.

---

**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

---

**service\_tier** string

When `group_by=service_tier` , this field provides the service tier of the grouped usage result.

---

**user\_id** string

When `group_by=user_id` , this field provides the user ID of the grouped usage result.

OBJECT Completions usage object

```
{  
    "object": "organization.usage.completions.result",  
    "input_tokens": 5000,  
    "output_tokens": 1000,  
    "input_cached_tokens": 4000,  
    "input_audio_tokens": 300,  
    "output_audio_tokens": 200,  
    "num_model_requests": 5,  
    "project_id": "proj_abc",  
    "user_id": "user-abc",  
    "api_key_id": "key_abc",  
    "model": "gpt-4o-mini-2024-07-18",  
    "batch": false,  
    "service_tier": "default"  
}
```

## Embeddings



GET <https://api.openai.com/v1/organization/usage/embeddings>

## Get embeddings usage details for the organization.

### Query parameters

---

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

---

**api\_key\_ids** array Optional

Return only usage for these API keys.

---

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

---

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

---

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`, `user_id`, `api_key_id`, `model` or any combination of them.

---

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

---

**models** array Optional

Return only usage for these models.

---

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

---

**project\_ids** array Optional

Return only usage for these projects.

---

**user\_ids** array Optional

Return only usage for these users.

---

## Returns

A list of paginated, time bucketed [Embeddings usage](#) objects.

---

### Example request

```
curl "https://api.openai.com/v1/organization/usage/embeddings?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "page",  
  "data": [  
    {  
      "object": "bucket",  
      "start_time": 1730419200,  
      "end_time": 1730505600,  
      "results": [  
        {  
          "object": "organization.usage.embeddings.result",  
          "input_tokens": 16,  
          "num_model_requests": 2,  
          "project_id": null,  
          "user_id": null,  
          "api_key_id": null,  
          "model": null  
        }  
      ]  
    }  
  ],  
  "has_more": false,  
  "next_page": null  
}
```

# Embeddings usage object



The aggregated embeddings usage details of the specific time bucket.

---

**api\_key\_id** string

When `group_by=api_key_id` , this field provides the API key ID of the grouped usage result.

---

**input\_tokens** integer

The aggregated number of input tokens used.

---

**model** string

When `group_by=model` , this field provides the model name of the grouped usage result.

---

**num\_model\_requests** integer

The count of requests made to the model.

---

**object** string

**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

---

**user\_id** string

When `group_by=user_id` , this field provides the user ID of the grouped usage result.

OBJECT Embeddings usage object

```
{  
    "object": "organization.usage.embeddings.result",  
    "input_tokens": 20,  
    "num_model_requests": 2,  
    "project_id": "proj_abc",  
    "user_id": "user-abc",  
    "api_key_id": "key_abc",  
    "model": "text-embedding-ada-002-v2"  
}
```

# Moderations



GET <https://api.openai.com/v1/organization/usage/moderations>

Get moderations usage details for the organization.

## Query parameters

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

---

**api\_key\_ids** array Optional

Return only usage for these API keys.

---

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

---

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

---

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`, `user_id`, `api_key_id`, `model` or any combination of them.

---

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

**models** array Optional

Return only usage for these models.

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

**project\_ids** array Optional

Return only usage for these projects.

**user\_ids** array Optional

Return only usage for these users.

## Returns

A list of paginated, time bucketed Moderations usage objects.

Example request

```
curl "https://api.openai.com/v1/organization/usage/moderations?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

Response

```
{  
  "object": "page",  
  "data": [  
    {  
      "object": "bucket",  
      "start_time": 1730419200,  
      "end_time": 1730505600,  
      "results": [  
        {  
          "object": "organization.usage.moderations.result",  
          "input_tokens": 16,  
          "num_model_requests": 2,  
          "project_id": null,  
          "user_id": null,  
          "api_key_id": null,  
          "model": null  
        }  
      ]  
    }  
  ],  
  "has_more": false,  
  "next_page": null  
}
```

# Moderations usage object



The aggregated moderations usage details of the specific time bucket.

---

**api\_key\_id** string

When `group_by=api_key_id` , this field provides the API key ID of the grouped usage result.

---

**input\_tokens** integer

The aggregated number of input tokens used.

---

**model** string

When `group_by=model` , this field provides the model name of the grouped usage result.

---

**num\_model\_requests** integer

The count of requests made to the model.

---

**object** string

**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

---

**user\_id** string

When `group_by=user_id` , this field provides the user ID of the grouped usage result.

OBJECT `Moderations usage object`

```
{  
  "object": "organization.usage.moderations.result",  
  "input_tokens": 20,  
  "num_model_requests": 2,  
  "project_id": "proj_abc",  
  "user_id": "user-abc",  
  "api_key_id": "key_abc",  
  "model": "text-moderation"  
}
```

# Images



GET <https://api.openai.com/v1/organization/usage/images>

Get images usage details for the organization.

## Query parameters

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

**api\_key\_ids** array Optional

Return only usage for these API keys.

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`, `user_id`, `api_key_id`, `model`, `size`, `source` or any combination of them.

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

**models** array Optional

Return only usage for these models.

---

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

---

**project\_ids** array Optional

Return only usage for these projects.

---

**sizes** array Optional

Return only usages for these image sizes. Possible values are `256x256` , `512x512` , `1024x1024` , `1792x1792` , `1024x1792` or any combination of them.

---

**sources** array Optional

Return only usages for these sources. Possible values are `image.generation` , `image.edit` , `image.variation` or any combination of them.

---

**user\_ids** array Optional

Return only usage for these users.

---

## Returns

A list of paginated, time bucketed Images usage objects.

#### Example request

```
curl "https://api.openai.com/v1/organization/usage/images?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

#### Response

```
{
  "object": "page",
  "data": [
    {
      "object": "bucket",
      "start_time": 1730419200,
      "end_time": 1730505600,
      "results": [
        {
          "object": "organization.usage.images.result",
          "images": 2,
          "num_model_requests": 2,
          "size": null,
          "source": null,
          "project_id": null,
          "user_id": null,
          "api_key_id": null,
        }
      ]
    }
  ]
}
```

```
        "model": null
    }
]
}
],
"has_more": false,
"next_page": null
}
```

## Images usage object



The aggregated images usage details of the specific time bucket.

**api\_key\_id** string

When `group_by=api_key_id` , this field provides the API key ID of the grouped usage result.

**images** integer

The number of images processed.

**model** string

When `group_by=model` , this field provides the model name of the grouped usage result.

**num\_model\_requests** integer

The count of requests made to the model.

**object** string**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

**size** string

When `group_by=size` , this field provides the image size of the grouped usage result.

**source** string

When `group_by=source` , this field provides the source of the grouped usage result, possible values are `image.generation` , `image.edit` , `image.variation` .

**user\_id** string

When `group_by=user_id` , this field provides the user ID of the grouped usage result.

OBJECT Images usage object

```
{  
  "object": "organization.usage.images.result",  
  "images": 2,  
  "num_model_requests": 2,  
  "size": "1024x1024",
```

```
"source": "image.generation",
"project_id": "proj_abc",
"user_id": "user-abc",
"api_key_id": "key_abc",
"model": "dall-e-3"
}
```

# Audio speeches



GET [https://api.openai.com/v1/organization/usage/audio\\_speeches](https://api.openai.com/v1/organization/usage/audio_speeches)

Get audio speeches usage details for the organization.

## Query parameters

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

**api\_key\_ids** array Optional

Return only usage for these API keys.

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

---

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

---

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`, `user_id`, `api_key_id`, `model` or any combination of them.

---

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

---

**models** array Optional

Return only usage for these models.

---

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

---

**project\_ids** array Optional

Return only usage for these projects.

**user\_ids** array Optional

Return only usage for these users.

## Returns

A list of paginated, time bucketed Audio speeches usage objects.

### Example request

```
curl "https://api.openai.com/v1/organization/usage/audio_speeches?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "page",
  "data": [
    {
      "object": "bucket",
      "start_time": 1730419200,
      "end_time": 1730505600,
      "results": [
```

```
{  
    "object": "organization.usage.audio_speeches.result",  
    "characters": 45,  
    "num_model_requests": 1,  
    "project_id": null,  
    "user_id": null,  
    "api_key_id": null,  
    "model": null  
}  
]  
}  
],  
"has_more": false,  
"next_page": null  
}
```

## Audio speeches usage object



The aggregated audio speeches usage details of the specific time bucket.

**api\_key\_id** string

When `group_by=api_key_id`, this field provides the API key ID of the grouped usage result.

**characters** integer

The number of characters processed.

**model** string

When `group_by=model` , this field provides the model name of the grouped usage result.

**num\_model\_requests** integer

The count of requests made to the model.

**object** string**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

**user\_id** string

When `group_by=user_id` , this field provides the user ID of the grouped usage result.

OBJECT Audio speeches usage object

```
{  
  "object": "organization.usage.audio_speeches.result",  
  "characters": 45,  
  "num_model_requests": 1,  
  "project_id": "proj_abc",  
  "user_id": "user-abc",
```

```
"api_key_id": "key_abc",  
"model": "tts-1"  
}
```

# Audio transcriptions



```
GET https://api.openai.com/v1/organization/usage/audio_transcriptions
```

Get audio transcriptions usage details for the organization.

## Query parameters

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

**api\_key\_ids** array Optional

Return only usage for these API keys.

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

---

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`, `user_id`, `api_key_id`, `model` or any combination of them.

---

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

---

**models** array Optional

Return only usage for these models.

---

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

---

**project\_ids** array Optional

Return only usage for these projects.

---

**user\_ids** array Optional

Return only usage for these users.

## Returns

A list of paginated, time bucketed [Audio transcriptions usage](#) objects.

### Example request

```
curl "https://api.openai.com/v1/organization/usage/audio_transcriptions?start_time=1730419200&limit=10" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "page",
  "data": [
    {
      "object": "bucket",
      "start_time": 1730419200,
      "end_time": 1730505600,
      "results": [
        {
          "object": "organization.usage.audio_transcriptions.result",
          "seconds": 20,
          "num_model_requests": 1,
          "model": "text-davinci-003"
        }
      ]
    }
  ]
}
```

```
        "project_id": null,  
        "user_id": null,  
        "api_key_id": null,  
        "model": null  
    }  
]  
}  
],  
"has_more": false,  
"next_page": null  
}
```

## Audio transcriptions usage object



The aggregated audio transcriptions usage details of the specific time bucket.

**api\_key\_id** string

When `group_by=api_key_id` , this field provides the API key ID of the grouped usage result.

**model** string

When `group_by=model` , this field provides the model name of the grouped usage result.

**num\_model\_requests** integer

The count of requests made to the model.

**object** string

**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

**seconds** integer

The number of seconds processed.

**user\_id** string

When `group_by=user_id` , this field provides the user ID of the grouped usage result.

OBJECT Audio transcriptions usage object

```
{  
  "object": "organization.usage.audio_transcriptions.result",  
  "seconds": 10,  
  "num_model_requests": 1,  
  "project_id": "proj_abc",  
  "user_id": "user-abc",  
  "api_key_id": "key_abc",  
  "model": "tts-1"  
}
```

# Vector stores



```
GET https://api.openai.com/v1/organization/usage/vector_stores
```

Get vector stores usage details for the organization.

## Query parameters

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`.

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

**project\_ids** array Optional

Return only usage for these projects.

## Returns

A list of paginated, time bucketed Vector stores usage objects.

### Example request

```
curl "https://api.openai.com/v1/organization/usage/vector_stores?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{  
  "object": "page",  
  "data": [  
    {  
      "object": "bucket",  
      "start_time": 1730419200,  
      "end_time": 1730505600,  
      "results": [  
        {  
          "object": "organization.usage.vector_stores.result",  
          "usage_bytes": 1024,  
          "project_id": null  
        }  
      ]  
    }  
  ],  
  "has_more": false,  
  "next_page": null  
}
```

## Vector stores usage object



The aggregated vector stores usage details of the specific time bucket.

**object** string

**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

**usage\_bytes** integer

The vector stores usage in bytes.

OBJECT Vector stores usage object

```
{  
  "object": "organization.usage.vector_stores.result",  
  "usage_bytes": 1024,  
  "project_id": "proj_abc"  
}
```

## Code interpreter sessions



GET [https://api.openai.com/v1/organization/usage/code\\_interpreter\\_sessions](https://api.openai.com/v1/organization/usage/code_interpreter_sessions)

## Get code interpreter sessions usage details for the organization.

### Query parameters

---

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

---

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently `1m`, `1h` and `1d` are supported, default to `1d`.

---

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

---

**group\_by** array Optional

Group the usage data by the specified fields. Support fields include `project_id`.

---

**limit** integer Optional

Specifies the number of buckets to return.

`bucket_width=1d` : default: 7, max: 31

`bucket_width=1h` : default: 24, max: 168

`bucket_width=1m` : default: 60, max: 1440

---

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

### **project\_ids** array Optional

Return only usage for these projects.

## Returns

A list of paginated, time bucketed [Code interpreter sessions usage](#) objects.

### Example request

```
curl "https://api.openai.com/v1/organization/usage/code_interpreter_sessions?start_time=1730419200&end_time=1730505600" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

### Response

```
{
  "object": "page",
  "data": [
    {
      "object": "bucket",
      "start_time": 1730419200,
      "end_time": 1730505600,
      "results": [
```

```
        {
          "object": "organization.usage.code_interpreter_sessions.result",
          "num_sessions": 1,
          "project_id": null
        }
      ]
    }
  ],
  "has_more": false,
  "next_page": null
}
```

## Code interpreter sessions usage object



The aggregated code interpreter sessions usage details of the specific time bucket.

---

**num\_sessions** integer

The number of code interpreter sessions.

---

**object** string

**project\_id** string

When `group_by=project_id` , this field provides the project ID of the grouped usage result.

OBJECT Code interpreter sessions usage object

```
{  
  "object": "organization.usage.code_interpreter_sessions.result",  
  "num_sessions": 1,  
  "project_id": "proj_abc"  
}
```

# Costs



GET <https://api.openai.com/v1/organization/costs>

Get costs details for the organization.

## Query parameters

**start\_time** integer Required

Start time (Unix seconds) of the query time range, inclusive.

**bucket\_width** string Optional Defaults to 1d

Width of each time bucket in response. Currently only `1d` is supported, default to `1d`.

**end\_time** integer Optional

End time (Unix seconds) of the query time range, exclusive.

**group\_by** array Optional

Group the costs by the specified fields. Support fields include `project_id`, `line_item` and any combination of them.

**limit** integer Optional Defaults to 7

A limit on the number of buckets to be returned. Limit can range between 1 and 180, and the default is 7.

**page** string Optional

A cursor for use in pagination. Corresponding to the `next_page` field from the previous response.

**project\_ids** array Optional

Return only costs for these projects.

## Returns

A list of paginated, time bucketed Costs objects.

### Example request

```
curl "https://api.openai.com/v1/organization/costs?start_time=1730419200&limit=1" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json"
```

## Response

```
{
  "object": "page",
  "data": [
    {
      "object": "bucket",
      "start_time": 1730419200,
      "end_time": 1730505600,
      "results": [
        {
          "object": "organization.costs.result",
          "amount": {
            "value": 0.06,
            "currency": "usd"
          },
          "line_item": null,
          "project_id": null
        }
      ]
    },
    "has_more": false,
```

```
"next_page": null  
}
```

# Costs object



The aggregated costs details of the specific time bucket.

## amount object

The monetary value in its associated currency.

> Show properties

## line\_item string

When `group_by=line_item` , this field provides the line item of the grouped costs result.

## object string

## project\_id string

When `group_by=project_id` , this field provides the project ID of the grouped costs result.

OBJECT Costs object

```
{  
  "object": "organization.costs.result",  
  "amount": {  
    "value": 0.06,  
    "currency": "usd"  
  },  
  "line_item": "Image models",  
  "project_id": "proj_abc"  
}
```

## Certificates Beta



Manage Mutual TLS certificates across your organization and projects.

[Learn more about Mutual TLS.](#)

## Upload certificate



```
POST https://api.openai.com/v1/organization/certificates
```

Upload a certificate to the organization. This does **not** automatically activate the certificate.

Organizations can upload up to 50 certificates.

## Request body

**content** string Required

The certificate content in PEM format

---

**name** string Optional

An optional name for the certificate

## Returns

A single [Certificate](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/certificates \  
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \  
-H "Content-Type: application/json" \  
-d '{  
  "name": "My Example Certificate",
```

```
"certificate": "-----BEGIN CERTIFICATE-----\\nMIIDeT...\\n-----END CERTIFICATE-----"  
}'
```

## Response

```
{  
  "object": "certificate",  
  "id": "cert_abc",  
  "name": "My Example Certificate",  
  "created_at": 1234567,  
  "certificate_details": {  
    "valid_at": 12345667,  
    "expires_at": 12345678  
  }  
}
```

# Get certificate



```
GET https://api.openai.com/v1/organization/certificates/{certificate_id}
```

Get a certificate that has been uploaded to the organization.

You can get a certificate regardless of whether it is active or not.

## Path parameters

---

**certificate\_id** string Required

Unique ID of the certificate to retrieve.

## Query parameters

---

**include** array Optional

A list of additional fields to include in the response. Currently the only supported value is `content` to fetch the PEM content of the certificate.

## Returns

---

A single [Certificate](#) object.

### Example request

```
curl "https://api.openai.com/v1/organization/certificates/cert_abc?include[]=content" \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY"
```

## Response

```
{  
  "object": "certificate",  
  "id": "cert_abc",  
  "name": "My Example Certificate",  
  "created_at": 1234567,  
  "certificate_details": {  
    "valid_at": 1234567,  
    "expires_at": 12345678,  
    "content": "-----BEGIN CERTIFICATE-----MIIDeT...-----END CERTIFICATE-----"  
  }  
}
```

## Modify certificate



```
POST https://api.openai.com/v1/organization/certificates/{certificate_id}
```

Modify a certificate. Note that only the name can be modified.

## Request body

**name** string Required

The updated name for the certificate

## Returns

The updated [Certificate](#) object.

### Example request

```
curl -X POST https://api.openai.com/v1/organization/certificates/cert_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{'
  "name": "Renamed Certificate"
}'
```

### Response

```
{
  "object": "certificate",
  "id": "cert_abc",
```

```
"name": "Renamed Certificate",
"created_at": 1234567,
"certificate_details": {
    "valid_at": 12345667,
    "expires_at": 12345678
}
}
```

# Delete certificate



```
DELETE https://api.openai.com/v1/organization/certificates/{certificate_id}
```

Delete a certificate from the organization.

The certificate must be inactive for the organization and all projects.

## Returns

A confirmation object indicating the certificate was deleted.

### Example request

```
curl -X DELETE https://api.openai.com/v1/organization/certificates/cert_abc \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY"
```

## Response

```
{  
  "object": "certificate.deleted",  
  "id": "cert_abc"  
}
```

# List organization certificates



GET <https://api.openai.com/v1/organization/certificates>

List uploaded certificates for this organization.

## Query parameters

**after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include after=obj\_foo in order to fetch the next page of

the list.

**limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

**order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

## Returns

A list of [Certificate](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/certificates \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY"
```

### Response

```
{
  "object": "list",
  "data": [
    {
      "object": "organization.certificate",
      "id": "cert_abc",
```

```
"name": "My Example Certificate",
"active": true,
"created_at": 1234567,
"certificate_details": {
    "valid_at": 12345667,
    "expires_at": 12345678
},
],
{
"first_id": "cert_abc",
"last_id": "cert_abc",
"has_more": false
}
```

# List project certificates



GET [https://api.openai.com/v1/organization/projects/{project\\_id}/certificates](https://api.openai.com/v1/organization/projects/{project_id}/certificates)

List certificates for this project.

## Path parameters

**project\_id** string Required

The ID of the project.

## Query parameters

---

### **after** string Optional

A cursor for use in pagination. `after` is an object ID that defines your place in the list. For instance, if you make a list request and receive 100 objects, ending with obj\_foo, your subsequent call can include `after=obj_foo` in order to fetch the next page of the list.

---

### **limit** integer Optional Defaults to 20

A limit on the number of objects to be returned. Limit can range between 1 and 100, and the default is 20.

---

### **order** string Optional Defaults to desc

Sort order by the `created_at` timestamp of the objects. `asc` for ascending order and `desc` for descending order.

---

## Returns

---

A list of [Certificate](#) objects.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/certificates \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY"
```

## Response

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "organization.project.certificate",  
      "id": "cert_abc",  
      "name": "My Example Certificate",  
      "active": true,  
      "created_at": 1234567,  
      "certificate_details": {  
        "valid_at": 12345667,  
        "expires_at": 12345678  
      }  
    },  
  ],  
  "first_id": "cert_abc",  
  "last_id": "cert_abc",  
  "has_more": false  
}
```

# Activate certificates for organization



```
POST https://api.openai.com/v1/organization/certificates/activate
```

Activate certificates at the organization level.

You can atomically and idempotently activate up to 10 certificates at a time.

## Request body

`certificate_ids` array Required

## Returns

A list of [Certificate](#) objects that were activated.

### Example request

```
curl https://api.openai.com/v1/organization/certificates/activate \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
  "data": ["cert_abc", "cert_def"]
}'
```

### Response

```
{  
  "object": "organization.certificate.activation",  
  "data": [  
    {  
      "object": "organization.certificate",  
      "id": "cert_abc",  
      "name": "My Example Certificate",  
      "active": true,  
      "created_at": 1234567,  
      "certificate_details": {  
        "valid_at": 12345667,  
        "expires_at": 12345678  
      }  
    },  
    {  
      "object": "organization.certificate",  
      "id": "cert_def",  
      "name": "My Example Certificate 2",  
      "active": true,  
      "created_at": 1234567,  
      "certificate_details": {  
        "valid_at": 12345667,  
        "expires_at": 12345678  
      }  
    },  
  ],  
}
```

# Deactivate certificates for organization



```
POST https://api.openai.com/v1/organization/certificates/deactivate
```

Deactivate certificates at the organization level.

You can atomically and idempotently deactivate up to 10 certificates at a time.

## Request body

---

**certificate\_ids** array Required

## Returns

---

A list of [Certificate](#) objects that were deactivated.

### Example request

```
curl https://api.openai.com/v1/organization/certificates/deactivate \  
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \  
-X POST
```

```
-H "Content-Type: application/json" \
-d '{
  "data": ["cert_abc", "cert_def"]
}'
```

## Response

```
{
  "object": "organization.certificate.deactivation",
  "data": [
    {
      "object": "organization.certificate",
      "id": "cert_abc",
      "name": "My Example Certificate",
      "active": false,
      "created_at": 1234567,
      "certificate_details": {
        "valid_at": 12345667,
        "expires_at": 12345678
      }
    },
    {
      "object": "organization.certificate",
      "id": "cert_def",
      "name": "My Example Certificate 2",
      "active": false,
      "created_at": 1234567,
      "certificate_details": {
```

```
        "valid_at": 12345667,  
        "expires_at": 12345678  
    },  
},  
],  
}
```

# Activate certificates for project



POST [https://api.openai.com/v1/organization/projects/{project\\_id}/certificates/activate](https://api.openai.com/v1/organization/projects/{project_id}/certificates/activate)

Activate certificates at the project level.

You can atomically and idempotently activate up to 10 certificates at a time.

## Path parameters

**project\_id** string Required

The ID of the project.

## Request body

**certificate\_ids** array Required

## Returns

A list of [Certificate](#) objects that were activated.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/certificates/activate \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
  "data": ["cert_abc", "cert_def"]
}'
```

### Response

```
{
  "object": "organization.project.certificate.activation",
  "data": [
    {
      "object": "organization.project.certificate",
      "id": "cert_abc",
      "name": "My Example Certificate",
      "active": true,
```

```
"created_at": 1234567,  
"certificate_details": {  
    "valid_at": 12345667,  
    "expires_at": 12345678  
}  
,  
{  
    "object": "organization.project.certificate",  
    "id": "cert_def",  
    "name": "My Example Certificate 2",  
    "active": true,  
    "created_at": 1234567,  
    "certificate_details": {  
        "valid_at": 12345667,  
        "expires_at": 12345678  
    }  
,  
],  
}
```

## Deactivate certificates for project



POST [https://api.openai.com/v1/organization/projects/{project\\_id}/certificates/deactivate](https://api.openai.com/v1/organization/projects/{project_id}/certificates/deactivate)

Deactivate certificates at the project level. You can atomically and idempotently deactivate up to 10 certificates at a time.

## Path parameters

---

**project\_id** string Required

The ID of the project.

## Request body

---

**certificate\_ids** array Required

## Returns

---

A list of [Certificate](#) objects that were deactivated.

### Example request

```
curl https://api.openai.com/v1/organization/projects/proj_abc/certificates/deactivate \
-H "Authorization: Bearer $OPENAI_ADMIN_KEY" \
-H "Content-Type: application/json" \
-d '{
  "data": ["cert_abc", "cert_def"]
}'
```

## Response

```
{  
  "object": "organization.project.certificate.deactivation",  
  "data": [  
    {  
      "object": "organization.project.certificate",  
      "id": "cert_abc",  
      "name": "My Example Certificate",  
      "active": false,  
      "created_at": 1234567,  
      "certificate_details": {  
        "valid_at": 12345667,  
        "expires_at": 12345678  
      }  
    },  
    {  
      "object": "organization.project.certificate",  
      "id": "cert_def",  
      "name": "My Example Certificate 2",  
      "active": false,  
      "created_at": 1234567,  
      "certificate_details": {  
        "valid_at": 12345667,  
        "expires_at": 12345678  
      }  
    }  
  ]  
}
```

```
    }
  },
],
}
```

# The certificate object



Represents an individual `certificate` uploaded to the organization.

**active** boolean

Whether the certificate is currently active at the specified scope. Not returned when getting details for a specific certificate.

**certificate\_details** object

> Show properties

**created\_at** integer

The Unix timestamp (in seconds) of when the certificate was uploaded.

**id** string

The identifier, which can be referenced in API endpoints

**name** string

The name of the certificate.

**object** string

The object type.

If creating, updating, or getting a specific certificate, the object type is `certificate`.

If listing, activating, or deactivating certificates for the organization, the object type is `organization.certificate`.

If listing, activating, or deactivating certificates for a project, the object type is `organization.project.certificate`.

#### OBJECT The certificate object

```
{  
  "object": "certificate",  
  "id": "cert_abc",  
  "name": "My Certificate",  
  "created_at": 1234567,  
  "certificate_details": {  
    "valid_at": 1234567,  
    "expires_at": 12345678,  
    "content": "-----BEGIN CERTIFICATE----- MIIGAjCCA...6znF10W+ -----END CERTIFICATE-----"  
  }  
}
```

# Completions

Legacy



Given a prompt, the model will return one or more predicted completions along with the probabilities of alternative tokens at each position. Most developer should use our [Chat Completions API](#) to leverage our best and newest models.

---

## Create completion

Legacy



```
POST https://api.openai.com/v1/completions
```

Creates a completion for the provided prompt and parameters.

### Request body

**model** string Required

ID of the model to use. You can use the [List models API](#) to see all of your available models, or see our [Model overview](#) for descriptions of them.

---

**prompt** string or array Required

The prompt(s) to generate completions for, encoded as a string, array of strings, array of tokens, or array of token arrays.

Note that <|endoftext|> is the document separator that the model sees during training, so if a prompt is not specified the model will generate as if from the beginning of a new document.

---

**best\_of** integer or null Optional Defaults to 1

Generates `best_of` completions server-side and returns the "best" (the one with the highest log probability per token).

Results cannot 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`.

---

**echo** boolean or null Optional Defaults to false

Echo back the prompt in addition to the completion

---

**frequency\_penalty** number or null Optional Defaults to 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.

[See more information about frequency and presence penalties.](#)

---

**logit\_bias** map Optional Defaults to 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](#) to convert text to token IDs. Mathematically, the bias is added to the logits

generated by the model prior to sampling. The exact effect will vary 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.

---

**logprobs** integer or null Optional Defaults to null

Include the log probabilities on the `logprobs` most likely output tokens, as well the chosen tokens. For example, if `logprobs` is 5, the API will return a list of the 5 most likely tokens. The API will always return the `logprob` of the sampled token, so there may be up to `logprobs+1` elements in the response.

The maximum value for `logprobs` is 5.

---

**max\_tokens** integer or null Optional Defaults to 16

The maximum number of tokens that can be generated in the completion.

The token count of your prompt plus `max_tokens` cannot exceed the model's context length. [Example Python code](#) for counting tokens.

---

**n** integer or null Optional Defaults to 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`.

---

**presence\_penalty** number or null Optional Defaults to 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.

[See more information about frequency and presence penalties.](#)

---

**seed** integer or null Optional

If specified, our system will make a best effort to sample deterministically, such that repeated requests with the same `seed` and parameters should return the same result.

Determinism is not guaranteed, and you should refer to the `system_fingerprint` response parameter to monitor changes in the backend.

---

**stop** string / array / null Optional Defaults to null

Not supported with latest reasoning models `o3` and `o4-mini`.

Up to 4 sequences where the API will stop generating further tokens. The returned text will not contain the stop sequence.

---

**stream** boolean or null Optional Defaults to false

Whether to stream back partial progress. If set, tokens will be sent as data-only [server-sent events](#) as they become available, with the stream terminated by a `data: [DONE]` message. [Example Python code](#).

---

**stream\_options** object Optional Defaults to null

Options for streaming response. Only set this when you set `stream: true`.

> Show properties

---

**suffix** string or null Optional Defaults to null

The suffix that comes after a completion of inserted text.

This parameter is only supported for `gpt-3.5-turbo-instruct`.

---

**temperature** number or null Optional Defaults to 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.

---

**top\_p** number or null Optional Defaults to 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.

---

**user** string Optional

A unique identifier representing your end-user, which can help OpenAI to monitor and detect abuse. [Learn more](#).

## Returns

Returns a [completion](#) object, or a sequence of completion objects if the request is streamed.

No streaming

Streaming

Example request

```
curl https://api.openai.com/v1/completions \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $OPENAI_API_KEY" \
```

```
-d '{  
    "model": "gpt-3.5-turbo-instruct",  
    "prompt": "Say this is a test",  
    "max_tokens": 7,  
    "temperature": 0  
}'
```

## Response

```
{  
    "id": "cmpl-uqkv1QyYK7bGYrRHQ0eXlWi7",  
    "object": "text_completion",  
    "created": 1589478378,  
    "model": "gpt-3.5-turbo-instruct",  
    "system_fingerprint": "fp_44709d6fcb",  
    "choices": [  
        {  
            "text": "\n\nThis is indeed a test",  
            "index": 0,  
            "logprobs": null,  
            "finish_reason": "length"  
        }  
    ],  
    "usage": {  
        "prompt_tokens": 5,  
        "completion_tokens": 7,  
        "total_tokens": 12  
    }  
}
```

# The completion object

Legacy



Represents a completion response from the API. Note: both the streamed and non-streamed response objects share the same shape (unlike the chat endpoint).

---

## **choices** array

The list of completion choices the model generated for the input prompt.

> Show properties

---

## **created** integer

The Unix timestamp (in seconds) of when the completion was created.

---

## **id** string

A unique identifier for the completion.

---

## **model** string

The model used for completion.

---

## **object** string

The object type, which is always "text\_completion"

### system\_fingerprint string

This fingerprint represents the backend configuration that the model runs with.

Can be used in conjunction with the `seed` request parameter to understand when backend changes have been made that might impact determinism.

### usage object

Usage statistics for the completion request.

> Show properties

OBJECT The completion object

```
{  
  "id": "cmpl-uqkv1QyYK7bGYrRHQ0eXlWi7",  
  "object": "text_completion",  
  "created": 1589478378,  
  "model": "gpt-4-turbo",  
  "choices": [  
    {  
      "text": "\n\nThis is indeed a test",  
      "index": 0,  
      "logprobs": null,  
      "finish_reason": "length"  
    }  
  ],  
  "usage": {
```

```
"prompt_tokens": 5,  
"completion_tokens": 7,  
"total_tokens": 12  
}  
}
```

## Realtime Beta Legacy



Communicate with a multimodal model in real time over low latency interfaces like WebRTC, WebSocket, and SIP. Natively supports speech-to-speech as well as text, image, and audio inputs and outputs. [Learn more about the Realtime API.](#)

## Realtime Beta session tokens



REST API endpoint to generate ephemeral session tokens for use in client-side applications.

# Create session



```
POST https://api.openai.com/v1/realtime/sessions
```

Create an ephemeral API token for use in client-side applications with the Realtime API. Can be configured with the same session parameters as the `session.update` client event.

It responds with a session object, plus a `client_secret` key which contains a usable ephemeral API token that can be used to authenticate browser clients for the Realtime API.

## Request body

**client\_secret** object Required

Ephemeral key returned by the API.

> Show properties

**input\_audio\_format** string Optional

The format of input audio. Options are `pcm16`, `g711_ulaw`, or `g711_alaw`.

**input\_audio\_transcription** object Optional

Configuration for input audio transcription, defaults to off and can be set to `null` to turn off once on. Input audio transcription is not native to the model, since the model consumes audio directly. Transcription runs asynchronously and should be treated as rough guidance rather than the representation understood by the model.

> Show properties

---

**instructions** string Optional

The default system instructions (i.e. system message) prepended to model calls. This field allows the client to guide the model on desired responses. The model can be instructed on response content and format, (e.g. "be extremely succinct", "act friendly", "here are examples of good responses") and on audio behavior (e.g. "talk quickly", "inject emotion into your voice", "laugh frequently"). The instructions are not guaranteed to be followed by the model, but they provide guidance to the model on the desired behavior. Note that the server sets default instructions which will be used if this field is not set and are visible in the `session.created` event at the start of the session.

---

**max\_response\_output\_tokens** integer or "inf" Optional

Maximum number of output tokens for a single assistant response, inclusive of tool calls. Provide an integer between 1 and 4096 to limit output tokens, or `inf` for the maximum available tokens for a given model. Defaults to `inf`.

---

**modalities** Optional

The set of modalities the model can respond with. To disable audio, set this to `["text"]`.

---

**output\_audio\_format** string Optional

The format of output audio. Options are `pcm16`, `g711_ulaw`, or `g711_alaw`.

---

**prompt** object Optional

Reference to a prompt template and its variables. [Learn more](#).

> Show properties

---

**speed** number Optional Defaults to 1

The speed of the model's spoken response. 1.0 is the default speed. 0.25 is the minimum speed. 1.5 is the maximum speed. This value can only be changed in between model turns, not while a response is in progress.

---

**temperature** number Optional

Sampling temperature for the model, limited to [0.6, 1.2]. Defaults to 0.8.

---

**tool\_choice** string Optional

How the model chooses tools. Options are `auto`, `none`, `required`, or specify a function.

---

**tools** array Optional

Tools (functions) available to the model.

> Show properties

---

**tracing** "auto" or object Optional

Configuration options for tracing. Set to null to disable tracing. Once tracing is enabled for a session, the configuration cannot be modified.

`auto` will create a trace for the session with default values for the workflow name, group id, and metadata.

> Show possible types

---

**truncation** string or object Optional

Controls how the realtime conversation is truncated prior to model inference. The default is `auto`.

> Show possible types

---

**turn\_detection** object Optional

Configuration for turn detection. Can be set to `null` to turn off. Server VAD means that the model will detect the start and end of speech based on audio volume and respond at the end of user speech.

> Show properties

---

#### **voice** string Optional

The voice the model uses to respond. Voice cannot be changed during the session once the model has responded with audio at least once. Current voice options are `alloy`, `ash`, `ballad`, `coral`, `echo`, `sage`, `shimmer`, and `verse`.

## Returns

---

The created Realtime session object, plus an ephemeral key

#### Example request

```
curl -X POST https://api.openai.com/v1/realtime/sessions \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "model": "gpt-realtime",
  "modalities": ["audio", "text"],
  "instructions": "You are a friendly assistant."
}'
```

#### Response

```
{  
    "id": "sess_001",  
    "object": "realtime.session",  
    "model": "gpt-realtime-2025-08-25",  
    "modalities": ["audio", "text"],  
    "instructions": "You are a friendly assistant.",  
    "voice": "alloy",  
    "input_audio_format": "pcm16",  
    "output_audio_format": "pcm16",  
    "input_audio_transcription": {  
        "model": "whisper-1"  
    },  
    "turn_detection": null,  
    "tools": [],  
    "tool_choice": "none",  
    "temperature": 0.7,  
    "max_response_output_tokens": 200,  
    "speed": 1.1,  
    "tracing": "auto",  
    "client_secret": {  
        "value": "ek_abc123",  
        "expires_at": 1234567890  
    }  
}
```

# Create transcription session



```
POST https://api.openai.com/v1/realtime/transcription_sessions
```

Create an ephemeral API token for use in client-side applications with the Realtime API specifically for realtime transcriptions. Can be configured with the same session parameters as the `transcription_session.update` client event.

It responds with a session object, plus a `client_secret` key which contains a usable ephemeral API token that can be used to authenticate browser clients for the Realtime API.

## Request body

---

**include** array Optional

The set of items to include in the transcription. Current available items are: `item.input_audio_transcription.logprobs`

---

**input\_audio\_format** string Optional Defaults to `pcm16`

The format of input audio. Options are `pcm16`, `g711_ulaw`, or `g711_alaw`. For `pcm16`, input audio must be 16-bit PCM at a 24kHz sample rate, single channel (mono), and little-endian byte order.

---

**input\_audio\_noise\_reduction** object Optional Defaults to `null`

Configuration for input audio noise reduction. This can be set to `null` to turn off. Noise reduction filters audio added to the input audio buffer before it is sent to VAD and the model. Filtering the audio can improve VAD and turn detection accuracy

(reducing false positives) and model performance by improving perception of the input audio.

> Show properties

---

#### **input\_audio\_transcription** object Optional

Configuration for input audio transcription. The client can optionally set the language and prompt for transcription, these offer additional guidance to the transcription service.

> Show properties

---

#### **turn\_detection** object Optional

Configuration for turn detection. Can be set to `null` to turn off. Server VAD means that the model will detect the start and end of speech based on audio volume and respond at the end of user speech.

> Show properties

---

## Returns

---

The created [Realtime transcription session object](#), plus an ephemeral key

### Example request

```
curl -X POST https://api.openai.com/v1/realtime/transcription_sessions \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{}'
```

### Response

```
{  
  "id": "sess_BBwZc7cFV3XizEyKGDCGL",  
  "object": "realtime.transcription_session",  
  "modalities": ["audio", "text"],  
  "turn_detection": {  
    "type": "server_vad",  
    "threshold": 0.5,  
    "prefix_padding_ms": 300,  
    "silence_duration_ms": 200  
  },  
  "input_audio_format": "pcm16",  
  "input_audio_transcription": {  
    "model": "gpt-4o-transcribe",  
    "language": null,  
    "prompt": ""  
  },  
  "client_secret": null  
}
```

# The session object



A Realtime session configuration object.

**audio** object

Configuration for input and output audio for the session.

> Show properties

---

**expires\_at** integer

Expiration timestamp for the session, in seconds since epoch.

---

**id** string

Unique identifier for the session that looks like `sess_1234567890abcdef`.

---

**include** array

Additional fields to include in server outputs.

null.

---

**instructions** string

The default system instructions (i.e. system message) prepended to model calls. This field allows the client to guide the model on desired responses. The model can be instructed on response content and format, (e.g. "be extremely succinct", "act friendly", "here are examples of good responses") and on audio behavior (e.g. "talk quickly", "inject emotion into your voice", "laugh frequently"). The instructions are not guaranteed to be followed by the model, but they provide guidance to the model on the desired behavior.

Note that the server sets default instructions which will be used if this field is not set and are visible in the `session.created` event at the start of the session.

---

**max\_output\_tokens** integer or "inf"

Maximum number of output tokens for a single assistant response, inclusive of tool calls. Provide an integer between 1 and 4096 to limit output tokens, or `inf` for the maximum available tokens for a given model. Defaults to `inf`.

---

**model** string

The Realtime model used for this session.

---

**object** string

The object type. Always `realtime.session`.

---

**output\_modalities**

The set of modalities the model can respond with. To disable audio, set this to `["text"]`.

---

**tool\_choice** string

How the model chooses tools. Options are `auto`, `none`, `required`, or specify a function.

---

**tools** array

Tools (functions) available to the model.

> Show properties

---

**tracing** "auto" or object

Configuration options for tracing. Set to null to disable tracing. Once tracing is enabled for a session, the configuration cannot be modified.

`auto` will create a trace for the session with default values for the workflow name, group id, and metadata.

> Show possible types

**turn\_detection** object

Configuration for turn detection. Can be set to `null` to turn off. Server VAD means that the model will detect the start and end of speech based on audio volume and respond at the end of user speech.

> Show properties

**OBJECT** The session object

```
{  
  "id": "sess_001",  
  "object": "realtime.session",  
  "expires_at": 1742188264,  
  "model": "gpt-realtime",  
  "output_modalities": ["audio"],  
  "instructions": "You are a friendly assistant.",  
  "tools": [],  
  "tool_choice": "none",  
  "max_output_tokens": "inf",  
  "tracing": "auto",  
  "truncation": "auto",  
  "prompt": null,  
  "audio": {  
    "input": {  
      "format": {  
        "type": "audio/pcm",  
        "rate": 24000  
      },  
      "transcription": { "model": "whisper-1" },  
      "noise_reduction": null,  
      "volume_normalization": null  
    }  
  }  
}
```

```
        "turn_detection": null
    },
    "output": {
        "format": {
            "type": "audio/pcm",
            "rate": 24000
        },
        "voice": "alloy",
        "speed": 1.0
    }
}
```

# The transcription session object



A new Realtime transcription session configuration.

When a session is created on the server via REST API, the session object also contains an ephemeral key.

Default TTL for keys is 10 minutes. This property is not present when a session is updated via the WebSocket API.

## `client_secret` object

Ephemeral key returned by the API. Only present when the session is created on the server via REST API.

> Show properties

### **input\_audio\_format** string

The format of input audio. Options are `pcm16`, `g711_ulaw`, or `g711_alaw`.

### **input\_audio\_transcription** object

Configuration of the transcription model.

> Show properties

### **modalities**

The set of modalities the model can respond with. To disable audio, set this to `["text"]`.

### **turn\_detection** object

Configuration for turn detection. Can be set to `null` to turn off. Server VAD means that the model will detect the start and end of speech based on audio volume and respond at the end of user speech.

> Show properties

OBJECT The transcription session object

```
{  
  "id": "sess_BBwZc7cFV3XizEyKGDCGL",  
  "object": "realtime.transcription_session",  
  "expires_at": 1742188264,  
  "modalities": ["audio", "text"],  
  "turn_detection": {  
    "type": "server_vad",  
    "threshold": 0.5,  
  },  
}
```

```
"prefix_padding_ms": 300,  
"silence_duration_ms": 200  
},  
"input_audio_format": "pcm16",  
"input_audio_transcription": {  
    "model": "gpt-4o-transcribe",  
    "language": null,  
    "prompt": ""  
},  
"client_secret": null  
}
```

## Realtime Beta client events



These are events that the OpenAI Realtime WebSocket server will accept from the client.



# session.update



Send this event to update the session's default configuration. The client may send this event at any time to update any field, except for `voice`. However, note that once a session has been initialized with a particular `model`, it can't be changed to another model using `session.update`.

When the server receives a `session.update`, it will respond with a `session.updated` event showing the full, effective configuration. Only the fields that are present are updated. To clear a field like `instructions`, pass an empty string.

---

## `event_id` string

Optional client-generated ID used to identify this event.

---

## `session` object

A new Realtime session configuration, with an ephemeral key. Default TTL for keys is one minute.

> Show properties

---

## `type` string

The event type, must be `session.update`.

```
OBJECT session.update
```

```
{
```

```
  "type": "session.update",
```

```
"session": {  
    "tools": [  
        {  
            "type": "function",  
            "name": "display_color_palette",  
            "description": "\nCall this function when a user asks for a color palette.\n",  
            "parameters": {  
                "type": "object",  
                "strict": true,  
                "properties": {  
                    "theme": {  
                        "type": "string",  
                        "description": "Description of the theme for the color scheme."  
                    },  
                    "colors": {  
                        "type": "array",  
                        "description": "Array of five hex color codes based on the theme.",  
                        "items": {  
                            "type": "string",  
                            "description": "Hex color code"  
                        }  
                    }  
                },  
                "required": [  
                    "theme",  
                    "colors"  
                ]  
            }  
        }  
    ]  
}
```

```
    ],
    "tool_choice": "auto"
},
"event_id": "5fc543c4-f59c-420f-8fb9-68c45d1546a7",
"timestamp": "2:30:32 PM"
}
```



## input\_audio\_buffer.append



Send this event to append audio bytes to the input audio buffer. The audio buffer is temporary storage you can write to and later commit. In Server VAD mode, the audio buffer is used to detect speech and the server will decide when to commit. When Server VAD is disabled, you must commit the audio buffer manually.

The client may choose how much audio to place in each event up to a maximum of 15 MiB, for example streaming smaller chunks from the client may allow the VAD to be more responsive. Unlike made other client events, the server will not send a confirmation response to this event.

**audio** string

Base64-encoded audio bytes. This must be in the format specified by the `input_audio_format` field in the session configuration.

**event\_id** string

Optional client-generated ID used to identify this event.

**type** string

The event type, must be `input_audio_buffer.append`.

OBJECT `input_audio_buffer.append`

```
{  
  "event_id": "event_456",  
  "type": "input_audio_buffer.append",  
  "audio": "Base64EncodedAudioData"  
}
```

## input\_audio\_buffer.commit



Send this event to commit the user input audio buffer, which will create a new user message item in the conversation. This event will produce an error if the input audio buffer is empty. When in Server VAD mode, the client does not need to send this event, the server will commit the audio buffer automatically.

Committing the input audio buffer will trigger input audio transcription (if enabled in session configuration), but it will not create a response from the model. The server will respond with an `input_audio_buffer.committed` event.

**event\_id** string

Optional client-generated ID used to identify this event.

**type** string

The event type, must be `input_audio_buffer.commit`.

```
OBJECT input_audio_buffer.commit
```

```
{  
  "event_id": "event_789",  
  "type": "input_audio_buffer.commit"  
}
```

## input\_audio\_buffer.clear



Send this event to clear the audio bytes in the buffer. The server will respond with an

`input_audio_buffer.cleared` event.

**event\_id** string

Optional client-generated ID used to identify this event.

**type** string

The event type, must be `input_audio_buffer.clear`.

```
OBJECT input_audio_buffer.clear
```

```
{  
  "event_id": "event_012",  
  "type": "input_audio_buffer.clear"  
}
```



## conversation.item.create



Add a new Item to the Conversation's context, including messages, function calls, and function call responses. This event can be used both to populate a "history" of the conversation and to add new items mid-stream, but

has the current limitation that it cannot populate assistant audio messages.

If successful, the server will respond with a `conversation.item.created` event, otherwise an `error` event will be sent.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**item** object

A single item within a Realtime conversation.

> Show possible types

---

**previous\_item\_id** string

The ID of the preceding item after which the new item will be inserted. If not set, the new item will be appended to the end of the conversation. If set to `root`, the new item will be added to the beginning of the conversation. If set to an existing ID, it allows an item to be inserted mid-conversation. If the ID cannot be found, an error will be returned and the item will not be added.

---

**type** string

The event type, must be `conversation.item.create`.

```
OBJECT conversation.item.create
```

```
{  
  "type": "conversation.item.create",  
  "item": {
```

```
"type": "message",
"role": "user",
"content": [
  {
    "type": "input_text",
    "text": "hi"
  }
],
"event_id": "b904fba0-0ec4-40af-8bbb-f908a9b26793",
}
```

## conversation.item.retrieve



Send this event when you want to retrieve the server's representation of a specific item in the conversation history. This is useful, for example, to inspect user audio after noise cancellation and VAD. The server will respond with a `conversation.item.retrieved` event, unless the item does not exist in the conversation history, in which case the server will respond with an error.

### `event_id` string

Optional client-generated ID used to identify this event.

**item\_id** string

The ID of the item to retrieve.

**type** string

The event type, must be `conversation.item.retrieve`.

OBJECT `conversation.item.retrieve`

```
{  
  "event_id": "event_901",  
  "type": "conversation.item.retrieve",  
  "item_id": "msg_003"  
}
```

## conversation.item.truncate



Send this event to truncate a previous assistant message's audio. The server will produce audio faster than realtime, so this event is useful when the user interrupts to truncate audio that has already been sent to the client but not yet played. This will synchronize the server's understanding of the audio with the client's playback.

Truncating audio will delete the server-side text transcript to ensure there is not text in the context that hasn't been heard by the user.

If successful, the server will respond with a `conversation.item.truncated` event.

---

**audio\_end\_ms** integer

Inclusive duration up to which audio is truncated, in milliseconds. If the `audio_end_ms` is greater than the actual audio duration, the server will respond with an error.

---

**content\_index** integer

The index of the content part to truncate. Set this to 0.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**item\_id** string

The ID of the assistant message item to truncate. Only assistant message items can be truncated.

---

**type** string

The event type, must be `conversation.item.truncate`.

---

OBJECT `conversation.item.truncate`

{

`"event_id": "event_678",`  
`"type": "conversation.item.truncate",`

```
"item_id": "msg_002",
"content_index": 0,
"audio_end_ms": 1500
}
```

# conversation.item.delete



Send this event when you want to remove any item from the conversation history. The server will respond with a `conversation.item.deleted` event, unless the item does not exist in the conversation history, in which case the server will respond with an error.

**event\_id** string

Optional client-generated ID used to identify this event.

**item\_id** string

The ID of the item to delete.

**type** string

The event type, must be `conversation.item.delete`.

OBJECT `conversation.item.delete`

```
{  
  "event_id": "event_901",  
  "type": "conversation.item.delete",  
  "item_id": "msg_003"  
}
```



## response.create



This event instructs the server to create a Response, which means triggering model inference. When in Server VAD mode, the server will create Responses automatically.

A Response will include at least one Item, and may have two, in which case the second will be a function call. These Items will be appended to the conversation history.

The server will respond with a `response.created` event, events for Items and content created, and finally a `response.done` event to indicate the Response is complete.

The `response.create` event can optionally include inference configuration like `instructions`, and `temperature`. These fields will override the Session's configuration for this Response only.

Responses can be created out-of-band of the default Conversation, meaning that they can have arbitrary input, and it's possible to disable writing the output to the Conversation. Only one Response can write to the default Conversation at a time, but otherwise multiple Responses can be created in parallel.

Clients can set `conversation` to `none` to create a Response that does not write to the default Conversation. Arbitrary input can be provided with the `input` field, which is an array accepting raw Items and references to existing Items.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**response** object

Create a new Realtime response with these parameters

> Show properties

---

**type** string

The event type, must be `response.create`.

OBJECT `response.create`

```
// Trigger a response with the default Conversation and no special parameters
{
  "type": "response.create",
```

```
}

// Trigger an out-of-band response that does not write to the default Conversation
{
  "type": "response.create",
  "response": {
    "instructions": "Provide a concise answer.",
    "tools": [], // clear any session tools
    "conversation": "none",
    "output_modalities": ["text"],
    "input": [
      {
        "type": "item_reference",
        "id": "item_12345",
      },
      {
        "type": "message",
        "role": "user",
        "content": [
          {
            "type": "input_text",
            "text": "Summarize the above message in one sentence."
          }
        ]
      }
    ],
  }
}
```

# response.cancel



Send this event to cancel an in-progress response. The server will respond with a `response.done` event with a status of `response.status=cancelled`. If there is no response to cancel, the server will respond with an error.

**event\_id** string

Optional client-generated ID used to identify this event.

**response\_id** string

A specific response ID to cancel - if not provided, will cancel an in-progress response in the default conversation.

**type** string

The event type, must be `response.cancel`.

OBJECT `response.cancel`

```
{  
  "event_id": "event_567",  
  "type": "response.cancel"  
}
```



## transcription\_session.update



Send this event to update a transcription session.

---

**event\_id** string

Optional client-generated ID used to identify this event.

---

**session** object

Realtime transcription session object configuration.

> Show properties

---

**type** string

The event type, must be `transcription_session.update`.

OBJECT `transcription_session.update`

```
{  
  "type": "transcription_session.update",  
  "session": {  
    "input_audio_format": "pcm16",  
    "input_audio_transcription": {  
      "model": "gpt-4o-transcribe",  
      "prompt": "",  
      "language": ""  
    },  
    "turn_detection": {  
      "type": "server_vad",  
      "threshold": 0.5,  
      "prefix_padding_ms": 300,  
      "silence_duration_ms": 500,  
      "create_response": true,  
    },  
    "input_audio_noise_reduction": {  
      "type": "near_field"  
    },  
    "include": [  
      "item.input_audio_transcription.logprobs",  
    ]  
  }  
}
```



# output\_audio\_buffer.clear



**WebRTC Only:** Emit to cut off the current audio response. This will trigger the server to stop generating audio and emit a `output_audio_buffer.cleared` event. This event should be preceded by a `response.cancel` client event to stop the generation of the current response. [Learn more.](#)

**event\_id** string

The unique ID of the client event used for error handling.

**type** string

The event type, must be `output_audio_buffer.clear`.

OBJECT `output_audio_buffer.clear`

```
{  
  "event_id": "optional_client_event_id",  
  "type": "output_audio_buffer.clear"  
}
```

## Realtime Beta server events



These are events emitted from the OpenAI Realtime WebSocket server to the client.

### error



Returned when an error occurs, which could be a client problem or a server problem. Most errors are recoverable and the session will stay open, we recommend to implementors to monitor and log error messages by default.

#### error object

Details of the error.

> Show properties

**event\_id** string

The unique ID of the server event.

**type** string

The event type, must be `error`.

OBJECT error

```
{  
    "event_id": "event_890",  
    "type": "error",  
    "error": {  
        "type": "invalid_request_error",  
        "code": "invalid_event",  
        "message": "The 'type' field is missing.",  
        "param": null,  
        "event_id": "event_567"  
    }  
}
```



# session.created



Returned when a Session is created. Emitted automatically when a new connection is established as the first server event. This event will contain the default Session configuration.

---

**event\_id** string

The unique ID of the server event.

---

**session** object

Realtime session object for the beta interface.

> Show properties

---

**type** string

The event type, must be `session.created`.

OBJECT `session.created`

```
{  
  "type": "session.created",  
  "event_id": "event_C9G5RJeJ2gF77mV7f2B1j",  
  "session": {  
    "object": "realtime.session",  
    "id": "sess_C9G5QPteg4UIbotdKLoYQ",  
    "model": "gpt-realtime-2025-08-28",  
    "modalities": [  
      "text",  
      "audio",  
      "image",  
      "video",  
      "file"  
    ]  
  }  
}
```

```
        "audio"
    ],
    "instructions": "Your knowledge cutoff is 2023-10. You are a helpful, witty, and friendly AI. /",
    "tools": [],
    "tool_choice": "auto",
    "max_response_output_tokens": "inf",
    "tracing": null,
    "prompt": null,
    "expires_at": 1756324625,
    "input_audio_format": "pcm16",
    "input_audio_transcription": null,
    "turn_detection": {
        "type": "server_vad",
        "threshold": 0.5,
        "prefix_padding_ms": 300,
        "silence_duration_ms": 200,
        "idle_timeout_ms": null,
        "create_response": true,
        "interrupt_response": true
    },
    "output_audio_format": "pcm16",
    "voice": "marin",
    "include": null
}
```

# session.updated



Returned when a session is updated with a `session.update` event, unless there is an error.

**event\_id** string

The unique ID of the server event.

**session** object

Realtime session object for the beta interface.

> Show properties

**type** string

The event type, must be `session.updated`.

OBJECT `session.updated`

```
{  
  "event_id": "event_5678",  
  "type": "session.updated",  
  "session": {  
    "id": "sess_001",  
    "object": "realtime.session",  
    "model": "gpt-realtime",  
    "modalities": ["text"],  
    "instructions": "New instructions",  
  },  
}
```

```
"voice": "sage",
"input_audio_format": "pcm16",
"output_audio_format": "pcm16",
"input_audio_transcription": {
    "model": "whisper-1"
},
"turn_detection": null,
"tools": [],
"tool_choice": "none",
"temperature": 0.7,
"max_response_output_tokens": 200,
"speed": 1.1,
"tracing": "auto"
}
```



## transcription\_session.created



Returned when a transcription session is created.

**event\_id** string

The unique ID of the server event.

### session object

A new Realtime transcription session configuration.

When a session is created on the server via REST API, the session object also contains an ephemeral key. Default TTL for keys is 10 minutes. This property is not present when a session is updated via the WebSocket API.

> Show properties

### type string

The event type, must be `transcription_session.created`.

```
OBJECT transcription_session.created

{
  "event_id": "event_5566",
  "type": "transcription_session.created",
  "session": {
    "id": "sess_001",
    "object": "realtime.transcription_session",
    "input_audio_format": "pcm16",
    "input_audio_transcription": {
      "model": "gpt-4o-transcribe",
      "prompt": "",
      "language": ""
    },
    "turn_detection": {
      "type": "server_vad",
    }
}
```

```
        "threshold": 0.5,  
        "prefix_padding_ms": 300,  
        "silence_duration_ms": 500  
    },  
    "input_audio_noise_reduction": {  
        "type": "near_field"  
    },  
    "include": []  
}  
}
```

## transcription\_session.updated



Returned when a transcription session is updated with a `transcription_session.update` event, unless there is an error.

**event\_id** string

The unique ID of the server event.

**session** object

A new Realtime transcription session configuration.

When a session is created on the server via REST API, the session object also contains an ephemeral key. Default TTL for keys is 10 minutes. This property is not present when a session is updated via the WebSocket API.

> Show properties

**type** string

The event type, must be `transcription_session.updated`.

OBJECT `transcription_session.updated`

```
{  
  "event_id": "event_5678",  
  "type": "transcription_session.updated",  
  "session": {  
    "id": "sess_001",  
    "object": "realtime.transcription_session",  
    "input_audio_format": "pcm16",  
    "input_audio_transcription": {  
      "model": "gpt-4o-transcribe",  
      "prompt": "",  
      "language": ""  
    },  
    "turn_detection": {  
      "type": "server_vad",  
      "threshold": 0.5,  
      "prefix_padding_ms": 300,  
      "silence_duration_ms": 500,  
      "create_response": true,  
      // "interrupt_response": false -- this will NOT be returned  
    },  
  },  
}
```

```
"input_audio_noise_reduction": {  
    "type": "near_field"  
,  
    "include": [  
        "item.input_audio_transcription.avg_logprob",  
    ],  
}  
}
```



## conversation.item.created



Returned when a conversation item is created. There are several scenarios that produce this event:

The server is generating a Response, which if successful will produce either one or two Items, which will be of type `message` (role `assistant`) or type `function_call`.

The input audio buffer has been committed, either by the client or the server (in `server_vad` mode). The server will take the content of the input audio buffer and add it to a new user message Item.

The client has sent a `conversation.item.create` event to add a new Item to the Conversation.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**previous\_item\_id** string

The ID of the preceding item in the Conversation context, allows the client to understand the order of the conversation. Can be

`null` if the item has no predecessor.

**type** string

The event type, must be `conversation.item.created`.

OBJECT `conversation.item.created`

```
{  
  "event_id": "event_1920",  
  "type": "conversation.item.created",  
  "previous_item_id": "msg_002",  
  "item": {  
    "id": "msg_003",  
    "object": "realtime.item",  
    "type": "message",  
    "status": "completed",  
  },  
}
```

```
        "role": "user",
        "content": []
    }
}
```

## conversation.item.retrieved



Returned when a conversation item is retrieved with `conversation.item.retrieve`.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**type** string

The event type, must be `conversation.item.retrieved`.

OBJECT `conversation.item.retrieved`

```
{  
  "event_id": "event_1920",  
  "type": "conversation.item.created",  
  "previous_item_id": "msg_002",  
  "item": {  
    "id": "msg_003",  
    "object": "realtime.item",  
    "type": "message",  
    "status": "completed",  
    "role": "user",  
    "content": [  
      {  
        "type": "input_audio",  
        "transcript": "hello how are you",  
        "audio": "base64encodedaudio=="  
      }  
    ]  
  }  
}
```



# conversation.item.input\_audio\_transcription.complete

## d



This event is the output of audio transcription for user audio written to the user audio buffer. Transcription begins when the input audio buffer is committed by the client or server (in `server_vad` mode). Transcription runs asynchronously with Response creation, so this event may come before or after the Response events.

Realtime API models accept audio natively, and thus input transcription is a separate process run on a separate ASR (Automatic Speech Recognition) model. The transcript may diverge somewhat from the model's interpretation, and should be treated as a rough guide.

---

**content\_index** integer

The index of the content part containing the audio.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the user message item containing the audio.

---

**logprobs** array

The log probabilities of the transcription.

> Show properties

**transcript** string

The transcribed text.

**type** string

The event type, must be `conversation.item.input_audio_transcription.completed`.

**usage** object

Usage statistics for the transcription.

> Show possible types

OBJECT `conversation.item.input_audio_transcription.completed`

```
{  
  "event_id": "event_2122",  
  "type": "conversation.item.input_audio_transcription.completed",  
  "item_id": "msg_003",  
  "content_index": 0,  
  "transcript": "Hello, how are you?",  
  "usage": {  
    "type": "tokens",  
    "total_tokens": 48,  
    "input_tokens": 38,  
    "input_token_details": {  
      "text_tokens": 10,  
      "audio_tokens": 28,  
    },  
  },  
}
```

```
        "output_tokens": 10,  
    }  
}
```

## conversation.item.input\_audio\_transcription.delta



Returned when the text value of an input audio transcription content part is updated.

**content\_index** integer

The index of the content part in the item's content array.

**delta** string

The text delta.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**logprobs** array

The log probabilities of the transcription.

> Show properties

**type** string

The event type, must be `conversation.item.input_audio_transcription.delta`.

OBJECT `conversation.item.input_audio_transcription.delta`

```
{  
  "type": "conversation.item.input_audio_transcription.delta",  
  "event_id": "event_001",  
  "item_id": "item_001",  
  "content_index": 0,  
  "delta": "Hello"  
}
```

## **conversation.item.input\_audio\_transcription.segment**

🔗

Returned when an input audio transcription segment is identified for an item.

**content\_index** integer

The index of the input audio content part within the item.

**end** number

End time of the segment in seconds.

**event\_id** string

The unique ID of the server event.

**id** string

The segment identifier.

**item\_id** string

The ID of the item containing the input audio content.

**speaker** string

The detected speaker label for this segment.

**start** number

Start time of the segment in seconds.

**text** string

The text for this segment.

**type** string

The event type, must be `conversation.item.input_audio_transcription.segment`.

OBJECT conversation.item.input\_audio\_transcription.segment

```
{  
    "event_id": "event_6501",  
    "type": "conversation.item.input_audio_transcription.segment",  
    "item_id": "msg_011",  
    "content_index": 0,  
    "text": "hello",  
    "id": "seg_0001",  
    "speaker": "spk_1",  
    "start": 0.0,  
    "end": 0.4  
}
```

## conversation.item.input\_audio\_transcription.failed



Returned when input audio transcription is configured, and a transcription request for a user message failed.

These events are separate from other `error` events so that the client can identify the related Item.

`content_index` integer

The index of the content part containing the audio.

**error** object

Details of the transcription error.

> Show properties

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the user message item.

**type** string

The event type, must be `conversation.item.input_audio_transcription.failed`.

OBJECT `conversation.item.input_audio_transcription.failed`

```
{  
  "event_id": "event_2324",  
  "type": "conversation.item.input_audio_transcription.failed",  
  "item_id": "msg_003",  
  "content_index": 0,  
  "error": {  
    "type": "transcription_error",  
    "code": "audio_unintelligible",  
    "message": "The audio could not be transcribed.",  
    "param": null  
  }  
}
```

# conversation.item.truncated



Returned when an earlier assistant audio message item is truncated by the client with a `conversation.item.truncate` event. This event is used to synchronize the server's understanding of the audio with the client's playback.

This action will truncate the audio and remove the server-side text transcript to ensure there is no text in the context that hasn't been heard by the user.

---

## `audio_end_ms` integer

The duration up to which the audio was truncated, in milliseconds.

---

## `content_index` integer

The index of the content part that was truncated.

---

## `event_id` string

The unique ID of the server event.

---

## `item_id` string

The ID of the assistant message item that was truncated.

**type** string

The event type, must be `conversation.item.truncated`.

OBJECT `conversation.item.truncated`

```
{  
  "event_id": "event_2526",  
  "type": "conversation.item.truncated",  
  "item_id": "msg_004",  
  "content_index": 0,  
  "audio_end_ms": 1500  
}
```

## conversation.item.deleted



Returned when an item in the conversation is deleted by the client with a `conversation.item.delete` event.

This event is used to synchronize the server's understanding of the conversation history with the client's view.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item that was deleted.

**type** string

The event type, must be `conversation.item.deleted`.

OBJECT `conversation.item.deleted`

```
{  
  "event_id": "event_2728",  
  "type": "conversation.item.deleted",  
  "item_id": "msg_005"  
}
```



## input\_audio\_buffer.committed



Returned when an input audio buffer is committed, either by the client or automatically in server VAD mode.

The `item_id` property is the ID of the user message item that will be created, thus a `conversation.item.created` event will also be sent to the client.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the user message item that will be created.

---

**previous\_item\_id** string

The ID of the preceding item after which the new item will be inserted. Can be `null` if the item has no predecessor.

---

**type** string

The event type, must be `input_audio_buffer.committed`.

OBJECT `input_audio_buffer.committed`

```
{  
  "event_id": "event_1121",  
  "type": "input_audio_buffer.committed",  
  "previous_item_id": "msg_001",  
  "item_id": "msg_002"  
}
```

## input\_audio\_buffer.cleared



Returned when the input audio buffer is cleared by the client with a `input_audio_buffer.clear` event.

**event\_id** string

The unique ID of the server event.

**type** string

The event type, must be `input_audio_buffer.cleared`.

OBJECT `input_audio_buffer.cleared`

```
{  
  "event_id": "event_1314",
```

```
"type": "input_audio_buffer.cleared"
```

## input\_audio\_buffer.speech\_started



Sent by the server when in `server_vad` mode to indicate that speech has been detected in the audio buffer. This can happen any time audio is added to the buffer (unless speech is already detected). The client may want to use this event to interrupt audio playback or provide visual feedback to the user.

The client should expect to receive a `input_audio_buffer.speech_stopped` event when speech stops. The `item_id` property is the ID of the user message item that will be created when speech stops and will also be included in the `input_audio_buffer.speech_stopped` event (unless the client manually commits the audio buffer during VAD activation).

---

### `audio_start_ms` integer

Milliseconds from the start of all audio written to the buffer during the session when speech was first detected. This will correspond to the beginning of audio sent to the model, and thus includes the `prefix_padding_ms` configured in the Session.

---

### `event_id` string

The unique ID of the server event.

**item\_id** string

The ID of the user message item that will be created when speech stops.

**type** string

The event type, must be `input_audio_buffer.speech_started`.

OBJECT `input_audio_buffer.speech_started`

```
{  
  "event_id": "event_1516",  
  "type": "input_audio_buffer.speech_started",  
  "audio_start_ms": 1000,  
  "item_id": "msg_003"  
}
```

## input\_audio\_buffer.speech\_stopped



Returned in `server_vad` mode when the server detects the end of speech in the audio buffer. The server will also send an `conversation.item.created` event with the user message item that is created from the audio buffer.

**audio\_end\_ms** integer

Milliseconds since the session started when speech stopped. This will correspond to the end of audio sent to the model, and thus includes the `min_silence_duration_ms` configured in the Session.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the user message item that will be created.

**type** string

The event type, must be `input_audio_buffer.speech_stopped`.

OBJECT `input_audio_buffer.speech_stopped`

```
{  
  "event_id": "event_1718",  
  "type": "input_audio_buffer.speech_stopped",  
  "audio_end_ms": 2000,  
  "item_id": "msg_003"  
}
```

# input\_audio\_buffer.timeout\_triggered



Returned when the Server VAD timeout is triggered for the input audio buffer. This is configured with `idle_timeout_ms` in the `turn_detection` settings of the session, and it indicates that there hasn't been any speech detected for the configured duration.

The `audio_start_ms` and `audio_end_ms` fields indicate the segment of audio after the last model response up to the triggering time, as an offset from the beginning of audio written to the input audio buffer. This means it demarcates the segment of audio that was silent and the difference between the start and end values will roughly match the configured timeout.

The empty audio will be committed to the conversation as an `input_audio` item (there will be a `input_audio_buffer.committed` event) and a model response will be generated. There may be speech that didn't trigger VAD but is still detected by the model, so the model may respond with something relevant to the conversation or a prompt to continue speaking.

---

## `audio_end_ms` integer

Millisecond offset of audio written to the input audio buffer at the time the timeout was triggered.

---

## `audio_start_ms` integer

Millisecond offset of audio written to the input audio buffer that was after the playback time of the last model response.

---

## `event_id` string

The unique ID of the server event.

**item\_id** string

The ID of the item associated with this segment.

**type** string

The event type, must be `input_audio_buffer.timeout_triggered`.

OBJECT `input_audio_buffer.timeout_triggered`

```
{  
  "type": "input_audio_buffer.timeout_triggered",  
  "event_id": "event_CEKKrf1KTGvemCPyiJTJ2",  
  "audio_start_ms": 13216,  
  "audio_end_ms": 19232,  
  "item_id": "item_CEKKrWH0GiwN0ET97NUZc"  
}
```



## response.created



Returned when a new Response is created. The first event of response creation, where the response is in an initial state of `in_progress`.

**event\_id** string

The unique ID of the server event.

**response** object

The response resource.

> Show properties

**type** string

The event type, must be `response.created`.

OBJECT `response.created`

```
{  
  "type": "response.created",  
  "event_id": "event_C9G8pqbTEddBSIxBN6Os",  
  "response": {  
    "object": "realtime.response",  
    "id": "resp_C9G8p7IH2WxLbkgPNouYL",  
    "status": "in_progress",  
    "status_details": null,  
    "output": [],  
    "conversation_id": "conv_C9G8mmBkLhQJwCon3hoJN",  
    "output_modalities": [  
      "audio"  
    ]  
  }  
}
```

```
],
  "max_output_tokens": "inf",
  "audio": {
    "output": {
      "format": {
        "type": "audio/pcm",
        "rate": 24000
      },
      "voice": "marin"
    }
  },
  "usage": null,
  "metadata": null
},
"timestamp": "2:30:35 PM"
}
```

## response.done



Returned when a Response is done streaming. Always emitted, no matter the final state. The Response object included in the `response.done` event will include all output Items in the Response but will omit the raw audio data.

**event\_id** string

The unique ID of the server event.

**response** object

The response resource.

> Show properties

**type** string

The event type, must be `response.done`.

OBJECT `response.done`

```
{  
  "event_id": "event_3132",  
  "type": "response.done",  
  "response": {  
    "id": "resp_001",  
    "object": "realtime.response",  
    "status": "completed",  
    "status_details": null,  
    "output": [  
      {  
        "id": "msg_006",  
        "object": "realtime.item",  
        "type": "message",  
        "status": "completed",  
        "role": "assistant",  
        "content": "Hello! How can I assist you today?"  
      }  
    ]  
  }  
}
```

```
        "content": [
            {
                "type": "text",
                "text": "Sure, how can I assist you today?"
            }
        ],
        "usage": {
            "total_tokens": 275,
            "input_tokens": 127,
            "output_tokens": 148,
            "input_token_details": {
                "cached_tokens": 384,
                "text_tokens": 119,
                "audio_tokens": 8,
                "cached_tokens_details": {
                    "text_tokens": 128,
                    "audio_tokens": 256
                }
            },
            "output_token_details": {
                "text_tokens": 36,
                "audio_tokens": 112
            }
        }
    }
}
```



# response.output\_item.added



Returned when a new Item is created during Response generation.

---

## **event\_id** string

The unique ID of the server event.

---

## **item** object

A single item within a Realtime conversation.

> Show possible types

---

## **output\_index** integer

The index of the output item in the Response.

---

## **response\_id** string

The ID of the Response to which the item belongs.

**type** string

The event type, must be `response.output_item.added`.

OBJECT `response.output_item.added`

```
{  
    "event_id": "event_3334",  
    "type": "response.output_item.added",  
    "response_id": "resp_001",  
    "output_index": 0,  
    "item": {  
        "id": "msg_007",  
        "object": "realtime.item",  
        "type": "message",  
        "status": "in_progress",  
        "role": "assistant",  
        "content": []  
    }  
}
```

## response.output\_item.done



Returned when an Item is done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

**event\_id** string

The unique ID of the server event.

**item** object

A single item within a Realtime conversation.

> Show possible types

**output\_index** integer

The index of the output item in the Response.

**response\_id** string

The ID of the Response to which the item belongs.

**type** string

The event type, must be `response.output_item.done`.

OBJECT `response.output_item.done`

```
{  
  "event_id": "event_3536",  
  "type": "response.output_item.done",  
  "response_id": "resp_001",
```

```
"output_index": 0,  
"item": {  
    "id": "msg_007",  
    "object": "realtime.item",  
    "type": "message",  
    "status": "completed",  
    "role": "assistant",  
    "content": [  
        {  
            "type": "text",  
            "text": "Sure, I can help with that."  
        }  
    ]  
}
```



## response.content\_part.added



Returned when a new content part is added to an assistant message item during response generation.

**content\_index** integer

The index of the content part in the item's content array.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item to which the content part was added.

**output\_index** integer

The index of the output item in the response.

**part** object

The content part that was added.

> Show properties

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.content_part.added`.

OBJECT `response.content_part.added`

{

`"event_id": "event_3738",`  
  `"type": "response.content_part.added",`

```
"response_id": "resp_001",
"item_id": "msg_007",
"output_index": 0,
"content_index": 0,
"part": {
    "type": "text",
    "text": ""
}
}
```

## response.content\_part.done



Returned when a content part is done streaming in an assistant message item. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**part** object

The content part that is done.

> Show properties

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.content_part.done`.

OBJECT `response.content_part.done`

{

```
"event_id": "event_3940",
"type": "response.content_part.done",
"response_id": "resp_001",
"item_id": "msg_007",
"output_index": 0,
"content_index": 0,
"part": {
```

```
        "type": "text",
        "text": "Sure, I can help with that."
    }
}
```



## response.output\_text.delta



Returned when the text value of an "output\_text" content part is updated.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**delta** string

The text delta.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_text.delta`.

OBJECT `response.output_text.delta`

```
{  
    "event_id": "event_4142",  
    "type": "response.output_text.delta",  
    "response_id": "resp_001",  
    "item_id": "msg_007",  
    "output_index": 0,  
    "content_index": 0,  
    "delta": "Sure, I can h"  
}
```

# response.output\_text.done



Returned when the text value of an "output\_text" content part is done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

---

**text** string

The final text content.

**type** string

The event type, must be `response.output_text.done`.

OBJECT `response.output_text.done`

```
{  
    "event_id": "event_4344",  
    "type": "response.output_text.done",  
    "response_id": "resp_001",  
    "item_id": "msg_007",  
    "output_index": 0,  
    "content_index": 0,  
    "text": "Sure, I can help with that."  
}
```



## **response.output\_audio\_transcript.delta**



Returned when the model-generated transcription of audio output is updated.

**content\_index** integer

The index of the content part in the item's content array.

**delta** string

The transcript delta.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_audio_transcript.delta`.

OBJECT `response.output_audio_transcript.delta`

{

`"event_id": "event_4546",`

```
"type": "response.output_audio_transcript.delta",
"response_id": "resp_001",
"item_id": "msg_008",
"output_index": 0,
"content_index": 0,
"delta": "Hello, how can I a"
}
```

## response.output\_audio\_transcript.done



Returned when the model-generated transcription of audio output is done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

### content\_index integer

The index of the content part in the item's content array.

---

### event\_id string

The unique ID of the server event.

---

### item\_id string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**transcript** string

The final transcript of the audio.

**type** string

The event type, must be `response.output_audio_transcript.done`.

OBJECT `response.output_audio_transcript.done`

```
{  
  "event_id": "event_4748",  
  "type": "response.output_audio_transcript.done",  
  "response_id": "resp_001",  
  "item_id": "msg_008",  
  "output_index": 0,  
  "content_index": 0,  
  "transcript": "Hello, how can I assist you today?"  
}
```



# response.output\_audio.delta



Returned when the model-generated audio is updated.

---

**content\_index** integer

The index of the content part in the item's content array.

---

**delta** string

Base64-encoded audio data delta.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_audio.delta`.

OBJECT `response.output_audio.delta`

```
{  
  "event_id": "event_4950",  
  "type": "response.output_audio.delta",  
  "response_id": "resp_001",  
  "item_id": "msg_008",  
  "output_index": 0,  
  "content_index": 0,  
  "delta": "Base64EncodedAudioDelta"  
}
```

## response.output\_audio.done



Returned when the model-generated audio is done. Also emitted when a Response is interrupted, incomplete, or cancelled.

**content\_index** integer

The index of the content part in the item's content array.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.output_audio.done`.

OBJECT `response.output_audio.done`

{

```
"event_id": "event_5152",
"type": "response.output_audio.done",
"response_id": "resp_001",
"item_id": "msg_008",
```

```
"output_index": 0,  
"content_index": 0  
}
```



## response.function\_call\_arguments.delta



Returned when the model-generated function call arguments are updated.

---

**call\_id** string

The ID of the function call.

---

**delta** string

The arguments delta as a JSON string.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the function call item.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.function_call_arguments.delta`.

OBJECT `response.function_call_arguments.delta`

```
{  
  "event_id": "event_5354",  
  "type": "response.function_call_arguments.delta",  
  "response_id": "resp_002",  
  "item_id": "fc_001",  
  "output_index": 0,  
  "call_id": "call_001",  
  "delta": "{\"location\": \"San\\""}  
}
```

# response.function\_call\_arguments.done



Returned when the model-generated function call arguments are done streaming. Also emitted when a Response is interrupted, incomplete, or cancelled.

---

**arguments** string

The final arguments as a JSON string.

---

**call\_id** string

The ID of the function call.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the function call item.

---

**output\_index** integer

The index of the output item in the response.

---

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.function_call_arguments.done`.

OBJECT `response.function_call_arguments.done`

```
{  
    "event_id": "event_5556",  
    "type": "response.function_call_arguments.done",  
    "response_id": "resp_002",  
    "item_id": "fc_001",  
    "output_index": 0,  
    "call_id": "call_001",  
    "arguments": "{\"location\": \"San Francisco\""}  
}
```



## **response.mcp\_call\_arguments.delta**



Returned when MCP tool call arguments are updated during response generation.

**delta** string

The JSON-encoded arguments delta.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP tool call item.

**obfuscation** string

If present, indicates the delta text was obfuscated.

**output\_index** integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.mcp_call_arguments.delta`.

OBJECT `response.mcp_call_arguments.delta`

{

`"event_id": "event_6201",`

```
"type": "response.mcp_call_arguments.delta",
"response_id": "resp_001",
"item_id": "mcp_call_001",
"output_index": 0,
"delta": "{\"partial\":true}"
}
```

## response.mcp\_call\_arguments.done



Returned when MCP tool call arguments are finalized during response generation.

### arguments string

The final JSON-encoded arguments string.

### event\_id string

The unique ID of the server event.

### item\_id string

The ID of the MCP tool call item.

### output\_index integer

The index of the output item in the response.

**response\_id** string

The ID of the response.

**type** string

The event type, must be `response.mcp_call_arguments.done`.

OBJECT `response.mcp_call_arguments.done`

```
{  
  "event_id": "event_6202",  
  "type": "response.mcp_call_arguments.done",  
  "response_id": "resp_001",  
  "item_id": "mcp_call_001",  
  "output_index": 0,  
  "arguments": "{\"q\":\"docs\"}"  
}
```



# response.mcp\_call.in\_progress



Returned when an MCP tool call has started and is in progress.

---

**event\_id** string

The unique ID of the server event.

---

**item\_id** string

The ID of the MCP tool call item.

---

**output\_index** integer

The index of the output item in the response.

---

**type** string

The event type, must be `response.mcp_call.in_progress`.

OBJECT `response.mcp_call.in_progress`

```
{  
  "event_id": "event_6301",  
  "type": "response.mcp_call.in_progress",  
  "output_index": 0,  
  "item_id": "mcp_call_001"  
}
```

## response.mcp\_call.completed



Returned when an MCP tool call has completed successfully.

### event\_id string

The unique ID of the server event.

### item\_id string

The ID of the MCP tool call item.

### output\_index integer

The index of the output item in the response.

**type** string

The event type, must be `response.mcp_call.completed`.

OBJECT `response.mcp_call.completed`

```
{  
  "event_id": "event_6302",  
  "type": "response.mcp_call.completed",  
  "output_index": 0,  
  "item_id": "mcp_call_001"  
}
```

## response.mcp\_call.failed



Returned when an MCP tool call has failed.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP tool call item.

**output\_index** integer

The index of the output item in the response.

**type** string

The event type, must be `response.mcp_call.failed`.

OBJECT `response.mcp_call.failed`

```
{  
  "event_id": "event_6303",  
  "type": "response.mcp_call.failed",  
  "output_index": 0,  
  "item_id": "mcp_call_001"  
}
```



## mcp\_list\_tools.in\_progress



Returned when listing MCP tools is in progress for an item.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP list tools item.

**type** string

The event type, must be `mcp_list_tools.in_progress`.

OBJECT `mcp_list_tools.in_progress`

```
{  
  "event_id": "event_6101",  
  "type": "mcp_list_tools.in_progress",  
  "item_id": "mcp_list_tools_001"  
}
```

## **mcp\_list\_tools.completed**



Returned when listing MCP tools has completed for an item.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP list tools item.

**type** string

The event type, must be `mcp_list_tools.completed`.

OBJECT `mcp_list_tools.completed`

```
{  
  "event_id": "event_6102",  
  "type": "mcp_list_tools.completed",  
  "item_id": "mcp_list_tools_001"  
}
```

## **mcp\_list\_tools.failed**



Returned when listing MCP tools has failed for an item.

**event\_id** string

The unique ID of the server event.

**item\_id** string

The ID of the MCP list tools item.

**type** string

The event type, must be `mcp_list_tools.failed`.

OBJECT `mcp_list_tools.failed`

```
{  
  "event_id": "event_6103",  
  "type": "mcp_list_tools.failed",  
  "item_id": "mcp_list_tools_001"  
}
```



## rate\_limits.updated



Emitted at the beginning of a Response to indicate the updated rate limits. When a Response is created some tokens will be "reserved" for the output tokens, the rate limits shown here reflect that reservation, which is then adjusted accordingly once the Response is completed.

**event\_id** string

The unique ID of the server event.

**rate\_limits** array

List of rate limit information.

> Show properties

**type** string

The event type, must be `rate_limits.updated`.

OBJECT `rate_limits.updated`

```
{  
  "event_id": "event_5758",  
  "type": "rate_limits.updated",  
  "rate_limits": [  
    {  
      "name": "requests",  
      "limit": 1000,  
      "remaining": 999,  
      "reset_seconds": 60  
    },
```

```
{  
    "name": "tokens",  
    "limit": 50000,  
    "remaining": 49950,  
    "reset_seconds": 60  
}  
]  
}
```