- 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. Contact sales to learn more about Priority processing.
- 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 or null Optional Defaults to true

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

`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.

`stream_options` object or null Optional Defaults to null

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

Show properties

`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.

`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.

The two categories of tools you can provide the model are:

```
  "has_more": false
}
```

Streaming

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",
```

```json
    "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.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
```

```json
{
  "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 uni
            "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
        "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

```json
{
  "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 name
    "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
  "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.