OpenAI Platform

Models

Copy page

Flagship models

GPT models

GPT models are fast, versatile, cost-efficient, and customizable.

Use GPT-4o

Use GPT-4o mini

Reasoning models

Reasoning models use chain-of-thought reasoning to excel at complex tasks.

Use o1

Use o3-mini

Model pricing details ›

Models overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT models	Our fast, versatile, high-intelligence flagship models
Reasoning models	Our o-series reasoning models excel at complex, multi-step tasks
GPT-4o Realtime	GPT-4o models capable of realtime text and audio inputs and outputs
GPT-4o Audio	GPT-4o models capable of audio inputs and outputs via REST API
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio

Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
Deprecated	A full list of models that have been deprecated along with the suggested replacement

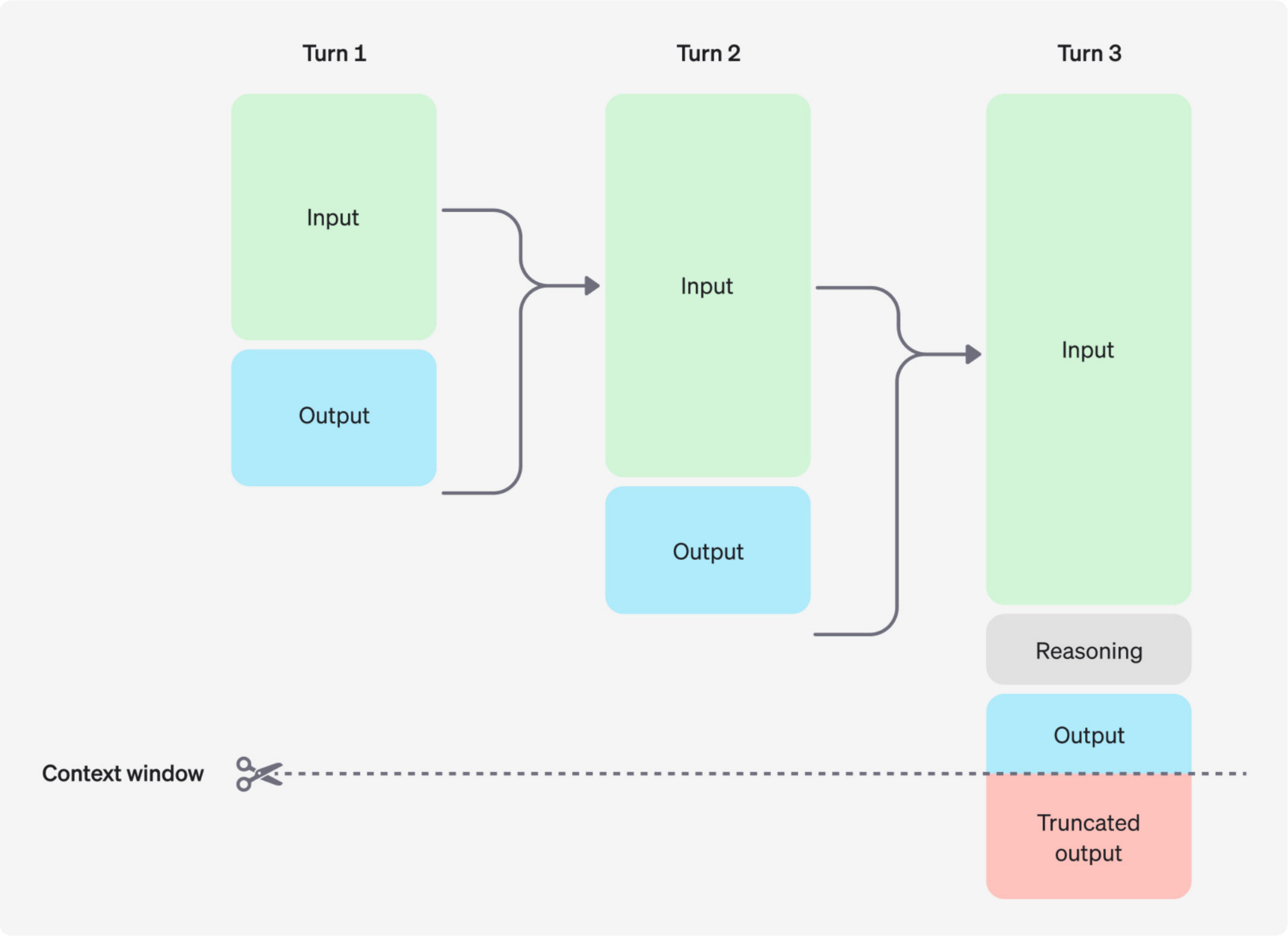
We have also published open source models including Point-E, Whisper, Jukebox, and CLIP.

Context window

Models on this page will list a **context window**, which refers to the maximum number of tokens that can be used in a single request, inclusive of both input, output, and reasoning tokens. For example, when making an API request to chat completions with the o1 model, the following token counts will apply toward the context window total:

- Input tokens (inputs you include in the `messages` array with chat completions)
- Output tokens (tokens generated in response to your prompt)
- Reasoning tokens (used by the model to plan a response)

Tokens generated in excess of the context window limit may be truncated in API responses.



You can estimate the number of tokens your messages will use with the tokenizer tool.

Model ID aliases and snapshots

In the tables below, you will see **model IDs** that can be used in REST APIs like chat completions to generate outputs. Some of these model IDs are **aliases** which point to specific **dated snapshots**.

For example, the `gpt-4o` model ID is an alias that points to a specific dated snapshot of GPT-4o. The dated snapshots that these aliases point to are periodically updated to newer snapshots a few months after a newer snapshot becomes available. Model IDs that are aliases note the model ID they currently point to in the tables below.

API request using a model alias

javascript

```javascript
import OpenAI from "openai";
const openai = new OpenAI();

const completion = await openai.chat.completions.create({
    model: "gpt-4o",
    messages: [
        { role: "developer", content: "You are a helpful assistant." },
        {
            role: "user",
            content: "Write a haiku about recursion in programming.",
        },
    ],
    store: true,
});

console.log(completion.choices[0].message);
```

In API requests where an alias was used as a model ID, the body of the response will contain the actual model ID used to generate the response.

```json
{
  "id": "chatcmpl-Af6LFgbOPpqu2fhGsVktc9xFaYUVh",
  "object": "chat.completion",
  "created": 1734359189,
  "model": "gpt-4o-2024-08-06",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "Code within a loop,  \nFunction calls itself again,  \nInfinite echoes."
        "refusal": null
      },
      "logprobs": null,
      "finish_reason": "stop"
    }
  ],
  "usage": {}
}
```

Current model aliases

Below, please find current model aliases, and guidance on when they will be updated to new versions (if guidance is available).

ALIAS	POINTS TO
`gpt-4o`	`gpt-4o-2024-08-06`
`chatgpt-4o-latest`	Latest used in ChatGPT
`gpt-4o-mini`	`gpt-4o-mini-2024-07-18`
`o1`	`o1-2024-12-17`
`o1-mini`	`o1-mini-2024-09-12`
`o3-mini`	`o3-mini-2025-01-31`
`o1-preview`	`o1-preview-2024-09-12`
`gpt-4o-realtime-preview`	`gpt-4o-realtime-preview-2024-12-17`
`gpt-4o-mini-realtime-preview`	`gpt-4o-mini-realtime-preview-2024-12-17`
`gpt-4o-audio-preview`	`gpt-4o-audio-preview-2024-12-17`

ⓘ In production applications, **it is a best practice to use dated model snapshot IDs** instead of aliases, which may change periodically.

GPT-4o

GPT-4o ("o" for "omni") is our versatile, high-intelligence flagship model. It accepts both text and image inputs, and produces text outputs (including Structured Outputs). Learn how to use GPT-4o in our text generation guide.

The `chatgpt-4o-latest` model ID below continuously points to the version of GPT-4o used in ChatGPT. It is updated frequently, when there are significant changes to ChatGPT's GPT-4o model.

The knowledge cutoff for GPT-4o models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`gpt-4o` ↳ `gpt-4o-2024-08-06`	128,000 tokens	16,384 tokens
`gpt-4o-2024-11-20`	128,000 tokens	16,384 tokens
`gpt-4o-2024-08-06`	128,000 tokens	16,384 tokens
`gpt-4o-2024-05-13`	128,000 tokens	4,096 tokens
`chatgpt-4o-latest` ↳ `GPT-4o used in ChatGPT`	128,000 tokens	16,384 tokens

GPT-4o mini

GPT-4o mini ("o" for "omni") is a fast, affordable small model for focused tasks. It accepts both text and image inputs, and produces text outputs (including Structured Outputs). It is ideal for fine-tuning, and model outputs from a larger model like GPT-4o can be distilled to GPT-4o-mini to produce similar results at lower cost and latency.

The knowledge cutoff for GPT-4o-mini models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`gpt-4o-mini` ↳ `gpt-4o-mini-2024-07-18`	128,000 tokens	16,384 tokens
`gpt-4o-mini-2024-07-18`	128,000 tokens	16,384 tokens

o1 and o1-mini

The **o1 series** of models are trained with reinforcement learning to perform complex reasoning. o1 models think before they answer, producing a long internal chain of thought before responding to the user. Learn about the capabilities of o1 models in our reasoning guide.

The **o1** reasoning model is designed to solve hard problems across domains. **o1-mini** is a faster and more affordable reasoning model, but we recommend using the newer **o3-mini** model that features higher intelligence at the same latency and price as o1-mini.

The latest o1 model supports both text and image inputs, and produces text outputs (including Structured Outputs). o1-mini currently only supports text inputs and outputs.

The knowledge cutoff for o1 and o1-mini models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`o1` ↳ `o1-2024-12-17`	200,000 tokens	100,000 tokens
`o1-2024-12-17`	200,000 tokens	100,000 tokens
`o1-mini` ↳ `o1-mini-2024-09-12`	128,000 tokens	65,536 tokens
`o1-mini-2024-09-12`	128,000 tokens	65,536 tokens
`o1-preview` ↳ `o1-preview-2024-09-12`	128,000 tokens	32,768 tokens
`o1-preview-2024-09-12`	128,000 tokens	32,768 tokens

o3-mini

o3-mini is our most recent small reasoning model, providing high intelligence at the same cost and latency targets of o1-mini. o3-mini also supports key developer features, like Structured Outputs, function calling, Batch API, and more. Like other models in the o-series, it is designed to excel at science, math, and coding tasks.

The knowledge cutoff for o3-mini models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`o3-mini` ↳ `o3-mini-2025-01-31`	200,000 tokens	100,000 tokens
`o3-mini-2025-01-31`	200,000 tokens	100,000 tokens

GPT-4o and GPT-4o-mini Realtime Beta

This is a preview release of the GPT-4o and GPT-4o-mini Realtime models. These models are capable of responding to audio and text inputs in realtime over WebRTC or a WebSocket interface. Learn more in the Realtime API guide.

The knowledge cutoff for GPT-4o Realtime models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`gpt-4o-realtime-preview` ↳ `gpt-4o-realtime-preview-2024-12-17`	128,000 tokens	4,096 tokens
`gpt-4o-realtime-preview-2024-12-17`	128,000 tokens	4,096 tokens
`gpt-4o-realtime-preview-2024-10-01`	128,000 tokens	4,096 tokens
`gpt-4o-mini-realtime-preview` ↳ `gpt-4o-mini-realtime-preview-2024-12-17`	128,000 tokens	4,096 tokens
`gpt-4o-mini-realtime-preview-2024-12-17`	128,000 tokens	4,096 tokens

GPT-4o and GPT-4o-mini Audio Beta

This is a preview release of the GPT-4o Audio models. These models accept audio inputs and outputs, and can be used in the Chat Completions REST API. Learn more.

The knowledge cutoff for GPT-4o Audio models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`gpt-4o-audio-preview` ↳ `gpt-4o-audio-preview-2024-12-17`	128,000 tokens	16,384 tokens
`gpt-4o-audio-preview-2024-12-17`	128,000 tokens	16,384 tokens
`gpt-4o-audio-preview-2024-10-01`	128,000 tokens	16,384 tokens
`gpt-4o-mini-audio-preview` ↳ `gpt-4o-mini-audio-preview-2024-12-17`	128,000 tokens	16,384 tokens
`gpt-4o-mini-audio-preview-2024-12-17`	128,000 tokens	16,384 tokens

GPT-4 Turbo and GPT-4

GPT-4 is an older version of a high-intelligence GPT model, usable in Chat Completions. Learn more in the text generation guide. The knowledge cutoff for the latest GPT-4 Turbo version is **December, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
`gpt-4-turbo` ↳ `gpt-4-turbo-2024-04-09`	128,000 tokens	4,096 tokens
`gpt-4-turbo-2024-04-09`	128,000 tokens	4,096 tokens
`gpt-4-turbo-preview` ↳ `gpt-4-0125-preview`	128,000 tokens	4,096 tokens
`gpt-4-0125-preview`	128,000 tokens	4,096 tokens
`gpt-4-1106-preview`	128,000 tokens	4,096 tokens
`gpt-4` ↳ `gpt-4-0613`	8,192 tokens	8,192 tokens
`gpt-4-0613`	8,192 tokens	8,192 tokens
`gpt-4-0314`	8,192 tokens	8,192 tokens

GPT-3.5 Turbo

GPT-3.5 Turbo models can understand and generate natural language or code and have been optimized for chat using the Chat Completions API but work well for non-chat tasks as well.

ⓘ As of July 2024, `gpt-4o-mini` should be used in place of `gpt-3.5-turbo`, as it is cheaper, more capable, multimodal, and just as fast. `gpt-3.5-turbo` is still available for use in the API.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS	KNOWLEDGE CUTOFF
`gpt-3.5-turbo-0125` The latest GPT-3.5 Turbo model with higher accuracy at responding in requested formats and a fix for a bug which caused a text encoding issue for non-English language function calls. Learn more.	16,385 tokens	4,096 tokens	Sep 2021
`gpt-3.5-turbo` Currently points to `gpt-3.5-turbo-0125`.	16,385 tokens	4,096 tokens	Sep 2021
`gpt-3.5-turbo-1106` GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Learn more.	16,385 tokens	4,096 tokens	Sep 2021

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS	KNOWLEDGE CUTOFF
`gpt-3.5-turbo-instruct` Similar capabilities as GPT-3 era models. Compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	4,096 tokens	Sep 2021

DALL·E

DALL·E is a AI system that can create realistic images and art from a description in natural language. DALL·E 3 currently supports the ability, given a prompt, to create a new image with a specific size. DALL·E 2 also support the ability to edit an existing image, or create variations of a user provided image.

DALL·E 3 is available through our Images API along with DALL·E 2. You can try DALL·E 3 through ChatGPT Plus.

MODEL	DESCRIPTION
`dall-e-3`	The latest DALL·E model released in Nov 2023. Learn more.
`dall-e-2`	The previous DALL·E model released in Nov 2022. The 2nd iteration of DALL·E with more realistic, accurate, and 4x greater resolution images than the original model.

TTS

TTS is an AI model that converts text to natural sounding spoken text. We offer two different model variates, `tts-1` is optimized for real time text to speech use cases and `tts-1-hd` is optimized for quality. These models can be used with the Speech endpoint in the Audio API.

MODEL	DESCRIPTION
`tts-1`	The latest text to speech model, optimized for speed.
`tts-1-hd`	The latest text to speech model, optimized for quality.

Whisper

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. The Whisper v2-large model is currently available through our API with the `whisper-1` model name.

Currently, there is no difference between the open source version of Whisper and the version available through our API. However, through our API, we offer an optimized inference process which makes running Whisper through our API much faster than doing it through other means. For more technical details on Whisper, you can read the paper.

Embeddings

Embeddings are a numerical representation of text that can be used to measure the relatedness between two pieces of text. Embeddings are useful for search, clustering, recommendations, anomaly detection, and classification tasks. You can read more about our latest embedding models in the announcement blog post.

MODEL	OUTPUT DIMENSION
`text-embedding-3-large` Most capable embedding model for both english and non-english tasks	3,072
`text-embedding-3-small` Increased performance over 2nd generation ada embedding model	1,536
`text-embedding-ada-002` Most capable 2nd generation embedding model, replacing 16 first generation models	1,536

Moderation

The Moderation models are designed to check whether content complies with OpenAI's usage policies. The models provide classification capabilities that look for content in categories like hate, self-harm, sexual content, violence, and others. Learn more about moderating text and images in our moderation guide.

MODEL	MAX TOKENS
`omni-moderation-latest` Currently points to `omni-moderation-2024-09-26`.	32,768
`omni-moderation-2024-09-26` Latest pinned version of our new multi-modal moderation model, capable of analyzing both text and images.	32,768
`text-moderation-latest` Currently points to `text-moderation-007`.	32,768
`text-moderation-stable` Currently points to `text-moderation-007`.	32,768
`text-moderation-007` Previous generation text-only moderation. We expect `omni-moderation-*` models to be the best default moving forward.	32,768

GPT base

GPT base models can understand and generate natural language or code but are not trained with instruction following. These models are made to be replacements for our original GPT-3 base models and use the legacy Completions API. Most customers should use GPT-3.5 or GPT-4.

MODEL	MAX TOKENS	KNOWLEDGE CUTOFF
babbage-002 Replacement for the GPT-3 ada and babbage base models.	16,384 tokens	Sep 2021
davinci-002 Replacement for the GPT-3 curie and davinci base models.	16,384 tokens	Sep 2021

How we use your data

Your data is your data.

As of March 1, 2023, data sent to the OpenAI API is not used to train or improve OpenAI models (unless you explicitly opt in to share data with us).

To help identify abuse, API data may be retained for up to 30 days, after which it will be deleted (unless otherwise required by law). For trusted customers with sensitive applications, zero data retention may be available. With zero data retention, request and response bodies are not persisted to any logging mechanism and exist only in memory in order to serve the request.

Note that this data policy does not apply to OpenAI's non-API consumer services like ChatGPT or DALL·E Labs.

Default usage policies by endpoint

ENDPOINT	DATA USED FOR TRAINING	DEFAULT RETENTION	ELIGIBLE FOR ZERO RETENTION
/v1/chat/completions*	No	30 days	Yes, except (a) image inputs, (b) schemas provided for Structured Outputs, or (c) audio outputs. *
/v1/assistants	No	30 days **	No
/v1/threads	No	30 days **	No
/v1/threads/messages	No	30 days **	No
/v1/threads/runs	No	30 days **	No
/v1/vector_stores	No	30 days **	No
/v1/threads/runs/steps	No	30 days **	No
/v1/images/generations	No	30 days	No
/v1/images/edits	No	30 days	No
/v1/images/variations	No	30 days	No

ENDPOINT	DATA USED FOR TRAINING	DEFAULT RETENTION	ELIGIBLE FOR ZERO RETENTION
`/v1/embeddings`	No	30 days	Yes
`/v1/audio/transcriptions`	No	Zero data retention	-
`/v1/audio/translations`	No	Zero data retention	-
`/v1/audio/speech`	No	30 days	Yes
`/v1/files`	No	Until deleted by customer	No
`/v1/fine_tuning/jobs`	No	Until deleted by customer	No
`/v1/batches`	No	Until deleted by customer	No
`/v1/moderations`	No	Zero data retention	-
`/v1/completions`	No	30 days	Yes
`/v1/realtime` (beta)	No	30 days	Yes

*** Chat Completions:**

> Image inputs via the `o1`, `gpt-4o`, `gpt-4o-mini`, `chatgpt-4o-latest`, or `gpt-4-turbo` models (or previously `gpt-4-vision-preview`) are not eligible for zero retention.
>
> Audio outputs are stored for 1 hour to enable multi-turn conversations, and are not currently eligible for zero retention.
>
> When Structured Outputs is enabled, schemas provided (either as the `response_format` or in the function definition) are not eligible for zero retention, though the completions themselves are.
>
> When using Stored Completions via the `store: true` option in the API, those completions are stored for 30 days. Completions are stored in an unfiltered form after an API response, so please avoid storing completions that contain sensitive data.

**** Assistants API:**

> Objects related to the Assistants API are deleted from our servers 30 days after you delete them via the API or the dashboard. Objects that are not deleted via the API or dashboard are retained indefinitely.

Evaluations:

> Evaluation data: When you create an evaluation, the data related to that evaluation is deleted from our servers 30 days after you delete it via the dashboard. Evaluation data that is not deleted via the dashboard is retained indefinitely.

For details, see our API data usage policies. To learn more about zero retention, get in touch with our sales team.

Model endpoint compatibility

ENDPOINT	LATEST MODELS
`/v1/assistants`	All o-series, all GPT-4o (except `chatgpt-4o-latest`), GPT-4o-mini, GPT-4, and GPT-3.5 Turbo models. The `retrieval` tool requires `gpt-4-turbo-preview` (and subsequent dated model releases) or `gpt-3.5-turbo-1106` (and subsequent versions).
`/v1/audio/transcriptions`	`whisper-1`
`/v1/audio/translations`	`whisper-1`
`/v1/audio/speech`	`tts-1`, `tts-1-hd`
`/v1/chat/completions`	All o-series, GPT-4o (except for Realtime preview), GPT-4o-mini, GPT-4, and GPT-3.5 Turbo models and their dated releases. `chatgpt-4o-latest` dynamic model. Fine-tuned versions of `gpt-4o`, `gpt-4o-mini`, `gpt-4`, and `gpt-3.5-turbo`.
`/v1/completions (Legacy)`	`gpt-3.5-turbo-instruct`, `babbage-002`, `davinci-002`
`/v1/embeddings`	`text-embedding-3-small`, `text-embedding-3-large`, `text-embedding-ada-002`
`/v1/fine_tuning/jobs`	`gpt-4o`, `gpt-4o-mini`, `gpt-4`, `gpt-3.5-turbo`
`/v1/moderations`	`text-moderation-stable`, `text-moderation-latest`
`/v1/images/generations`	`dall-e-2`, `dall-e-3`
`/v1/realtime (beta)`	`gpt-4o-realtime-preview`, `gpt-4o-realtime-preview-2024-10-01`

This list excludes all of our deprecated models.