OpenAI Platform

‹ Models

GPT-4 Turbo

Default

An older high-intelligence GPT model

Compare

Try in Playground

Intelligence	● ●
Speed	⚡ ⚡ ⚡
Price	$10 · $30
Input	T 🖼 ☒ ☒
Output	T ☒ ☒ ☒

GPT-4 Turbo is the next generation of GPT-4, an older high-intelligence GPT model. It was designed to be a cheaper, better version of GPT-4. Today, we recommend using a newer model like GPT-4o.

128,000 context window

4,096 max output tokens

Dec 01, 2023 knowledge cutoff

Pricing

Pricing is based on the number of tokens used, or other metrics based on the model type. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the pricing page.

Text tokens

Per 1M tokens · Batch API price ○

Input

$10.00

Output

$30.00

Quick comparison

Input Output

GPT-4 Turbo	$10.00
o3-mini	$1.10
GPT-4o mini	$0.15

Modalities

Text
Input and output

Image
Input only

Audio
Not supported

Video
Not supported

Endpoints

Chat Completions
v1/chat/completions

Responses
v1/responses

Realtime
v1/realtime

Assistants
v1/assistants

Batch
v1/batch

Fine-tuning
v1/fine-tuning

Embeddings
v1/embeddings

Image generation
v1/images/generations

Videos
v1/videos

Image edit
v1/images/edits

Speech generation
v1/audio/speech

Transcription
v1/audio/transcriptions

Translation
v1/audio/translations

Moderation
v1/moderations

Completions (legacy)
v1/completions

Features

Streaming
Supported

Function calling
Supported

Structured outputs
Not supported

Fine-tuning
Not supported

Distillation
Not supported

Predicted outputs
Not supported

Snapshots

Snapshots let you lock in a specific version of the model so that performance and behavior remain consistent. Below is a list of all available snapshots and aliases for GPT-4 Turbo.

`gpt-4-turbo`
↳ `gpt-4-turbo-2024-04-09`

`gpt-4-turbo-2024-04-09`

`gpt-4-turbo-preview` Deprecated
↳ `gpt-4-0125-preview`

`gpt-4-0125-preview`

`gpt-4-1106-vision-preview`

Rate limits

Rate limits ensure fair and reliable access to the API by placing specific caps on requests or tokens used within a given time period. Your usage tier determines how high these limits are set and automatically increases as you send more requests and spend more on the API.

TIER	RPM	TPM	BATCH QUEUE LIMIT
Free	Not supported		
Tier 1	500	30,000	90,000
Tier 2	5,000	450,000	1,350,000
Tier 3	5,000	600,000	40,000,000
Tier 4	10,000	800,000	80,000,000
Tier 5	10,000	2,000,000	300,000,000