OpenAI Platform

‹ Models

GPT-5.1

Default

The best model for coding and agentic tasks with configurable reasoning effort.

Compare   Try in Playground

REASONING	SPEED	PRICE	INPUT	OUTPUT
●●●●	⚡⚡⚡⚡	$1.25 · $10		
Higher	Fast	Input · Output	Text, image	Text, image

GPT-5.1 is our flagship model for coding and agentic tasks with configurable reasoning and non-reasoning effort. Learn more in our GPT-5.1 usage guide.

- 400,000 context window
- 128,000 max output tokens
- Sep 30, 2024 knowledge cutoff
- Reasoning token support

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	Cached input	Output
$1.25	$0.125	$10.00

Quick comparison

Input Cached input Output

GPT-5.1	$1.25
GPT-5	$1.25
GPT-5 mini	$0.25

Modalities

Text
Input and output

Image
Input and output

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
Supported

Fine-tuning
Not supported

Distillation
Supported

Tools

Tools supported by this model when using the Responses API.

Web search
Supported

File search
Supported

Image generation
Supported

Code interpreter
Supported

Computer use
Not supported

MCP
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-5.1.

gpt-5.1
↳ gpt-5.1-2025-11-13

gpt-5.1-2025-11-13

gpt-5
↳ gpt-5-2025-08-07

gpt-5-2025-08-07

gpt-5-mini
↳ gpt-5-mini-2025-08-07

gpt-5-mini-2025-08-07

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	500,000	1,500,000
Tier 2	5,000	1,000,000	3,000,000
Tier 3	5,000	2,000,000	100,000,000
Tier 4	10,000	4,000,000	200,000,000
Tier 5	15,000	40,000,000	15,000,000,000