OpenAI Platform

< Models

GPT-4.1

Default

Smartest non-reasoning model

Compare

Try in Playground

Intelligence	●●●●
Speed	⚡⚡⚡
Price	$2 · $8
Input	T 🖼
Output	T

GPT-4.1 excels at instruction following and tool calling, with broad knowledge across domains. It features a 1M token context window, and low latency without a reasoning step.

Note that we recommend starting with GPT-5 for complex tasks.

- 1,047,576 context window
- 32,768 max output tokens
- Jun 01, 2024 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	**$2.00**
Cached input	**$0.50**
Output	**$8.00**

Quick comparison

	Input	Cached input	Output
GPT-4o	$2.50		
GPT-4.1	$2.00		
o3-mini	$1.10		

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
Supported

Fine-tuning
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-4.1.

gpt-4.1
↳ `gpt-4.1-2025-04-14`

`gpt-4.1-2025-04-14`

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.

Long Context

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	800,000	50,000,000
Tier 4	10,000	2,000,000	200,000,000
Tier 5	10,000	30,000,000	5,000,000,000