OpenAI Platform

< Models

GPT-3.5 Turbo

Default

Legacy GPT model for cheaper chat and non-chat tasks

Compare

Try in Playground

Intelligence	●
Speed	⚡⚡
Price	$0.5 • $1.5
Input	T
Output	T

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, use gpt-4o-mini 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.

- 16,385 context window
- 4,096 max output tokens
- Sep 01, 2021 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
$0.50

Output
$1.50

Quick comparison Input Output

o3-mini	$1.10
GPT-3.5 Turbo	$0.50
GPT-4o mini	$0.15

Modalities

Text
Input and output

Image
Not supported

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

Transcription

v1/audio/speech

v1/audio/transcriptions

Translation
v1/audio/translations

Moderation
v1/moderations

Completions (legacy)
v1/completions

Features

Streaming
Not supported

Function calling
Not supported

Structured outputs
Not supported

Fine-tuning
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-3.5 Turbo.

gpt-3.5-turbo
↳ gpt-3.5-turbo-0125

gpt-3.5-turbo-0125

gpt-3.5-turbo-1106

gpt-3.5-turbo-instruct

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	RPD	TPM	BATCH QUEUE LIMIT
Free			Not supported	
Tier 1	500	10,000	200,000	2,000,000
Tier 2	5,000	–	2,000,000	5,000,000
Tier 3	5,000	–	4,000,000	50,000,000
Tier 4	10,000	–	10,000,000	1,000,000,000
Tier 5	10,000	–	50,000,000	10,000,000,000