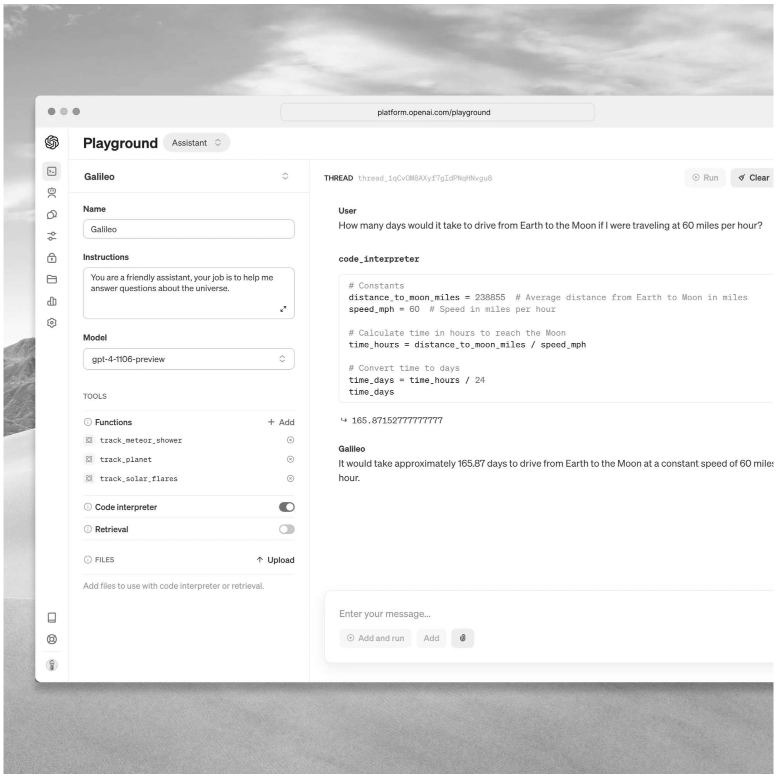
Blog

New models and developer products announced at DevDay

GPT-4 Turbo with 128K context and lower prices, the new Assistants API, GPT-4 Turbo with Vision, DALL·E 3 API, and more.



Name

Galileo

Instructions

You are a friendly assistant, your job is to help me answer questions about the universe.

Model

gpt-4-1106-preview

User

How many days would it take to drive from Earth to the Moon if I were traveling at 60 miles per hour?

code_interpreter

```
# Constants
distance_to_moon_miles = 238855  # Average distance from Earth to Moon in miles
speed_mph = 60  # Speed in miles per hour

# Calculate time in hours to reach the Moon
time_hours = distance_to_moon_miles / speed_mph

# Convert time to days
time_days = time_hours / 24
time_days
```

```
↳ 165.87152777777777
```

Galileo

It would take approximately 165.87 days to drive from Earth to the Moon at a constant speed of 60 miles hour.

November 6, 2023

Authors

OpenAI ↓

Announcements, Product

Update: We previously stated that applications using the `gpt-3.5-turbo` name will automatically be upgraded to the new model version on December 11. We have edited the blogpost to remove this line since this will no longer be happening.

Today, we shared dozens of new additions and improvements, and reduced pricing across many parts of our platform. These include:

- New GPT-4 Turbo model that is more capable, cheaper and supports a 128K context window
- New Assistants API that makes it easier for developers to build their own assistive AI apps that have goals and can call models and tools
- New multimodal capabilities in the platform, including vision, image creation (DALL·E 3), and text-to-speech (TTS)

We'll begin rolling out new features to OpenAI customers starting at 1pm PT today.

Learn more about OpenAI DevDay announcements for ChatGPT.

GPT-4 Turbo with 128K context

We released the first version of GPT-4 in March and made GPT-4 generally available to all developers in July. Today we're launching a preview of the next generation of this model, GPT-4 Turbo.

GPT-4 Turbo is more capable and has knowledge of world events up to April 2023. It has a 128k context window so it can fit the equivalent of more than 300 pages of text in a single prompt. We also optimized its performance so we are able to offer GPT-4 Turbo at a 3x cheaper price for input tokens and a 2x cheaper price for output tokens compared to GPT-4.

GPT-4 Turbo is available for all paying developers to try by passing `gpt-4-1106-preview` in the API and we plan to release the stable production-ready model in the coming weeks.

Function calling updates

Function calling lets you describe functions of your app or external APIs to models, and have the model intelligently choose to output a JSON object containing arguments to call those functions. We're releasing several improvements today, including the ability to call multiple functions in a single message: users can send one message requesting multiple actions, such as "open the car window and turn off the A/C", which would previously require multiple roundtrips with the model (learn more). We are also improving function calling accuracy: GPT-4 Turbo is more likely to return the right function parameters.

Improved instruction following and JSON mode

GPT-4 Turbo performs better than our previous models on tasks that require the careful following of instructions, such as generating specific formats (e.g., “always respond in XML”). It also supports our new JSON mode, which ensures the model will respond with valid JSON. The new API parameter `response_format` enables the model to constrain its output to generate a syntactically correct JSON object. JSON mode is useful for developers generating JSON in the Chat Completions API outside of function calling.

Reproducible outputs and log probabilities

The new `seed` parameter enables **reproducible outputs** by making the model return consistent completions most of the time. This beta feature is useful for use cases such as replaying requests for debugging, writing more comprehensive unit tests, and generally having a higher degree of control over the model behavior. We at OpenAI have been using this feature internally for our own unit tests and have found it invaluable. We’re excited to see how developers will use it. Learn more.

We’re also launching a feature to return the **log probabilities** for the most likely output tokens generated by GPT-4 Turbo and GPT-3.5 Turbo in the next few weeks, which will be useful for building features such as autocomplete in a search experience.

Updated GPT-3.5 Turbo

In addition to GPT-4 Turbo, we are also releasing a new version of GPT-3.5 Turbo that supports a 16K context window by default. The new 3.5 Turbo supports improved instruction following, JSON mode, and parallel function calling. For instance, our internal evals show a 38% improvement on format following tasks such as generating JSON, XML and YAML. Developers can access this new model by calling `gpt-3.5-turbo-1106` in the API. Older models will continue to be accessible by passing `gpt-3.5-turbo-0613` in the API until June 13, 2024. Learn more.

Assistants API, Retrieval, and Code Interpreter

Today, we’re releasing the Assistants API, our first step towards helping developers build agent-like experiences within their own applications. An assistant is a purpose-built AI that has specific instructions, leverages extra knowledge, and can call models and tools to perform tasks. The new Assistants API provides new capabilities such as Code Interpreter and Retrieval as well as function calling to handle a lot of the heavy lifting that you previously had to do yourself and enable you to build high-quality AI apps.

This API is designed for flexibility; use cases range from a natural language-based data analysis app, a coding assistant, an AI-powered vacation planner, a voice-controlled DJ, a smart visual canvas—the list goes on. The Assistants API is built on the same capabilities that enable our new GPTs product: custom instructions and tools such as Code interpreter, Retrieval, and function calling.

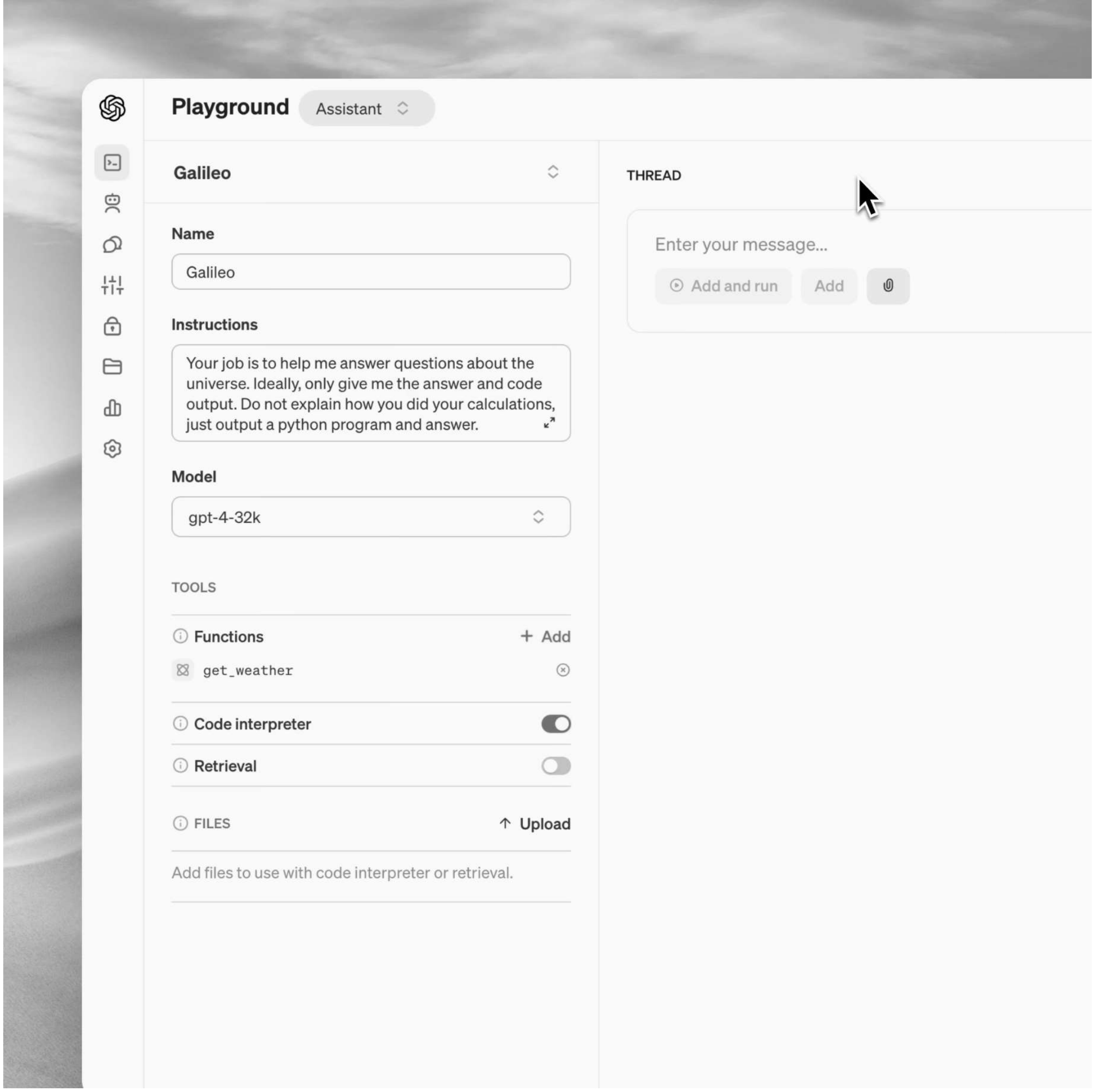
A key change introduced by this API is **persistent and infinitely long threads**, which allow developers to hand off thread state management to OpenAI and work around context window constraints. With the Assistants API, you simply add each new message to an existing `thread`.

Assistants also have access to call new tools as needed, including:

- **Code Interpreter**: writes and runs Python code in a sandboxed execution environment, and can generate graphs and charts, and process files with diverse data and formatting. It allows your assistants to run code iteratively to solve challenging code and math problems, and more.
- **Retrieval**: augments the assistant with knowledge from outside our models, such as proprietary domain data, product information or documents provided by your users. This means you don’t need to compute and store embeddings for your documents, or implement chunking and search algorithms. The Assistants API optimizes what retrieval technique to use based on our experience building knowledge retrieval in ChatGPT.
- **Function calling**: enables assistants to invoke functions you define and incorporate the function response in their messages.

As with the rest of the platform, data and files passed to the OpenAI API are never used to train our models and developers can delete the data when they see fit.

You can try the Assistants API beta without writing any code by heading to the Assistants playground.



Galileo

Name

Galileo

Instructions

Your job is to help me answer questions about the universe. Ideally, only give me the answer and code output. Do not explain how you did your calculations, just output a python program and answer.

Model

gpt-4-32k

Functions

`get_weather`

Code interpreter ●

Retrieval ○

Use the Assistants playground to create high quality assistants without code.

The Assistants API is in beta and available to all developers starting today. Please share what you build with us (@OpenAI) along with your feedback which we will incorporate as we continue building over the coming weeks. Pricing for the Assistants APIs and its tools is available on our pricing page.

New modalities in the API

GPT-4 Turbo with vision

GPT-4 Turbo can accept images as inputs in the Chat Completions API, enabling use cases such as generating captions, analyzing real world images in detail, and reading documents with figures. For example, BeMyEyes uses this technology to help people who are blind or have low vision with daily tasks like identifying a product or navigating a store. Developers can access this feature by using `gpt-4-vision-preview` in the API. We plan to roll out vision support to the main GPT-4 Turbo model as part of its stable release. Pricing depends on the input image size. For instance, passing an image with 1080×1080 pixels to GPT-4 Turbo costs $0.00765. Check out our vision guide.

DALL·E 3

Developers can integrate DALL·E 3, which we recently launched to ChatGPT Plus and Enterprise users, directly into their apps and products through our Images API by specifying `dall-e-3` as the model. Companies like Snap, Coca-Cola, and Shutterstock have used DALL·E 3 to programmatically generate images and designs for their customers and campaigns. Similar to the previous version of DALL·E, the API incorporates built-in moderation to help developers protect their applications against misuse. We offer different format and quality options, with prices starting at $0.04 per image generated. Check out our guide to getting started with DALL·E 3 in the API.

Text-to-speech (TTS)

Developers can now generate human-quality speech from text via the text-to-speech API. Our new TTS model offers six preset voices to choose from and two model variants, `tts-1` and `tts-1-hd.` `tts` is optimized for real-time use cases and `tts-1-hd` is optimized for quality. Pricing starts at $0.015 per input 1,000 characters. Check out our TTS guide to get started.

Listen to voice samples

Select text Scenic ⌄

As the golden sun dips below the horizon, casting long shadows across the tranquil meadow, the world seems to hush, and a sense of calmness envelops the Earth, promising a peaceful night's rest for all living beings.

Select voice Alloy ⌄

0:00 / 0:13

Model customization

GPT-4 fine tuning experimental access

We're creating an experimental access program for **GPT-4 fine-tuning**. Preliminary results indicate that GPT-4 fine-tuning requires more work to achieve meaningful improvements over the base model compared to the substantial gains realized with GPT-3.5 fine-tuning. As quality and safety for GPT-4 fine-tuning improves, developers actively using GPT-3.5 fine-tuning will be presented with an option to apply to the GPT-4 program within their fine-tuning console.

Custom models

For organizations that need even more customization than fine-tuning can provide (particularly applicable to domains with extremely large proprietary datasets—billions of tokens at minimum), we're also launching a **Custom Models program**, giving selected organizations an opportunity to work with a dedicated group of OpenAI researchers to train custom GPT-4 to their specific domain. This includes modifying every step of the model training process, from doing additional domain specific pre-training, to running a custom RL post-training process tailored for the specific domain. Organizations will have exclusive access to their custom models. In keeping with our existing enterprise privacy policies, custom models will not be served to or shared with other customers or used to train other models. Also, proprietary data provided to OpenAI to train custom models will not be reused in any other context. This will be a very limited (and expensive) program to start—interested orgs can apply here.

Lower prices and higher rate limits

Lower prices

We're decreasing several prices across the platform to pass on savings to developers (all prices below are expressed per 1,000 tokens):

- GPT-4 Turbo input tokens are 3x cheaper than GPT-4 at $0.01 and output tokens are 2x cheaper at $0.03.
- GPT-3.5 Turbo input tokens are 3x cheaper than the previous 16K model at $0.001 and output tokens are 2x cheaper at $0.002. Developers previously using GPT-3.5 Turbo 4K benefit from a 33% reduction on input tokens at $0.001. Those lower prices only apply to the new GPT-3.5 Turbo introduced today.
- Fine-tuned GPT-3.5 Turbo 4K model input tokens are reduced by 4x at $0.003 and output tokens are 2.7x cheaper at $0.006. Fine-tuning also supports 16K context at the same price as 4K with the new GPT-3.5 Turbo model. These new prices also apply to fine-tuned `gpt-3.5-turbo-0613` models.

	Older models	**New models**
GPT-4 Turbo	GPT-4 8K Input: $0.03 Output: $0.06 GPT-4 32K Input: $0.06 Output: $0.12	GPT-4 Turbo 128K Input: $0.01 Output: $0.03
GPT-3.5 Turbo	GPT-3.5 Turbo 4K Input: $0.0015 Output: $0.002 GPT-3.5 Turbo 16K Input: $0.003 Output: $0.004	GPT-3.5 Turbo 16K Input: $0.001 Output: $0.002
GPT-3.5 Turbo fine-tuning	GPT-3.5 Turbo 4K fine-tuning Training: $0.008 Input: $0.012 Output: $0.016	GPT-3.5 Turbo 4K and 16K fine-tuning Training: $0.008 Input: $0.003 Output: $0.006

Higher rate limits

To help you scale your applications, we're doubling the tokens per minute limit for all our paying GPT-4 customers. You can view your new rate limits in your rate limit page. We've also published our usage tiers that determine automatic rate limits increases, so you know what to expect in how your usage limits will automatically scale. You can now request increases to usage limits from your account settings.

Copyright Shield

OpenAI is committed to protecting our customers with built-in copyright safeguards in our systems. Today, we're going one step further and introducing Copyright Shield—we will now step in and defend our customers, and pay the costs incurred, if you face legal claims around copyright infringement. This applies to generally available features of ChatGPT Enterprise and our developer platform.

Whisper v3 and Consistency Decoder

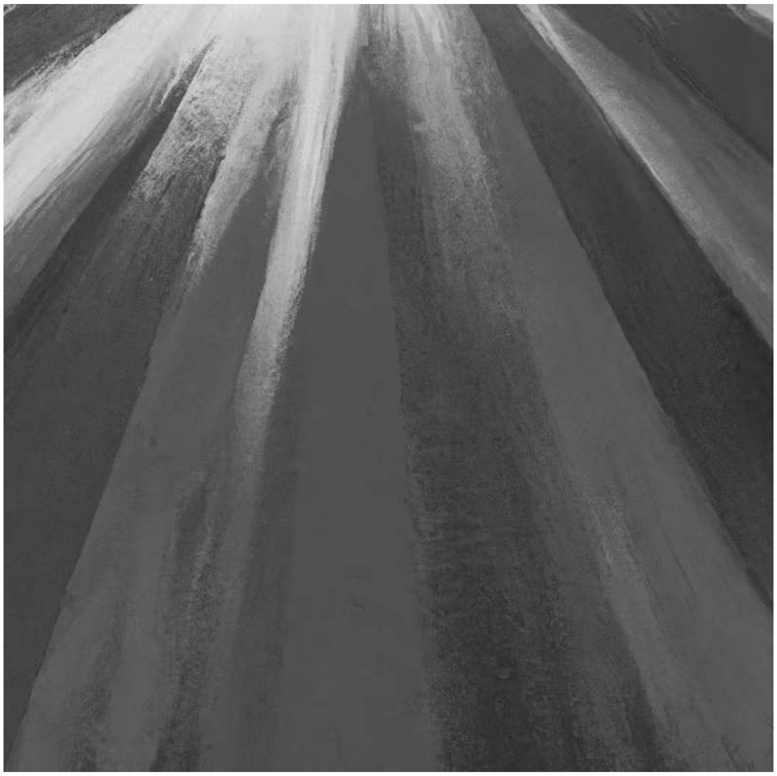
We are releasing Whisper large-v3, the next version of our open source automatic speech recognition model (ASR) which features improved performance across languages. We also plan to support Whisper v3 in our API in the near future.

We are also open sourcing the Consistency Decoder, a drop in replacement for the Stable Diffusion VAE decoder. This decoder improves all images compatible with the by Stable Diffusion 1.0+ VAE, with significant improvements in text, faces and straight lines.

Learn more about our OpenAI DevDay announcements for ChatGPT.

Related research

View all research

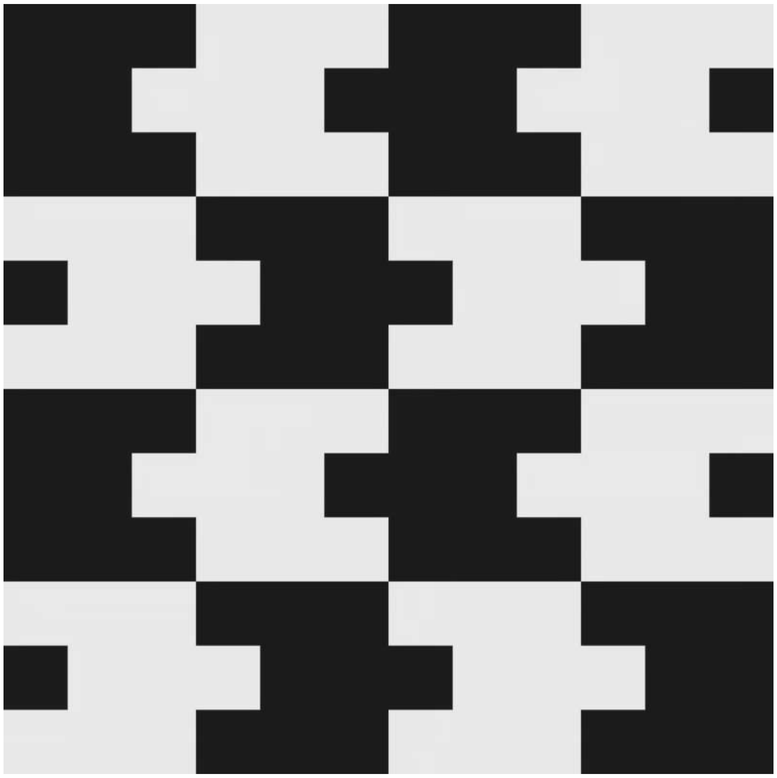


Weak-to-strong generalization

Practices for Governing Agentic AI Systems

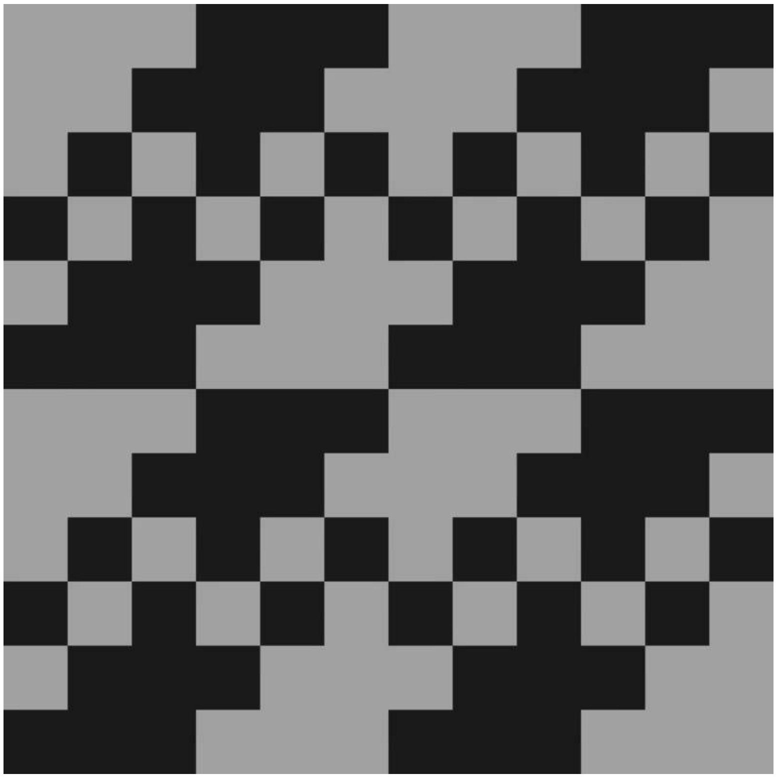
Dec 14, 2023

Dec 14, 2023



DALL·E 3 system card

Oct 3, 2023



GPT-4V(ision) system card

Sep 25, 2023