

GPT-4.1: Three new million token input models from OpenAI, including their cheapest model yet

14th April 2025

OpenAI [introduced](#) three new models this morning: GPT-4.1, GPT-4.1 mini and GPT-4.1 nano. These are API-only models right now, not available through the ChatGPT interface (though you can try them out in OpenAI's [API playground](#)). All three models can handle 1,047,576 tokens of input and 32,768 tokens of output, and all three have a May 31, 2024 cut-off date (their previous models were mostly September 2023).

The models score higher than GPT-4o *and* GPT-4.5 on coding benchmarks, and do very well on long context benchmarks as well. They also claim improvements in instruction following—following requested formats, obeying negative instructions, sorting output and obeying instructions to say “I don’t know”.

I released a new version of my [llm-openai](#) plugin supporting the new models. This is a new thing for the [LLM](#) ecosystem: previously OpenAI models were only supported in core, which meant I had to ship a full LLM release to add support for them.

You can run the new models like this:

```
llm install llm-openai-plugin -U
llm -m openai/gpt-4.1 "Generate an SVG of a pelican riding a bicycle"
```

The other model IDs are openai/gpt-4.1-mini and openai/gpt-4.1-nano.

Here's the [pelican riding a bicycle](#) I got from full sized GPT-4.1:



I'm particularly excited by [GPT-4.1 nano](#), which handles image and text input up to a million tokens and is priced lower than any other previous OpenAI model: \$0.10/million for input and \$0.40/million for output, less than previous cheapest OpenAI model GPT-4o-mini (\$0.15/\$0.60). I've updated my [LLM pricing table](#) to include the new models.

They're not the cheapest overall though: Gemini 2.0 Flash Lite and, Gemini 1.5 Flash 8B, Amazon Nova Lite and Nova Micro and Mistral's 3B, 8B and Small 3.1 hosted models remain less expensive.

Model prices (per million tokens)

Model	Input cost	↑	Output cost
Amazon Nova Micro	\$0.035		\$0.14
Gemini 1.5 Flash-8B ≤128k	\$0.0375		\$0.15
Ministral 3B 24.10	\$0.04		\$0.04
Amazon Nova Lite	\$0.06		\$0.24
Gemini 2.0 Flash Lite	\$0.075		\$0.30
Gemini 1.5 Flash ≤128k	\$0.075		\$0.30
Gemini 1.5 Flash-8B >128k	\$0.075		\$0.30
Gemini 2.0 Flash	\$0.10		\$0.40
Mistral Small 3.1	\$0.10		\$0.30
Ministral 8B 24.10	\$0.10		\$0.10
GPT 4.1 Nano	\$0.10		\$0.40
Gemini 1.5 Flash >128k	\$0.15		\$0.60
GPT-4o Mini	\$0.15		\$0.60

As a demonstration of just how cheap GPT-4.1 nano is, consider this photograph I took of some pelicans:



I ran that through GPT-4.1 nano like this:

```
llm -m openai/gpt-4.1-nano describe \  
-a https://static.simonwillison.net/static/2025/two-pelicans.jpg
```

And got back this response:

The image shows two pelicans flying against a clear blue sky. They are positioned in flight, with their wings extended, showcasing their large wingspan and feather patterns. The pelican on the right appears slightly higher and facing slightly to the left, while the pelican on the left is oriented more horizontally with its beak pointing forward. Both birds display a combination of brown, tan, and white plumage, characteristic of pelicans, with their long beaks prominently visible.

Running `llm logs -c --usage` revealed that this used 1,933 input tokens and 96 output tokens. Passing those through [the pricing calculator](#) returns a total cost of \$0.000232, or 0.0232 cents.

That means I could use GPT-4.1 nano to generate descriptions of 4,310 images like this one for just shy of a dollar.

A few closing thoughts on these new models:

- The 1 million input token context thing is a *really big deal*. The huge token context has been a major competitive advantage for the Google Gemini models for a full year at this point—it's reassuring to see other vendors start to catch up. I'd like to see the same from Anthropic—Claude was the first model to hit 200,000 but hasn't shipped more than that yet (aside from a 500,000 token model that was restricted to their big enterprise partners).

- When I added [fragments support](#) to LLM last week the feature was mainly designed to help take advantage of longer context models. It's pleasing to see another one show up so shortly after that release.
- OpenAI really emphasized code performance for this model. They called out the [Aider benchmark](#) in their announcement post.
- [As expected](#), GPT-4.5 turned out to be not long for this world:

We will also begin deprecating GPT-4.5 Preview in the API, as GPT-4.1 offers improved or similar performance on many key capabilities at much lower cost and latency. GPT-4.5 Preview will be turned off in three months, on July 14, 2025, to allow time for developers to transition

- In [the livestream announcement](#) Michelle Pokrass [let slip](#) that the codename for the model was Quasar—that's the name of the stealth model that's [been previewing on OpenRouter](#) for the past two weeks. That has [now been confirmed](#) by OpenRouter.
- OpenAI shared a [GPT 4.1 Prompting Guide](#), which includes this tip about long context prompting:

Especially in long context usage, placement of instructions and context can impact performance. If you have long context in your prompt, ideally place your instructions at both the beginning and end of the provided context, as we found this to perform better than only above or below. If you'd prefer to only have your instructions once, then above the provided context works better than below.

Adding instructions before the content is incompatible with [prompt caching](#)—I always keep user instructions at the end since doing so means multiple prompts can benefit from OpenAI's prefix cache.

They also recommend XML-style delimiters over JSON for long context, suggesting this format (complete with the XML-invalid unquoted attribute) that's similar to the format [recommended by Anthropic](#) for Claude:

```
<doc id=1 title="The Fox">The quick brown fox jumps over the lazy dog</doc>
```

There's an extensive section at the end [describing their recommended approach to applying file diffs](#): “we open-source here one recommended diff format, on which the model has been extensively trained”.

- One thing notably absent from the GPT-4.1 announcement is any mention of audio support. The “o” in GPT-4o stood for “omni”, because it was a multi-modal model with image and audio input and output. The 4.1 models appear to be text and image input and text output only.

Posted [14th April 2025](#) at 6:12 pm · Follow me on [Mastodon](#), [Bluesky](#), [Twitter](#) or [subscribe to my newsletter](#)

More recent articles

- [CaMeL offers a promising new direction for mitigating prompt injection attacks](#) - 11th April 2025
- [Model Context Protocol has prompt injection security problems](#) - 9th April 2025

ai 1214

openai 277

generative-ai 1042

llms 1030

llm 168

vision-llms 51

llm-pricing 26

pelican-riding-a-bicycle 23

long-context 13

llm-release 93

Previous: [CaMeL offers a promising new direction for mitigating prompt injection attacks](#)