GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why

We got a first look at the much-anticipated big new language model from OpenAI. But this time how it works is even more deeply under wraps.

Will Douglas Heaven

OpenAI has finally unveiled GPT-4, a next-generation large language model that was rumored to be in development for much of last year. The San Francisco-based company's last surprise hit, ChatGPT, was always going to be a hard act to follow, but OpenAI has made GPT-4 even bigger and better.

Yet how much bigger and why it's better, OpenAI won't say. GPT-4 is the most secretive release the company has ever put out, marking its full

transition from nonprofit research lab to for-profit tech firm.

“That’s something that, you know, we can’t really comment on at this time,” said OpenAI’s chief scientist, Ilya Sutskever, when I spoke to members of the GPT-4 team in a video call an hour after the announcement. “It’s pretty competitive out there.”

GPT-4 is a multimodal large language model, which means it can respond to both text and images. Give it a photo of the contents of your fridge and ask it what you could make, and GPT-4 will try to come up with recipes that use the pictured ingredients. It's also great at explaining jokes, says Sutskever: "If you show it a meme, it can tell you why it's funny."

Access to GPT-4 will be available to users who sign up to the waitlist and for subscribers of the premium paid-for ChatGPT Plus in a limited, text-only capacity.

“The continued improvements along many dimensions are remarkable,” says Oren Etzioni at the Allen Institute for AI. “GPT-4 is now the standard by which all foundation models will be evaluated.”

“A good multimodal model has been the holy grail of many big tech labs for the past couple of years,” says Thomas Wolf, cofounder of Hugging Face, the AI startup behind the open-source large language model BLOOM. “But it has remained elusive.”

In theory, combining text and images could allow multimodal models to understand the world better. “It might be able to tackle traditional weak points of language models, like spatial reasoning,” says Wolf.

It is not yet clear if that’s true for GPT-4. OpenAI’s new model appears to be better at some basic reasoning than ChatGPT, solving simple puzzles such

as summarizing blocks of text in words that start with the same letter. In my demo during the call, I was shown GPT-4 summarizing the announcement blurb from OpenAI's website using words that begin with g: "GPT-4, groundbreaking generational growth, gains greater grades. Guardrails, guidance, and gains garnered. Gigantic, groundbreaking, and globally gifted." In another demo, GPT-4 took in a document about taxes and answered questions about it, citing reasons for its responses.

It also outperforms ChatGPT on human tests, including the Uniform Bar Exam (where GPT-4 ranks in the 90th percentile and ChatGPT ranks in the 10th) and the Biology Olympiad (where GPT-4 ranks in the 99th percentile and ChatGPT ranks in the 31st). "It's exciting how evaluation is now starting to be conducted on the very same benchmarks that humans use for themselves," says Wolf. But he adds that without seeing the technical details, it's hard to judge how impressive these results really are.

According to OpenAI, GPT-4 performs better than ChatGPT—which is based on GPT-3.5, a version of the firm's previous technology—because it is a larger model with more parameters (the values in a neural network that get tweaked during training). This follows an important trend that the company discovered with its previous models. GPT-3 outperformed GPT-2 because it was more than 100 times larger, with 175 billion parameters to GPT-2's 1.5 billion. "That fundamental formula has not really changed much for years," says Jakub Pachocki, one of GPT-4's developers. "But it's still like building a spaceship, where you need to get all these little components right and make sure none of it breaks."

But OpenAI has chosen not to reveal how large GPT-4 is. In a departure from its previous releases, the company is giving away nothing about how GPT-4 was built—not the data, the amount of computing power, or the training techniques. "OpenAI is now a fully closed company with scientific

communication akin to press releases for products," says Wolf.

OpenAI says it spent six months making GPT-4 safer and more accurate. According to the company, GPT-4 is 82% less likely than GPT-3.5 to respond to requests for content that OpenAI does not allow, and 60% less likely to make stuff up.

OpenAI says it achieved these results using the same approach it took with ChatGPT, using reinforcement learning via human feedback. This involves asking human raters to score different responses from the model and using those scores to improve future output.

The team even used GPT-4 to improve itself, asking it to generate inputs that led to biased, inaccurate, or offensive responses and then fixing the model so that it refused such inputs in future.

GPT-4 may be the best multimodal large language model yet built. But it is not in a league of its own, as GPT-3 was when it first appeared in 2020. A lot has happened in the last three years. Today GPT-4 sits alongside other multimodal models, including Flamingo from DeepMind. And Hugging Face is working on an open-source multimodal model that will be free for others to use and adapt, says Wolf.

Faced with such competition, OpenAI is treating this release more as a product tease than a research update. Early versions of GPT-4 have been shared with some of OpenAI's partners, including Microsoft, which confirmed today that it used a version of GPT-4 to build Bing Chat. OpenAI is also now working with Stripe, Duolingo, Morgan Stanley, and the government of Iceland (which is using GPT-4 to help preserve the Icelandic language), among others.

Many other companies are waiting in line: "The costs to bootstrap a model of

this scale is out of reach for most companies, but the approach taken by OpenAI has made large language models very accessible to startups," says Sheila Gulati, cofounder of the investment firm Tola Capital. "This will catalyze tremendous innovation on top of GPT-4."

Never before has powerful new AI gone from lab to consumer-facing products so fast. (In other news today, Google announced that it is making its own large language model PaLM available to third party developers and rolling out chatbot features in Google Docs and Gmail; and AI firm Anthropic announced a new large language model called Claude, which is already being tried out by several companies, including Notion and Quora.)

And yet large language models remain fundamentally flawed. GPT-4 can still generate biased, false, and hateful text; it can also still be hacked to bypass its guardrails. Though OpenAI has improved this technology, it has not fixed it by a long shot. The company claims that its safety testing has been sufficient for GPT-4 to be used in third-party apps. But it is also braced for surprises.

"Safety is not a binary thing; it is a process," says Sutskever. "Things get complicated any time you reach a level of new capabilities. A lot of these capabilities are now quite well understood, but I'm sure that some will still be surprising."

Even Sutskever suggests that going slower with releases might sometimes be preferable: "It would be highly desirable to end up in a world where companies come up with some kind of process that allows for slower releases of models with these completely unprecedented capabilities**."**