GPT-4 IS HERE: WHAT SCIENTISTS THINK

Researchers are excited about the AI, but frustrated by the secrecy surrounding its underlying engineering.

By Katharine Sanderson

Artificial intelligence (AI) company OpenAI this week unveiled GPT-4, the latest incarnation of the large language model that powers its popular chatbot ChatGPT. The company says GPT-4 contains big improvements — it has already stunned people with its ability to create text resembling that written by humans and generate images and computer code from almost any prompt. Researchers say these abilities have the potential to transform science — but some are frustrated that they cannot yet access the technology, its underlying code or information on how it was trained. That raises concern about the technology's safety and makes it less useful for research, say scientists.

GPT-4 was released on 14 March, and one upgrade is that it can now handle images as well as text. And as a demonstration of its language prowess, OpenAI, which is based in San Francisco, California, says that it passed the US bar legal exam with results in the ninetieth centile, compared with the tenth centile for the previous version of ChatGPT. But the technology is not yet widely accessible — only paying subscribers so far have access.

"There's a waiting list at the moment so you cannot use it right now," says Evi-Anne van Dis, a psychologist at the University of Amsterdam Medical Centers. But she has seen demos of GPT-4. "We watched some videos in which they demonstrated capacities and it's mind-blowing," she says. One instance, she recounts, was a hand-drawn doodle of a website, which GPT-4 used to produce the computer code needed to build that website, as a demonstration of the ability to handle images as inputs.

Black box

But there is frustration in the science community over OpenAI's secrecy around how the model was trained and what data were used, and how GPT-4 actually works. "All of these closed-source models, they are essentially dead ends in science," says Sasha Luccioni, a research scientist specializing in climate at HuggingFace, an open-source AI cooperative. "They [OpenAI] can keep building upon their research, but for the community at large, it's a dead end."

JAAP ARRIENS/NURPHOTO VIA GETTY

The GPT-4 artificial-intelligence model is not yet widely available.

Andrew White, a chemical engineer at the University of Rochester, New York, has had privileged access to GPT-4 as a 'red-teamer': a person paid by OpenAI to test the platform to try and make it do something bad. He has had access to GPT-4 for the past six months, he says. "Early on in the process, it didn't seem that different," compared with previous iterations.

He put to the bot queries about what chemical reaction steps were needed to make a compound, predict the reaction yield and choose a catalyst. "At first, I was actually not that impressed," White says. "It was really surprising because it would look so realistic, but it would hallucinate an atom here. It would skip a step there," he adds. But when, as part of his red-team work, he gave GPT-4 access to scientific papers, things changed drastically. "It made us realize that these models maybe aren't so great just alone. But when you start connecting them to the Internet to tools like a retrosynthesis planner, or a calculator, all of a sudden, new kinds of abilities emerge."

Danger prevention

And with those abilities come concerns. For instance, could GPT-4 allow dangerous chemicals to be made? With input from people such as White, OpenAI engineers fed back into their model to discourage GPT-4 from creating dangerous, illegal or damaging content, White says.

Outputting false information is another problem. Luccioni says that models such as GPT-4, which exist to predict the next word in a sentence, can't be cured of coming up with fake facts — known as hallucinating. "You can't rely on these kinds of models because there's so much hallucination," she says. And this remains a concern in the latest version, she says, although OpenAI says that it has improved safety in GPT-4.

Without access to the data used for training, OpenAI's assurances about safety fall short for Luccioni. "You don't know what the data is. So you can't improve it. I mean, it's just completely impossible to do science with a model like this," she says.

The mystery about how GPT-4 was trained is also a concern for van Dis's colleague at Amsterdam, psychologist Claudi Bockting. "It's very hard as a human being to be accountable for something that you cannot oversee," she says. "One of the concerns is they could be far more biased than, for instance, the bias that human beings have by themselves." Without being able to access the code behind GPT-4, it is impossible to see where the bias might have originated, or to remedy it, Luccioni explains.

Ethics discussions

Bockting and van Dis are also concerned that these AI systems are increasingly owned by big tech companies. The researchers want to make sure the technology is properly tested and verified by scientists. "This is also an opportunity because collaboration with big tech can, of course, speed up processes," she adds.

Van Dis, Bockting and colleagues argued earlier this year that there is an urgent need to develop a set of 'living' guidelines to govern how AI and tools such as GPT-4 are used and developed. They are concerned that any legislation around AI technologies will struggle to keep up with the pace of development. Bockting and van Dis have convened a summit of invited participants at the University of Amsterdam on 11 April to discuss these concerns, with representatives from organizations including the science-ethics committee of UNESCO, the United Nations' scientific and cultural agency, the Organisation for Economic Co-operation and Development and the World Economic Forum.

Despite the concern, GPT-4 and its future iterations will shake up science, says White. "I think it's actually going to be a huge infrastructure change in science, almost like the Internet was a big change," he says. It won't replace scientists, he adds, but could help with some tasks. "I think we're going to start realizing we can connect papers, data programs, libraries that we use and computational work or even robotic experiments."