ILLUSTRATION BY PAWEŁ JONCA

THE PROMISE AND PERIL OF GENERATIVE AI

Researchers are excited but apprehensive about how tools such as ChatGPT could transform science and society. **By Chris Stokel-Walker and Richard Van Noorden**

In December, computational biologists Casey Greene and Milton Pividori embarked on an unusual experiment: they asked an assistant who was not a scientist to help them improve three of their research papers. Their assiduous aide suggested revisions to sections of documents in seconds; each manuscript took about five minutes to review. In one biology manuscript, their helper even spotted a mistake in a reference to an equation. The trial didn't always run smoothly, but the final manuscripts were easier to read – and the fees were modest, at less than US$0.50 per document.

This assistant, as Greene and Pividori reported in a preprint[1] on 23 January, is not a person but an artificial-intelligence (AI) algorithm called GPT-3, first released in 2020. It is one of the much-hyped generative AI chatbot-style tools that can churn out convincingly fluent text, whether asked to produce prose, poetry, computer code or – as in the scientists' case – to edit research papers.

The most famous of these tools, also known as large language models, or LLMs, is ChatGPT, a version of GPT-3 that shot to fame after its release in November last year because it was made free and easily accessible. Other generative AIs can produce images, or sounds.

"I'm really impressed," says Pividori, who

works at the University of Pennsylvania in Philadelphia. "This will help us be more productive as researchers." Other scientists say they now regularly use LLMs not only to edit manuscripts, but also to help them write or check code and to brainstorm ideas. "I use LLMs every day now," says Hafsteinn Einarsson, a computer scientist at the University of Iceland in Reykjavik. He started with GPT-3, but has since switched to ChatGPT, which helps him to write presentation slides, student exams and coursework problems, and to convert student theses into papers. "Many people are using it as a digital secretary or assistant," he says.

LLMs form part of search engines, code-writing assistants and even a chatbot that negotiates with other companies' chatbots to get better prices on products. ChatGPT's creator, OpenAI in San Francisco, California, has announced a subscription service for $20 per month, promising faster response times and priority access to new features (although its trial version remains free). And tech giant Microsoft, which had already invested in OpenAI, announced a further investment in January, reported to be around $10 billion. LLMs are destined to be incorporated into general word- and data-processing software. Generative AI's future ubiquity in society seems assured, especially because today's tools represent the technology in its infancy.

But LLMs have also triggered widespread concern – from their propensity to return falsehoods, to worries about people passing off AI-generated text as their own (see page 224). When *Nature* asked researchers about the potential uses of chatbots such as ChatGPT, particularly in science, their excitement was tempered with apprehension. "If you believe that this technology has the potential to be transformative, then I think you have to be nervous about it," says Greene, at the University of Colorado School of Medicine in Aurora. Much will depend on how future regulations and guidelines might constrain AI chatbots' use, researchers say.

Fluent but not factual

Some researchers think LLMs are well-suited to speeding up tasks such as writing papers or grants, as long as there's human oversight. "Scientists are not going to sit and write long introductions for grant applications any more," says Almira Osmanovic Thunström, a neurobiologist at Sahlgrenska University Hospital in Gothenburg, Sweden, who has co-authored a manuscript[2] using GPT-3 as an experiment. "They're just going to ask systems to do that."

Tom Tumiel, a research engineer at InstaDeep, a London-based software consultancy firm, says he uses LLMs every day as assistants to help write code. "It's almost like a better Stack Overflow," he says, referring to the popular community website where coders answer each others' queries.

But researchers emphasize that LLMs are fundamentally unreliable at answering questions, sometimes generating false responses. "We need to be wary when we use these systems to produce knowledge," says Osmanovic Thunström.

This unreliability is baked into how LLMs are built. ChatGPT and its competitors work by learning the statistical patterns of language in enormous databases of online text – including any untruths, biases or outmoded knowledge. When LLMs are then given prompts (such as Greene and Pividori's carefully structured requests to rewrite parts of manuscripts), they simply spit out, word by word, any way to continue the conversation that seems stylistically plausible.

The result is that LLMs easily produce errors and misleading information, particularly for technical topics that they might have had little data to train on. LLMs also can't show the origins of their information; if asked to write an academic paper, they make up fictitious citations. "The tool cannot be trusted to get facts right or produce reliable references," noted a January editorial on ChatGPT in the journal *Nature Machine Intelligence*[3].

With these caveats, ChatGPT and other LLMs can be effective assistants for researchers who have enough expertise to directly spot problems or to easily verify answers, such as whether an explanation or suggestion of computer code is correct.

But the tools might mislead naive users. In December, for instance, Stack Overflow temporarily banned the use of ChatGPT, because site moderators found themselves flooded with a high rate of incorrect but seemingly persuasive LLM-generated answers sent in by enthusiastic users. This could be a nightmare for search engines.

Can shortcomings be solved?

Some search-engine tools, such as the researcher-focused Elicit, get around LLMs' attribution issues by using their capabilities first to guide queries for relevant literature, and then to briefly summarize each of the websites or documents that the engines find – so producing an output of apparently referenced content (although an LLM might still mis-summarize each individual document).

Companies building LLMs are also well aware of the problems. In September last year, Google subsidiary DeepMind published a paper[4] on a 'dialogue agent' called Sparrow, which the firm's chief executive and co-founder Demis Hassabis later told *TIME* magazine would be released in private beta this year; the magazine reported that Google aimed to work on features including the ability to cite sources. Other competitors, such as Anthropic, say that they have solved some of ChatGPT's issues (Anthropic, OpenAI and DeepMind declined interviews for this article).

For now, ChatGPT is not trained on sufficiently specialized content to be helpful in technical topics, some scientists say. Kareem Carr, a biostatistics PhD student at Harvard University in Cambridge, Massachusetts, was underwhelmed when he trialled it for work. "I think it would be hard for ChatGPT to attain the level of specificity I would need," he says. (Even so, Carr says that when he asked ChatGPT for 20 ways to solve a research query, it spat back gibberish and one useful idea – a statistical term he hadn't heard of that pointed him to a new area of academic literature.)

Some tech firms are training chatbots on specialized scientific literature – although they have run into their own issues. In November last year, Meta – the tech giant that owns Facebook – released an LLM called Galactica, which was trained on scientific abstracts, with the intention of making it particularly good at producing academic content and answering research questions. The demo was pulled from public access (although its code remains available) after users got it to produce inaccuracies and racism. "It's no longer possible to have some fun by casually misusing it. Happy?," Meta's chief AI scientist, Yann LeCun, tweeted in a response to critics. (Meta did not respond to a request, made through their press office, to speak to LeCun.)

Safety and responsibility

Galactica had hit a familiar safety concern that ethicists have been pointing out for years: without output controls LLMs can easily be used to generate hate speech and spam, as well as racist, sexist and other harmful associations that might be implicit in their training data.

Besides directly producing toxic content, there are concerns that AI chatbots will embed historical biases or ideas about the world from their training data, such as the superiority of particular cultures, says Shobita Parthasarathy, director of a science, technology and public-policy programme at the University of Michigan in Ann Arbor. Because the firms that are creating big LLMs are mostly in, and from, these cultures, they might make little attempt to overcome such biases, which are systemic and hard to rectify, she adds.

OpenAI tried to skirt many of these issues when deciding to openly release ChatGPT. It restricted its knowledge base to 2021, prevented it from browsing the Internet and installed filters to try to get the tool to refuse to produce content for sensitive or toxic prompts. Achieving that, however, required human moderators to label screeds of toxic text. Journalists have reported that these workers are poorly paid and some have suffered trauma. Similar concerns over worker exploitation have also been raised about social-media firms that have employed people to train automated bots for flagging toxic content.

OpenAI's guardrails have not been wholly

successful. In December last year, computational neuroscientist Steven Piantadosi at the University of California, Berkeley, tweeted that he'd asked ChatGPT to develop a Python program for whether a person should be tortured on the basis of their country of origin. The chatbot replied with code inviting the user to enter a country; and to print "This person should be tortured" if that country was North Korea, Syria, Iran or Sudan. (OpenAI subsequently closed off that kind of question.)

Last year, a group of academics released an alternative LLM, called BLOOM. The researchers tried to reduce harmful outputs by training it on a smaller selection of higher-quality, multilingual text sources. The team involved also made its training data fully open (unlike OpenAI). Researchers have urged big tech firms to responsibly follow this example — but it's unclear whether they'll comply.

Some researchers say that academics should refuse to support large commercial LLMs altogether. Besides issues such as bias, safety concerns and exploited workers, these computationally intensive algorithms also require a huge amount of energy to train, raising concerns about their ecological footprint. A further worry is that by offloading thinking to automated chatbots, researchers might lose the ability to articulate their own thoughts. "Why would we, as academics, be eager to use and advertise this kind of product?" wrote Iris van Rooij, a computational cognitive scientist at Radboud University in Nijmegen, the Netherlands, in a blogpost urging academics to resist their pull.

A further confusion is the legal status of some LLMs, which were trained on content scraped from the Internet with sometimes less-than-clear permissions. Copyright and licensing laws currently cover direct copies of pixels, text and software, but not imitations in their style. When those imitations — generated through AI — are trained by ingesting the originals, this introduces a wrinkle. The creators of some AI art programs, including Stable Diffusion and Midjourney, are currently being sued by artists and photography agencies; OpenAI and Microsoft (along with its subsidiary tech site GitHub) are also being sued for software piracy over the creation of their AI coding assistant Copilot. The outcry might force a change in laws, says Lilian Edwards, a specialist in Internet law at Newcastle University, UK.

Enforcing honest use

Setting boundaries for these tools, then, could be crucial, some researchers say. Edwards suggests that existing laws on discrimination and bias (as well as planned regulation of dangerous uses of AI) will help to keep the use of LLMs honest, transparent and fair. "There's loads of law out there," she says, "and it's just a matter of applying it or tweaking it very slightly."

At the same time, there is a push for LLM use to be transparently disclosed. Scholarly publishers (including the publisher of *Nature*) have said that scientists should disclose the use of LLMs in research papers (see also *Nature* **613**, 612; 2023); and teachers have said they expect similar behaviour from their students. The journal *Science* has gone further, saying that no text generated by ChatGPT or any other AI tool can be used in a paper[5].

One key technical question is whether AI-generated content can be spotted easily. Many researchers are working on this, with the central idea to use LLMs themselves to spot the output of AI-created text. Last December, for instance, Edward Tian, a computer-science undergraduate at Princeton University in New Jersey, published GPTZero. This AI-detection tool analyses text in two ways. One is 'perplexity', a measure of how familiar the text seems to an LLM. Tian's tool uses an earlier model, called GPT-2; if it finds most of the words and sentences predictable, then text is likely to have been AI-generated. The tool also examines variation in text, a measure known as 'burstiness': AI-generated text tends to be more consistent in tone, cadence and perplexity than does that written by humans.

"Why would we, as academics, be eager to use and advertise this kind of product?"

Many other products similarly aim to detect AI-written content. OpenAI itself had already released a detector for GPT-2, and it released another detection tool in January. For scientists' purposes, a tool that is being developed by the firm Turnitin, a developer of anti-plagiarism software, might be particularly important, because Turnitin's products are already used by schools, universities and scholarly publishers worldwide. The company says it's been working on AI-detection software since GPT-3 was released in 2020, and expects to launch it in the first half of this year.

However, none of these tools claims to be infallible, particularly if AI-generated text is subsequently edited. Also, the detectors could falsely suggest that some human-written text is AI-produced, says Scott Aaronson, a computer scientist at the University of Texas at Austin and guest researcher with OpenAI. The firm said that in tests, its latest tool incorrectly labelled human-written text as AI-written 9% of the time, and only correctly identified 26% of AI-written texts. Further evidence might be needed before, for instance, accusing a student of hiding their use of an AI solely on the basis of a detector test, Aaronson says.

A separate idea is that AI content would come with its own watermark. Last November, Aaronson announced that he and OpenAI were working on a method of watermarking ChatGPT output. It has not yet been released, but a 24 January preprint[6] from a team led by computer scientist Tom Goldstein at the University of Maryland in College Park, suggested one way of making a watermark. The idea is to use random-number generators at particular moments when the LLM is generating its output, to create lists of plausible alternative words that the LLM is instructed to choose from. This leaves a trace of chosen words in the final text that can be identified statistically but are not obvious to a reader. Editing could defeat this trace, but Goldstein suggests that edits would have to change more than half the words.

An advantage of watermarking is that it rarely produces false positives, Aaronson points out. If the watermark is there, the text was probably produced with AI. Still, it won't be infallible, he says. "There are certainly ways to defeat just about any watermarking scheme if you are determined enough." Detection tools and watermarking only make it harder to deceitfully use AI — not impossible.

Meanwhile, LLM creators are busy working on more sophisticated chatbots built on larger data sets (OpenAI is expected to release GPT-4 this year) — including tools aimed specifically at academic or medical work. In late December, Google and DeepMind published a preprint about a clinically-focused LLM it called Med-PaLM[7]. The tool could answer some open-ended medical queries almost as well as the average human physician could, although it still had shortcomings and unreliabilities.

Eric Topol, director of the Scripps Research Translational Institute in San Diego, California, says he hopes that, in the future, AIs that include LLMs might even aid diagnoses of cancer, and the understanding of the disease, by cross-checking text from academic literature against images of body scans. But this would all need judicious oversight from specialists, he emphasizes.

The computer science behind generative AI is moving so fast that innovations emerge every month. How researchers choose to use them will dictate their, and our, future. "To think that in early 2023, we've seen the end of this, is crazy," says Topol. "It's really just beginning."

Chris Stokel-Walker is a freelance journalist in Newcastle, UK. **Richard Van Noorden** is a features editor for *Nature* in London.

1. Pividori, M. & Greene, C. S. Preprint at bioRxiv https://doi.org/10.1101/2023.01.21.525030 (2023).
2. GPT, Osmanovic Thunström, A. & Steingrimsson, S. Preprint at HAL https://hal.science/hal-03701250 (2022).
3. *Nature Mach. Intell.* **5**, 1 (2023).
4. Glaese, A. *et al.* Preprint at https://arxiv.org/abs/2209.14375 (2022).
5. Thorp, H. H. *Science* **379**, 313 (2023).
6. Kirchenbauer, J. *et al.* Preprint at https://arxiv.org/abs/2301.10226 (2023).
7. Singhal, K. *et al.* Preprint at https://arxiv.org/abs/2212.13138 (2022).

Correction

This News feature misrepresented Scott Aaronson's views on the accuracy of watermarking in identifying AI-produced text. Human-produced text might also be flagged as having a watermark, but the probability is extremely low.

Corrected 8 February 2023