ILLUSTRATION: RAFAŁ KWICZOR

AI IS TRANSFORMING MATHEMATICS. WILL IT CONQUER THE FIELD?

Thanks to some surprising advances, mathematicians are starting to realize that artificial intelligence could radically alter their profession. **By Davide Castelvecchi**

Liam Price has no formal training in mathematics and has yet to attend university, but last month, he managed to break new ground in mathematical research — with the help of ChatGPT.

From his home in southwest England, Price got the popular artificial-intelligence tool to solve what is known as Erdős problem #1196, one of more than 1,000 puzzles that Hungarian mathematician Paul Erdős (1913–1996) collected throughout his life. Unlike other AI-generated solutions to mathematical problems, this one used a strategy that surprised specialists (B. Alexeev *et al.* Preprint at arXiv https://doi.org/q6p7; 2026).

Posting on the social-media site X, mathematician Jared Duker Lichtman at Stanford University in California drew an analogy with chess. It was, he wrote, as if AI had discovered an opening no one had thought of before because of "human aesthetics and convention".

This is one of the more remarkable examples in a string of successes for AI in mathematics. Researchers in academia and at AI companies have been making a major push to see how far the systems can go. Computers are now contributing not just brute-force calculations, but also the type of logically sound reasoning that has been the province of mathematicians since Euclid more than 2,300 years ago.

In many cases, advances have come from systems that are based on general-purpose large language models (LLMs), such as GPT, Gemini and Claude, without any special mathematical training. And — as with many areas of AI — the progress has been astoundingly fast.

The systems are still mostly rehashing techniques they absorbed from the existing literature, and that was the case with some of the solutions to other Erdős problems that Price first achieved with his collaborator, Kevin

Barreto, a mathematics undergraduate student at Cambridge University, UK.

But in cases such as Erdős problem #1196, mathematicians have started to spot glimpses of original 'thought' in the models' outputs – with the tools making surprising connections between subfields. "It is incredible," says Sébastien Bubeck, a mathematician at OpenAI in San Francisco, California. "A year ago, people thought maybe there would be some fundamental obstruction – that LLMs could never go beyond their training data."

Bubeck and others now think that it is only a matter of time before AI autonomously makes contributions at the level of the greatest mathematicians – and beyond. "I hope that perhaps by 2030, AI and mathematicians can jointly win a Fields Medal," says Thang Luong, who heads the Superhuman Reasoning team at Google DeepMind in Mountain View, California.

Innovative approaches

Erdős posed problem #1196 in 1966, and it concerns 'primitive' sets of whole numbers – meaning that none of the numbers evenly divides any of the others. (Prime numbers are the prototypical example of primitive sets.)

According to several commenters on various platforms, those who had tried solving problem #1196 had used the language of probability theory, so their efforts began by rephrasing the problem that way. GPT instead solved the problem in the original language in which it was formulated, and yet its solution implicitly established a link between numbers and probability, says Terence Tao, a mathematician at the University of California, Los Angeles.

Daniel Litt, a mathematician at the University of Toronto, Canada, says the result is "reasonably interesting", unlike previous examples in recent months of AI solutions to Erdős problems. He is relatively unimpressed by the results AI has achieved so far – and critical of the hype surrounding them. But Litt says that when it comes to future potential, it's the sceptics who have it wrong.

In fact, he says that he is puzzled that the AI systems are not already making big discoveries. Their knowledge of existing mathematics is superhuman and they have shown strong reasoning capabilities. Plus, they don't get tired or demotivated.

"Part of the mystery is, we don't know what makes a human mathematician good at math," Litt says, adding that it is unclear whether humans have some "secret sauce" that makes them uniquely creative.

Proof positive

As in many areas of AI, scaling up – particularly by adding computing power – and improving the efficiency of the algorithms will continue to make the models more powerful. One of the main limitations of AI-produced mathematics is that current models can produce proofs that are at most three or four pages long. Models tested internally at Google can already do better, says Luong, and might reach ten pages soon.

"One hundred is not within their capabilities right now, but we are working towards there, and we see improvements," Luong says – but that will be a mixed blessing, he adds. Already, human referees are stretched thin when it comes to evaluating the correctness of human-written mathematics papers, and scores of AI-generated ones are making the problem worse. "AI models can be capable of producing something that looks pretty convincing, and it takes a lot of time to figure out if there is a mistake," says Lauren Williams, a mathematician at Harvard University in Cambridge, Massachusetts.

Like many researchers in every discipline, she is worried about the proliferation of 'AI slop'. "You can find several editors at math journals who can tell you horror stories," Williams says.

Many researchers anticipate possible technological solutions for this problem as well. One common strategy is to enter the text into an LLM (which might or might not be the same one that generated the text) and ask it to check the accuracy of the proof. Price and Barreto, for example, routinely feed ChatGPT's proposed solutions back into the chatbot, have it find its own errors and make it try again until it says the proof looks correct. Many mathematicians now do this for their own text as well. Google's team has developed a specialized multi-agent AI system called Aletheia, which includes a 'verifier' module aimed at mathematical text. However, although they can be useful, LLMs still miss many errors and detect some that don't exist.

"Part of the mystery is, we don't know what makes a human mathematician good at math."

The safer strategy, say researchers, is to translate mathematics into Lean, an open-source formal (programming) language. When proofs are translated into Lean, researchers can use it to verify them. Bin Dong, a computational mathematician at Peking University in Beijing, and his collaborators demonstrated this approach in the solution of an algebra problem (H. Ju *et al*. Preprint at arXiv https://doi.org/q6p8; 2026). Meanwhile, a start-up in California called Math, Inc. used such a translator to help speed up the formalization of the Fields Medal-winning work of Maryna Viazovska, the first high-profile result to be turned into Lean.

Another strategy is to have an AI tool that formulates proofs directly in Lean or a similar language, a technique pioneered by a Google DeepMind system called AlphaProof.

The big catch is that so far, the range of mathematics that can be written in, or translated into, Lean is limited. (ChatGPT's solution to Erdős problem #1196 was a rare one that could be formalized and certified automatically, which Barreto did using Math, Inc.'s software.) Expanding Lean is a painstaking process. And teams of volunteers have been working on this. For now, says Luong, "there's only a handful of problems you can formalize, and for the rest, you need natural language".

Those limitations were in evidence in early February, when researchers did a test run of First Proof, a benchmark for AI in mathematics (M. Abouzaid *et al*. Preprint at arXiv https://doi.org/hbn2sp; 2026). Specialists in various subfields provided questions that only they knew the answers to – meaning that their unpublished work had shown the statements to be either true or false. Anybody could submit AI-generated solutions. Nearly all were produced in natural language, and only one was verified in Lean. Some solutions were verified manually, but for others, it is still unclear whether they are correct.

In June, the organizers of First Proof will put a fresh batch of questions to a variety of AI systems and will manually verify the solutions – a test that researchers say they are awaiting eagerly. The benchmark will focus on publicly accessible models because they are available to most mathematicians, says Williams, who is one of the organizers of the initiative. "We are hoping what we do will be a service to the community of mathematicians."

The consensus among researchers is that human mathematicians will be in the driver's seat, for a while at least. "What problems to study is more a judgement call. For a while, it will be humans doing it," says Mark Sellke, a mathematician at OpenAI.

"I now don't even dare to think what the future will be in five years," says Javier Gómez-Serrano, a mathematician at Brown University in Providence, Rhode Island. "Things are moving so quickly, at this point anything can happen."

Many researchers stress that throughout the upcoming disruption, it will be crucial to keep the field human-centric. They say that it would be pointless, or even dangerous, to have machines develop ideas that are beyond the grasp of even the smartest humans. "Ultimately, the goal of mathematics is to understand mathematical phenomena. For that, we need to be in the loop," says Jeremy Avigad, a mathematician at Carnegie Mellon University in Pittsburgh, Pennsylvania. "We don't want AI just outputting stuff and saying 'Yes, the theorem is true'."

Davide Castelvecchi is a senior reporter with *Nature* in London.