nature

NEWS | 12 June 2026

Humans outperform AI at this highly rigorous mathematics test

A new benchmark pitting AI against previously unseen maths problems shows systems still fall short of top human expertise.

By Davide Castelvecchi

The top-performing artificial-intelligence model scored 6 out of 10 in the First Proof set of mathematical challenges. Credit: vitacopS/Getty

Artificial intelligence has undergone its most scrupulous maths test yet. The results are in, and the AI models that took part didn't live up to the problem-solving skills of top mathematicians.

The test — part of a project called First Proof, which aims to evaluate the ability of AI to solve complex questions in mathematics — posed ten research-level maths problems to four AI systems. A jury of anonymous human specialists in the relevant mathematical fields then assessed the models' answers. This test was the first of its kind to satisfy three key conditions simultaneously: first, it consisted of research-level maths questions; second, it involved problems that did not appear in the training data; and third, it was formally graded by mathematicians. The results were unveiled on the First Proof website on 10 June.

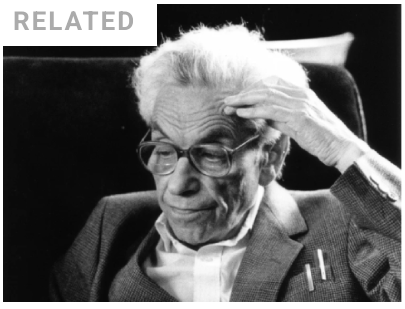
These findings follow recent AI breakthroughs in solving maths problems. Last month, for example, a chatbot made by the technology firm OpenAI, in San Francisco, California, solved an 80-year-old maths challenge set by the late mathematician Paul Erdős. The First Proof team says that future iterations of the test could help researchers to judge how useful AI models could be for mathematicians; for example, in solving problems autonomously, checking proofs or acting as research assistants.

Prove this

One important innovation of the First Proof test was that the questions had not previously been mentioned anywhere in the published literature or on the Internet — cutting the risk that the models could simply be regurgitating information they had learnt during their training. Instead, ten researchers from a broad range of mathematical specialities each provided a question that they had solved in the course of their own research but had not yet published.

First Proof ran a trial test in February with a different batch of novel problems. In that round, anyone could try their own favourite AI systems on the problems, and many groups did — but the results were not officially verified by the First Proof team. There was also no way to independently check that the AIs had not received any help from humans.

RELATED



AI cracks 80-year-old mathematics challenge — researchers are astonished

This time, First Proof ran the test itself: the team asked the models to solve problems in an entirely autonomous way, and had a group of 30 mathematicians vet the answers. "The organizers have clearly thought through the second batch more carefully to make it more controlled and systematic," says mathematician Jeremy Avigad, who heads the Institute for Computer-Aided Reasoning in Mathematics at Carnegie Mellon University in Pittsburgh, Pennsylvania.

Another rule was that the participating models had to be publicly available. This meant that Google's Aletheia — a system designed specifically for solving maths problems — and the full, unreleased version of Claude Mythos, a model made by Anthropic in San Francisco, California, could not be used. OpenAI was the only big company that participated, with its model ChatGPT 5.5 Pro.

The other systems were provided by three academic groups, from the University of California, Los Angeles (UCLA); Princeton University in New Jersey; and the Swiss Federal Institute of Technology (ETH) in Zurich. All three built 'harnesses' on top of existing chatbots, such as ChatGPT, Google's Gemini and the publicly available version of Anthropic's Claude. (A harness is an automated system that asks a chatbot a question and has the answer checked by another chatbot, often with repeated back-and-forth.)

Maths results

The ETH team's model performed the best, solving six out of ten problems with a system in which ChatGPT's answers were vetted or improved on by an 'advisory council' made up of all three major chatbots. The UCLA team, which built a harness on top of ChatGPT, was the second best, followed by the OpenAI team (ChatGPT with no harness) and Princeton (a harness using mainly Gemini 3.1 Pro as its backend).

Mathematician Johannes Schmitt, who was part of the ETH team, says that to fine-tune their system before the competition, he and his collaborators reached out to the broader mathematical community and asked for problems. "The response was amazing: within days we received 30 submitted problems, from a variety of areas of mathematics, and people were very curious and open-minded."

The ETH team also conducted a preliminary investigation into why three First Proof problems could not be solved by any of the four competitors. In some cases, it seems that the systems were "missing one more critical and unexpected idea that the human solution uses to close the last gap", Schmitt says. "For other problems, the basic

architecture of the approach was right but the systems didn't quite manage to pull through all the details."

"It is unclear whether the unsolved problems were necessarily harder than the others," says Lauren Williams, a mathematician at Harvard University in Cambridge, Massachusetts, and a member of the First Proof team. "I think the problems that weren't solved tended to just be further away — either in subject matter or proof ideas — from the sorts of things that had appeared in the literature before," she adds.

RELATED

'It is incredible': How AI is transforming mathematics

The reasoning models were also prone to hallucinating (producing factually incorrect outputs), even when told explicitly to check their references — a known issue with large language models. Willams says she was surprised by an "egregious" lack of citations in all of the AI models' answers — particularly in the case of problem 2, which several models cracked by adapting the way a similar problem had been solved by humans in the past. "Several solutions were, in some places, copying phrases from the previous paper line by line, and reusing precise notations and terminology — but never cited that paper anywhere."

Now that the First Proof problems have been published, companies that did not officially participate will probably be using them to test their own systems informally. Kevin Barreto, a mathematician at the University of Cambridge, UK, who has run his own informal maths benchmarks for AI, says he "personally would have enjoyed seeing internal models tested from the three labs, just to see where the actual frontier currently is".

doi: https://doi.org/10.1038/d41586-026-01888-9

Latest on:

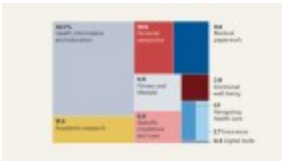
Technology Mathematics and computing

An innovative technology boosts image quality for protein structures

TECHNOLOGY FEATURE | 12 JUN 26

Gen Z scepticism towards AI is a wake-up call — universities must take it seriously

WORLD VIEW | 10 JUN 26



People are turning to AI chatbots to plug gaps in health information

NEWS & VIEWS | 09 JUN 26

Nature (*Nature*) | ISSN 1476-4687 (online) | ISSN 0028-0836 (print)