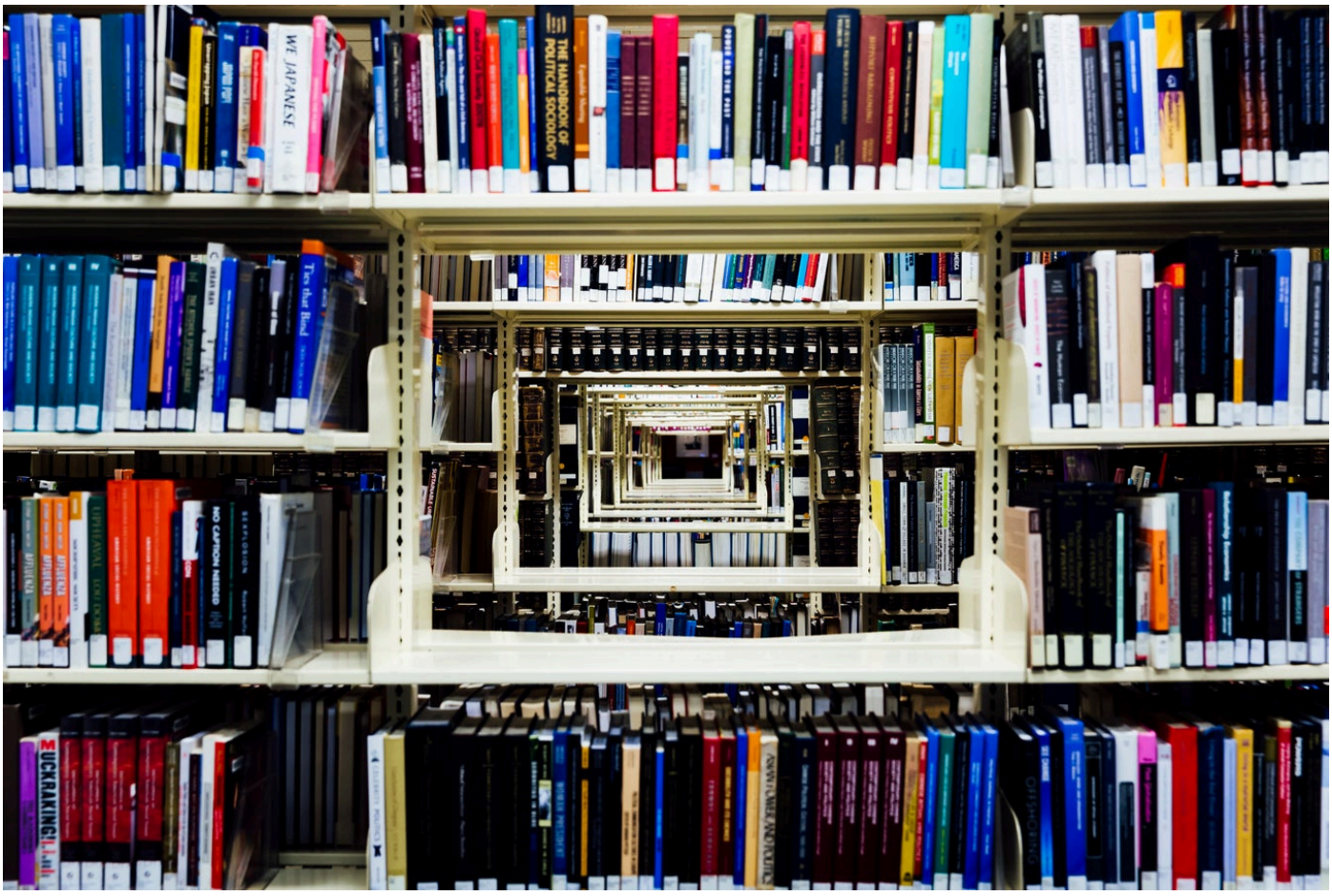
As AI keeps improving, mathematicians struggle to foretell their own future

scientificamerican.com/article/as-ai-keeps-improving-mathematicians-struggle-to-foretell-their-own-future

Joseph Howlett March 16, 2026

5 min read

First Proof is an effort to see whether LLMs can contribute meaningfully to pure mathematics research. The dust has settled on round one, and the results are surprising



Mathematicians are imagining a far-off future in which every proof they could ever invent is already right at their fingertips, provided by superintelligent AI. But how remote is that reality? And is it one they want?
Michael Kai/Getty Images

In the ongoing campaign by artificial intelligence companies to take over pure mathematics, another round is commencing.

The team behind First Proof, an effort to benchmark the ability of large language models (LLMs) to contribute to research-level mathematics, has announced its next exam. For this second round, which it plans to roll out over the next few months, the team is requiring access and transparency from any AI company that wants to participate.

This is occurring amid a sea change in mathematics research. In just the past few months, the best publicly available models have begun generating valid proofs for minor theorems of actual use for working mathematicians. To some experts, the opening round of First Proof was a pivotal moment in this ongoing story.

“We were quite impressed with how the AI models did,” says Lauren Williams, a Harvard University mathematician and First Proof team member. “The problems that we proposed really are on the forefront of what AI models—perhaps together with experts—can solve.”

First Proof grew out of its 11-person team’s own eye-opening—if sometimes frustrating—experiences with AI. No preexisting benchmarks seemed sufficient for testing LLMs as a mathematician’s assistant. In principle, an LLM could save time by proving smaller “lemmas”—intermediate propositions along a mathematician’s path to developing larger theorems of greater interest. In practice, however, such AI assists have tended to go awry.

So for their initial, “experimental” test, the First Proof team decided on 10 lemmas from papers that members had written but not yet released and then set a one-week deadline for AI companies (and anyone else) to try proving these propositions using their favorite models.

Groups from both OpenAI and Google posted their LLMs’ responses to all of the problems. Five of the OpenAI model’s proofs appeared to be correct. And Google Deepmind’s Aletheia agent seemed to get six (although experts aren’t unanimous on the validity of one of these proofs). Comparing the two models’ performances, Williams was surprised to find each had solved multiple problems that the other couldn’t. “It’s interesting to see that their capabilities are different,” she says.

“The performance was higher than I expected,” says Daniel Litt, a mathematician at the University of Toronto, who isn’t directly involved in the First Proof effort. All in all, as many as eight of the 10 problems appear to have been solved at least partially by AI. “It’s clear that capabilities have been improving really rapidly,” Litt says.

A Hazy but Hopeful Future

Litt isn’t afraid of AI’s growing mathematical prowess. “I don’t expect, five years from now, to be useless,” he says. “I actually expect to be doing the best work I’ve ever done, because I’ll have these amazing tools.” In fact, the First Proof results inspired him to pen an essay, which was widely circulated among mathematicians over the past few weeks. It presents a speculative, optimistic view of the field’s AI-infused future.

For the sake of argument, Litt imagines a hypothetical library generated by superintelligent AIs and containing every proof possible in the mathematical universe. A mere human mathematician wandering among its innumerable shelves could peruse all its volumes but could create no novel proof themself.

But that doesn't mean mathematicians would be crippled with ennui, Litt says. Far from it. "They would be unbelievably excited, and immediately get to work," he wrote in the essay. The mathematical universe is so vast, he says, that the joy is in exploring it, whether by reading and digesting a proof or writing a new one. "My job wouldn't even change at all," he says. "The job now is to try to understand things."

Even if all mathematicians agreed with Litt's decidedly utopian take on this thought experiment, the current situation is far from that lofty ideal—as evidenced by First Proof's first round. "Combined, the models solved maybe eight of the problems," he says. "But they also produced thousands and thousands of pages of garbage."

Current AIs, it turns out, are frequently wrong but convincingly confident. They'll cite a result in the literature but pretend it's stronger than it is. Or they'll bury a crucial mistake deep inside a tedious calculation, where it's easy to miss. "Students make errors, but they're definitely not *trying* to make errors," Litt says. "The models are not very honest."

This qualitative difference in the types of quantitative errors LLMs produce can make judging their answers very challenging. "One of the things we learned from this first round is how difficult it can be to check the correctness of the results," says Mohammed Abouzaid, a First Proof team member and mathematician at Stanford University. "You would almost say, 'No human who would know what all these words mean would make this mistake!'"

For round two, the team plans to outsource the task of evaluating each entry to mathematicians hired as anonymous reviewers, funded with a mix of grant money and donations from AI companies. But with no sign of the en masse mathematical onslaught slowing down, a deluge of LLM-written, subtly wrong proofs may soon overwhelm human resources. "People need to start thinking about this," Litt says. "Our institutions and the profession are not adapting to what's coming down the line."

An Unexplained Gap

The first round apparently revealed a glaring chasm between public and proprietary efforts. This would seem to challenge the notion that AI usurping human skills will democratize them—for instance, by broadening who is able to contribute meaningfully to math's advancement.

In the team's internal tests prior to posting the first round's 10 lemmas, even the best publicly available models were only able to prove two. In the weeklong test period, various groups of amateurs and professional mathematicians tried to do better by building "scaffolds," collaborative

networks of LLMs that talked to one another to suss out mistakes. But all these efforts only solved one additional problem.

A few different factors could explain why Google and OpenAI were able to (at least partially) solve eight problems versus the public's three. The companies could be using improved, unreleased versions of their LLMs or some more robust, internal scaffolds. Or the answers could rely on some undisclosed input from human mathematicians. (Google's team posted an explanation of its methodology. The team said this approach included "absolutely no human intervention"—the sort of claim that First Proof's new requirements would verify in the second round.)

That's what the second round is meant to sort out, Williams says. "This was an experiment," she says, "to get feedback from the community to figure out how to do a more formal round."

In addition to more robust human judging, this round will require that participants package models so the First Proof team can prompt them directly. "If it is not a public model, then we need to run it," Abouzaid says, "because otherwise, it's not clear what we're testing."

It remains to be seen whether OpenAI and Google will comply—or if the many other LLM companies and AI-for-math start-ups that were conspicuously absent during the first round will do so.

In the coming months, First Proof and other AI benchmarks might help foretell the still-hazy fate of mathematics—a tiny niche of the scientific world that suddenly has some of the Earth's wealthiest eyes trained upon it.

"One of our main motivations is to make sure that we can tell young people what we expect the field to look like in a few years," Abouzaid says. "And that requires understanding what these systems are actually capable of."

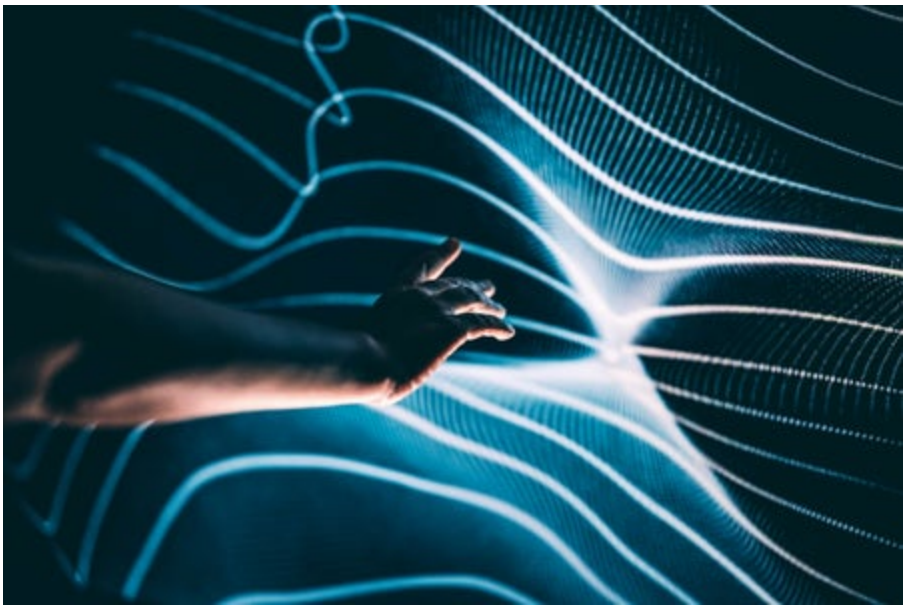
Popular Stories

PsychologyMarch 28, 2026

How to build self-control, according to psychologists

Exercising self-control doesn’t need to be unpleasant, research shows

Francine Russo

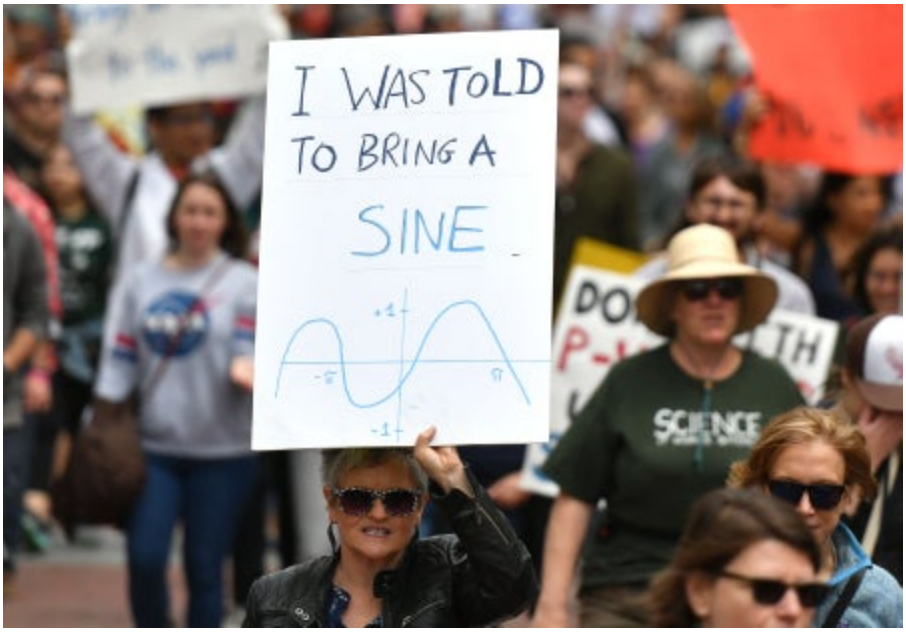


CosmologyMarch 27, 2026

We thought we knew the shape of the universe. We were wrong

Decades of data have suggested the universe is flat, much like an infinite plane. But a new analysis reveals deep flaws in that simple conclusion

Paul M. Sutter

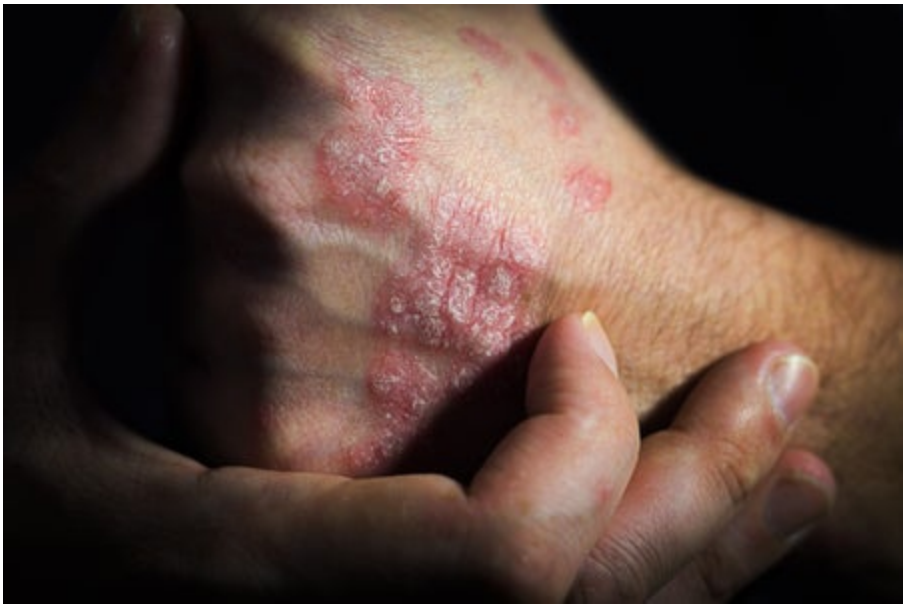


MathematicsMarch 26, 2026

Why mathematicians are boycotting their biggest conference

More than 1,500 mathematicians are demanding that their field's most prestigious meeting be moved from the U.S.

Joseph Howlett

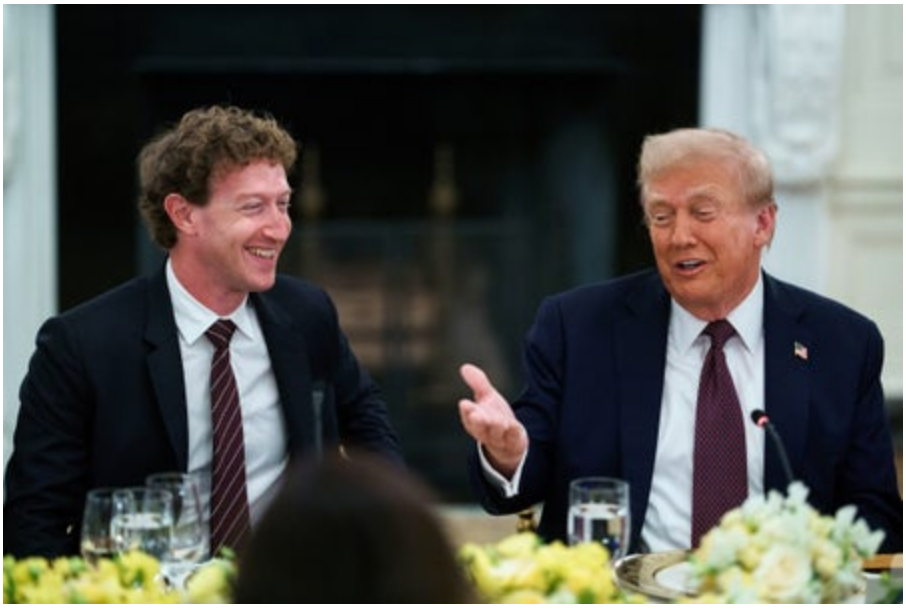


BiologyMarch 26, 2026

Skin cells remember inflammation for life. Here's why

Skin conditions such as psoriasis often flare up in the same spots throughout one's life. Now scientists think they know why

Claire Cameron



PoliticsMarch 27, 2026

Trump’s new science panel includes 9 tech billionaires—and just one scientist

There’s a glaring hole in the president’s new science and tech council

Dan Garisto, Nature magazine

Artificial IntelligenceMarch 27, 2026

An AI-authored paper just passed peer review. The scientific community isn’t ready

The arrival of AI-generated research papers marks a turning point that could radically accelerate discovery—or drown it in automated mediocrity

Jacek Krywko