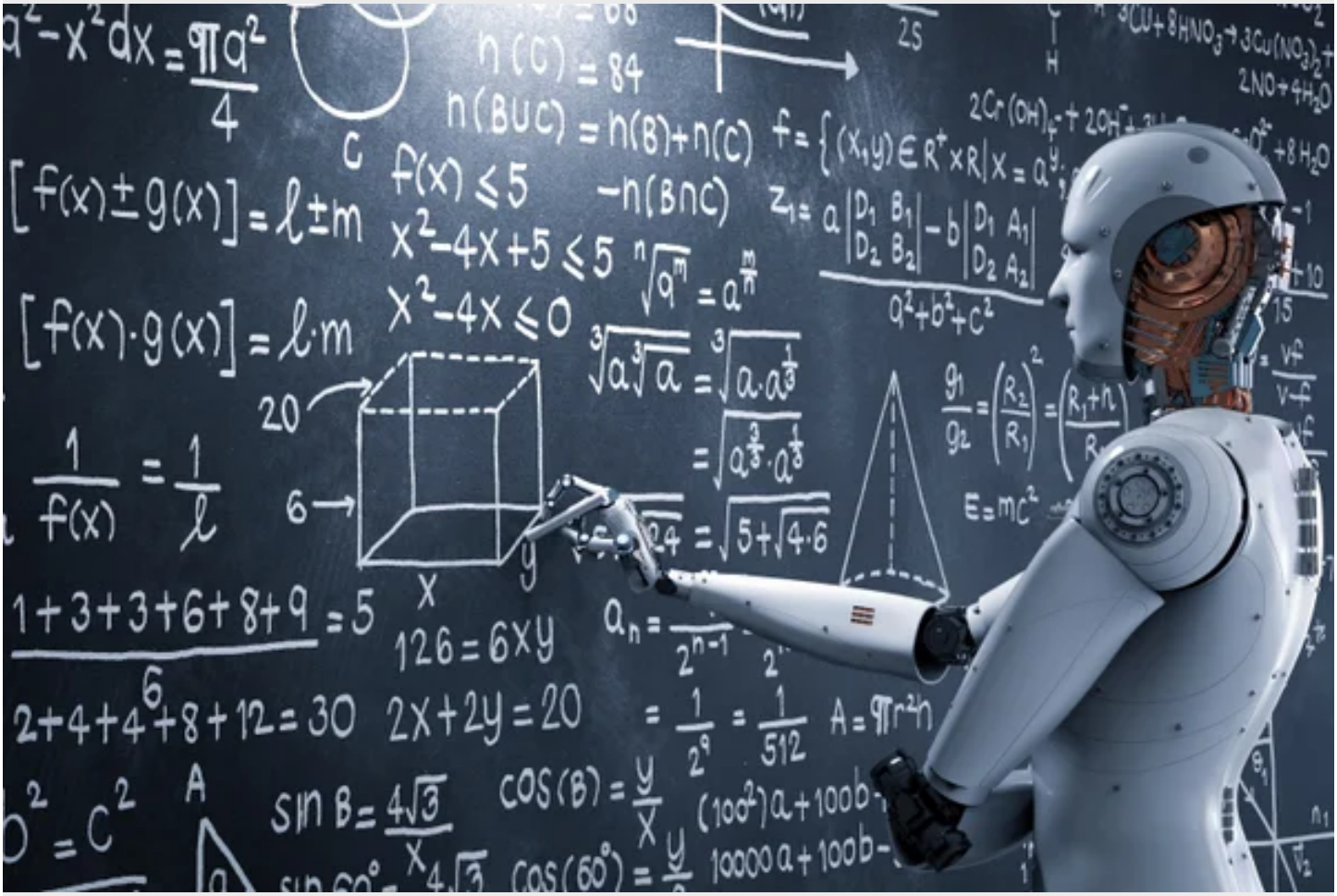
June 10, 2026 | 4 min read

AI scores a ‘C–’ on its hardest math test yet

The second batch of “First Proof” problems is meant to evaluate AI’s usefulness for research-level math. The best model got six or seven of the 10 questions basically right

BY JOSEPH HOWLETT EDITED BY LEE BILLINGS



PhonlamaiPhoto/Getty Images

Mathematics

The best-yet test of artificial intelligence’s mathematical mettle has released its first official round of results. The verdict is that large language models (LLMs) are emerging as useful—albeit deeply flawed—assistants for math research.

Organized by a team of top mathematicians, the “First Proof” project is a response to AI companies’ growing fixation on using advanced math as a benchmark for their products—regardless of whether those metrics reflect the problems professional mathematicians actually care about. Results of a pilot round in February were mixed, with companies’ opaque, internal efforts vastly outperforming their public models.

This latest batch of tests involves a broader range of math problems and more rigorous protocols for its participants—to which only OpenAI and a trio of academic groups agreed. The results were again mixed, with six to seven of the 10 problems answered essentially correctly by at least one AI. Although peak performance continues to improve, the models also churn out copious amounts of garbage as a by-product, requiring heroic interventions to sift sense from slop.

“We felt very strongly that if we’re going to be doing a public service for the greater community, we need to test publicly available models,” says Lauren Williams, a mathematician at Harvard University and member of the First Proof team. That limited the entrants to OpenAI’s ChatGPT-5.5 Pro and three models built by groups at the Swiss Federal Institute of Technology Zurich (ETH Zurich) and Aarhus University in Denmark, the University of California, Los Angeles, and Princeton University.

The team solicited problems from mathematicians across a great breadth of subject areas. It also employed expert graders who were paid to evaluate the AIs’ responses. “Grading an AI-generated solution is kind of a painful, thankless task,” Williams says. The graders assembled last week at Harvard’s Center of Mathematical Sciences and Applications for two days of intensive “peer” review—accelerating a process that, for a typical math proof, takes half a year or more.

The team considered a proof basically correct if its flaws were minor and likely to be easily patched—a standard commonly applied by math journals under the phrase “accept with minor revisions.” Some answers, though, fell on the edge of this somewhat murky threshold—thus the slight toss-up in final scores.

The results mirrored recent trends from AI’s ongoing push into math. To solve any given problem, the models are particularly adept at digging up obscure references from the literature and tirelessly mulling over well-worn mathematical techniques for possible new applications. In one case, the AI employed a strategy that the problem’s authors had identified but found too tedious to pursue, says Mohammed Abouzaid, a mathematician at Stanford University and a member of the First Proof team. But thanks to the LLM’s unbridled stamina—fueled, of course, by an expensive and unseen computing infrastructure—it powered right through.

Much of the latest progress comes from clever tricks behind the scenes. A state-of-the-art model tuned for math, such as ChatGPT-5.5 Pro (which got four to five problems right), isn’t really one model at all. It’s actually several models combined in an opaque, unified framework. A basic LLM, given an unsolved math problem, will simply evade by saying it’s too hard or will instead hallucinate a nonsensical solution or citation. Even LLMs, it turns out, can be lazy. Companies and academics counter this by using other LLMs to automatically check the base model’s work, provide feedback and push it to try harder. “You’re making the AI persist, continuing to work toward the problem,” Abouzaid says.

This “scaffolding” makes a difference. IMProofBench, built by scientists at ETH Zurich and Aarhus University, has the same ChatGPT model at its core. But that model, when stuck, can consult a “council” of other LLMs that includes Anthropic’s Claude and Google’s Gemini. This Frankenstein of models got the best score of the bunch, six or seven out of 10.

But the cost is also significant. In some cases, Abouzaid says, the legions of overlapping LLMs racked up almost $1,000 in query charges—just to get the wrong answer. Abouzaid worries about a future where grant proposals contain big budget lines for purchasing tokens from tech companies. “I truly believe this is an economic question—about research funding and research productivity,” he says.

The models also persisted in their flagrant violation of academic norms. “There were a lot of missing citations,” Williams says. “If it was a human, one might call it plagiarism.” She hopes the math community can pressure AI companies to bring their products in line with scientific ethics.

Funding for this round of tests came from philanthropic foundations, as well as unrestricted donations from major AI companies—including Anthropic, though it did not submit its model for testing.

The team is planning to release additional problems over the next several weeks for amateurs and professionals alike to try their hands and their favorite models at. They say the next official round will be in the fall.

“I’m really just excited about the fact that we have, you know, we’ve now executed something that’s much closer to a being a proper benchmark, as opposed to an experiment,” Williams says. “We tried very hard to be as objective and transparent as possible, and I think we’ve done a pretty good job.”

JOSEPH HOWLETT is a staff reporter at *Scientific American* covering physics, math, astronomy and more. He was previously a math staff writer at *Quanta Magazine*, and holds a Ph.D. in particle physics from Columbia University.

More by Joseph Howlett