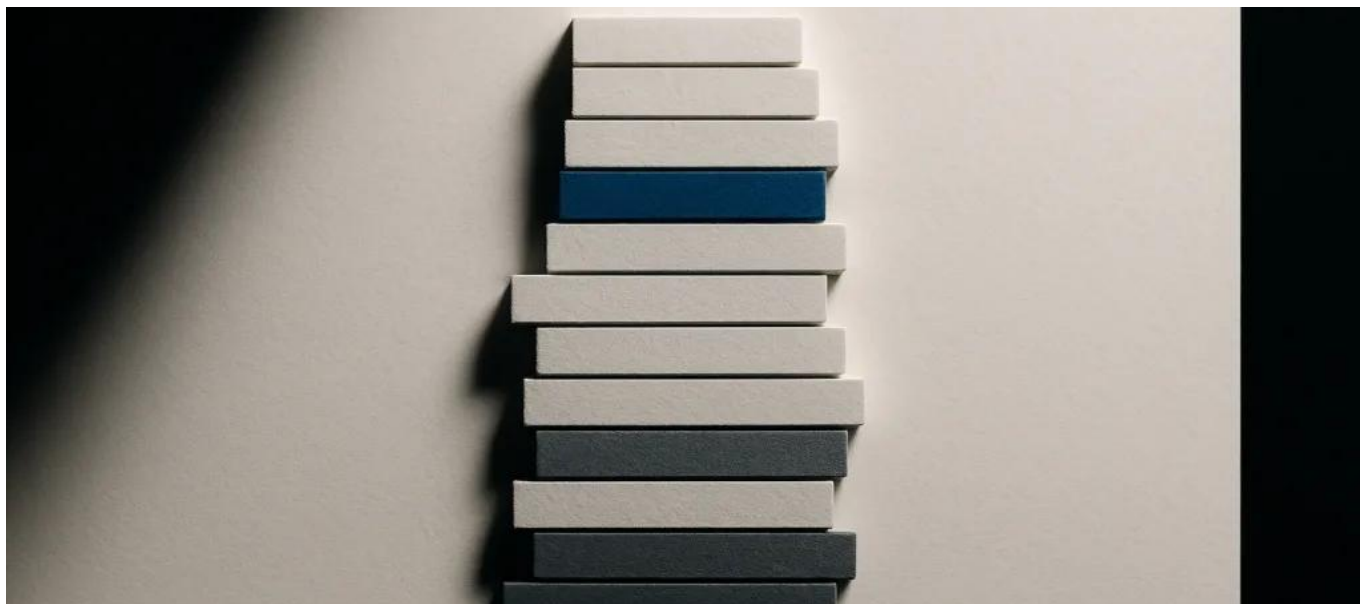




FrontierMath Erdős by the Numbers: 68 Open Problems, 2 Solved,
98% on the Benchmark Next Door



FrontierMath Erdős by the Numbers: 68 Open Problems, 2 Solved, 98% on the Benchmark Next Door

Epoch AI calls FrontierMath Tier 4 “approaching saturation” for GPT-6 Astra at 98%. The same model, handed 68 problems nobody had solved, got 2.

On September 1, 2026, Epoch AI released FrontierMath Erdős: 68 Erdős problems still open as of August 2026, each formalized in Lean, each requiring a complete machine-checked proof or disproof to count. A pre-release GPT-6 Astra solved two, a score of 3%. GPT-5.6 Sol, GPT-5.5, Claude Fable 5.1 and Claude Fable 5 each scored zero. That is the cleanest measure we have of the distance between saturating a math benchmark and doing mathematics.



FrontierMath Erdős by the Numbers: 68 Open Problems, 2 Solved, 98% on the Benchmark Next Door

2 / 68

best score on FrontierMath Erdős (GPT-6 Astra, pre-release)
Epoch AI, Sept 1 2026

98%

same model on FrontierMath Tier 4
The Epoch Brief, Sept 12 2026

0 / 68

GPT-5.6 Sol, GPT-5.5, Claude Fable 5.1, Claude Fable 5
arXiv 2609.25050, Sept 2026

~\$222

compute cost of the one disproof (problem 74), ~10 hours
Epoch AI, Sept 1 2026

The two solves were not partial credit on a graded rubric. Astra [disproved problem 74 with a counterexample and proved problem 126](#). Both results passed Lean. There is nothing to argue about in the scoring, which is why Epoch built the thing in Lean in the first place.

68 problems kept out of a pool of 652, one attempt each

Tom Adamczewski and Greg Burnham built the benchmark at Epoch AI, with problem curation by Thomas Bloom at the University of Manchester. They started from a pool of 652 open Erdős problems and kept 68, roughly 10%. Each model got one attempt per problem under identical autonomous conditions, capped at \$300 and 72 hours of working time.

Those caps matter more than they look. The winning attempts came in well under budget: problem 74 at roughly \$222 and 10 hours, problem 126 at roughly \$172 and 10 hours. So the failures on the other 66 were not failures to buy enough compute. A model that was one GPU-hour away from a proof had 62 hours left on the clock and did not get there.



The catch Epoch states itself

The 68 problems were curated for difficulty and interest, not randomly sampled from the 652. And because scoring runs through Lean, the formalization can be the failure point rather than the mathematics: a model that has the idea but cannot write the proof term scores zero, same as a model with nothing.

That second caveat cuts both ways, and I cannot resolve it from the published data. If Lean is the bottleneck, 3% understates mathematical ability and the number will move fast as formalization tooling improves. If Lean is mostly incidental and the models simply do not



have the ideas, 3% is close to the real figure. Epoch does not publish a breakdown of near-misses, so both readings survive the evidence.

Why does the same model score 98% on one math benchmark and 3% on another?

Because the two benchmarks measure different things, and only one of them was ever a capability measure in the strong sense people use it for.

FrontierMath Tier 4 problems are hard, and they are also solved. Solutions exist somewhere in the mathematical corpus, in the structure of how such problems get approached, in the distribution the model was trained on. FrontierMath Erdős problems had no solutions anywhere in August 2026, by construction. The gap between 98% and 3% is the gap between recovering a known result under difficult conditions and producing one that did not exist.

This is the same failure mode I wrote about in [AI Model Leaderboards: Myth vs Reality](#), where 78% of a Codeforces evaluation set turned out to be in training data, and in [the piece on decontamination detectors dropping to coin-flip accuracy after one round of GRPO](#). The structural problem here is not cheating. Any benchmark with published answers drifts toward being a retrieval test, and nobody can reliably tell you how far along that drift a given score sits. **Erdős sidesteps the argument entirely by using problems that had no answers to leak.**

A 4-point spread at the top, and a 2-problem spread that matters more

As of October 2026, the Epoch Capabilities Index puts GPT-6 Astra at 166, Claude Fable 5.1 at 164, and GPT-5.6 Sol at 162. Four points across the frontier. On Math-ECI, Astra leads at 170; on SWE-ECI, Claude Fable 5.1 leads at 167 against Astra's 164. OpenAI gave Epoch pre-release access to Astra, which set records across ECI, math, continual-learning and game-puzzle benchmarks. The aggregate picture is three models sitting close to each other.

Erdős tells a different story about the same three models.



FrontierMath Erdős by the Numbers: 68 Open Problems, 2 Solved, 98% on the Benchmark Next Door

MEASURE	GPT-6 ASTRA (PRE-RELEASE)	CLAUDE FABLE 5.1	GPT-5.6 SOL
Epoch Capabilities Index (Oct 2026)	166	164	162
Math-ECI	170	n/a	n/a
SWE-ECI	164	167	n/a
FrontierMath Erdős (of 68)	2	0	0
FrontierMath Tier 4	98%	n/a	n/a

Two points of ECI separate Astra from Fable 5.1. Two solved problems separate them on Erdős, which in percentage terms is the difference between something and nothing. I would not read that as Astra being dramatically stronger at mathematics. With 68 curated problems and one attempt each, 2 versus 0 is a thin result, and Epoch's own framing is careful here.

A second data point complicates the headline. Thomas Bloom [described a separate pre-release run](#) on X in September 2026: 172 attempts, roughly \$220,000 in token-equivalent cost, 5 solved questions, about 7%. Different experiment, different budget, not the 2/68 figure. But it hints at the shape of the curve. Throw roughly a thousand times the compute at the problem and the hit rate roughly doubles. That is a brutal scaling slope if it holds.

MY READ

I expect the Erdős number to climb faster than the underlying mathematical ability does, because Lean formalization is the kind of narrow, verifiable, high-signal task that post-training eats quickly. When a model hits 15 or 20 out of 68, the interesting question will be how much of the jump came from better proof-writing versus better proof-finding. Epoch is well placed to answer that, and I hope they publish the near-miss distribution.

Which of the two numbers your roadmap assumes

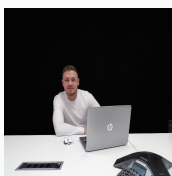
If you are making a build-or-wait decision on anything that depends on a model generating genuinely novel technical work, the 98% is the wrong number to anchor on and the 3% is closer to right. **The 3% measures output for which no answer existed before the model



produced it.** It is also a September 2026 snapshot of a pre-release model on 68 hand-picked problems, so treat it as a floor rather than a ceiling.

The practical translation for engineering work: tasks whose solutions are well represented in public code and literature are where the frontier is at 98%. Tasks that require an idea nobody has written down are where it is at 3%, costs around \$200 a shot, and runs for 10 hours. Most commercial software work sits much closer to the first category than engineers like to admit, which is why the Tier 4 numbers still predict useful productivity. But anyone pitching autonomous research, novel algorithm discovery or original proof generation as a near-term product should be asked which of the two numbers their roadmap assumes.

One more detail worth keeping. The attempts that succeeded cost \$222 and \$172 and took about 10 hours each, inside a \$300 and 72-hour envelope. Open mathematics at roughly two hundred dollars a result, when it works, is cheap by the standards of academic research funding. The problem is the 66 attempts next to them that cost real money and returned nothing, and there is currently no way to predict in advance which bucket a given problem lands in.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →