

Axiom's Agents Check the Finite Simple Groups Proof in Lean as Genie 4 Launches and a New Index Puts Open Weights 12 Points Behind

AI High Signal Digest

2026-10-28

Axiom's Agents Check the Finite Simple Groups Proof in Lean as Genie 4 Launches and a New Index Puts Open Weights 12 Points Behind

By AI High Signal Digest • October 28, 2026

AI agents machine-checked the biggest proof in mathematics, and Google DeepMind launched Genie 4. Artificial Analysis's private-heavy Index v5 shows open-weight models 12 points behind the leader. Anthropic disputes a rumor that attackers reached its Slack, and outside testers trace Qwen's Alibaba Cloud requests to RL training.

AI agents machine-check the biggest proof in mathematics

Axiom announced a complete Lean auto-formalization of the Classification of Finite Simple Groups. The proof runs about 15,000 pages and was built by more than 100 mathematicians over generations [1]. According to Omar Sar's summary, an agent swarm produced about 50 million lines of Lean in 24 days. The agents claimed tasks, wrote proofs, reviewed each other's work and merged it through GitHub. Along the way they fixed "hundreds of major errata" in the published proof [2]. The Rundown says the agents also caught a flaw in one of the original arguments [3].

This adds to the week's string of AI math claims. Anthropic says an internal Claude model found its own disproof of the Exponential Time Hypothesis on Oct. 16, separately from OpenAI. Psi says its AI made progress on five open quantum physics problems, each with a computer-checked proof [3]. On a smaller scale, Vals wrote up how ten Claude agents proved in Lean the optimal arrangement for the N=7 Thomson problem [4].

Looking ahead, After42 tested whether AI can *propose* research questions. Experts rated 45 of its 135 final questions genuinely interesting, with wide variation across subfields. Model judges matched expert consensus but not individual experts’ preferences [5].

Genie 4: a world model with real-time steering and native audio

Google DeepMind’s Genie 4 is built on Gemini Omni. DeepMind highlights: - advanced physics and live prompting to change the scene or interact with characters; - native audio, including dialogue; - Gemini reasoning that plans multi-step events so scenes stay coherent.

DeepMind presents it as a step toward safe virtual environments for testing how agents and robots navigate [6]. The research preview is rolling out to trusted testers now. It is coming “soon” to Google AI Ultra subscribers in Project Genie and to early-access partners in the Gemini API and AI Studio [7].

New benchmarks widen the open-weights gap

Artificial Analysis launched Intelligence Index v5. 60% of it now runs on private test sets or held-out answers [8]. It adds Cognition’s FrontierCode 2.0, graded on whether a maintainer would merge the change, and Terminal-Bench-Science with 70 tasks [8]. Results: - Claude Opus 5.5 leads at 58, ahead of GPT-6 Astra at 57 and GPT-6.1 Sol at 55 [8]. - The open-weights gap grew from 3 points to 12. The top open-weights model, Kimi K3.1 Pro, scores 46, but its weights are not yet released. The best models with released weights, DeepSeek V4.1 Flash 1017 and GLM-5.3, score 32, below Claude Haiku 5.5 [8]. - Effort settings matter more than in v4.3. Opus 5.5 scores 35 at low effort and 58 at max, at \$0.98 versus \$10.04 per task [9]. Among the top 10 models, GPT-6.1 Terra is cheapest at \$0.96 per task [8].

FrontierCode 2.0 deliberately gives models underspecified prompts to test their judgment. The top model scores only 45% [10]. On Cognition’s own price-performance chart, Kimi K3 and GLM 5.3 score 17% and the US open models Beam and Inkling score 3% [11]. It also reorders the field. Muse Wonder 1.0 and Gemini 4 Argon top DeepSWE v1.1 but land mid-pack here [12].

Anthropic’s security incident: corrections, and a dispute over Slack

Anthropic’s Logan Graham disputed parts of the Washington Post account covered in the last brief. He says the intrusion took several days of attempts, not 12 hours. The attackers then exploited two third-party vulnerabilities, and “nothing came close” to model weights or Claude conversations [13]. He warns that models 0.5–1 generations behind the frontier can already run attacks like this around the clock [13].

Will Depue, citing what he is “hearing,” claims the attackers (whom he calls Cathedral) got full admin access to Anthropic’s Slack, where research secrets live [14]. Graham replied that Depue’s facts are wrong but declined to go point by point [15]. Depue notes that Graham’s correction never actually mentions Slack [16]. Whether Slack was accessed is still unresolved.

Qwen’s Alibaba Cloud requests may have been learned during RL training

Following Cohere’s report, Vals found that Qwen3.8-Max and Qwen3.8-27B made unwanted requests to Alibaba Cloud, showing “no obvious attempts at data exfiltration” [17]. The requests went to an aliyuncs.com domain during ordinary webpage-fetch steps on law and finance tasks [18]. Vals’ best guess is that the server served as a web-fetch proxy during RL training and the habit carried over into the released models. It judges the behavior concerning but apparently not malicious [19]. Qwen3.7 does not show the behavior, and the dates the models hallucinated in these requests line up with likely RL training timing [20].

Anthropic moves into drug discovery

Anthropic says it has run research partnerships with six drug developers and has started its own drug discovery work [21]. Posters report that it acquired the one-person startup Mithril Medicine. Its founder, using Claude and 11 external labs, reportedly took a muscular dystrophy candidate from idea to promising mouse results in under 18 months [22].

One commentator cautions that the partner case studies were validated in vitro and that the clinical-trial planning work with Bristol Myers Squibb is very early. He also notes that Claude often orchestrates specialized tools such as protein-folding models rather than doing the science itself [23].

Compute and infrastructure

- **Epoch AI:** using customs data, Epoch estimates the US took 72% of global AI compute since 2022 and China about 10%. The US share of new compute reached 78% in the first half of 2026 [24]. China holds 18% of chips but only 10% of compute because much of its hardware is lower-performance [25].
- **Oracle and Google:** Google TPUs are coming to Oracle Cloud. Oracle says this is the first time they will be offered on a cloud outside Google [26].
- **Anthropic S-1:** SemiAnalysis cites the filing as showing fleet-wide inference revenue per megawatt rising from about \$16M to more than \$60M. Q2 revenue grew 1,368% while compute and infrastructure expenses grew 300% [27].

Models and launches

- **Liquid AI LFM3:** six roughly 2B models covering text, vision, audio, omni and decision tasks. They can be fine-tuned on one GPU and are optimized for 14 processor classes [28].
 - **Microsoft MAI-Image-2.7-Preview:** ranks #2 on AA-Image-Editing, close behind GPT Image 2.5 Sunburst. It is available now for editing only [29].
 - **Giotto:** this Swiss lab launched Alto and Solo, which it bills as “sovereign” European models [30]. Its terms reportedly bar users from publishing benchmarks without written consent [31].
-

Sources

1. X post by @axiommathai
2. X post by @omarsar0
3. X post by @TheRunDownAI
4. X post by @ValsAI
5. X post by @FanNie1208
6. X post by @GoogleDeepMind
7. X post by @Google
8. X post by @ArtificialAnlys
9. X post by @ArtificialAnlys
10. X post by @cognition
11. X post by @cognition
12. X post by @cognition
13. X post by @logangraham
14. X post by @willdepue
15. X post by @logangraham
16. X post by @willdepue
17. X post by @ValsAI
18. X post by @ValsAI
19. X post by @ValsAI
20. X post by @ValsAI
21. X post by @AnthropicAI
22. X post by @kimmonismus
23. X post by @iScienceLuvr
24. X post by @EpochAIResearch
25. X post by @EpochAIResearch
26. X post by @OracleCloud
27. X post by @SemiAnalysis__
28. X post by @liquidai
29. X post by @ArtificialAnlys
30. X post by @giotto_ai
31. X post by @Dorialexander