

OpenAI releases 722 AI-generated math papers as AI-driven cyber findings rise sharply

VC Tech Radar

2026-10-07

OpenAI releases 722 AI-generated math papers as AI-driven cyber findings rise sharply

By VC Tech Radar • October 7, 2026

OpenAI published a large set of math results from an unreleased model, Mistral previewed a 1T-parameter model it plans to open-weight, a16z data shows AI changing vulnerability discovery, and new rounds went to Valon, General Medicine and Multiply Labs.

OpenAI's math release: big claims, still under review

OpenAI published mathematical results from an internal frontier model it has not released. It says it consulted the Institute for Advanced Study's Advisory Group on Mathematics and AI on how to release them [1]. Reported figures:

- **Size:** 722 manuscripts, grouped into 372 families of related results.
- **Source:** an evaluation of about 4,000 research problems.
- **Compute:** roughly three hours of ChatGPT Pro thinking per result, on average [1].
- **What's included:** papers, proof artifacts and selected reasoning summaries [1].

Sam Altman called it “a new era of discovery” [1].

Reaction is split. Mathematician Levent Alpöge called it “the most significant moment in mathematical history” [1]. One analysis estimates about 20% of the results are disproofs or counterexamples [1]. Will Depue expects some results not to survive scrutiny. François Chollet asks whether progress in verifiable domains like math and code carries over to domains that still depend on human data [1]. Individual highlights, such as integer multiplication faster than $n \log n$, have not been independently verified [1].

Why it matters for investors: treat this as a claim about scientific capability to watch over the coming weeks, not a settled result. The open question for AI-for-science companies is whether these gains extend beyond domains where answers can be checked automatically.

Mistral Large 4: a strong preview with disputed benchmarks

Mistral previewed Large 4:

- **Size:** 1T total parameters, 49B active, natively multimodal [2].
- **Availability:** API now, open weights promised for the end of October. Mistral says it is “forged in Europe end-to-end” and is working privately with cybersecurity partners [2].
- **Training:** about 3,800 Grace Blackwells in Europe. The RL run is still going and shows no sign of saturation, and a larger model is already training [1].
- **Pricing:** \$1.36 per million input tokens and \$4.18 per million output tokens [1].

Independent results are mixed:

- **Artificial Analysis:** an Intelligence Index score of 38, the top score from outside the US and China. But cost per task is more than 4x that of open models with similar scores [1].
- **Cyber benchmarks:** Cline attributes much of the cyber lead to the model refusing fewer tasks [1].
- **Comparison with GLM:** critics note it trails GLM-5.3 on Artificial Analysis’s index [1].
- **Open-weight label:** Clem Delangue pointed out that it can’t be called the best open-weight model until the weights actually ship [3].

For investors, the takeaway is that Europe now has a credible sovereign model built on its own compute. Its cost competitiveness against Chinese open models is not yet established.

AI is producing far more security findings

a16z data shows reported critical vulnerabilities at 21 major software companies, including Apple, AWS, Microsoft and Google. They never exceeded 100 a month in four years, but have topped 600 a month since spring [4]. About 87% of exploited bugs are now attacked on or before the day they become public, up from 23% in 2020 [5].

Armadin. Kevin Mandia, who built Mandiant, says Armadin has found more than 90 zero-day vulnerabilities at customer sites since January, all in production systems. It works as an outside “black box” attacker with no source code, using models post-trained with real exploit developers [6]. By his own account,

Armadin has no sales force or go-to-market strategy yet, and needs funding to build them [7].

OpenAI’s internal effort. Greg Brockman said OpenAI moved 25% of its production engineers onto security, using models to find holes. That work eventually “saturated” at the critical issues Astra could find. OpenAI is now building an automated “defense factory” to rerun the process with each new model [8].

The investable idea is continuous, model-driven offense and defense that repeats with every new capability release.

Agents meeting businesses: a protocol, and a privacy problem

Meta and Sierra announced the Personal Agent Protocol, an open standard for how personal agents interact with businesses. Partners include Genesys, Instinct, Shopify, Stripe and Walmart [9]. Sriram Krishnan compared it to OAuth. He asked how Amazon, airlines and other aggregators will respond if they can no longer own the final user experience [10]. This picks up the question from the last brief about where network effects end up in agents.

In the same period, Time reported that Meta’s Muse keeps hourly-updated dossiers on users and the people they mention. Lessons learned are shared across Muse agent instances, even though Meta describes each user’s virtual machine as isolated [11].

How engineering work is changing

The Pragmatic Engineer reports that the most productive engineers now run 5–10 agent sessions in parallel and have stopped writing code by hand [12]. Agent-authored pull requests on GitHub outnumbered human-authored ones in August [12]. At Linear, agents have created more issues than humans since July [12].

Costs are a counterweight. Uber’s token usage rose while its costs stayed flat, thanks to open models and smart routing between models [12]. On capacity, Google’s Amin Vahdat says Google aims to double effective token-serving capacity every six months, with as much or more of that coming from software as from hardware [13].

Support-agent benchmark. Gorgias, a vendor that ranks fifth in its own test, benchmarked 13 AI support vendors on 212 live stores. The top five resolve 64–75% of conversations with no human involved, and the median resolves 48% [14].

Deals

- **Valon** raised a \$150M Series D at a \$2.3B valuation. It signed more than \$200M in deals within six months of starting software sales. It first ran its own mortgage servicer until it reached 3x industry efficiency. One of the largest US servicers is moving 4M loans onto its platform, nearly 10% of the market [15].
- **General Medicine:** a16z led a \$120M Series B for this “general store for healthcare.” It combines pricing, clinical guidance and a clinician marketplace, and is also building for enterprise customers. It reunites the PillPack team [16].
- **Multiply Labs (YC S16)** raised a \$75M Series B for robots that manufacture gene therapies and mRNA drugs. Its robots run inside AstraZeneca and Legend Biotech facilities [17].
- **Type One Energy** raised \$200M for fusion. Its CEO says this covers about half of a 400-MW plant targeted for 2034, built through a project-specific supplier network [18].
- **Outer Space** launched with an \$8M pre-seed led by Upfront. It builds outdoor structures that generate and store solar power for homes [19].

Investor debate and signals

Seed checks as options. Leo Polovets responded to Menlo’s view that seed checks are options to be sized up later: “No seed investment should be an option bet.” His fund does little follow-on investing by design [20].

Software moats. Naval argues models are software’s last moat, as AI decompiles and recodes products. He expects more software to retreat to the server [21]. Amjad Masad predicts all software will soon be “de facto open-source” [22].

Robotics. Generalist says its Gen 1.5 robot model improvised with unfamiliar tools, using a banana as a brush and a dustpan with both hands. The company says it was not explicitly trained to do this [23].

Sources

1. [AINews] Quasi-Riemann-Hypothesis: OpenAI publishes 722 math papers solving 90 of the top 500 open math problems; “the most significant moment” in >100 years of mathematics
2. X post by @MistralAI
3. X post by @ClementDelangue
4. X post by @a16z
5. X post by @a16z
6. X post by @a16z
7. Building Cyber Defense for the Agentic Era

8. X post by @a16z
9. X post by @btaylor
10. X post by @sriramk
11. r/artificial post by u/SpiritRealistic8174
12. The state of the tech industry in 2026
13. Google's AI Infrastructure Chief, Amin Vahdat, on the Physics & Economics of Frontier AI
14. How Much of Your Customer Support Can AI Really Resolve? The Best Get About 70%. The Median Is 48%. Here's the Real Data From 13 Vendors
15. X post by @a16z
16. X post by @a16z
17. X post by @ycombinator
18. r/Futurology comment by u/Gari_305
19. X post by @jiakeliu
20. X post by @lpolovets
21. X post by @naval
22. X post by @amasad
23. Building AI for the Physical World with Generalist