

OpenAI Publishes Hundreds of Math Results From an Unreleased Model; Mistral Large 4 Enters the Open-Model Race

AI High Signal Digest

2026-10-07

OpenAI Publishes Hundreds of Math Results From an Unreleased Model; Mistral Large 4 Enters the Open-Model Race

By AI High Signal Digest • October 7, 2026

OpenAI released a large set of mathematical manuscripts produced by an internal model, including a reported zero-free strip for the Riemann zeta function. Mistral previewed Large 4, Google shipped EmbeddingGemma 2, and new findings put agent oversight tools under scrutiny.

OpenAI releases a large body of math from an unreleased model

OpenAI published “a broad range of new mathematical results produced by an internal frontier model” on GitHub. It says it consulted the Institute for Advanced Study’s independent Advisory Group on Mathematics and AI on how to release them [1]. One widely shared summary gives these figures: 722 manuscripts in 372 families of related results, drawn from about 4,000 research problems. It says the standard procedure averaged three hours of ChatGPT Pro thinking compute per result. The release includes papers, proof artifacts and selected reasoning summaries. The model itself is not released [2].

Two results drew the most attention from specialists: - **Riemann zeta**. Mathematician Alex Kontorovich says the previous best zero-free region got thinner higher up the imaginary axis, and that the release instead proves a zero-free *strip*. He says a human who did this would get “an instant Fields Medal” [3]. This is a partial result, not a proof of RH. - **Integer multiplication**. A linked preprint in the repo is titled “Integer multiplication below $n \log n$ ” [4].

One reader estimates that about 20% of the results are disproofs or counterexamples. He argues this undercuts the view that AI math progress is mostly brute-force search [5]. Others played down the compute figure. @teortaxesTex says three hours of Pro thinking “is not much” and concludes that frontier math is “about as hard as SWE hobby projects” for this model [6]. SWE-bench co-author Ofir Press says coding has not yet had a comparable moment. His test would be agents rewriting ffmpeg or SQLite to run $5\times$ faster on half the memory [7].

Mistral Large 4: real gains, disputed superlatives

Mistral announced Large 4 (“Le Chonk”). It is a natively multimodal model with 1T total and 49B active parameters, available by API now, with open weights due at the end of October. Mistral calls it the best open-weights model from the US or Europe on aggregated benchmarks and says it beats closed frontier models on visual grounding [8].

Independent numbers are mixed: - **Artificial Analysis** scores the preview at 38 on its Intelligence Index, close to GPT-6 Luna (38) and DeepSeek V4.1 Flash (39). That makes it the strongest model from outside the US and China. It also scores 82% on CyberGym-E2E-AA. But cost per task is \$1.13, more than $4\times$ similar open models such as GLM-5.3-Flash at \$0.25. A launch discount halves that to \$0.57 for two weeks [9]. - **Yuchen Jin** points out it trails GLM-5.3, and even GLM-5.3 Flash, on that index, and calls this an “eval crisis” [10]. - **Clément Delangue** notes the model “can’t be the best open-weight model if you’re not open-weight yet” [11]. - **Surge’s blind human evaluation**, commissioned by Mistral, has professional engineers rank coding quality. It placed Large 4 first among open-weight models and second of five overall, behind Opus 5 [12]. - **Cline** says Large 4 beats Opus 5.5 and GPT-6 Astra on cyber benchmarks mostly because it refuses fewer tasks. About 40% of Opus and Astra tasks were blocked by their safety filters [13].

Coming a day after Reflection’s Beam, this prompted Jin to say both Western models reached roughly GLM-5.2 level. He speculates the remaining gap exists because Chinese labs can distill Anthropic and OpenAI models and Western labs can’t [14]. In China, Ant Group’s Ling 3.1 Flash (560B total, 25B active, 1M context, weights “soon”) scored 41 on Artificial Analysis’s index, up from 20 for its predecessor [15]. Its price is higher, at \$0.99 per task [15].

Cheap “decision models” become a product category

OpenAI put its Decisions API in public beta for choosing a model, tool or action in near real time. It says decisions are up to $10\times$ faster than calling GPT-6 Luna through the Responses API [16]. Pricing starts at \$0.10 per million input tokens, with no output or cache charges [17]. Theo sees value in using it to expose the right tools to an agent, but calls using it to decide what “level of intelligence” a task needs “absolutely useless” [18].

Competitors moved the same day. Perplexity’s open-weights pplx-decider-v1.1-27b tops Hugging Face’s new Decision Index 0.3 at \$0.02 per million input tokens, half the price of v1 [19]. Vals AI tested TypeSafe’s Jev on SEC-filing claim verification. Jev tied GPT-6 Astra at 97.5% accuracy for about \$0.025 per 1,000 cases versus \$12.29 [20]. But it ranked last on a LegalBench slice [21].

Google ships on-device embeddings and a cheaper image model

EmbeddingGemma 2 is Google’s first natively multimodal open embedding model, released under Apache 2.0. It maps images, video, audio and code into one space [22]. At 740M parameters it uses about 191–567MB of active RAM and has an 8K context window [23]. Nano Banana 2.1 is billed as beating the previous Pro image model at \$0.034 per image versus \$0.134 [24].

Can we trust agent oversight and evals?

- **METR** found a bug in the transcript viewer of Inspect, a popular evaluation framework. It would have let an agent show reviewers a fake or edited transcript [25]. The attack was a MathJax JavaScript injection that could be triggered from the agent’s reasoning. METR has not seen it exploited, and Inspect was patched within a day [26]. METR’s lesson: treat every AI output as untrusted input to the systems that supervise it [27].
- **CI-aware bench** measures whether models notice AI control interventions in their text. Most models were near chance in spring. GPT-6 Astra “nearly saturates it now” [28].
- **AutomationBench Verified** audited all 600 public tasks in Zapier’s benchmark. Reviewers confirmed and fixed 206 buggy verifiers. Re-grading 1,235 Kimi K3 runs changed 27.9% of grades [29].
- **PZero Research** says frontier models’ experimental research taste has doubled about every 3 months since December 2025, and that Opus 5.5 now beats its expert baseline. Most of those experts have not worked at a frontier lab [30].

Anthropic expanded its Cyber Verification Program. Verified security professionals get access to Claude Mythos 5.1, Opus 5.5 and Sonnet 5.5, and new tiers cover authorized offensive work such as penetration testing [31].

Epoch AI points to a commercial problem for open-weight labs. After Zhipu released GLM 5.3 Flash, 26 other OpenRouter providers were serving it within three weeks. Zhipu’s share of the model’s tokens fell from 88% to 22% [32].

Sources

1. X post by @OpenAI

2. X post by @kimmonismus
3. X post by @AlexKontorovich
4. X post by @AcerFur
5. X post by @nrehiew__
6. X post by @teortaxesTex
7. X post by @OfirPress
8. X post by @MistralAI
9. X post by @ArtificialAnlys
10. X post by @Yuchenj_UW
11. X post by @ClementDelangue
12. X post by @HelloSurgeAI
13. X post by @cline
14. X post by @Yuchenj_UW
15. X post by @ArtificialAnlys
16. X post by @OpenAIDevs
17. X post by @reach_vb
18. X post by @theo
19. X post by @perplexitydevs
20. X post by @ValsAI
21. X post by @ValsAI
22. X post by @Google
23. X post by @Google
24. X post by @_philschmid
25. X post by @METR_Evals
26. X post by @METR_Evals
27. X post by @METR_Evals
28. X post by @jonasgeiping
29. X post by @shaped
30. X post by @pzeroresearch
31. X post by @AnthropicAI
32. X post by @EpochAIResearch