

OpenAI Publishes 722 Math Manuscripts From an Unreleased Model—Not All Verified

2026-10-07 / 00:20:31

“The fraction you can check is the fraction you can trust, and the rest is fluent prose.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

OpenAI posted 722 mathematics manuscripts from a model nobody outside the company can run, and some come with machine-checked proofs while others don't. That split sets up the day: who gets to verify, who gets to set the rules for agents, and who gets to say stop to a data center.

- [OpenAI's announcement](#) and [its post on sharing AI progress in mathematics](#) give the 722 manuscripts and 372 result families, with Lean proofs on many and others unverified; [The Verge](#) covers the reaction, and [Ethan Mollick's post](#) looks beyond math.
- [Jeff Weinstein](#), [Sierra](#), [CNBC via Techmeme](#) and [SiliconANGLE](#) cover the Personal Agent Protocol from Meta, Sierra, Stripe, Shopify, Walmart and others, with the spec itself still to be read.
- [Axios](#) and [Techmeme](#) report Mistral Large 4's specs and staged release, weights due October 27, while [Nathan Lambert's Interconnects piece](#) argues GLM-5.3's open release hasn't produced the attacks Anthropic warned about.

- [OpenAI's Decisions API video](#), [Strands Decider](#) and [Simon Willison's plugin](#) show two small, constrained-choice models arriving the same day.
- [CNBC](#) and [Techmeme](#) report Finland's stop order to Google by October 23; [the FT](#) via [Techmeme](#) reports SpaceX's roughly \$40 billion debt raise for Nvidia chips; [CNBC](#) carries Temasek's warning.
- [Amjad Masad](#), [Naval](#) and [eigenrobot](#) on AI decompilation, including an unconfirmed Adobe claim.
- [Reuters](#) on South Korea's bank-hack statement and [Toby Walsh in The Guardian](#) on OpenAI in Australia.
- [The Verge](#) via [Techmeme](#) on Common Sense Media's rating of ChatGPT for Teens and [NBC News](#) on Oversight Board advice.
- Preprints: [power draw and covert compute](#), [auditable claims about AI agents](#) and [AdvSim2Real](#).
- AI Engineer talks: [Stigg on metering](#), [Resolve AI](#), [Langfuse](#) and [Tessl](#).

SEGMENTS

[00:00:04](#) 722 manuscripts, some checked and some not

[00:04:47](#) The Personal Agent Protocol

[00:07:46](#) Mistral Large 4 and the staged release

[00:11:06](#) Models that only pick from a list

[00:13:41](#) Finland, Google, and borrowed money

[00:15:43](#) Shorter items: decompilation, bank hacks, teen safety, and papers

Transcript

1. Lenar Kess 00:00:04

Say you're a mathematician, and you open your laptop this morning to find that a lab has posted 722 manuscripts overnight, grouped into 372 families of related results, all written by a model that nobody outside the lab can use. How long does it take you to read them? Who tells you which ones are right? And if some of them come with proofs a computer has checked, and some don't, how do you decide which pile to open first?

- [x.com](#)
- [openai.com](#)
- [theverge.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [techmeme.com](#)
- [siliconangle.com](#)
- [techmeme.com](#)
- [axios.com](#)
- [techmeme.com](#)
- [youtube.com](#)
- [strandsagents.com](#)
- [x.com](#)
- [cnbc.com](#)
- [techmeme.com](#)
- [techmeme.com](#)
- [cnbc.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [reuters.com](#)
- [theguardian.com](#)
- [techmeme.com](#)
- [nbcnews.com](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [youtube.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [youtube.com](#)

That's the arithmetic that interests me. The lab can produce these faster than the field can referee them. So the count matters less than the sorting: which ones have a proof a machine verified, and which ones are a very fluent author telling you it's right.

3. Lenar Kess 00:00:45

That's OpenAI's release from last night, and it's our lead. After it we've got a protocol from Meta, Stripe, Shopify, Walmart and Sierra for how personal agents deal with businesses. Mistral's trillion-parameter open-weight model comes next, with a delayed weights date. Then two small models that only pick from a list, and Finland telling Google to stop building, set beside a forty-billion-dollar debt raise for Nvidia chips. We finish with shorter items on AI decompilation, South Korea's bank hacks, and a teen-safety rating. There's also a paper about measuring hidden compute from power draw.

4. Damra Vol 00:01:22

Start with what's in the release, because the headlines will reach for the word breakthrough.

5. Lenar Kess 00:01:26

OpenAI's post is titled "Sharing AI progress in mathematics." The numbers are 722 manuscripts in 372 result families, produced by an unreleased frontier model. Many come with proofs written in Lean, which is a proof language a computer can check mechanically, and others are still unverified, which is OpenAI's own description. They also say they consulted an independent advisory group on mathematics and AI. The Verge, in Robert Hart's piece, calls it another batch, and says it extends a run of results that have both impressed and unsettled parts of the mathematical community while raising questions about research ethics.

6. Damra Vol 00:02:06

Lean changes the character of the claim. A Lean proof is the one place in all of this where nobody has to trust anybody. If the file compiles, the theorem follows from the axioms, and whether the model was brilliant or lucky stops mattering. The unverified pile is the opposite case.

7. Lenar Kess 00:02:24

Even a compiled file only certifies that the statement as formalized got proven. Whether that formal statement matches what a mathematician means by the problem is a separate check, and a person still has to make it.

8. Damra Vol 00:02:36

Which is where the advisory group comes in, and consulted is a weak verb. It tells you somebody outside looked at the process. It doesn't tell you anybody refereed 722 manuscripts. I wouldn't read it as a certificate.

9. Lenar Kess 00:02:51

Consulted is the word OpenAI used, and nothing in the material says that group checked any individual result. So I'd call the unverified manuscripts what OpenAI calls them, unverified, and leave breakthrough for the ones with proofs behind them.

10. Damra Vol 00:03:06

The other detail is the model itself. It's unreleased, so nobody outside can rerun it, probe it, or ask it what it meant on line forty of a proof. The artifacts are public and the process that made them is sealed.

11. Lenar Kess 00:03:21

That's a little different from yesterday. On Tuesday we had the preprint crediting Claude with an algorithm that beats the barriers for 3SUM and all-pairs shortest paths, and that was one result with named human authors who could be pushed on. This is a batch with a lab's name on it.

12. Damra Vol 00:03:36

One result you can argue about. Seven hundred and twenty-two you can only triage. And triage has politics. My guess is the famous problems get read first, and everything else sits there with a logo on it, waiting for a volunteer with a free semester.

13. Lenar Kess 00:03:52

Mollick's reaction pointed outside math, toward fields like economic sociology. I'd keep that separate from OpenAI's claims, since it's a reaction and not a result, but it points at what comes next.

14. Damra Vol 00:04:05

It gets stranger outside math, because math is the one field with a verifier. Swap in sociology and the same batch of manuscripts has no compiler to tell you which ones hold. So the split in this release is a preview of the general problem: the fraction you can check is the fraction you can trust, and the rest is fluent prose.

15. Lenar Kess 00:04:25

The Verge mentions research ethics, though the sources don't go far on it.

16. Damra Vol 00:04:30

Right, so these are my questions and not anyone's findings. Who's the author on a paper a model wrote? What does a mathematician owe a reader when she posts something she hasn't verified herself? Those norms are going to be written by whoever posts the most manuscripts, and today that's a company.

17. Lenar Kess 00:04:47

On Tuesday afternoon a group of companies published the Personal Agent Protocol, PAP. Meta and Sierra are in it, along with Genesys, Instinct and Rocket, and Shopify, Stripe and Walmart round out the names in the announcements. SiliconANGLE says Bret Taylor, Sierra's co-founder, introduced it. CNBC's description, via Techmeme, is that it's meant to standardize and secure how AI bots interact with businesses. It arrives about a month after Meta's Muse personal agent turned into a viral hit.

18. Damra Vol 00:05:19

Which way does the arrow point? A personal agent is mine, acting for me, talking to a retailer or a payment processor. So this is a protocol for the buyer's robot knocking on the seller's door, and the people writing it are mostly the sellers.

19. Lenar Kess 00:05:34

Shopify, Walmart and Stripe are on the receiving end, and Meta ships the personal agent. Sierra is described as an enterprise AI firm, so it sits on the business side too.

20. Damra Vol 00:05:45

Then my first question is who's missing. Anthropic, OpenAI and Google aren't among the names in anything I've read. Meta's agent plus the merchants is a coherent bloc, but the labs behind most of the agents people actually use aren't in the announcement.

21. Lenar Kess 00:06:01

It also has founding members and a draft, so it isn't a settled standard. Jeff Weinstein's post and Sierra's own announcement both name the group, and neither spells out what PAP requires of an agent: how it identifies itself, what a business can demand of it, what happens when something goes wrong.

22. Damra Vol 00:06:18

Yesterday's story about the MCP agent-to-agent injection flaw ended on the line that anything passed from a language model to your tool should be treated like input from a stranger on the

internet. A protocol where my agent talks to a merchant's agent is that stranger, in both directions. So when they say secure, I want to know what that means.

23. Lenar Kess 00:06:40

It could be three different jobs. Authenticating tells the merchant which agent is calling. Bounding limits what the agent can commit you to, like a purchase or a refund. Sanitizing is about text on a product page redirecting your agent, which is the problem one of today's papers goes after.

24. Damra Vol 00:06:57

There's a plain commercial angle too. A merchant that helps define how agents identify themselves also helps decide which agents are welcome. Standards in commerce tend to start as interoperability and can end up as gatekeeping. I'm not saying that's the plan. I'm saying a protocol co-written by Walmart and Shopify naturally carries their interest in knowing who's shopping.

25. Lenar Kess 00:07:20

The last time agent commerce came up here it was PayPal's projections and the mess of product catalog data. A protocol doesn't fix a catalog that lists the wrong size.

26. Damra Vol 00:07:30

[chuckle] A standard can't make the shoe fit. But it gives you somewhere to file the complaint.

27. Lenar Kess 00:07:36

What exists today is a real group, a month-old consumer agent that gives it a reason to exist, and a spec we should read before anyone says anything about its security model.

28. Lenar Kess 00:07:46

Mistral announced Mistral Large 4, nicknamed "le Chonk." According to Axios, it's a one-trillion-parameter multimodal model with 49 billion parameters active at a time, which implies a mixture-of-experts design. It was trained on 4,000 Nvidia Grace Blackwell GPUs over two months in Mistral's own European data centers. The release comes in tiers. A moderated API comes first, then a version with fewer restrictions and broader cybersecurity capabilities shared with select partners for testing, and finally weights on October 27 after more reinforcement learning and safety testing.

29. Damra Vol 00:08:25

The staging is what stands out. Three tiers: public API with moderation, a partner build with the cyber limits loosened, and weights later. It's an open-weights release run the way a closed lab runs a rollout.

30. Lenar Kess 00:08:38

Pierre Stock, Mistral's VP of science, told Axios the company believes it's the world's best open-weight model and that it can outperform closed models on certain tasks. Then he said, quote, "We're not there yet on the frontier."

31. Damra Vol 00:08:53

I like that he said it. It also means every number here comes from Mistral by way of Axios, and there are no independent evals before the weights go out. Today we have a spec sheet and a claim.

32. Lenar Kess 00:09:04

The pitch is ownership. Reflection's CEO Misha Laskin, in the same Axios piece, said, "Once you're spending that amount on intelligence, you want to move from renting it to owning it yourself." Stock listed what customers keep asking about: control over data and intellectual property, business continuity, customization and cost. Reflection's Beam was yesterday's story, and this is the same argument from a second company.

33. Damra Vol 00:09:31

Stock also said, "I don't want to live in the future in which any oligopoly controls closed access to this type of intelligence." Notice the training happened in Mistral's own European data centers. For this company, ownership is geography as well as license.

34. Lenar Kess 00:09:47

Against that, Nathan Lambert's Interconnects piece argues that GLM-5.3's open release hasn't produced major attacks despite Anthropic's warnings about its Mythos-level cyber risk, and that this undercuts calls to ban open models. His subtitle is about open weights, ideology and acknowledging trade-offs.

35. Damra Vol 00:10:07

That's one model and a short window, so an absence of attacks is thin evidence in either direction. It's still evidence. And Mistral's own staging reads like a hedge against that argument being wrong, because the cyber-capable build stays with select partners until the weights go out.

36. Lenar Kess 00:10:24

Stock's defense is that openness helps defenders more. In his words, "If we open weight as many models as possible, including ML4 — which is among the best models in the world on cyber — then you accelerate the defense part way more."

37. Damra Vol 00:10:39

Maybe. Attackers need one working exploit and defenders need to cover everything, so I can't tell you which side that asymmetry favors, and Mistral can't either until someone measures it. What I'd like to see is the actual cyber evaluation on the restricted build, run by somebody who isn't Mistral.

38. Lenar Kess 00:10:58

October 27 is when the weights, the license and the first outside benchmarks arrive, and that's when the Axios numbers get tested.

39. Lenar Kess 00:11:06

OpenAI introduced a Decisions API running on GPT-6 Luna. You give it text or an image, a question, and a fixed list of allowed answers, and it returns one of them. OpenAI's video claims roughly a tenfold latency reduction and server-side responses under 100 milliseconds. Those are OpenAI's numbers from its own demo.

40. Damra Vol 00:11:28

Offering only A, B or C sounds like a downgrade until you see what it removes, which is the generation loop. There are no reasoning tokens and no prose. It's a classification with a multimodal front end.

41. Lenar Kess 00:11:41

The demos cover a lot of ground. It sorts inbound text into sales leads or support tickets, and it turns frames from a driving simulation into steering decisions. It pairs with GPT-Live-1 so a voice conversation keeps flowing while the decision happens separately, and a Hugging Face Microduck robot uses it to turn camera frames into directional commands. The presenter says it complements full reasoning models and doesn't replace them.

42. Damra Vol 00:12:07

The driving simulation clarifies what it's for. At a high frame rate, a reasoning model is still thinking about the last frame while the car is already in the ditch. A fixed menu lets you run at the speed of the camera.

43. Lenar Kess 00:12:20

The same day, Strands published Decider, a small open-source decision model of two billion parameters. It got 194 points and 48 comments on Hacker News. All I've seen of it is the listing, so I won't quote benchmarks. Simon Willison also shipped a plugin for his command-line tool that talks to the OpenAI one, the same day.

44. Damra Vol 00:12:42

Two independent releases with the same design, one hosted and one small and open. On Monday we talked about auditing small decision models for routing savings, and this is that idea turning into a product category with names on it.

45. Lenar Kess 00:12:58

The trade-off between them is plain enough. The hosted one gives you OpenAI's multimodal understanding and language coverage without running anything. A two-billion-parameter model you can put next to the robot, without a network hop in between.

46. Damra Vol 00:13:12

Most steps in an agent loop are already menu choices. Which tool, did the test pass, is this page a login form. Today people pay reasoning-model prices for them. My bet is the first thing builders do with these is carve those steps out. But a constrained answer is only as good as the menu: if the right answer isn't on the list, the model still picks one, and it picks confidently.

47. Lenar Kess 00:13:36

So the first test I'd run is whether a none-of-these option gets chosen when it should.

48. Lenar Kess 00:13:41

Finland has ordered Google's subsidiary Tuusula Oy to halt construction on data centers in Muhos and Kajaani by October 23, pending environmental impact assessments. CNBC's headline calls Finland the Texas of Europe, and the story notes that Google had committed to the buildout about a month earlier.

49. Damra Vol 00:14:01

A month from commitment to stop order. The only reason in the reporting I've seen is the environmental impact assessments, so I'll stay there. But look at what an assessment does to a schedule. It turns a construction timeline into a regulatory one, and the date stops being Google's.

50. Lenar Kess 00:14:18

Yesterday it was grid deals and political backlash over power demand. Finland's version is a permitting office.

51. Damra Vol 00:14:25

Which is a different kind of brake. Chips and electricity are supply problems, and money can lean on them. An agency saying finish the paperwork before you pour more concrete is a process problem, and money leans on it much less.

52. Lenar Kess 00:14:39

On the money side, the Financial Times reports, citing sources, that SpaceX is seeking about \$40 billion to buy Nvidia chips, roughly \$10 billion in bank loans and \$30 billion in investment-grade debt, in a deal led by Apollo. It's sourced reporting and not a confirmed deal.

53. Damra Vol 00:14:58

Thirty billion of it investment-grade means somebody is underwriting GPUs the way you'd underwrite heavy equipment. And the Finland story is a reminder that the building the equipment goes in can be stopped by an agency.

54. Lenar Kess 00:15:10

Singapore's Temasek said this morning that an unwinding of the AI trade is the biggest risk facing markets.

55. Damra Vol 00:15:17

Put a sovereign investor's warning next to a thirty-billion-dollar debt deal and you get the same week saying two things. One side is lending more, and another is saying that what they're lending against could unwind.

56. Lenar Kess 00:15:29

These are two separate stories, and the only link is that building AI capacity takes both permission and borrowed money. Today we got a pause on one and a very large ask on the other, with Finland's deadline on October 23.

57. Lenar Kess 00:15:43

A few shorter ones. Amjad Masad wrote that AI-powered reverse engineering and decompilation is moving fast enough that all software will be de facto open source. Naval argued that models are the last moat, since AI can decompile and recode software but a model resists being distilled. And eigenrobot relayed a report that Adobe's creative suite was decompiled, rebuilt in Rust and released as open source.

58. Damra Vol 00:16:10

Flag the last one. It's a secondhand report in a tweet, and there's no repo in front of us, so it's unconfirmed. These are all tweets, which means opinions. The interesting angle is legal. A decompiled-then-rewritten codebase might be an independent reimplementation or it might be a derivative of the binary, and courts had views on that long before models existed. The models just make the cheap path cheaper.

59. Lenar Kess 00:16:35

Reuters reports that South Korea's president says AI agents appear to have been used in attacks on the country's banks. The wording is appear to, and the evidence behind it isn't in the item I've seen. Yesterday we had the Swarmchasers scanning report and the Australian hearings, and a head of state attributing bank attacks to agents is a step up in who's saying it.

60. Damra Vol 00:16:57

And the Guardian ran Toby Walsh's column on OpenAI in Australia. He says Altman was asked to appear before the parliamentary committee and sent Jason Kwon instead. He writes, "We don't accept drugs killing a few people because they cure most patients. Or a few planes crashing because most land safely. AI should be no different." That's an op-ed, and the committee reports at the end of November.

61. Lenar Kess 00:17:22

Also this morning, Common Sense Media rated ChatGPT for Teens an unacceptable risk, saying its protections fall short of OpenAI's promises: it doesn't send alerts to parents when it should, and it still does kids' homework. NBC separately reports that members of Meta's Oversight Board have advice for AI companies setting up their own independent safety committees.

62. Damra Vol 00:17:43

Common Sense is an advocacy group with its own rating method, so that's an outside score and not an audit. The homework complaint is the product behaving like the product. The parent-alert one is a specific promise with a specific miss, and that's the one OpenAI will have to answer.

63. Lenar Kess 00:18:01

Three preprints this morning. The power-draw paper asks how much hidden computation a GPU's power trace can exclude. Measurements on NVIDIA A100 chips put the worst-case bound at 1.16 of declared machine capacity. Adversarial strategies that match the energy profile can hide at least 0.41. A verifier who can re-execute the declared work gets that down to 0.059, but only under the most restrictive threat model.

64. Damra Vol 00:18:29

A number above one means the power trace can't even rule out hiding another whole machine's worth of compute. So power draw alone is a weak witness, and any treaty on frontier compute would need much more than a wattmeter.

65. Lenar Kess 00:18:42

The second, “Auditable Claims about AI Agents,” says, “to be checked, a claim about an agent must first name its policy, its scope, the records that would settle it, and who writes them.” The third, AdvSim2Real, trains a four-billion-parameter web agent against adaptive prompt injection. It reports a 33.6 percent relative gain in completion against an adversary it never trained on, over 150 web tasks.

66. Damra Vol 00:19:10

The audit paper's sentence is also a decent test for the PAP announcement from earlier. Name the policy and the records, and say who writes them.

67. Lenar Kess 00:19:19

Last, four AI Engineer talks, all vendor talks, so each is also a pitch. The concrete one is Dor Sasson of Stigg, who argues that entitlement checks happen after consumption and after the invoice instead of synchronously before the spend, so thousands of agents hitting a shared credit pool overspend it. His fix borrows from banking: reserve capacity before inference, settle actuals asynchronously, and keep a double-entry ledger with holds. Resolve AI described a six-pillar harness for production operations. Langfuse showed a human-in-the-loop self-improvement loop with a changelog-writing agent, and Tessl an inner, outer and meta loop model for software factories.

68. Damra Vol 00:20:01

Stigg's is the one I'd steal from, because anyone running agents against a budget has had a runaway loop. Check and reserve first, reconcile later, and the overspend problem has an old answer.

69. Lenar Kess 00:20:14

That covers it. The dates on the board are Finland's October 23 deadline and Mistral's weights on October 27, plus whenever PAP's spec text turns up somewhere readable. This was Braid, from Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic

