Ten Problems, No List

2026-08-02 / 00:27:45

> *“A claim that big should arrive with a list attached, and this one arrived with a screenshot and a lot of enthusiasm.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

An OpenAI researcher says an internal model closed ten open problems in mathematics and theoretical computer science. Nobody has published the ten problems. Today's episode is about what the industry does in the gap between a claim and its evidence — and about four other stories where the evidence is a screenshot, a vendor chart, or a sarcastic Reddit post.

- Noam Brown's post on the Astra results — the source of the claim, relayed to us through a Hacker News submission rather than any first-party writeup from OpenAI.
- Theo Jaffee relaying a model's assessment that this was "plausibly the most significant single day in the history of mathematics," and Victor Taelin noting that the mathematics subreddit is unimpressed.
- Miles Brundage arguing the effort would do more good pointed at verifiably secure software than at open math.
- Mario Zechner running OpenAI's security evaluation against his own copy protection, undefeated for six years, and reporting what GPT-5.6 Sol did to it.
- The EU AI Act transparency date — today, August 2 — discussed in the local-models community with a clown emoji in the title.

- Simon Willison cataloguing what ChatGPT Work can actually do, and Brett Bauman using it as a scheduled job.
- The 105× total-cost figure for DeepSeek V4-Flash, with Aravind Srinivas calling two orders of magnitude rare.
- Elvis's four ways agent state disappears and Cameron Wolfe on why scoring a trajectory is harder than scoring an answer.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Here's a question to sit with before we get into any of it. If somebody told you that ten open problems in mathematics and theoretical computer science had been closed in a single day, what would you want to see before you believed it? A name, a proof, or a mathematician you trust saying, yes, I checked one of them? Yesterday afternoon Noam Brown posted that an internal OpenAI model called Astra had done exactly that. Ten problems. This morning the post is on Hacker News, Kevin Roose has weighed in, and the timeline has been arguing about it for about twenty hours.

   - twitter.com
   - x.com
   - x.com

- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- reddit.com
- youtube.com
- reddit.com
- bbc.com
- x.com
- x.com
- x.com
- x.com
- x.com
- reddit.com
- x.com
- x.com
- reddit.com
- costperprompt.com
- x.com
- x.com
- x.com
- youtube.com
- youtube.com
- reddit.com
- x.com
- x.com
- x.com
- wafer.ai
- trufflesecurity.com
- x.com
- alexiglad.github.io
- youtube.com

2. Damra Vol 00:00:38

   And none of the things you just listed exist yet. There's no list of the ten problems, no proof artifact anybody can open, and no first-party writeup from OpenAI. What we have is a researcher's post and

a very large volume of people reacting to it.

3. Lenar Kess 00:00:55

   Let's start there, because the reaction is all any of us has to look at. Theo Jaffee relayed a line that's been travelling on its own — that this was, quote, plausibly the most significant single day in the history of mathematics. And that's a model's phrasing, relayed by Theo, not Theo's own assessment.

4. Damra Vol 00:01:14

   [tsk] That detail matters more than it sounds. A model was asked to evaluate a claim about a model, produced a superlative, and the superlative is now doing the rounds as if a person said it. That's a new kind of citation chain and I don't think anybody has decided how to handle it.

5. Lenar Kess 00:01:31

   Meanwhile Victor Taelin's contribution was to point out that the mathematics subreddit seems unimpressed. Which is a useful counterweight, but it's also just one community's first read on a Saturday.

6. Damra Vol 00:01:43

   Mathematicians are unamused as a professional posture. That's not evidence either. <soft>Though I'd note it's the community best positioned to check, and the ten proofs they'd need to check haven't been handed to them.</soft>

7. Lenar Kess 00:01:54

   So let me lay out where we're going today. This is the lead, and we'll spend real time on it. After that: Mario Zechner ran OpenAI's own security evaluation against a piece of copy protection he wrote, which has held for six years, and reported what happened. The European AI Act's transparency obligations take effect today, August second, which isn't something we get to schedule around. People who ship software are using ChatGPT Work to run recurring work overnight. And there's a cost number on DeepSeek's V4-Flash that's worth taking apart because it isn't the number you think it is.

8. Damra Vol 00:02:32

   Back to Astra for a second, because there's a version of this that's ordinary and a version that isn't, and they look identical from outside. Ordinary: a model with a lot of reinforcement learning behind it and a lot of compute at inference time grinds through problems that were open in the sense of nobody having bothered, not open in the sense of hard.

9. Lenar Kess 00:02:52

Right. And what counts as open in that sentence decides the whole story.

10. Damra Vol 00:02:56

    There are open problems that three generations of people have broken themselves on, and there are open problems that are open because they sit in a corner of a subfield with eleven active researchers and nobody had time. Both are technically open. Only one of them is the most significant day in the history of mathematics.

11. Lenar Kess 00:03:14

    Yo Shavit was in the same conversation making a more structural point — he's speculating about why OpenAI might have an edge here specifically, and his answer is the combination of reinforcement learning, test-time compute, and formal proof machinery. Which is a coherent guess. Proof is one of the few domains where you can check the answer mechanically.

12. Damra Vol 00:03:34

    That's what makes me take the claim more seriously than I otherwise would. If you're doing formal proof against a checker, you have a reward signal that doesn't lie to you. You can sample a thousand attempts and the checker tells you which one survived. That's a very different training situation from asking a model to write a good essay.

13. Lenar Kess 00:03:53

    And it means the ten results, if they exist, might be verifiable in a way that most model claims aren't. Somebody could just run the proofs.

14. Damra Vol 00:04:01

    So publish them. That's the whole ask. If the results are machine-checkable, the verification cost is nearly zero and the credibility payoff is enormous. Sitting on them while the timeline works itself into a state is a choice somebody made.

15. Lenar Kess 00:04:16

    Miles Brundage came at it from a different angle entirely, and I keep coming back to his line. His view is roughly: if you've got a system that can do this, he'd rather see that effort concentrated on verifiably secure software than on random math stuff. He said it more bluntly than that.

16. Damra Vol 00:04:34

    [chuckle] Random math stuff is a great phrase from someone who spent years inside OpenAI policy. But he's making a real allocation argument. There's a finite amount of frontier capability being

pointed somewhere, and pointing it at proof automation is a prestige play. Pointing it at making software that doesn't fall over changes what happens to ordinary people.

17. Lenar Kess 00:04:56

    Although those aren't as separate as they sound. Formal verification of software is proof. It's the same machinery.

18. Damra Vol 00:05:03

    It is, and that's the strongest reading of the day. If the proof capability is what they say it is, the near path isn't more theorems. It's verified compilers, verified kernels, and verified cryptographic implementations — code that's provably correct and that nobody has had the labor to prove, because the proofs run longer than the programs.

19. Lenar Kess 00:05:24

    That's the version of this story I'd want to be true. It's also the version nobody's announcing, because announcing ten theorems gets you a day of timeline and announcing a verified TLS implementation gets you a polite nod from four hundred people.

20. Damra Vol 00:05:38

    Four hundred very important people. <laugh-speak>But yes — the incentives point at theorems. </laugh-speak>

21. Lenar Kess 00:05:43

    So the standing question is narrow and answerable. Somebody at OpenAI publishes the list of ten, or a mathematician outside the building says they verified one. Until one of those happens, what we have is a Saturday afternoon post that a lot of serious people found credible enough to amplify — which is information, just not the kind that settles anything.

22. Damra Vol 00:06:05

    And the gap between how big the claim is and how thin the artifact is stays the most interesting number in the story.

23. Lenar Kess 00:06:12

    Here's a capability report with an artifact attached, which is a nice change. Mario Zechner — badlogicgames, he's been writing game tooling for a very long time — took OpenAI's cybersecurity evaluation and ran it against his own digital rights management scheme. His own product. Something he wrote, that he says nobody has defeated in over six years.

24. Damra Vol 00:06:34

    He picked the right target. He isn't testing a synthetic benchmark. He's testing the code he'd be embarrassed to lose. And he reports two different outcomes from two different models, which is more informative than a single score.

25. Lenar Kess 00:06:47

    Kimi K3 cracked localized defenses — chipped away at pieces of it. GPT-5.6 Sol did something he describes differently: it combined techniques and saw, his words, the entire shit cake, and it got there faster.

26. Damra Vol 00:07:01

    The entire cake being the architecture. Not one check defeated, but the structure of how all the checks relate to each other. That's the qualitative jump. Defeating a check is a puzzle. Understanding why the checks are arranged the way they are is reverse engineering.

27. Lenar Kess 00:07:17

    And his own conclusion is more measured than the excitement around it. You still can't hand it a binary and say crack this thing and get a result. But if you already know what you're doing, the agent takes over the tedious parts.

28. Damra Vol 00:07:30

    Which is the accurate version of every agentic coding claim right now, and it's the version that changes who does the work rather than whether the work is possible. Security research has always been ninety percent grinding — tracing, annotating, and trying the boring approach eleven times. If that compresses, the number of people who can do serious reverse engineering goes up a lot, and the ones who benefit most are the people who already had the taste and not the hours.

29. Lenar Kess 00:07:58

    He also mentions the agent generating machine code directly at points, which is a detail I'd like more of.

30. Damra Vol 00:08:04

    That's the one I'd want to see the transcript for. Writing correct machine code by hand is a skill that maybe a few thousand people still have, and it doesn't have a friendly error surface. Either the bytes execute or the process dies.

31. Lenar Kess 00:08:18

Separately, and on the same model, there's an r slash OpenAI post from overnight claiming GPT-5.6 Sol's raw reasoning surfaced during a failed tool call. A screenshot. Zero points, no discussion.

32. Damra Vol 00:08:33

    Unverified screenshot on a dead thread, so hold it loosely. But the mechanism it describes is plausible and that's why it's interesting. The reasoning trace is generated either way. You see a summary the product layer produces on top of it. When the tool call fails, the code path that does the summarizing sometimes isn't the code path that runs.

33. Lenar Kess 00:08:54

    So it isn't hidden in the model. It's hidden by a wrapper that has error cases.

34. Damra Vol 00:08:59

    Every lab has decided that the reasoning trace is proprietary and that users see a sanitized version. That decision is enforced by ordinary software with ordinary bugs. It leaks the way anything leaks — at the seam nobody tested.

35. Lenar Kess 00:09:14

    There's a third data point on this model, and it cuts the other direction. Nate Jones has a short video where his summary is that 5.6 is a dumber model and he loves it. That's a strange sentence until you hear the rest. It set a new high on an evaluation covering long-running professional work across fifty-five different fields.

36. Damra Vol 00:09:34

    Dumber meaning less inclined to perform intelligence at you. Fewer flourishes, less hedging, and more finishing. If the evaluation rewards completing long professional tasks, then a model that stops narrating and just works is going to score well on it.

37. Lenar Kess 00:09:50

    Zechner's report and that evaluation are describing the same behavior from opposite sides. Take over the tedious part, don't editorialize about it.

38. Damra Vol 00:09:59

    And a model that grinds without commentary is exactly what you want pointed at six years of somebody's copy protection.

39. Lenar Kess 00:10:06

Today is August second, and the European Union's AI Act transparency obligations take effect. That's not a prediction or a debate — it's a calendar date that arrived this morning while most of the people affected by it were asleep.

40. Damra Vol 00:10:20

    And what I find strange is where we're hearing about it. The source that surfaced today is a post in the local-models community with a clown emoji in the title. That's the coverage. A major regulatory deadline for a trading bloc of four hundred and fifty million people, and the discussion is a sarcastic thread among people running models on their own hardware.

41. Lenar Kess 00:10:41

    The obligation in broad terms is marking. Generated images, audio, video, and text carry disclosure requirements. I'll stay conservative about the scope, because I'm not reading the text of the regulation to you and I'd rather not invent legal specifics.

42. Damra Vol 00:10:57

    The scope is just what the sarcasm is about, though. Take the person in that thread. They're running an open-weights model on a machine in their apartment in Lisbon, generating images for a hobby project. Does the obligation reach them? Who would know? What would enforcement even look like — someone comparing metadata on a forum post?

43. Lenar Kess 00:11:17

    I'd expect the enforcement pressure to land on the platforms and the API providers, not the person with the graphics card. That's how these regimes usually work. You regulate the chokepoints that have a legal department and an address in Brussels.

44. Damra Vol 00:11:31

    Which produces a two-tier outcome that nobody voted for. Commercial generation gets marked, self-hosted generation doesn't, and the marking becomes a signal of provenance in the wrong direction. The labelled content is the content from companies you could already identify.

45. Lenar Kess 00:11:47

    That's the uncomfortable version, yes. Though the counter is that most generated media people encounter comes through commercial pipes, so marking those covers most of the volume even if it doesn't cover the hard cases.

46. Damra Vol 00:11:59

Volume is a fair defense. It's just not the defense the law is usually sold with.

47. Lenar Kess 00:12:04

    There's a liability story running alongside it this week that I think is the more consequential one. The BBC has the boss of a company that got hacked saying AI firms must answer for rogue bots. Which is a demand that the model provider carry some of the responsibility when an agent does damage.

48. Damra Vol 00:12:22

    And the AI Act doesn't answer who pays. Marking tells you something was generated. It says nothing about who covers the damage when an agent with credentials does something nobody sanctioned. We spent yesterday on sandbox escapes — Joseph Thacker posted that he'd found them in his own logs — and the liability for that is currently nowhere.

49. Lenar Kess 00:12:42

    Nathan Calvin was making an adjacent point about how these incidents get described. His complaint is with the language companies reach for when their systems misbehave.

50. Damra Vol 00:12:52

    The vocabulary problem is downstream of the liability problem. If nobody owes anything, then the incident report is a communications exercise, and communications exercises produce the word unexpected a lot.

51. Lenar Kess 00:13:05

    So today the marking rule is live and the ownership question isn't even drafted. That's the state of it as of this morning.

52. Lenar Kess 00:13:13

    Greg Brockman spent yesterday and this morning telling people to ask ChatGPT Work to do any recurring task. And people who write software took him up on it — Brett Bauman published a movie-tracking page built on it and called ChatGPT Work the new scheduled job.

53. Damra Vol 00:13:29

    Before the substance, there's a naming problem Simon Willison flagged that I can't get past. The same product name means two different things in two different OpenAI surfaces. The capabilities he found — a browser that takes screenshots, plus deployment — are in the mobile and web app. The desktop thing with the same name is something else.

54. Lenar Kess 00:13:50

    Simon's been cataloguing it in public for a couple of days now, which is how most of us are learning what's in it. There's no feature list going around; there's a person poking at it and posting what he finds.

55. Damra Vol 00:14:02

    That's a shipping decision, not an accident. You put capability into the product and let developers discover it, and the discovery becomes the marketing. It works right up until somebody discovers a capability you didn't intend to ship.

56. Lenar Kess 00:14:16

    On the substance — I don't love the scheduled-job comparison, and the mismatch is where it gets interesting. A scheduled job is a line in a file. It runs a command at a time. It does the same thing every time, and when it fails it fails the same way, and you can read the command and know what it'll do before it runs.

57. Damra Vol 00:14:35

    Whereas this one has a browser. Brockman's own post makes a point of the cloud browser being something you can watch and step into while it's running. That's not a scheduled job. That's a colleague you're supervising who works at three in the morning.

58. Lenar Kess 00:14:49

    And the failure modes are different in kind. A cron line that breaks leaves a non-zero exit code. An agent with a browser that breaks might have clicked something.

59. Damra Vol 00:14:59

    Might have clicked something, or filled a form, or decided that the way to complete a task was to sign up for an account. The scheduled-job metaphor imports an expectation of determinism that isn't there, and the people adopting it fastest are the ones who most trust that expectation.

60. Lenar Kess 00:15:15

    There's a companion post in the agents community that's blunt about the adoption side — the argument is that people underestimate how easy automating your own machine has become. It's enthusiastic rather than rigorous, but the enthusiasm is a real signal.

61. Damra Vol 00:15:31

It's the same energy as the early days of scripting your own machine, and I mean that warmly. Somebody automating their movie list isn't building infrastructure, they're playing. Play is where the interesting patterns come from. It's just that this particular toy has a browser and your session cookies.

62. Lenar Kess 00:15:48

    Bauman's page is a personal demo and reads like one. But a personal demo built on a product that shipped days ago tells you the surface is usable, which isn't true of every agent product that gets announced.

63. Damra Vol 00:16:01

    And I'd rather have a working movie tracker from a real developer than another enterprise deployment case study nobody can inspect.

64. Lenar Kess 00:16:08

    There's a number moving around on DeepSeek's V4-Flash, relayed by Chubby from an Artificial Analysis chart: 105 times lower total cost than Fable 5 on the same set of benchmark tasks. Aravind Srinivas's response was four sentences long and the useful part was, two orders of magnitude improvements are quite rare, this is a big deal.

65. Damra Vol 00:16:31

    Total cost is the operative word and most people repeating the number are dropping it. That figure isn't a per-token price ratio. Per-token pricing between those two isn't anywhere near a hundred to one.

66. Lenar Kess 00:16:44

    So where does the rest of it come from?

67. Damra Vol 00:16:46

    Token count. To finish the same task, one model burns a great deal more reasoning than the other. Multiply a cheaper token by far fewer tokens and you get two orders of magnitude without either factor being extreme on its own. It's the compounding that produces the headline.

68. Lenar Kess 00:17:03

    Which means the number is task-shaped. Change the benchmark and the ratio moves, sometimes by a lot.

69. Damra Vol 00:17:09

    It moves with the task mix, yes. And it's relayed secondhand off a chart rather than read out of a methodology section, so I'd hold the specific figure loosely while accepting the direction it points.

70. Lenar Kess 00:17:22

    The direction being that verbosity is now a line item. For a couple of years the reasoning models were rewarded for thinking longer, and nobody costed the thinking.

71. Damra Vol 00:17:31

    Somebody's costing it now. There's a Show HN today — CostPerPrompt, live pricing with real-workload calculators — and the sharpest contribution in the thread isn't the tool, it's a commenter asking what happens to the estimate when you lose your prompt cache partway through a session.

72. Lenar Kess 00:17:48

    Which wrecks any of these calculators.

73. Damra Vol 00:17:50

    Completely. Your cost model assumes a warm cache and your actual session bounces between machines, or you idle past the cache window, and suddenly you're paying full freight on a context you already paid for. All of these calculators model the good case.

74. Lenar Kess 00:18:06

    The other V4-Flash item today is smaller and more useful to anyone running it locally: the tool-calling fix for the 0731 build went into llama.cpp, which clears up the looping behavior people were hitting.

75. Damra Vol 00:18:21

    That's the item that makes the cost story reachable for people outside an API contract. A model that's cheap on someone's hosted endpoint is interesting. One that's cheap and calls tools correctly on hardware you own is a different proposition, and until yesterday the second one didn't work.

76. Lenar Kess 00:18:39

    We went deep on V4-Flash pricing yesterday so I don't want to re-litigate it. The new material is the multiple and the local fix, and the multiple is smaller than it sounds and more interesting than it sounds, both at once.

77. Damra Vol 00:18:52

Fewer tokens to finish is a better property than cheaper tokens. It's just harder to put on a chart.

78. Lenar Kess 00:18:58

    Two posts from yesterday afternoon describe the same hole from opposite sides. Elvis laid out how working state disappears when you're running teams of Claude Code agents over a long stretch. His list is specific: state vanishes when a terminal closes, compaction condenses the conversation away, and the team can't be resumed.

79. Damra Vol 00:19:18

    Each of those is a person losing hours of accumulated context to a window closing. And what's striking is how ordinary the failures are. This isn't a frontier capability problem. It's that the working memory of a multi-hour session lives in a process, and processes end.

80. Lenar Kess 00:19:35

    Compaction is the one that bothers me most, because it's the system doing its job. It condenses the conversation to fit, and what it discards is whatever you didn't know you'd need.

81. Damra Vol 00:19:47

    Compaction is lossy on purpose and it's lossy according to a heuristic nobody tuned for your task. The summarizer making the call doesn't know which three lines from hour two turn out to matter in hour six.

82. Lenar Kess 00:19:59

    EJ Campbell was asking the obvious follow-on this morning — why doesn't Anthropic ship a multiplexer itself, given how many people are building one?

83. Damra Vol 00:20:08

    Because the moment you ship persistent multi-instance workspaces, you own state. You own storage and retention and deletion. You own whose data sits in whose workspace, and you own what happens when a session gets subpoenaed. Right now that all belongs to the user's terminal, and a terminal is nobody's compliance problem.

84. Lenar Kess 00:20:29

    So the missing feature is missing for institutional reasons rather than technical ones.

85. Damra Vol 00:20:34

Partly. It's also that the moment state is durable, agents can run for weeks, and a long-running agent is a much scarier product to support than a session that dies politely at the end of the day.

86. Lenar Kess 00:20:45

    Cameron Wolfe posted the other half of it, on evaluation. His point is structural: evaluating a language model means scoring one response to one prompt, and evaluating an agent means scoring a trajectory through an environment with tools and observations along the way.

87. Damra Vol 00:21:01

    And a trajectory can be right in the wrong way. An agent reaches the correct answer having read six files it had no business reading, or gets there by trying something destructive and recovering. Score only the endpoint and both of those pass.

88. Lenar Kess 00:21:16

    Which connects back to yesterday's material without me having to force it — the escapes people are finding are trajectory problems. The output looked fine.

89. Damra Vol 00:21:25

    The output always looks fine. That's what makes a bad trajectory hard to catch.

90. Lenar Kess 00:21:30

    On the mechanics side, there's a good explanation going around of how skills actually get selected — the agent reads only the name and description first, then loads the instructions if the task matches, and only pulls supporting material when directed. Lazy loading, so your context doesn't fill up with every template you've ever installed.

91. Damra Vol 00:21:49

    Which means a vague description is a silent failure. Your skill never triggers and nothing tells you. Or it triggers on everything and you spend your context budget on edge cases for a task you're not doing. Most people collecting skills from public repositories have no idea that the description field is the whole matching mechanism.

92. Lenar Kess 00:22:09

    That explanation comes packaged with someone selling a skill builder, so take the pitch separately from the mechanism.

93. Damra Vol 00:22:15

The mechanism stands on its own. And it explains something people complain about constantly, which is skills that never fire and never say why.

94. Lenar Kess 00:22:24

    A few things that don't need a segment each. Codeberg's policy banning projects primarily generated by large language models keeps generating discussion, and the number that does the most work in it isn't philosophical — solid state drive prices went from seven hundred euros to thirty-five hundred. They're a self-hosted non-profit absorbing high-volume unreviewed commits, and that's a hosting bill, not a position on copyright.

95. Damra Vol 00:22:49

    The vote figures are worth knowing too: about two-thirds of voting members approved, and voting members are roughly half the user base. And the unresolved problem is the word mostly. Nobody has a technical definition of mostly generated, and without one you're back to maintainer judgment, which without a reputation system means new contributors get treated as suspects.

96. Lenar Kess 00:23:12

    The alternative floated in that discussion is a usage fee — internalize the infrastructure cost instead of banning the category.

97. Damra Vol 00:23:20

    Which is clearer about where the cost lands. But what I keep turning over is the psychology they describe: everyone rates their own use of these tools as responsible and everyone else's as slop. That bias is what turns a cost problem into a community fracture, and it's operating on all of us right now including in this conversation.

98. Lenar Kess 00:23:41

    Sitting right next to it, someone in the Claude community posted about ten months of building a full kitchen management system with Claude Code — meal planning, the whole thing. That's the person a categorical ban catches.

99. Damra Vol 00:23:53

    Ten months isn't a weekend of generated slop. That's someone who kept showing up. Any policy written around the word mostly has to decide what to do with them, and right now it doesn't.

100. Lenar Kess 00:24:05

Two release notes. MiniMax H3, also appearing as Hailuo H3, came out this morning with video generation priced at roughly thirty percent of Seedance 2.0. Pierrick Chevallier's hands-on read is that dialogue scenes are where it's strongest — facial expressions, timing, camera movement, and keeping a character looking like themselves across shots.

101. Damra Vol 00:24:28

    Both posts read promotional and the price ratio is a vendor-adjacent figure, so that's release-note altitude. Dialogue scenes being the strong axis is a specific enough claim to check, though, and it's the axis that matters if you're trying to make something with people talking in it. Luma's Ray 3.2 also picked up the ability to recut into any aspect ratio while keeping the action intact, which sounds small and isn't if you're delivering the same cut to four platforms.

102. Lenar Kess 00:24:58

    On hardware: wafer dot ai published benchmarks running Kimi K3 on AMD's MI355X and claims better performance per dollar than Nvidia's B300. Hundred and forty-five points on Hacker News, sixty-one comments, and the top comment isn't about the numbers.

103. Damra Vol 00:25:16

    [laugh] Of course it isn't. The top comment objects that the writeup itself reads as generated, which makes the work hard to take seriously. And that's a live problem for anyone publishing technical results right now — your prose style is being read as evidence about your rigor, fairly or not.

104. Lenar Kess 00:25:35

    We don't have the methodology, so the performance-per-dollar claim stays a claim. Independent numbers comparing AMD and Nvidia inference economics on a current open model are rare enough that I'd like it to hold up.

105. Damra Vol 00:25:47

    I'd like to read it in prose that doesn't make me squint at it first.

106. Lenar Kess 00:25:51

    Truffle Security scanned seven point six petabytes of Hugging Face training data looking for live credentials and published what they found, organized around keys that actually open something rather than raw counts.

107. Damra Vol 00:26:04

And the property that makes that bad is that a credential in a public training corpus isn't in one place. It's in every fine-tune, every derived dataset, and every mirror somebody pulled last year. Rotating it fixes your service and does nothing about the copies.

108. Lenar Kess 00:26:20

    Last one. Alex Glad published a writeup on a training method he calls explorative modeling — sample K generations, then train on the best one. Ninety-nine points on Hacker News, and one commenter's reaction was that if the numbers hold up on a large training run, every image model trained before this is obsolete.

109. Damra Vol 00:26:40

    That's a commenter's conditional, not the author's claim, and the conditional is carrying everything. But the idea underneath it rhymes with something Grant Sanderson said on Dwarkesh's show this week about autoregression — that predicting one token at a time rewards the statistically likely continuation and penalizes the unlikely-but-necessary one.

110. Lenar Kess 00:27:01

    Two separate observations about what the training objective rewards, and I don't want to weld them into a thesis about a paradigm shift. They just happen to rhyme.

111. Damra Vol 00:27:10

    They rhyme, and one of them has ninety-nine points and no replication.

112. Lenar Kess 00:27:15

    Which is where the day ends up, more or less. The lead is a claim without its attachments, the security result is one skilled person testing his own work, the cost multiple is relayed off a chart, and the reasoning leak is a screenshot on a dead thread. Any of those could be exactly what it says it is. A list of ten problems would settle the biggest one, and publishing it takes about a minute.

113. Damra Vol 00:27:38

    And if it doesn't show up in the next few days, that absence tells us something about which of the ten were the interesting kind of open.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic