DISPATCH 116 · 2026-08-14 | GSV THE INVOICE TOLD THE TRUTH

Same Score, Third of the Price

2026-08-14 / 00:25:21

> *"Two models scored identically and one of them cost two and a half times more. Nobody won on capability. Somebody won on the invoice."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Friday's news was almost entirely about price. Grok 4.6 posted the same score as Claude Fable 5 on Perplexity's WANDR benchmark at a third of the cost, Gemini 3.7 Flash shipped at half the price of 3.6 Flash, and OpenAI previewed a Sol variant whose pitch is latency rather than intelligence. Underneath that, two Apache-2.0 open-weight releases and a mathematical bound that a model broke and then refused to believe.

- Perplexity ran Grok 4.6 and Claude Fable 5 on its WANDR agentic-search benchmark: identical 0.496 scores, $7.58 per task versus $20.30. Perplexity has no stake in xAI, which is what makes the number usable.
- Elon Musk and X Freeze flagged Grok 4.6 at the top of CursorBench 3.2 for real-world coding tasks.
- Daniel McKinnon reported Grok 4.6 topping RareBench for rare-disease diagnosis ahead of Claude Opus 5 — a research benchmark, not a clinical result.
- The AI Daily Brief put the list price at $2 per million input tokens and $6 per million output, roughly $0.84 per benchmark task, and noted community reports of occasional truncated outputs.

- Cerebras published the engineering write-up behind OpenAI's GPT-5.6 Sol Ultrafast preview, up to 14x faster and gated to a small set of API customers.
- dots studio shipped dots3-note preview — 280 billion parameters, 16 billion active, 512K context, multimodal, Apache 2.0 — the same night Z.ai released GLM 5.3 and Paul Graham noted that tuning open weights has swung back into fashion.
- Harrison Chase shipped cron scheduling into Managed Deep Agents, Vercel wired nine agent CLIs behind one gateway, and Perplexity turned Sonar into an Agent API.
- The Rails team published its first agent benchmark: eight models, 21 atomic tasks, three runs each.
- Two Minute Papers covered an unreleased Claude improving a prime-distribution bound past the human record after roughly 650 prompts — and labeling its own result "too strong to be new."

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

Here's a number before I tell you where it came from. Two models ran against the same benchmark, with the same harness and the same tasks. Both score 0.496. Not close — <emphasis>identical</emphasis>, to three decimal places. One of them costs $7.58 per task. The other costs $20.30. [pause] So what do you call that? On capability, nothing happened. Nobody won.

- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- youtube.com
- x.com
- x.com
- x.com
- x.com
- cerebras.ai
- youtube.com
- x.com
- x.com
- z.ai
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- youtube.com
- x.com
- x.com
- blog.google
- x.com
- x.com
- x.com

- x.com
- x.com
- youtube.com
- x.com
- cdn.openai.com
- x.com
- x.com
- i.redd.it
- reuters.com

2. Damra Vol 00:00:29

   You call it an invoice. And the reason it matters is that the party running the benchmark had no reason to flatter either model.

3. Lenar Kess 00:00:37

   Right — this is Perplexity's WANDR benchmark, their agentic search evaluation, published yesterday afternoon. Aravind Srinivas put the Pareto chart out himself. The cheap one is Grok 4.6. The expensive one is Claude Fable 5. And Perplexity doesn't own either model, which is the whole reason I'm reading the number out loud instead of ignoring it.

4. Damra Vol 00:00:59

   Yesterday we spent a long segment on what xAI <emphasis>hadn't</emphasis> published about Grok 4.6 — the missing model card, the disclosure gap. Today the model card is still incomplete, and it almost doesn't matter, because three different people who don't work for xAI ran the thing and posted their own numbers.

5. Lenar Kess 00:01:18

   Three harnesses, and let me lay them out, because they're measuring different things. Perplexity's WANDR is agentic web search — multi-step research where the model has to go get things. Then there's CursorBench 3.2, which is Cursor's own evaluation on real-world coding tasks, and Grok 4.6 took the top slot there overnight. Musk posted it around 5:40 this morning, and X Freeze had it three minutes earlier.

6. Damra Vol 00:01:46

   And the third one is the strange one. RareBench — rare-disease diagnosis in children. Daniel McKinnon reported Grok 4.6 at the top of it, ahead of Claude Opus 5. I want to be careful with that one, because a research benchmark on diagnostic reasoning is not a statement about clinical deployment, and anyone who reads it that way is going to hurt someone.

7. Lenar Kess 00:02:09

   Agreed. The reason RareBench interests me isn't the medicine — it's that this is a third unrelated task family. Coding, agentic search, and diagnostic reasoning. If a model were benchmark-gaming, you'd expect it to spike on the ones the lab targeted and sag everywhere else.

8. Damra Vol 00:02:27

   You'd expect exactly that, yeah. Three unrelated harnesses agreeing is a different kind of evidence than one harness agreeing three times. <soft>Though I'd note the sample is still two days old, and the number of people who've run this on real workloads is small.</soft>

9. Lenar Kess 00:02:42

   Let's get the list price on the table, because the per-task cost is downstream of it. The AI Daily Brief's roundup put Grok 4.6 at two dollars per million input tokens and six dollars per million output. Their own math works out to roughly eighty-four cents per benchmark task on the aggregate suite — call it 32% cheaper than GPT-5.6 Sonnet, and 73% cheaper than Fable 5. Artificial Analysis has its overall intelligence index at 61.

10. Damra Vol 00:03:12

    And some of that gap isn't the sticker price at all, which is the technically interesting bit. They're claiming token efficiency gains — the model uses fewer tokens to get to the same answer. So you're multiplying a cheaper rate by a smaller count. That compounds in a way a headline price cut doesn't.

11. Lenar Kess 00:03:30

    How much of that do you actually believe?

12. Damra Vol 00:03:33

    The direction, fully. The magnitude, provisionally. Token efficiency is one of the easiest things in the world to measure and one of the easiest to measure <emphasis>selectively</emphasis> — you pick the task family where your model doesn't ramble. But the WANDR number isn't self-reported, and it's the one carrying the cost claim.

13. Lenar Kess 00:03:51

    There's a caveat in the community testing I don't want to skip past. The same roundup flags occasional incomplete outputs — the model just stopping short — and concerns about how it handles security-sensitive material. Those are exactly the failures that a per-task cost number can't see, because a truncated answer is cheap.

14. Damra Vol 00:04:10

    [tsk] That asterisk sits on every cost-per-task metric, and it isn't specific to xAI. If your model bails at eighty percent of the work, your cost per task looks fantastic and your cost per <emphasis>completed</emphasis> task is unmeasured. Nobody publishes that second number.

15. Lenar Kess 00:04:28

    It shipped fast, though, which tells you something about how much the price mattered to the people integrating it. Perplexity had it in production the same day. Warp put it in their terminal and their Agent CLI within hours of the release.

16. Damra Vol 00:04:41

    Warp is the one I'd point at. Warp's entire product is a terminal where an agent runs a lot of turns on your behalf. That's a workload where cost-per-turn is the dominant line item on their bill, not a rounding error. When a company like that swaps models in a single afternoon, they're not making an editorial statement about quality. They're doing arithmetic.

17. Lenar Kess 00:05:02

    There was one more xAI item last night that's a different flavor entirely. They posted about reverse-engineering compiled binaries back into clean, readable C. Not decompiler output — actual structured C.

18. Damra Vol 00:05:16

    That one I want to see independently reproduced before I get excited, because nobody has said what "clean" means there. Decompilation to something that compiles is one problem. Decompilation to something a human maintainer can reason about is a much harder one, and the difference between them sits entirely with whoever's grading. It's a striking demo. It's a demo.

19. Lenar Kess 00:05:38

    So where does that leave the lead? My read is that yesterday xAI had a disclosure problem and today they have a pricing position, and those two facts are unrelated to each other. The model card is still thin. The cost-per-task number came from Perplexity's harness regardless.

20. Damra Vol 00:05:55

    And the pressure that creates lands on Anthropic more than on anyone else. If you're the vendor charging $20.30 for the score someone else delivers at $7.58, you need an argument for the difference that isn't a benchmark — reliability under load, or behavior on adversarial input, or

something else the harness doesn't capture. Anthropic may well have that argument. They haven't made it this week.

21. Lenar Kess 00:06:21

    OpenAI made a different argument on the same day, and it's also not a capability argument. They previewed an Ultrafast mode for GPT-5.6 Sol — up to 14 times faster, launching first in the API to a select group of customers as capacity grows. Cerebras wrote the engineering post behind it, which went to 626 points on Hacker News.

22. Damra Vol 00:06:43

    The Cerebras involvement is the substance. Cerebras builds wafer-scale chips — one enormous piece of silicon instead of many small ones networked together — and what that architecture is good at is inference latency, because you're not paying to move activations between chips. So a 14x claim from that pairing is at least mechanically plausible in a way a pure software claim wouldn't be.

23. Lenar Kess 00:07:08

    OpenAI's own pitch for it is a ninety-second video, and it's entirely about incident response. An on-call engineer describes what used to be a one-to-two-hour cycle of gathering logs, normalizing telemetry, and adding context — collapsing to ten or fifteen minutes, because the model runs the searches concurrently instead of one after another.

24. Damra Vol 00:07:29

    And there's a line in there I keep turning over. He says refactoring the codebase costs almost nothing now — in time <emphasis>or attention</emphasis>. The attention half is the real claim. He's not saying the model got smarter. He's saying that when the round trip is short enough, you stop deciding whether a question is worth asking.

25. Lenar Kess 00:07:48

    Does that hold, though? Because my instinct is that the bottleneck during an incident was never the log aggregation. It was somebody with enough context deciding what the logs mean.

26. Damra Vol 00:07:58

    It half holds. You're right that the judgment doesn't move. But there's a real thing underneath the pitch, which is that during an outage, sequential waiting is corrosive. You run a query, you wait ninety seconds, and in those ninety seconds you context-switch to Slack and lose the thread you were pulling. Killing the wait doesn't make you smarter. It stops fragmenting you.

27. Lenar Kess 00:08:21

That's a more modest claim than the video makes.

28. Damra Vol 00:08:24

    It is, and it's the one I'd defend. The video's claim — that the trade-off between depth and speed has been eliminated — is marketing. The narrower version, that latency was taxing your concentration all along and now taxes it less, is something I think a lot of people will recognize from their own week.

29. Lenar Kess 00:08:42

    The access story is the part to keep an eye on. This is a preview, gated to a select group of API customers, expanding as capacity grows. That's the language of something supply-constrained, and the supply in question is Cerebras silicon.

30. Damra Vol 00:08:57

    Which makes it a very different product from a price cut. Google can halve the price of Flash by decision. OpenAI can't hand you 14x latency by decision — they have to have the wafers. So one of these is a lever and the other is an inventory.

31. Lenar Kess 00:09:13

    Two open-weight releases came out within hours of each other overnight, and one spec sheet tells most of the story. dots studio shipped a preview of dots3-note: 280 billion total parameters in a mixture-of-experts design, with 16 billion active at any time. The context window runs to 512 thousand tokens, and it handles text, vision, audio, and video. It also ships with native FP8 and speculative decoding built in. All of it under Apache 2.0.

32. Damra Vol 00:09:43

    Sixteen billion active out of 280 is the number that changes who can run it. You need the memory to hold the whole thing, which is not nothing, but the compute per token is what a 16-billion-parameter dense model would cost you. That ratio is why mixture-of-experts became the default architecture for anybody trying to ship something large that people will actually serve.

33. Lenar Kess 00:10:06

    And Apache 2.0 on all of it. No research-only clause, no acceptable-use appendix, and no revenue threshold.

34. Damra Vol 00:10:14

Which means a company can fine-tune it, ship it inside a product, and never have a conversation with a lawyer about whether they're allowed to. That's the difference between a model you can experiment with and a model you can build a business on top of. Tiezhen Wang at Hugging Face flagged the multimodal-plus-long-context combination specifically, and he sees more of these releases than almost anyone.

35. Lenar Kess 00:10:38

    The other one is GLM 5.3 from Z.ai, positioned as frontier coding with — their phrase — emergent cyber capabilities.

36. Damra Vol 00:10:48

    [tsk] That's vendor language and I'd like it labeled as such. "Emergent cyber capabilities" could mean the model is good at reading vulnerable code, or it could mean it's good at writing exploits, and those are extremely different things to put in a launch headline. If it's the second one, saying it that breezily is a choice.

37. Lenar Kess 00:11:07

    Paul Graham posted something last night that gives both of these a frame. His observation is that startups tuning open-weight models went from common, to a period where the consensus was that it was a waste of time, and now back again. Three phases, and we're in the third.

38. Damra Vol 00:11:23

    So what changed? Because the reason it became a waste of time was real. Around the time the frontier labs pulled meaningfully ahead, you could spend three months tuning an open model and end up behind where a frontier API had gotten to on its own, for free, while you worked.

39. Lenar Kess 00:11:39

    That's the question I'd put to the whole segment. What's different?

40. Damra Vol 00:11:43

    My read is that the gap you were tuning to close got small enough that three months of work can actually close it. When the open model is two years behind, tuning is a losing race. When it's a few months behind and Apache-licensed, tuning is a product decision — you're buying control, unit economics, and the ability to run it somewhere specific. Those were always the reasons. They just weren't worth a capability deficit.

41. Lenar Kess 00:12:09

There's an adjacent item I'll flag and not lean on. Mikko Ohtamaa and Yishan were both circulating a read that Mistral is repositioning around hosting Chinese open-weight models in Europe. That's commentary, not a Mistral announcement, and I'm not going to treat it as one.

42. Damra Vol 00:12:26

    It's plausible as a business, and it would mean a European company monetizing Chinese weights under European data rules. If that becomes a real category, it's a new position on the board. But somebody has to announce it first.

43. Lenar Kess 00:12:40

    Three separate vendors shipped agent infrastructure yesterday, and none of them shipped a model. LangChain put first-class cron scheduling into Managed Deep Agents. Harrison Chase's framing is that agents running in the background is how this work scales past people prompting directly.

44. Damra Vol 00:12:57

    Wait — cron. That's the entire pitch and I think it's correct. Everything about how these systems have been built assumes a person is sitting there. The request comes in because someone typed. The moment you put a scheduler in front of it, you've got a thing that wakes up at six in the morning and does work nobody asked for that morning.

45. Lenar Kess 00:13:16

    Which raises an obvious question about what happens when it wakes up and gets it wrong.

46. Damra Vol 00:13:21

    It does, and I notice nobody shipped the answer to that yesterday. But I don't want to be sour about the scheduler, because it's the piece that was conspicuously absent. People have been faking this with their own cron jobs and their own retry logic for a year. Making it part of the product means the breakages get observed centrally instead of by each person separately.

47. Lenar Kess 00:13:42

    The Vercel one is my favorite artifact of the day, purely for what it says about how people actually work now. Chris Tate posted it: their command-line tool now wires Claude Code, Codex, Cursor, Cline, OpenCode, Pi, Hermes, Kilo, and OpenClaw into a single AI Gateway. One budget across all of them. Credentials in the macOS Keychain.

48. Damra Vol 00:14:07

Nine. Somebody at Vercel looked at how their own engineers were working and counted nine coding agents on one laptop, each with its own API key pasted into its own config file and its own billing relationship. <laugh-speak>That's a support ticket that got promoted to a product.</laugh-speak>

49. Lenar Kess 00:14:24

    The Keychain detail is the one I'd point at. That's an admission that people have been storing frontier-model API keys in dotfiles across nine tools, and at least one of those dotfiles is in a repo somewhere.

50. Damra Vol 00:14:37

    It absolutely is. And the budget line is the other half — a single spend cap across nine agents means you can find out you've spent four hundred dollars before you find out from your statement. Key storage and a spend cap don't sell a launch post, and they're exactly why people will turn this on.

51. Lenar Kess 00:14:55

    Perplexity's contribution is on the other side of the wire. They folded Sonar into an Agent API — multi-step research and code execution behind one endpoint, so instead of you orchestrating a browsing loop, they run it and hand you the result.

52. Damra Vol 00:15:10

    And Srinivas put numbers behind it, comparing the agent path against straight search. What I'd want to know is where the failure boundary sits. When you own the loop, you can see the model go down a bad path on step three and stop it. When the loop is hosted, you get an answer and a bill, and the reasoning about how it got there is somebody else's.

53. Lenar Kess 00:15:31

    There's a counterpoint floating in the same conversation, from a talk at the AI Engineer conference — the argument that computer-use agents will end up agentifying the existing web rather than everyone building agent-shaped APIs. I only have the title, so I'm not going to characterize the argument beyond that.

54. Damra Vol 00:15:50

    It's a live disagreement even without the talk. Perplexity's Agent API is a bet that the right move is a clean endpoint. Computer-use is a bet that the web is never going to reorganize itself for machines, so the machine should just learn to click. Both of those are being funded simultaneously right now, and one of them is going to look silly in eighteen months.

55. Lenar Kess 00:16:11

The Rails team published their first agent benchmark report yesterday. Eight models, 21 atomic tasks, and three runs each. The task types are the part I want to describe: a bug report, a security finding, and a feature request.

56. Damra Vol 00:16:27

    Three runs each is the design decision I'd point at. One run tells you what the model did. Three runs tells you whether it does the same thing twice, and variance across runs is what determines whether you can put an agent anywhere near your issue tracker. Most published benchmarks report a single pass and don't mention it.

57. Lenar Kess 00:16:45

    And the task taxonomy maps onto a maintainer's actual inbox rather than onto an abstract capability. A bug report and a security finding demand different things — one wants a fix, the other wants you to understand a threat model before you touch anything.

58. Damra Vol 00:17:02

    Which is why a framework team building this themselves is more useful than another vendor leaderboard. They know what a good Rails patch looks like. They know the idioms that a model trained on ten years of Stack Overflow will get subtly wrong. A general coding benchmark can't grade that.

59. Lenar Kess 00:17:19

    DHH read the results as OpenAI's Luna beating GLM 5.2 and DeepSeek V4 and nearly matching the top tier. That's his ranking. I'd take the methodology as the finding and treat the leaderboard as a snapshot with a short shelf life, given that half the models in it got replaced this week.

60. Damra Vol 00:17:38

    The ranking will be stale by next Friday. The 21 tasks won't be. If you maintain a framework, the transferable thing here is the method: pick your real task types, write atomic versions of them, and run everything three times.

61. Lenar Kess 00:17:54

    Elvis was circulating something adjacent — work on measuring what a bad skill costs an agent. Not whether the agent succeeds, but the price of having a wrong tool in the library.

62. Damra Vol 00:18:05

That's the same instinct pointed at a different layer, and almost nobody asks it. Everyone measures whether adding a capability helps. Very few people measure what a subtly wrong capability costs you across a thousand runs, and the answer probably isn't zero — a bad skill doesn't just fail, it gets <emphasis>selected</emphasis> when it shouldn't be.

63. Lenar Kess 00:18:25

    This one's a process story more than a result story. An unreleased version of Claude was pointed at the Rayman hypothesis — a longstanding conjecture about how prime numbers distribute. It didn't prove the hypothesis. It did improve a related bound past the previous human record.

64. Damra Vol 00:18:41

    Say the second part again, because the distinction is going to get flattened everywhere else. The conjecture stands. A bound associated with it moved. Those are enormously different achievements and only one of them happened.

65. Lenar Kess 00:18:55

    The bound moved. And the prompting was done by someone who says they're not a mathematician, over roughly 650 attempts, where the substantive input was largely variations on "keep going."

66. Damra Vol 00:19:06

    Six hundred and fifty. That's the number I can't get past. It means the thing that produced the result wasn't insight in a prompt — it was persistence applied to a system that fails in recoverable ways. The human contribution was refusing to stop.

67. Lenar Kess 00:19:22

    It had internet access and didn't use it during the critical stretch. It went down several wrong paths, recovered, and after about 37 minutes of computation produced the improved bound. And there's a formalized version of the proof, machine-verifiable, which is why we can talk about this as a result at all rather than as a claim.

68. Damra Vol 00:19:41

    The formalization is what makes it real. Otherwise you've got a model producing confident mathematics and a non-mathematician unable to check it, which is a horror story rather than a breakthrough. A machine-checkable proof means the trust question is settled by a proof assistant instead of by vibes.

69. Lenar Kess 00:19:59

And then the odd detail. When it output the improved bound, it flagged the result as — quote — "too strong to be new."

70. Damra Vol 00:20:07

    [breath] So it pattern-matched its own correct novel result as something that must already exist in the literature. Call that a prior rather than modesty. Somewhere in there is a learned belief that results of this strength are the kind of thing humans already found, and the model applied that belief to itself and got it wrong.

71. Lenar Kess 00:20:26

    Which suggests something measurable — that a model's self-assessment is calibrated to a version of the world where it wasn't capable of this.

72. Damra Vol 00:20:34

    And it's checkable, which is the part I like. You can't usually test whether a model's confidence is well-calibrated on novelty, because novelty is hard to adjudicate. Here you can. The proof either formalizes or it doesn't, and the literature either contains the bound or it doesn't.

73. Lenar Kess 00:20:51

    Lionel Levine posted the context I'd attach to this — that when labs pursue mathematics, a large part of the motivation is automating AI research itself. Math is the domain where you can verify a step without running an experiment.

74. Damra Vol 00:21:05

    Which is a much less romantic reason to care about prime distribution than the usual telling implies, and I think it's the accurate one. Mathematics is where you get a dense, cheap, unambiguous reward signal. If you're trying to build a system that improves systems, that's an engineering decision before it's a poetic one.

75. Lenar Kess 00:21:24

    A few smaller things that stand on their own. Google shipped Gemini 3.7 Flash yesterday afternoon. It's better than 3.6 Flash on debugging and issue resolution, it generates web layouts with fewer prompts, and its PDF understanding improved. Demis Hassabis says the introductory price is half what 3.6 Flash originally cost. The Hacker News thread hit 869 points.

76. Damra Vol 00:21:49

The PDF line is the one I'd actually use. Gemini has consistently been better at document understanding than its benchmark position suggests, and if that improved again, it changes somebody's afternoon more than a coding delta does. <soft>Everyone has a pile of PDFs.</soft>

77. Lenar Kess 00:22:05

    OpenAI also shipped something into the ChatGPT desktop app called Computer History. It retains what you did across the applications and websites on your machine, so future interactions require less explanation.

78. Damra Vol 00:22:17

    The announcement is one sentence long. It doesn't say how long it's retained, whether you can scope it per-application, whether it leaves the machine, or what happens to it when you close the app. I'm not going to guess at the implementation — but those are four questions with answers, and none of the answers shipped with the feature.

79. Lenar Kess 00:22:35

    OpenAI also put out a research paper on how organizations use ChatGPT — agentic workflow spread across industries and functions. 114 points on Hacker News, and the comments went straight at the causal story.

80. Damra Vol 00:22:50

    Will Brown had the shortest demolition of it: the top companies as ranked by AI usage use more AI than companies which use AI less. [chuckle] And a commenter made the serious version of the same point — the largest companies are also the ones with the resources to run structured rollouts, so you can't separate the adoption from the budget that made adoption possible.

81. Lenar Kess 00:23:12

    It's vendor research about vendor product, which doesn't make it worthless, but does mean the comment section is doing the peer review.

82. Damra Vol 00:23:19

    Ethan Mollick was circulating it too, and he's usually careful about exactly this confound. I'd read the descriptive parts — which functions, which industries — and skip the causal ones entirely.

83. Lenar Kess 00:23:31

    Last thing, and I want to hedge it properly. There's a figure going around attributed to SK Hynix, projecting the United States and China together at 95% of compute demand by the first quarter of

2027. It reached me as a screenshot on Reddit, so I'm reporting it as a claim someone attributed to Hynix, not as a Hynix filing I've read.

84. Damra Vol 00:23:52

    Treat it as directionally interesting and numerically unverified. The item sitting next to it is better sourced. Reuters reported this morning that Apple is training its own model for the China market with Alibaba's support. That would potentially make Apple the first foreign company approved to offer its own model there. It's a sources-say story, but it's Reuters and it's specific.

85. Lenar Kess 00:24:16

    Which is a different kind of fact than a demand forecast. A forecast is a projection. A regulatory approval is an event with a date on it.

86. Damra Vol 00:24:25

    And if it lands, the question I'd ask isn't about Apple's China revenue. I'd want the approval conditions, because whatever Apple agreed to in order to get through that door becomes the template every other foreign company gets handed.

87. Lenar Kess 00:24:39

    So — Friday. Four vendors moved on price or latency in a single week, and the one number I'll carry into next week is Perplexity's: 0.496 and 0.496, $7.58 and $20.30. If that holds up when more people run it, Anthropic owes the market an argument about what the extra twelve dollars and change is buying.

88. Damra Vol 00:25:03

    The second half of the Rayman story is still open — whether the formalized proof clears verification, and whether anyone finds the bound already sitting in a paper from 1997. The model called its own result too strong to be new. Someone should go check whether it was right.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic