

Anthropic Says Claude Refuted the Exponential Time Hypothesis, and Shipped a Lean Proof to Check

2026-10-29 / 00:30:37

“Lean checks the proofs, but no machine checks that this file says what the paper says; the file is kept short so that a reader can check this.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

Anthropic's repository says an internal Claude model found a subexponential algorithm for 3-SAT and shipped a Lean proof that rebuilds in eight minutes, and its own README admits that no machine checks that the short statement file matches the paper. That question of who checks the checkable thing runs through the rest of the day, from a self-hosted coding model at Revolut to four conference talks on verification.

- [Anthropic's formal-math repository, ETH folder](#) holds the paper, the Lean 4 formalization, the $2^{O(n \log \log n / \sqrt{\log n})}$ counting bound, and the README's statement-fidelity caveat.
- [The PrimeTime's explainer](#) is our secondhand route to OpenAI's parallel claim, and it leans on cryptography implications the Anthropic README does not support.
- [Axiom Math on the Classification of Finite Simple Groups](#) describes a roughly 43-million-line Lean formalization built by agents in 24 days, the nearest data point on machine-checked proof.

- [Google's Gemini 4 Carbon announcement](#) is the source for the index rank, cyber benchmark figures, pricing, and the staged rollout, all self-reported.
- [The Verge on Gemini 4 Carbon](#) reports the index ranking relative to Claude Opus 5.5 and GPT-6 Astra.
- [The r/Anthropic post](#) carries the Fable 5.5 speculation, which is a poster's guess and not reporting.
- [Revolut's AI Engineer talk](#) gives self-reported numbers on running a GLM model on 56 B200 GPUs for about 2,000 engineers.
- [Two Minute Papers on DeepSeek v4.1 Flash](#) separates price per token from price per task.
- [The Indian Express report](#) carries the headline claim that DeepSeek and Moonshot used GPT-6 Astra to train their latest models; only the headline was available to us.
- [The Claude Code teardown gist](#) measures background-agent summary calls at 16.4 percent of spend by the author's own telemetry.
- [Sonar and Anthropic on software factories](#) is the source for 30,000 daily merges and the light-versus-dark factory split.
- [Temporal's survey of 554 engineers](#) has the 80.8 percent daily use versus 49.1 percent in production gap.
- [Qodo on code review](#) argues for independent reviewer agents and a graph of cross-repository reach.
- [Sonar's Hunter Agent demo](#) targets logic-level vulnerabilities that static analysis misses.
- [Fireship's week-in-finance rundown](#) relays the neocloud debt mechanism, CoreWeave figures, and Musk's prediction.
- [The Wall Street Journal on Nvidia and Meta](#) reports a \$3 billion Nvidia investment in Meta's equity offering, known to us through TLDR's summary.

- [Nate B Jones on Anthropic's S-1](#) reads under-one-percent enterprise spend as headroom.
- [Anthropic's drug-development post](#) lists six partners, early preclinical results, and the Mithril Medicine acquisition.
- [The AI Daily Brief](#) covers reaction to the Mithril deal and OpenAI's handling of verification for its science results.
- [The Verge on xAI in Southaven](#) relays The Guardian's report of home buyouts conditioned on not joining a lawsuit.
- [The Verge's feature on Arizona's water](#) looks at the TSMC fabs and the desert.
- [Epoch AI's chip imports explorer](#) estimates AI compute arrivals by country since 2022.
- [OpenAI's API credits demo](#) describes monthly API credits bundled with ChatGPT Pro and Business.
- [OpenAI and ServiceNow](#) expand their partnership across enterprise workflows.
- [Latent Space with the Supabase founders](#) covers Compute, OrioleDB, MultiGres, and the Turso acquisition.
- [Qwenputer](#) describes a deterministic RISC-V CPU compiled into Qwen3.5's weights.

SEGMENTS

[00:00:04](#) Anthropic's Claude says it refuted the Exponential Time Hypothesis

[00:06:30](#) Gemini 4 Carbon

[00:10:07](#) Revolut self-hosts its coding model

[00:15:31](#) Generating code is cheap, verifying it isn't

[00:20:02](#) Who pays for the GPUs

00:22:52 Claude in the lab, and xAI in Southaven

00:26:38 Credits, databases, and a CPU inside Qwen

Transcript

1. Lenar Kess 00:00:04

Picture opening a folder on GitHub where the paper is titled A Two-to-the-little-o-of-n Time Algorithm for 3-SAT, and the author line says Claude. Not someone using Claude. Claude. The folder sits in Anthropic's formal-math repository, and the README opens by saying a deterministic algorithm solves 3-SAT in subexponential time, which would refute the Exponential Time Hypothesis, an assumption complexity theorists have leaned on for decades.

- github.com
- youtube.com
- axiommath.ai
- blog.google
- theverge.com
- reddit.com
- youtube.com
- youtube.com
- indianexpress.com
- gist.github.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- wsj.com
- youtube.com
- anthropic.com
- youtube.com
- theverge.com
- theverge.com
- epoch.ai
- youtube.com
- openai.com
- youtube.com

- divagr.com

2. Damra Vol 00:00:30

Okay, but who checked it? And what does checked even mean when the paper and the proof were both written by models?

3. Lenar Kess 00:00:36

We'll get there, and the repo has a more direct answer than I expected. Here's the day. This proof claim comes first, then Google's Gemini 4 Carbon, then a European bank that says it cut its coding-agent bill to a tenth by hosting an open model itself. Four conference talks follow, all about the gap between generating code and trusting it. Then the money behind the GPU buildout, and some shorter items, from Anthropic's drug-development work to a CPU running inside a language model. The claim comes first, though, since the README is precise about it.

4. Damra Vol 00:01:09

Right, because refuted a famous hypothesis is a big sentence to hang on one README.

5. Lenar Kess 00:01:14

It is. 3-SAT is the problem of deciding whether a formula made of three-literal clauses can be satisfied. Every known algorithm takes time exponential in the number of variables, and the hypothesis says that's unavoidable. The README states: "A deterministic algorithm solves 3-SAT on n variables in time two to the little- o of n , which refutes the Exponential Time Hypothesis. The algorithm in fact solves the harder problem, counting the satisfying assignments, in time two to the $n \log \log n$ over the square root of $\log n$." And it says the result "was discovered by an internal research Claude model."

6. Damra Vol 00:01:50

Little- o of n sounds like a collapse, but look at that exponent. Take a million variables. $\log n$ is about twenty, and the log of that is about four. The square root of twenty is about four and a half, so the factor in front of n is close to one. Ignoring the hidden constant, the discount at that size is close to nothing. That's my arithmetic, not theirs, but it says the practical picture hasn't moved.

7. Lenar Kess 00:02:14

The README says it outright: "The result is purely asymptotic and says nothing about solving formulas in practice." That matters because The PrimeTime's explainer, which is how a lot of people met this story, goes straight to RSA and SHA-256 falling over and protein folding getting solved. The repo makes no such claim.

8. Damra Vol 00:02:33

And there's a logic gap in that jump. The hypothesis is a stronger assumption than $P \neq NP$. Refuting it doesn't hand you $P = NP$, it says one specific lower bound people believed in is false. That's a big deal for the people who build conditional lower bounds on top of it, and almost nothing for your TLS handshake.

9. Lenar Kess 00:02:53

Now the Lean side, which answers your question better than I expected. There's a single file called `MainResults` that states two theorems, that the hypothesis is false and the counting algorithm with its time bound, on Mathlib's Turing machines. The README says the file “contains no proofs and imports nothing but Mathlib.” The proofs live elsewhere, and a tool called `Comparator` checks that the statements match a challenge file and re-checks the proofs with a second, independently written kernel called `nanoda`. The axioms it prints are the three standard ones.

10. Damra Vol 00:03:24

And the build time?

11. Lenar Kess 00:03:25

About eight minutes. The README gives the machine as twenty-four virtual CPUs, with the largest process using about seven gigabytes of memory.

12. Damra Vol 00:03:34

A formalization you can rebuild over a coffee. I like that. But read me the caveat, because I think that's the interesting sentence.

13. Lenar Kess 00:03:41

“Lean checks the proofs, but no machine checks that this file says what the paper says; the file is kept short so that a reader can check this.” Appendix E of the paper prints the file and lists how it differs from the statements printed in the paper.

14. Damra Vol 00:03:55

So the trust problem moved. Nobody has to believe the proof anymore, the kernel does that. What a human has to believe is that one short file means what it appears to mean. If a quantifier is slightly off, you get a flawless proof of the wrong theorem. Models wrote the paper, models wrote the formalization, and the human job is reading the one file that was kept short on purpose.

15. Lenar Kess 00:04:17

There's one more detail I liked. The README carries a condensed sketch of the key idea, built on tensor-network contraction, and it ends with the words "Prove it." The README says the sketch only rederives that 3-SAT can be decided in time $2^{o(n)}$ for every positive δ . It doesn't reach the sharper bounds at the top of the page. The repo itself is labeled "Research artifact. Not maintained and not accepting contributions."

16. Damra Vol 00:04:42

So it's a lab publishing a challenge, more or less. What about the OpenAI side? You said this showed up on the same day.

17. Lenar Kess 00:04:49

The TLDR summary says Anthropic's repo appeared hours after OpenAI's own refutation, and that's all I have on OpenAI's paper, plus The PrimeTime's explainer. That explainer gives the bound as $2^{o(n)}$, which is missing the $\log \log$ factor Anthropic has. It could be a different algorithm or a loose summary. The AI Daily Brief adds that OpenAI appointed expert groups to verify its recent math results, and that some academics were annoyed about findings OpenAI kept private while they were being formalized.

18. Damra Vol 00:05:19

Fireship described OpenAI's result as disproving a twenty-seven-year-old complexity theory conjecture, and now there's a second lab with a second proof inside a day. Either that's coincidence, or both labs are pointing very capable models at the same pile of open problems and collecting what falls out. I lean toward the second.

19. Lenar Kess 00:05:39

The nearby data point is Axiom Math. Per the TLDR summary of their post, Axiom released the first complete computer-checked proof of the Classification of Finite Simple Groups, a proof that ran to about fifteen thousand pages across more than five hundred articles by about a hundred mathematicians. It took dozens of agents twenty-four days and produced about 43 million lines of Lean.

20. Damra Vol 00:06:01

That's the mirror image. Anthropic's result is a model producing a new proof, Axiom's is models transcribing a human mountain into something a kernel can check. Forty-three million lines that no person will ever read, and the same question applies: somebody has to audit the statements, not the proofs.

21. Lenar Kess 00:06:19

Until outside complexity theorists have read the paper itself, what exists is a claim with an unusually inspectable attachment: eight minutes of build time and a statement file short enough to read over lunch.

22. Lenar Kess 00:06:30

Google's announcement post says Gemini 4 Carbon “ranks #1 on the recently updated v5.0 of Artificial Analysis Intelligence Index at 59.” That's Google's sentence about a third party's index, and per The Verge's writeup, Claude Opus 5.5 and GPT-6 Astra sit behind it. Google calls Carbon “our strongest model with state-of-the-art cybersecurity and long horizon capabilities.”

23. Damra Vol 00:06:55

So the rank is Google quoting a scoreboard it's winning. Fine. What else is in the post?

24. Lenar Kess 00:07:01

Coding and agentic work get the phrase “significant gains.” On WebDev Arena, the post says Carbon gained eighty-nine Elo points to reach number four, at 1768. Google also says internal teams watched it work for days on multi-step engineering problems, like tracking down subtle bugs in large codebases.

25. Damra Vol 00:07:21

Number one on the index and number four in the arena. [chuckle] I'd read the pitch as the days-long runs rather than the leaderboard position, since that's the claim a leaderboard can't capture.

26. Lenar Kess 00:07:31

The cyber section is where the post spends the most numbers. On external benchmarks, evaluated with the safety restrictions turned off, it reports 88.2 percent on SEC-bench Pro, which tests finding vulnerabilities. It gets 54.8 percent on ExploitGym for exploitation and 71 percent on CWE-bench for patching. Internally, Google claims more than a 2x uplift in generating secure code over Argon, the earlier Gemini 4 model. It also says the Chrome team found Carbon “more than 10X better than another lab's large frontier model” at producing quality patches.

27. Damra Vol 00:08:06

It finds the bug, exploits it, and patches it, all in one model. And look at the rollout. This week it goes to a subset of cyber defenders, then to trusted partners through something called the Fairwind Program, and general release is “as soon as possible” with no date. That's a release plan that admits what the model can do. Also, which other lab's model lost by ten times to it in Chrome's patch quality? The post doesn't say.

28. Lenar Kess 00:08:31

The safety numbers are unusual for a launch post. Successful cheating on Terminal-Bench 4.0 fell to 0.0 percent, with cheating attempts at 0.4 percent. High-severity misalignment alerts in internal live coding dropped by almost eighty percent. And on Gray Swan's prompt-injection benchmark, the attack success rate is 0.0 percent against Argon's 0.7.

29. Damra Vol 00:08:54

That a company now reports how often its model cheats on a coding benchmark, and treats it as a number to push down, tells you what the failure looked like in the previous generation. But every one of these figures is Google's own, and the post gives no independent verification.

30. Lenar Kess 00:09:09

The pricing is the one detail you can check later against the product page. The introductory rate is two dollars per million input tokens and ten per million output, with cached input at ninety-five percent off. A footnote says that after the introductory period it goes to four and twenty.

31. Damra Vol 00:09:26

So the number to remember is the footnote. The headline price is half the real one.

32. Lenar Kess 00:09:31

On the reaction side, the Hacker News thread on Google's post sits at 954 points and 763 comments. There's also a Reddit post on the Anthropic subreddit titled "Google has beaten anthropic for the first time," with a parenthetical guessing at a Fable 5.5 release soon, three question marks included. That's a poster's speculation, not reporting.

33. Damra Vol 00:09:53

Which is how every model release now goes: a rank, a rumor about the rival's answer, and access limited enough this week that almost nobody can test the cyber claims. For now Google's table is the only evidence on the page.

34. Lenar Kess 00:10:07

At the AI Engineer conference, David, Revolut's head of AI operations, described moving about two thousand engineers off hosted Gemini and Claude Code and onto an open GLM model the bank runs itself on 56 B200 GPUs. Revolut says its blended cost is about thirty cents per million tokens, and that its AI invoice dropped to roughly a tenth of what it had been paying providers like Anthropic and OpenAI. Every number there is Revolut's own, from a conference talk.

35. Damra Vol 00:10:35

A tenth is a headline number. I'd like to know what the denominator includes, because fifty-six B200s aren't free, and the people keeping them tuned aren't either. The talk gives a blended cost per token, and I can't tell how much of the hardware and staffing it counts.

36. Lenar Kess 00:10:51

What the talk does give is the engineering. Revolut's target is a time to first token under three seconds, and latency above five seconds hurts the experience. Untuned, the time to first token was over ten seconds. After tuning it got to 1.6 seconds, while throughput went from three requests per second to over seven. The talk says two hundred tokens per second is enough for interactive coding.

37. Damra Vol 00:11:14

So the tuning is the product. They run four-bit floating point on Blackwell GPUs, which the talk says saves about forty percent of the GPU footprint without precision loss, and they pick tensor parallelism around the limits of the key-value cache. They also aim for a cache hit rate above ninety percent on coding traffic, which makes sense when every request in a session re-reads mostly the same context.

38. Lenar Kess 00:11:38

Then there's the compliance argument, which is the one I wanted to hear. Revolut is running a Chinese open model inside a European bank. It says it keeps data residency inside equivalence frameworks covering the US, EU, UK, and Israel. It also enforces human-in-the-loop controls through sandboxes that restrict autonomous actions. Image tasks get routed to other models like Gemini, and GLM stays on text coding. On Monday we had IBM's CEO saying open models are about ninety percent as good at a fraction of the price. This is one bank's version of that claim, with GPUs attached.

39. Damra Vol 00:12:12

And the argument puts the risk in where the servers sit, not in the weights. That holds up, since a set of weights is a file that doesn't phone home unless you let it.

40. Lenar Kess 00:12:21

That's cost per token. Two Minute Papers, on DeepSeek v4.1 Flash, makes the other half of the point. The release is called 1017 and came five weeks after the previous one. It has strong terminal and coding performance, open weights, and a price a bit over a dollar per million output tokens. But the video says independent testing shows it using two to three times the tokens of Claude Opus on the same task. That makes the cost per task higher than the sticker suggests, though still favorable.

41. Damra Vol 00:12:50

Which is the Revolut number's weak spot. Thirty cents per million tokens only helps if a task doesn't take three times the tokens. Price per token and price per task are different bills, and the talk reports the first one.

42. Lenar Kess 00:13:03

The video credits the jump to the model learning from its own feedback, critiquing and refining its own outputs, since the architecture didn't change. That's the host's reading of the method rather than something DeepSeek published. Separately, the Indian Express ran a headline this morning: "DeepSeek, Moonshot used OpenAI's GPT-6 Astra to train latest models: Report." All I have is the headline, so I can't tell you who alleged what or on what evidence.

43. Damra Vol 00:13:29

There are two explanations for a big jump with no architecture change, self-feedback or training on a stronger model's outputs. They're not exclusive, but a video can only be confident about one of them. And Monday's episode flagged a Kimi K3.1 Pro weight release for November fourth, so we'll get to compare another open release on the same questions soon.

44. Lenar Kess 00:13:51

There's a cost in the other direction, in the tools themselves. A gist that reached 423 points and 155 comments on Hacker News documents that Claude Code takes the entire transcript of each background agent and sends it to that agent's own model, usually Opus, about every thirty seconds, to write a short status line in the agents panel. The reply never goes back into the conversation.

45. Damra Vol 00:14:13

A status line. By the gist author's telemetry, those calls made up 16.4 percent of Claude Code spend, more than the main conversation's 15.7 percent, and about one in five API requests. The gist cites an issue with four independent measurements between eighteen and twenty-six percent.

46. Lenar Kess 00:14:32

Caching softens it, since about ninety-eight percent of those calls read the transcript from cache, though the cold-cache two percent account for over a third of the cost. And these requests don't appear in the on-disk session logs, so cost tools that read transcripts miss them. Only the OpenTelemetry export shows them, under a query source labeled agent summary. The author says no documented setting turns it off, and that it's been on by default since version 2.1.232. Those numbers are the author's own measurement, and the cost figures are Claude Code's per-request estimates, not invoices.

47. Damra Vol 00:15:06

The workaround is a Bun preload that wraps the timer functions and swaps the summary timer for one that never fires, matched by source-text strings, so an update could break it with no warning. One commenter also reported the user's email shows up in the system prompt, which is a single commenter's claim. The practical loss on a subscription is rate-limit headroom, and I find it a little funny that the telemetry which exposed it was in the product the whole time.

48. Lenar Kess 00:15:31

Anthropic's Addy Osmani said in a talk at the AI Engineer conference, alongside Manish Kapoor of Sonar, that Anthropic has scaled from 2,500 to 30,000 daily code merges, with Claude doing a first-pass review. That figure comes from a conference talk rather than a filing. It's one of four talks posted this week that keep circling the same problem: generating code is cheap, and verifying and operating it isn't. Yesterday's episode was about agents failing at tasks. These are about teams coping with what the agents produce.

49. Damra Vol 00:16:01

Thirty thousand merges a day is the kind of number that makes review the whole job. What does Anthropic's review process look like at that volume?

50. Lenar Kess 00:16:09

Per the talk, Claude does the first pass, including end-to-end testing and linting. Automated hooks correct bugs found in diffs. Teams codify their practices in a Claude.md file, and for UI changes the agent produces verification evidence, like browser videos or screenshots. Osmani separates light factories from dark ones. In a light factory, human judgment stays central to planning, building, and shipping. In a dark factory the agents run the whole loop from scoping to shipping, and people step in only for exceptions or to define intent. He predicts most teams will be light factories with some dark corners by 2027. The talk also says three in ten merge attempts fail.

51. Damra Vol 00:16:50

Browser videos as evidence is the detail I like. The agent attaches a screen recording to the pull request the way a person would attach a screenshot to a bug report. That's verification evidence you can skim in ten seconds instead of re-deriving.

52. Lenar Kess 00:17:04

The Temporal talk has the most checkable numbers. A developer advocate named Shai presented a survey of 554 engineering professionals. Of those, 80.8 percent use agents daily, but only 49.1 percent have them in production. Trust is high at 85.5 percent, and yet 41.1 percent hit issues daily

and 16.4 percent hit them hourly. The top productivity blocker is state management at 35.7 percent, ahead of model performance. Another 22.2 percent have lost progress to crashes.

53. Damra Vol 00:17:37

Trust at eighty-five percent and daily problems at forty-one is an uncomfortable pair. The talk also says teams describing themselves as successful were 1.7 times more likely to face continuous issues. I read that as teams doing more with agents seeing more failures. Temporal sells durable execution, so keep in mind who is presenting the survey.

54. Lenar Kess 00:17:58

Their demo makes the pitch concrete. It's a nine-stage payout workflow with a language-model fraud-investigation agent in the middle. They kill the worker process on purpose, and the system keeps the agent's work in an event history and resumes from the last successful step, without redoing the model calls or losing reserved funds.

55. Damra Vol 00:18:16

Which is a reasonable thing to want. If an agent has spent a long run investigating something, a crash that throws that away is its own cost. Whether you need a vendor for it or a database table is a separate argument.

56. Lenar Kess 00:18:29

Then Qodo. Their example comes from Bitwarden: a one-line change to a data model that ripples out to the backend servers, the web and mobile clients, and the CLI tools, yet still passes unit tests and linters. To catch that, they hand review agents a graph of how repositories connect. They also add independent reviewers, so the implementing agent isn't grading its own work, possibly across multiple models and with adversarial reviewers. And they run checks while the agent is partway through, at twenty-five, fifty, and seventy-five percent of generation.

57. Damra Vol 00:19:01

Qodo's point that an agent grading its own work is a weak check is hard to argue with. But every one of these talks is also a pitch for the layer the speaker sells: Sonar, Qodo, and Temporal each describe a gap that happens to match their product. Doesn't mean they're wrong. It means the survey numbers are better evidence than the architecture diagrams.

58. Lenar Kess 00:19:21

The fourth is Sonar's Hunter Agent, from Taylor Luttrell. It targets logic-level and intent-based vulnerabilities that pattern-based static analysis misses. Examples include broken access control

and skipped workflow steps, and it also goes after missing rate limits and session fixation. It maps the repository, runs three specialized agents, and uses adversarial validation to filter out false positives. In the demo, a sample Python application turned up 46 open issues.

59. Damra Vol 00:19:50

A sample application, so forty-six tells you the demo works. It doesn't give you a hit rate on real code. What I'd want is the false-positive rate on a codebase nobody built to be vulnerable.

60. Lenar Kess 00:20:02

The Wall Street Journal reports, per TLDR's one-line summary of a paywalled story, that Nvidia will invest three billion dollars in Meta's equity offering as the two expand their AI partnership. TLDR adds that Meta intends to increase its development on Nvidia infrastructure, and that they're considering working together in other areas, like monitoring agent behavior. That's the most concrete data point in a week of anxious financing videos.

61. Damra Vol 00:20:27

And the anxious videos are the context. Fireship's rundown of the week has Meta issuing non-voting stock to fund GPU purchases and Microsoft offering voluntary exits. It also has CoreWeave's valuation at a third of its monthly high and a seven percent drop in SpaceX stock. Fireship is a comedic, fast format, so I'd treat those as relayed.

62. Lenar Kess 00:20:49

Fireship walks through the mechanism like this. Neoclouds are companies that borrow money to rent Nvidia GPUs out. In his telling, Nvidia lends to them, they borrow from Wall Street using future GPU rental income as collateral, and the loop reinforces itself until debt markets tighten. He says CoreWeave has borrowed fifty-two billion dollars and now faces fifteen percent interest. He adds that Elon Musk predicts more than two hundred neocloud failures and says debt markets have shut for AI-only data center companies.

63. Damra Vol 00:21:18

Those are the video's figures and Musk's prediction, relayed. If they're right, Nvidia putting three billion into Meta's raise is the same direction of cash seen from another angle. Meta isn't a neocloud, it has an enormous business paying for the GPUs, but a supplier funding part of its customer's purchase has the same circular structure in miniature.

64. Lenar Kess 00:21:39

The counterweight is Nate B Jones on Anthropic's S-1. We went through the filing on the twenty-seventh, so just his argument. He cites about sixteen billion dollars in first-half revenue, and says the average publicly traded customer spends less than a dollar per thousand dollars of expenses with Anthropic. Even the ten largest Fortune 500 customers, spending more than eight times the average, are under one percent of their expenses. He concludes this is headroom, not a ceiling.

65. Damra Vol 00:22:06

I find the headroom reading more persuasive for the labs themselves, because low penetration plus a customer base that keeps widening is how a market looks early. But it answers a different question than the neocloud one. Headroom is about whether revenue arrives eventually. The debt story is about whether the people who borrowed against rental contracts can make payments before that revenue shows up. Both can be true at once.

66. Lenar Kess 00:22:30

Fireship's own conclusion points the same way. The industry isn't collapsed, and for developers a correction brings subsidized token prices and open models that approach Opus at half the cost.

67. Damra Vol 00:22:41

Subsidized token prices are the bit I'd hold onto. A neocloud with a debt payment due has a reason to fill its racks at almost any price. That's my read of Fireship's point, not his words.

68. Lenar Kess 00:22:52

Anthropic published a research post on its life sciences team's deep research partnerships. The partners are BigHat, Bristol Myers Squibb, Divergence Labs, Enveda, Generate Biomedicines, and The Column Group. Claude was applied at different stages of drug development, not just molecule design. The post is explicit about the stage: "All of this research is still at an early stage." Most of it is preclinical, meaning not yet tested in humans.

69. Damra Vol 00:23:19

What did it actually do in each one?

70. Lenar Kess 00:23:21

At Generate, scientists used Claude Code to propose shorter genetic components for a CAR T redesign, one about seventy percent shorter than the standard version that performed better in lab tests, with about ten times as many receptors. At Divergence Labs, Claude turned goals into a design plan and produced about 35,000 antibody designs in about a month, and early tests found three IgE

binders and twenty designs that dislodged IgE. At BigHat, the lab tested about 1,010 variants Claude proposed, and the best bound more than ten times more tightly.

71. Damra Vol 00:23:55

Those are all cases where Claude proposes and a lab tests. The one with a checkable comparison is Enveda, in manufacturing and filings. The post says Claude “caught all 11 issues that human reviewers had spotted—and flagged seven they'd missed,” and review time dropped eighty-nine percent. Eleven out of eleven plus seven extras is the most direct comparison against human reviewers in the post.

72. Lenar Kess 00:24:18

There's also an acquisition in the post: “We've acquired Mithril Medicine, a one-person company.” The founder, Cody Tranbarger, used Claude to coordinate contracts, logistics, and experiments across eleven outside labs, going from an idea to promising results in mice in under eighteen months. He now works at Anthropic. If the candidates hold up, human trials could begin as soon as late 2028. Anthropic has also started its own programs on rare childhood forms of muscular dystrophy, where it says Claude handles much of the scientific and operational work.

73. Damra Vol 00:24:51

The mechanism that jumps out is the coordination. A single person managing eleven outside labs, contracts and logistics included, is Claude as project manager for contract research. The AI Daily Brief covered the reaction. Some people praised one person doing this, and some doubted the founder's scientific background. One reading called it a pre-IPO demonstration that Claude can solve problems beyond code, possibly with Anthropic owning outcomes instead of selling intelligence. Those are the Daily Brief's readings, and I'd add that mouse results are a long way from a medicine.

74. Lenar Kess 00:25:24

Different kind of story next. The Verge, relaying reporting from The Guardian, says a subsidiary of xAI is offering to buy homes in a Southaven, Mississippi neighborhood, on the condition that residents don't join a lawsuit targeting the company's nearby data center and power plant. The Verge's headline puts it as xAI buying out homeowners who promise not to sue over data center noise.

75. Damra Vol 00:25:46

The condition is the detail. A buyout is a fair deal for someone who wants to leave. A buyout conditional on staying out of the lawsuit is also a way to shrink the group of plaintiffs. I read it that way, though xAI's side isn't in what the Verge summary gives us.

76. Lenar Kess 00:26:01

Elsewhere in the physical world, The Verge has a feature by Gaby Del Valle called “AI has come for Arizona's water,” which opens on a TSMC fab as “a hulking gray giant” rising from the desert. And Epoch AI released an AI Chip Imports explorer, which estimates AI compute arrivals by country since 2022 and sits alongside their existing data on chip components, sales, and data centers.

77. Damra Vol 00:26:25

I'd keep those separate from Mississippi, since a lawsuit and a water feature don't add up to a theory of anything. The Epoch explorer is the reference tool of the three: if you want to know where the chips physically ended up, it's a place to look.

78. Lenar Kess 00:26:38

OpenAI says eligible ChatGPT Pro and Business subscribers can claim monthly API credits by linking an API organization in the Billing section of the ChatGPT app. In OpenAI's own demo video, the Pro 200 plan includes two hundred dollars in monthly API credits and the Pro 500 plan includes five hundred. The demo uses the OpenAI Developers plugin in Codex to build a Mac app that judges whether a message is ready to send, using the Decisions API, with GPT-6.1 Terra suggesting edits.

79. Damra Vol 00:27:09

The credits blur the line between a chat subscription and an API bill. You pay for a plan to talk to a model and get a prototyping budget attached, which is a decent nudge toward building something. OpenAI also expanded its ServiceNow partnership, bringing OpenAI models and agents into ServiceNow, and ServiceNow workflows into ChatGPT and Codex, which is a routine enterprise distribution item.

80. Lenar Kess 00:27:34

Latent Space interviewed the Supabase co-founders after Supabase Select, and the announcements fall into four. Supabase Compute is a serverless, long-running environment placed next to the database. It replaces the Deno-based Edge Functions with a container runtime that supports Rust and Python. OrioleDB is a drop-in replacement for the Postgres heap storage engine. Supabase claims it's 1.5 to 2 times faster because it removes vacuum overhead, and it's Apache 2 licensed. MultiGres is a rewrite of a Postgres operator, originally from the Vitess founder, aimed at zero data loss on failover and eventually sharding.

81. Damra Vol 00:28:10

And the fourth is the Turso acquisition. Turso is SQLite with multi-version concurrency control, which allows multiple writers, and it's being adapted to use object storage for durability. Supabase's pitch expects agents to launch billions of databases a month, and a SQLite-style file is a natural unit for that. It's a founder interview, so the speed figure is Supabase's own claim. But the databases-per-month jump is a prediction I find believable, since an agent that needs scratch state can create a database where a person would have created a spreadsheet.

82. Lenar Kess 00:28:44

Last one, a palate cleanser. A blog post called Qwenputer reached 522 points on Hacker News, describing a functioning RISC-V CPU implemented inside Qwen3.5. Qwen's Gated DeltaNet layers, which are a linear recurrent network, become RAM. With the gate and write strength both set to one and one-hot keys, each write “will overwrite exactly one row,” and a one-hot query reads back an exact value.

83. Damra Vol 00:29:10

No training involved?

84. Lenar Kess 00:29:12

None. A compiler sets the weights directly: attention heads do lookups, feed-forward units do arithmetic, and the DeltaNet heads hold memory. The CPU adds eleven layers in front of Qwen's thirty-two, and its instructions come out as a few generated tokens each, so the CPU runs as the model decodes. The post doesn't define deterministic. Its evidence is exact memory, eighteen million instructions run free without errors, and about 230,000 output positions replayed against a reference machine with zero mismatches.

85. Damra Vol 00:29:43

What can it run?

86. Lenar Kess 00:29:44

It's a 32-bit RV32IM core with a small bignum unit. It passes the official riscv-tests, runs C and MicroPython, and the author says it's “not a Linux-class chip yet.” On long-integer arithmetic, accuracy goes from 48.6 percent to 100 percent. It's slow, since each instruction costs a few decoded tokens: about 3,200 CPU tokens per second on an L4. The checkpoint is 18.8 gigabytes, roughly double base Qwen, and needs its own driver rather than a plain generate call.

87. Damra Vol 00:30:16

A language model that does exact long-integer arithmetic by being a computer, rather than by approximating arithmetic. It's a neat artifact and nobody should deploy it, which I mean as the highest praise.

88. Lenar Kess 00:30:28

Kimi's K3.1 Pro weights are due November fourth, and I'll read the license before the benchmark table. I'm Lenar.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic