The Table Arrived First

2026-08-04 / 00:22:24

> *“A Lean kernel doesn't care who trained the model that wrote the proof.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Alibaba promised Qwen3.8 weights and delivered a benchmark screenshot instead — which is the shape of most of today: claims arriving ahead of the artifacts that would let anyone check them. Astra's math results now have a problem list from OpenAI and a Reddit-sourced story about Lean 4 certificates. A leveraged AI fund got margin-called out of its entire public book. And a Hacker News argument about developer tools turns on whether large language models just made the freedom to modify real.

- A results screenshot for Qwen3.8-Max posted to r/LocalLLaMA puts a 2.4-trillion-parameter model level with Kimi K3 and DeepSeek V4-Flash, and ahead on coding — but it's an image on Reddit, not an Alibaba model card, and the weights still aren't up.
- Nathan Lambert's new Artifacts Hub and adoption dashboard concede that the release rate has outrun anyone's ability to track which releases mattered — the person best equipped to hold it in his head built a database instead.
- An anonymous post from someone claiming to work at a Chinese lab maps four labs onto four different bets — coverage, capability per training compute, agentic long-

horizon work, and serving cost — and Cloudflare's write-up on running Kimi and GLM at scale is the only outside measurement of the last one.

- OpenAI published the actual problem list for Astra — sphere packing, coding theory, lattice cryptography, and non-sofic groups — while the claims about machine-checkable Lean 4 certificates and a ~$2,000 inference bill come only from a Reddit thread.
- Leopold Aschenbrenner's fund was margin-called out of its public equities book and Citadel bought it at a discount — a thesis that worked, levered, and liquidated on timing rather than on being wrong.
- A 636-point argument that devtools must be open source collides with Elon Musk's claim that source code is on the verge of becoming like assembly — two incompatible readings of what models do to human-readable code.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

Yesterday we closed the show on a promise with a date attached to it — Alibaba saying the Qwen3.8 weights would be up next week. We were a little sour about it. Then overnight the sequence went sideways in an interesting way, because the weights still aren't up, but the benchmark table is. There's a post on the r-slash-LocalLLaMA subreddit from a user called davidthesong with a results screenshot for Qwen3.8-Max — two point four trillion parameters — and the claim is that it sits roughly level with Kimi K3 and DeepSeek V4-Flash, and ahead of both on the coding and software-engineering rows.

- i.redd.it
- x.com
- x.com
- x.com
- x.com
- x.com
- i.redd.it
- blog.cloudflare.com
- reddit.com
- reddit.com
- youtube.com
- youtube.com
- youtube.com
- theregister.com
- reddit.com
- x.com
- x.com
- reddit.com
- x.com
- x.com
- x.com
- hoplite.sh
- armature.tech
- x.com
- blog.cloudflare.com
- reddit.com
- blog.exe.dev
- x.com
- hamvocke.com
- x.com
- x.com

2. Damra Vol 00:00:43

   [tsk] A screenshot. That's the provenance story right there, before anything else. It's not an Alibaba model card, it isn't a post from the Qwen team, it's an image on Reddit with zero points and zero comments at the time it reached us. The numbers might be exactly right. Nobody in this conversation has seen the artifact they came from.

3. Lenar Kess 00:01:03

   Agreed, and hold that, because elvis — omarsar0 on X — picked it up the same afternoon, and the reason it travels is the comparison target. Not good for an open model. Level with the two systems people currently reach for when they want frontier behavior without a frontier bill.

4. Damra Vol 00:01:20

   The coding row is the one I keep going back to. Sitting level on general reasoning tells you about training compute. Being ahead on software engineering tells you about post-training data and how much agentic trajectory work somebody did. Those are different achievements, and inside the same building they usually come from different teams.

5. Lenar Kess 00:01:40

   We start there today. After that, an unusually specific post from someone who says they work at one of the Chinese labs, laying out how four of them are making structurally different bets. Then Astra — OpenAI put out the list of math problems, and there's a cost number attached now. Then the money, including a leveraged fund that got margin-called out of its entire public book. Then the White House meeting happening today, an argument about developer tools that pulled six hundred and thirty-six points on Hacker News, and a handful of agent shipments.

6. Damra Vol 00:02:12

   Before the labs, though — the other thing that shipped yesterday says something about the release rate itself. Nathan Lambert and the Interconnects team put up an Artifacts Hub, plus an adoption dashboard, on the same day the Qwen numbers went around.

7. Lenar Kess 00:02:27

   Same day, and that's what makes it interesting. Lambert's job for the last two years has been reading every open-weight release and writing up what changed. He's about as well-equipped as anyone alive to keep that in his head. He built a database instead.

8. Damra Vol 00:02:41

Florian Brand — xeophon — reposted it, and the piece he pulled out was the adoption dashboard, which he called the depressing half. I know exactly what he means. The release list is the fun page. The adoption page is where you find out that a large fraction of these models get downloaded a few thousand times and then nothing happens.

9. Lenar Kess 00:03:01

   So the release rate has outrun our ability to know which releases mattered.

10. Damra Vol 00:03:06

    Which is different from outrunning the ability to read them. I can read a model card in ten minutes. What I can't do is tell you six weeks later whether anybody built on it. That's a longitudinal question, and it needs a table, not a person.

11. Lenar Kess 00:03:21

    Does having the table change how you'd read a day like today?

12. Damra Vol 00:03:24

    It changes what I'd want from the Qwen3.8-Max row specifically. Not the benchmark column — the download and derivative count three weeks out. If a two point four trillion parameter model ships and almost nobody can serve it, the benchmark table is a statement about what Alibaba can do, and the adoption table is a statement about what anybody else can do with it.

13. Lenar Kess 00:03:47

    And the weights still have a date on them rather than a URL. When they show up, the license text is what I'll read before the numbers.

14. Lenar Kess 00:03:55

    The post I keep coming back to from that same subreddit is titled, roughly, the Chinese labs everyone lumps together are making four pretty different bets, I work at one of them. It comes from an account called AcanthisittaOk1699, it's anonymous, and it self-identifies as an employee without naming the employer. Treat it accordingly. What makes it readable anyway is that the claims are about design choices rather than impressions.

15. Damra Vol 00:04:23

    And the four aren't ranked, which I appreciated. He puts Alibaba on coverage — every size and every modality, be the default everywhere. DeepSeek on capability per unit of training compute.

Moonshot chasing agentic behavior and long-horizon tool use. And Ant, with the Ling models, optimizing for serving cost, which is the bet nobody writes headlines about.

16. Lenar Kess 00:04:46

    Serving cost is what I'd bet gets underrated for another year. Everybody grades these labs on the benchmark table because the benchmark table is public. Nobody publishes tokens per dollar per unit of quality on their own hardware.

17. Damra Vol 00:05:00

    Cloudflare sort of did, though. They put up a post yesterday — smaller, faster, safer, running Kimi and GLM at scale — that pulled two hundred and thirty points on Hacker News. That's the other half of the picture. One source is a claim about intent, and the other is a measurement of what serving those models actually costs somebody who serves a lot of them.

18. Lenar Kess 00:05:22

    So why does Cloudflare's number carry more with you than Ant's stated intent?

19. Damra Vol 00:05:27

    Because a lab telling you it optimized for serving cost is unfalsifiable until somebody outside the lab runs the model at volume and reports numbers. Cloudflare has the volume. When the people paying the electricity bill write up which models they chose and why, that's a stronger signal about the design bet than the design bet's own press release.

20. Lenar Kess 00:05:48

    There's a third data point in the same neighborhood, much smaller and much more human. Somebody on that subreddit posted, and I'm quoting the headline — "I can't believe I've got DeepSeek-V4-Flash-0731, a frontier model, running on my home PC. Insane." The can't is in capital letters.

21. Damra Vol 00:06:07

    [chuckle] And the capital letters are the actual content of that post. It isn't a benchmark. It's a person having the specific feeling of a floor moving under them. Eighteen months ago that model class needed a rack and a colocation contract.

22. Lenar Kess 00:06:21

    There's a companion post from a user called EmPips — V4-Flash-0731, vibes after the first weekend of use — and it names the quantization before it names the verdict, which is what more practitioner

posts should do.

23. Damra Vol 00:06:36

    Which matters because half the arguments about whether an open model is any good turn out to be arguments about which quantized build somebody ran. Second-weekend impressions aren't benchmarks, and I'd still rather read a hundred of them than one vendor chart.

24. Lenar Kess 00:06:51

    So the map is four bets, one anonymous cartographer, and one company with a serving bill big enough to check part of it.

25. Lenar Kess 00:06:59

    On Sunday we spent a chunk of the episode on OpenAI's claim that Astra had solved ten long-open math problems, and the complaint was that the claim was much bigger than the artifact. There was no list. Two days later there's a list. OpenAI posted specifics. The problems run through sphere packing, coding theory, and group theory, and they carry on into quantum complexity, lattice cryptography, and extremal combinatorics — one of them establishes the existence of non-sofic groups.

26. Damra Vol 00:07:30

    And a separate claim on the r-slash-singularity subreddit that the proofs are formalized in Lean 4, with machine-checkable certificates in a GitHub repository. That same post mentions a two hundred and forty-nine page manuscript, and says the inference bill for the whole run came to about two thousand dollars. Every part of that comes from Reddit, not from OpenAI. The problem list is first-party. The repository, the formalization, and the price aren't.

27. Lenar Kess 00:08:00

    The formalization is what changes the epistemics, if it holds up.

28. Damra Vol 00:08:04

    It's the only part that does. A Lean kernel doesn't care who trained the model that wrote the proof. If there's a certificate and it type-checks, you don't have to trust OpenAI's characterization of anything — you run the checker yourself. That's a completely different verification path from a lab saying its model did a hard thing.

29. Lenar Kess 00:08:23

    Which is what Sunday's complaint actually was.

30. Damra Vol 00:08:26

    Right, and I'll be even-handed about the two thousand dollars, because that number is going to get quoted a lot and it's the softest thing in the story. Inference cost for what? The successful runs only, or every search branch that went nowhere? Those differ by orders of magnitude, and nobody has published the accounting.

31. Lenar Kess 00:08:44

    Joscha Bach's read was that the interesting shift isn't the mathematics, it's that models are becoming the primary agents of breakthroughs rather than the instruments. Which is a large claim to hang on ten problems.

32. Damra Vol 00:08:57

    It's a large claim, and it's also exactly the claim these results are structured to support, which is why the Lean files interest me more than the manuscript does. Ten formal certificates settle a question that ten thousand words of narrative can't.

33. Lenar Kess 00:09:12

    Here's a very different item from the same day, and it belongs right next to this. Susan Zhang posted about a draft paper that cites another unpublished draft as support. Nothing to do with Astra or OpenAI. Just a piece of the current preprint ecology.

34. Damra Vol 00:09:28

    Citation rings made of documents that don't exist yet. [sigh] And you can see how it happens without anybody being a villain — generation got cheap, submission got cheap, and reviewing capacity didn't move at all.

35. Lenar Kess 00:09:41

    So in one week you've got machine-checkable mathematics and unverifiable preprints citing each other, and both are downstream of the same capability.

36. Damra Vol 00:09:51

    That's the pairing, and I don't think it resolves. The technology that lets you formalize a proof is the technology that lets you generate a plausible-looking one. The difference is entirely whether the artifact ships with something a machine can check.

37. Lenar Kess 00:10:05

Which is a reasonable thing to ask of the next lab that says it solved ten problems.

38. Lenar Kess 00:10:10

Now the money, and this one has an actual event in it rather than another bubble column. Nate Jones covered it on his channel: Leopold Aschenbrenner's fund — the one built on reasoning backward from compute requirements into the supply chain — got margin-called out of its entire public equities book, and Citadel bought the book at a discount.

39. Damra Vol 00:10:29

The setup matters. That thesis worked. Jones puts it at roughly twenty times last year and better than two times this year. Then it was levered. And what killed it wasn't the thesis being wrong — it was a sag after an SK Hynix offering, plus a Citadel investor note calling for Federal Reserve rate hikes, which made volatile assets less attractive at exactly the wrong moment.

40. Lenar Kess 00:10:53

And then Griffin buys the position from the person who just got liquidated.

41. Damra Vol 00:10:57

Which is one of the older stories in finance wearing new clothes. Being right about the direction and wrong about the timing is indistinguishable from being wrong, if you borrowed. He keeps the private startup book, apparently — the illiquid half survives precisely because nobody can margin-call it.

42. Lenar Kess 00:11:15

[chuckle] The illiquidity is the feature.

43. Damra Vol 00:11:17

This week, yes.

44. Lenar Kess 00:11:19

The reason it's more than a personality story is the market context around it. Semiconductor indices are off twenty-three percent from their June peaks. The Korean KOSPI dropped forty percent in a month, driven by retail margin liquidation across more than a million accounts.

45. Damra Vol 00:11:36

And the argument on the AI Daily Brief is that none of that is about AI demand. Their evidence is revenue: Anthropic's run rate tracked at seventy-one billion dollars, up from forty-seven billion in

May, with OpenAI near fifty billion and July alone reportedly beating the whole second quarter. All of that's podcast commentary rather than filings. Nobody has audited those figures in public.

46. Lenar Kess 00:12:01

    Their debt argument is the more interesting one to me, because it's specific. Roughly one point six five trillion dollars of data-center debt structured through special purpose vehicles, sitting off the hyperscalers' balance sheets.

47. Damra Vol 00:12:15

    And the comparison they're answering is 2008, which they say fails for three reasons: the hyperscalers have real cash flows, the paper is sold to private credit and pension funds rather than being mispriced as risk-free, and you'd need actual corporate defaults rather than equity drawdowns before anything breaks.

48. Lenar Kess 00:12:34

    How much of that do you buy?

49. Damra Vol 00:12:36

    The cash-flow point, mostly. The sold-to-private-credit point is the one I'd push on, because the risk is held by someone who knows it's risky is only comforting until you ask who those pension funds are underwriting for. It isn't a 2008 analogue. It's also not nothing.

50. Lenar Kess 00:12:53

    And then Dwarkesh Patel published a piece yesterday arguing something close to the opposite of the comfortable version — that smarter models could push compute prices up rather than down, potentially by a large multiple.

51. Damra Vol 00:13:07

    Which is the counterpoint that makes both stories true at once. If demand outruns supply for years, and better models make each unit of compute more valuable rather than less, then the capital spending is justified and the squeeze arrives for everybody who isn't a hyperscaler. Cheap inference would be a phase rather than a trend line.

52. Lenar Kess 00:13:27

    The Register ran a piece yesterday headlined that the bubble is already popping and we just don't know it yet. Sixteen points, five comments.

53. Damra Vol 00:13:36

    [tsk] And an opinion column. I'd rather have one dated forced sale than ten of those.

54. Lenar Kess 00:13:42

    Same. The forced sale is the one with a date on it.

55. Lenar Kess 00:13:45

    Today, in Washington, representatives from OpenAI, Anthropic, and Google are at the White House for a preview of the finished voluntary AI framework, the one that came out of the June second executive order. This reaches us through a Reddit gallery post rather than a wire story, so hold the details loosely.

56. Damra Vol 00:14:04

    The detail that survives the sourcing caveat is that the companies were lobbying over specific language, and one of the items they cared about is open source. The actual wording is going to decide this, because open source in a policy document is a definitional fight before it's a policy fight.

57. Lenar Kess 00:14:21

    Weights, or weights plus data, or weights plus training code.

58. Damra Vol 00:14:25

    And whether a release with a use restriction still counts. Every large lab has a different answer that happens to match what they already ship. Nobody outside the room has read the text, so I won't characterize provisions — but the lobbying tells you where the money thinks the sharp edge is.

59. Lenar Kess 00:14:41

    There's a small human story from the same day that describes the same fault line from the hiring side. Arnav Gupta posted about somebody scrubbing pro-open-source material off their online presence before an Anthropic interview.

60. Damra Vol 00:14:55

    Secondhand, about an unnamed person, and I'd hate for that to become a claim about Anthropic's hiring policy, because it isn't one. What it is, is a data point about what a candidate believed the room wanted. That's its own kind of information about the culture, even if the belief is wrong.

61. Lenar Kess 00:15:12

And there's an obvious tension in the fact that the same week, the strongest coding model in that benchmark screenshot is open-weight and Chinese.

62. Damra Vol 00:15:20

    That's the constraint the framework has to survive. A voluntary American framework that discourages open weights doesn't reduce the number of open-weight models in the world. It changes who publishes them.

63. Lenar Kess 00:15:32

    Watch item on the same beat, sourced from a single wire-style post: Apple is reportedly suing the UK over an encryption back-door demand. There's no filing and no document behind it yet.

64. Damra Vol 00:15:44

    Not an AI story today. It becomes one the moment there's a precedent about compelled access to encrypted user data, because on-device inference is going to be sitting inside that boundary within a couple of hardware generations.

65. Lenar Kess 00:15:58

    We pick that up when there's a document to read.

66. Lenar Kess 00:16:01

    A post called devtools must be open source pulled six hundred and thirty-six points and two hundred and eleven comments on Hacker News yesterday. The argument itself is old — you should be able to read and modify the tools you build with.

67. Damra Vol 00:16:15

    And the standard objection to it is older still. The freedom to modify was theoretical for almost everybody. You have the source to your editor. You've never opened it. What's in dispute in that thread is whether large language models just made that freedom usable, because the cost of reading and patching an unfamiliar codebase went from a weekend to an afternoon.

68. Lenar Kess 00:16:36

    Simon Willison is in the thread on that side, and it's the strongest version of the argument. If you can point a model at an unfamiliar repository and get a working patch, then nobody has time to exercise the freedom stops being a defense of closed tooling.

69. Damra Vol 00:16:51

I'd add the sharper version. It isn't only that you can patch it, it's that you'll actually try. What kept me out of other people's build systems was never ideology, it was the two hours of orientation before the first useful edit. Take that away and license terms start mattering to people who never cared about licenses.

70. Lenar Kess 00:17:11

    Elon Musk posted the opposite bet the same day — that source code is on the verge of becoming like assembly. Meaning models generate binaries directly and the human-readable layer stops being where the work lives.

71. Damra Vol 00:17:24

    [tsk] That's a tweet with nothing shipped behind it, and I hold it at exactly that weight. But it's a clean opposite. One camp says models make source more valuable because now you can actually use it, and the other says models turn source into an intermediate representation nobody reads. Those can't both be the direction.

72. Lenar Kess 00:17:43

    Which would you take?

73. Damra Vol 00:17:44

    The first, and not on principle. Debugging. Everything I've seen about agent-written code says the expensive part is figuring out what it did and why, and that requires an artifact a human can read. Compiling straight to a binary removes the only surface where you can ask that question.

74. Lenar Kess 00:18:02

    There was a smaller post the same day about using task runners for common coding tasks — sixty-six points, a Makefile-shaped argument for giving both humans and agents a named entry point for every routine job.

75. Damra Vol 00:18:15

    It's a small thing that changes agent behavior more than it should. If your repository has a named command for running the tests, an agent finds it. If that knowledge lives in a senior engineer's shell history, the agent invents something worse.

76. Lenar Kess 00:18:29

    Three separate launches yesterday, all pointed at roughly the same gap. Hoplite came up as a Launch HN, out of the Y Combinator summer batch, deploying cloud coding agents — seventy-five

points and sixty comments. Armature did a Show HN for product analytics and evaluations on agent sessions over MCP, the Model Context Protocol. And LangChain added bring-your-own-key support to the LangSmith gateway.

77. Damra Vol 00:18:57

    The Hoplite detail I liked is the primitive: one thread equals one virtual machine. That's a design decision rather than a feature. It makes the unit of isolation the conversation, so an agent that wrecks its environment wrecks exactly one environment. Commenters immediately started comparing it to Amp, which is fair.

78. Lenar Kess 00:19:16

    And Armature's thread has the obvious question sitting right at the top of it.

79. Damra Vol 00:19:20

    Which is how you actually capture the model's reasoning. You can log the tool calls, and tool calls are the easy part. The interesting failures happen in reasoning that never crossed a wire. Two comments on that post, so nobody's answered it yet.

80. Lenar Kess 00:19:35

    The LangChain piece is smaller and more practical. Bring your own key means you can put your existing model contracts behind their fallback and rate-limiting machinery instead of buying tokens twice.

81. Damra Vol 00:19:47

    And Cloudflare shipped a billable usage API the same day for programmatic cost visibility, where the top comment is the one you'd expect: it's still not a hard cost cap. Visibility after the fact and a limit that stops the spend aren't the same product, and everybody running agents in production has learned that the expensive way.

82. Lenar Kess 00:20:06

    Two more quick ones. Intology published results claiming their system, Locus, post-trained Qwen3 base models past Qwen's own instruct releases, measured on something they call PostTrainBench.

83. Damra Vol 00:20:19

    Vendor-published, self-named benchmark, single source, and the r-slash-singularity headline was models are now training models, which is considerably larger than the evidence. The claim itself is checkable, though, and that's why it's interesting — beating the original team's instruct tune on their

own base model is a specific thing. Give me the benchmark definition and one reproduction from outside Intology.

84. Lenar Kess 00:20:44

    And Cristóbal Valenzuela, Runway's chief executive, argued that with benchmarks saturating, model releases should ship real-world impact measures instead. His example was the number of diseases cured.

85. Damra Vol 00:20:57

    [chuckle] As stated, unworkable. No release ships with a cure count, and no methodology exists to attribute one. But he's pointing at something real, and today is the day to notice it. We spent the first twenty minutes on a benchmark screenshot we can't verify, for a model we can't download, in a table where the differences between three frontier systems are a couple of points.

86. Lenar Kess 00:21:20

    That table stopped discriminating a while ago.

87. Damra Vol 00:21:23

    And nothing replaced it. Miles Brundage was on an adjacent point yesterday about interpretability work aimed at human behavior rather than code. Different subject, same underlying complaint: the measurements we have describe what the model does on the test, not what changes in the world once it's deployed.

88. Lenar Kess 00:21:41

    So — weights promised and a table delivered, proofs claimed and certificates maybe, a fund liquidated for being early with borrowed money. The item I chase first tomorrow is the Qwen3.8-Max license text, because a two point four trillion parameter model with a use restriction and one without are different objects, and the benchmark row looks identical either way.

89. Damra Vol 00:22:04

    And I'd chase the Lean repository. If those certificates exist and they check, that's the one thing from today that stays true no matter who says what about it afterward.

90. Lenar Kess 00:22:14

    Then tomorrow we're checking two files instead of arguing about two claims. Good trade. That's Damra Vol, and I'm Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic