Whose numbers are these

2026-08-26 / 00:26:43

> *“Anybody with GB300 access can produce the Nvidia half of that comparison, and nobody can rent a Jalapeño. The claim is falsifiable in one direction only.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A day of performance claims made by the companies whose products are being measured, and what it takes to check any of them.

- SemiAnalysis put OpenAI's Jalapeño inference chip through the InferenceX benchmark against Nvidia, AMD and Google parts on open-weight models. OpenAI claims 1.5–1.9x better performance per watt and 1.7–3.6x lower latency. The Nvidia half is reproducible by anyone with the hardware; the Jalapeño half is not, because you can't rent one.
- Z.ai confirmed the stealth model Ox Alpha is a new GLM iteration, with weights shipping tonight after it topped OpenRouter's leaderboard at zero cost.
- Reuters reports Moonshot AI is in early talks with Microsoft, Amazon and Google over revenue sharing to host Kimi K3, seeking up to a 30% share — an open-weights lab asking for a cut of inference revenue it has no legal claim to.
- Meta drew up plans to cut many teams by as much as 60% to become "AI native", then pulled back after staff pushed and internal data showed the agents were ineffective. The plan preceded the evidence.

- Dylan Patel's per-megawatt arithmetic: base infrastructure at $10–15M per megawatt annually against frontier models generating up to $50M, with ASML tools and Zeiss mirrors as the binding physical constraint.
- Joshua Penman's Semantic Overlays paper adds an out-of-band annotation channel to a frozen model's residual stream. An overlay asserting a snippet is in a different programming language makes the model rewrite it faithfully in that language — the same power an attacker wants, held by the serving stack.
- Paritok-4B compresses coding-agent context to 25.7% of its size with an extractive adapter, and reports that using GPT-5 as your compressor costs more than the tokens it saves.
- The Guardian traced 124 reports and 560,000 words in nine days to a site badged with a thinktank that doesn't exist, built on a commercial platform that sells chatbot-citation optimization.
- A proposed EPA change would let air pollution permits issue without public input, which changes when neighbours find out rather than what gets built.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   SemiAnalysis published a teardown yesterday of Jalapeño, the inference chip OpenAI built with Broadcom, and for the first time there are numbers attached to it. Against Nvidia's GB200 and GB300 parts, OpenAI claims better performance per watt by a factor of one and a half to one point nine. On latency, the claimed improvement runs from one point seven times up to three point six. Those runs went through a public benchmark called InferenceX, on open-weight models, including DeepSeek R1 and Kimi. It's Wednesday, August twenty-sixth, and I'm here with Damra Vol. Today we've got those chip numbers, plus a wave of Chinese open-weight releases that includes one lab asking for thirty percent of hosted revenue. Then a Reuters story about a Meta headcount cut that got drawn up and shelved. Dylan Patel does per-megawatt arithmetic on the whole buildout, and there's a prompt-injection paper that annotates spans underneath the tokens.

   - techmeme.com
   - x.com
   - x.com
   - forbes.com
   - techmeme.com
   - bloomberg.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - youtube.com
   - x.com
   - youtube.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techcrunch.com
   - tomshardware.com
   - techmeme.com
   - theguardian.com
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - techmeme.com
   - x.com
   - x.com

- siliconangle.com
- x.com
- theguardian.com
- openai.com
- techmeme.com
- techmeme.com
- techmeme.com
- techmeme.com
- theguardian.com

2. Damra Vol 00:01:04

   Sixteen months got them from a standing start to a benchmarked part, working with Broadcom. Greg Brockman's entire public comment was five words: "ai for chip design is underrated." People following the project noted that OpenAI kept the chip team deliberately small and leaned on model-assisted iteration to get through the design cycles. If that holds up, the watts are almost the less interesting number. A small team with a model in the loop produced a competitive inference part in a domain where programs usually run years and staff in the hundreds.

3. Lenar Kess 00:01:38

   These are OpenAI's performance claims run through SemiAnalysis's harness. SemiAnalysis is a serious shop, InferenceX is public, and the models are open, which is a long way better than a slide with no methodology on it. But it's still the chip's owner supplying the runs. Nobody outside OpenAI has put independent silicon on an independent bench.

4. Damra Vol 00:02:00

   Re-runnable matters, though. If the models are DeepSeek R1 and Kimi, and the benchmark is published, then anybody with GB300 access can produce the Nvidia half of that comparison themselves and see whether it matches. The half nobody can reproduce is Jalapeño, because you can't rent a Jalapeño. So the claim is falsifiable in one direction only.

5. Lenar Kess 00:02:23

   The Techmeme summary of the piece mentions total cost of ownership, throughput per megawatt, and a comparison against Rubin, which is Nvidia's next generation. That's the comparison that decides anything. The alternative OpenAI is choosing between isn't today's rack, it's whatever it would otherwise be buying in eighteen months.

6. Damra Vol 00:02:43

And the timing sits right on top of Nvidia's earnings this week, with Vera Rubin prices going up. Paulo Carvão wrote a piece for Forbes yesterday setting the earnings up as a scorecard for the whole boom: data-center revenue, Rubin demand, customer concentration, and what he calls mounting infrastructure financing risks. Customer concentration is the line that connects back to Jalapeño. If your largest customers are all building their own inference silicon, concentration and substitution stop being two separate risks.

7. Lenar Kess 00:03:15

   The chip's main job right now, I think, is negotiating leverage. OpenAI buys Nvidia parts in quantities that move a quarter. Having a credible in-house alternative with published ratios changes what a purchase order looks like, whether or not a single Jalapeño ever displaces a GB300 in production.

8. Damra Vol 00:03:33

   The sixteen months is the number I'd most want confirmed from outside. If model-assisted design compressed a chip program that hard, then every hyperscaler with a Broadcom relationship just got a much shorter path to its own part, and Nvidia's problem in 2028 isn't AMD.

9. Lenar Kess 00:03:51

   Z.ai confirmed this morning that Ox Alpha is a new iteration of its GLM series. Ox Alpha is the model that appeared on OpenRouter's leaderboard at zero cost and went straight to the top with nobody knowing whose it was. Luz Ding has the confirmation at Bloomberg, and the weights ship tonight.

10. Damra Vol 00:04:10

    Free on a public leaderboard is a distribution strategy, and this one worked. A stealth entry gets evaluated on its outputs, because there's no lab name for anyone to react to. That's about the best evaluation conditions a Chinese lab can get from Western developers who wouldn't otherwise have picked it out of a dropdown.

11. Lenar Kess 00:04:28

    Alibaba also released Qwen3.8-Flash today. Open weights, 125 billion parameters, built on what Alibaba calls its next-generation Qwen 4 architecture. Alibaba says it rivals Opus 4.6 and V4-Flash. That is Alibaba's claim, on Alibaba's release day, with no third-party evaluation in front of me. We covered Qwen 3.8 long-context work on consumer hardware on Sunday, so file this as a separate model in the same family rather than a continuation.

12. Damra Vol 00:05:01

Moonshot is the one with structural news underneath it. Reuters reports, from sources, that Moonshot AI is in early talks with Microsoft, Amazon and Google about revenue sharing to host Kimi K3. The ask, according to those sources, is for up to a thirty percent share.

13. Lenar Kess 00:05:19

    It's early talks, according to sources, with no deal on paper, so hold it loosely. But the ask itself is the artifact I care about. An open-weights lab telling the three largest cloud providers that it wants a cut of hosted inference revenue is making a claim about who owns the weights economically after they've been released to everyone.

14. Damra Vol 00:05:39

    Because in the legal sense they don't belong to anyone at that point. If Kimi K3 ships under a permissive license, Amazon can host it and owe Moonshot nothing at all. So thirty percent has to be buying something other than permission to run the model. Support, tuning help, early access to the next checkpoint, or maybe an official-build badge that says this one is the real thing. Whether anyone pays comes down to whether serving a Moonshot model without Moonshot is measurably worse.

15. Lenar Kess 00:06:09

    And there's a live comparison for that, because DeepSeek has been running the pure open-weights version of this for two years. The Information reports DeepSeek did seventy point seven million dollars of revenue in the first seven months of this year, which is about four hundred and seventy-five million yuan. Against that, they posted a net loss of a hundred and six million. The revenue figure is roughly ten times its full-year 2025 number, which came alongside a hundred and thirty-nine million dollar loss.

16. Damra Vol 00:06:37

    Ten times the revenue and a smaller loss is the healthiest line in that sentence, and it's still seventy million dollars. Against the capital numbers we'll get to in a few minutes, that's a rounding error. The Chinese open-weights labs are winning distribution and losing money, which is what a land grab looks like from the inside.

17. Lenar Kess 00:06:56

    One housekeeping note on this whole segment. Every item here reached me through a Techmeme or Bloomberg summary. There's no primary lab post in front of me for Ox Alpha or for Qwen3.8-Flash, so every capability claim in the last five minutes belongs to the company that made it.

18. Damra Vol 00:07:13

Tonight is the checkable part. Ox Alpha topped a leaderboard as an anonymous free entry, and once the weights are public somebody will run it against GLM 5.3 and against Qwen3.8-Flash on the same hardware. Then we find out whether that leaderboard was measuring the model or measuring the price.

19. Lenar Kess 00:07:31

    Katie Paul at Reuters has been reading internal Meta documents. They show that Meta explored cutting many of its teams by as much as sixty percent, in two waves, to become what the documents call "AI native." Then Meta pulled back. Two reasons appear in the reporting: staff revolted, and Meta's own data showed the AI agents were ineffective.

20. Damra Vol 00:07:53

    "Ineffective" is Reuters characterizing Meta's internal data, and we don't have the metric behind it. So I won't turn that into a verdict on agents generally. What we can say concretely is that a very large company ran the substitution experiment against its own telemetry and decided not to proceed.

21. Lenar Kess 00:08:11

    The plan is the document I'd keep. Somebody at Meta wrote down sixty percent, in two waves, as a target, before the data came back. That ordering tells you something about how the decision was being made. The headcount number arrived first and the evidence arrived second.

22. Damra Vol 00:08:27

    Which is how most of these get done. And the counterexample from the same day is useful, with a large asterisk on it. OpenAI published a customer video with Mike Jones, the chief technology officer at loveholidays. That's an online travel agent operating in eight European countries, and Jones quotes something like sixty trillion package combinations a day. He says AI-assisted coding went from near zero to around eighty percent of all code production within a year, with a seventy-three percent increase in deployment frequency and stable headcount.

23. Lenar Kess 00:09:01

    Stable headcount is the figure that does the contradicting. Same technology, opposite staffing conclusion, on the same news day. Although the obvious thing applies: that's an OpenAI-published customer video, so those numbers are marketing until somebody outside loveholidays sources them.

24. Damra Vol 00:09:19

Agreed on the decimals. The direction is still checkable over time. Jones also says they doubled data-platform modifications while halving support-ticket volume, and he describes the operating principle as "everybody is a builder." Put that next to Meta's shelved plan and you get two theories of what this technology is for. One of them removes people. The other one gives the same people more surface area to work on.

25. Lenar Kess 00:09:45

    Ethan Mollick pointed at a wrinkle running underneath both, which is that OpenAI killed Agent Builder back in June while enterprises are still running it. If you're loveholidays and you've made Codex your unified control plane, that's the exposure you've taken on. Not the model getting worse, the product getting discontinued underneath a workflow you already rebuilt around it.

26. Damra Vol 00:10:07

    That's a procurement problem, and procurement problems have known answers. Keep the interface thin, hold on to an escape hatch, and keep somebody on staff who remembers how the old path worked. Meta's retreat and loveholidays' rollout might both come down to whether anyone drew that boundary before committing to it.

27. Lenar Kess 00:10:25

    Meta shelved the plan rather than deleting it. Reuters has the documents, and a target that got written down once can be picked back up in a quarter when the internal numbers look different.

28. Damra Vol 00:10:35

    And the staff revolt is part of the record now too, which changes the price of trying again.

29. Lenar Kess 00:10:41

    Dylan Patel did seventy-six minutes with Dwarkesh Patel, who is his twin brother, and the arithmetic in it is the most legible accounting of the buildout I've heard. Annual AI infrastructure capital spending already exceeds one trillion dollars, and he forecasts it past two trillion by 2028. Techmeme's summary puts eleven trillion of capital spending between 2024 and 2029.

30. Damra Vol 00:11:05

    The per-megawatt unit is what makes it usable. Base infrastructure runs ten to fifteen million dollars per megawatt annually. Frontier models generate up to fifty million dollars per megawatt. That spread is the entire investment case in one ratio, and it explains why nobody is slowing down while the totals get absurd.

31. Lenar Kess 00:11:25

On the manufacturing side, about six billion dollars of fab capital expenditure yields one gigawatt of annual compute capacity, which Patel maps to roughly a hundred billion dollars of downstream AI revenue. Those are his projections rather than disclosed figures. They're precise enough to sound like accounting, and they aren't accounting.

32. Damra Vol 00:11:45

    The concentration forecast is the number that will get quoted everywhere. OpenAI and Anthropic take about thirty percent of marginal compute today. Patel has that going to forty or fifty percent next year, and potentially seventy to eighty percent by 2028. Global capacity doubles annually while lab compute triples. Each of those two labs went from about two gigawatts early this year to over five by year end, and he has them trending toward tens of gigawatts each by late 2028.

33. Lenar Kess 00:12:16

    That's a forecast too, and he says as much in his own summary line, which is "Every force is screeching towards centralization." It's a directional claim about incentives, not an output from a model of the market.

34. Damra Vol 00:12:29

    The constraint he names is the most convincing thing in the interview, because it's physical. ASML's extreme-ultraviolet tools, and the mirrors Carl Zeiss makes for them, scaling toward roughly a hundred tools annually by 2030. You can raise money much faster than Zeiss can polish mirrors, and that's the ceiling everything else is pressed up against.

35. Lenar Kess 00:12:50

    He also says Anthropic turned a profit in the second quarter and expects OpenAI to in the third, driven by Codex and model version 5.6. If both of those hold, the economics stop being a matter of belief and start being a matter of disclosure.

36. Damra Vol 00:13:06

    On the other side, he expects compute pricing may inflect to twenty-five to fifty million dollars per megawatt, which squeezes anyone renting rather than owning. His note for everybody outside that circle is that open-weight models served through vLLM or SGLang stay economically viable at today's rates. That's the practical option for people who aren't buying gigawatts.

37. Lenar Kess 00:13:30

    The financing layer produced two stories today. The Financial Times has the US data-center buildout straining the lenders underwriting it, because a building full of accelerators is a new asset

class and nobody holds a long loss history on one. And Bloomberg's Saritha Rai reports an Indian firm, AM Intelligence, has ordered nine thousand Nvidia Vera Rubin systems, planning a gigawatt as part of an eight billion dollar project.

38. Damra Vol 00:13:57

    Meanwhile Jonathan Weil at the Wall Street Journal went after Microsoft's disclosure: capital spending, the OpenAI arrangement, and Azure buried inside the Intelligent Cloud segment. His line was "Cloud computing, meet cloudy numbers." When the per-megawatt math is this leveraged, segment reporting stops being an accounting preference and becomes the only way anyone outside the company can check the ratio.

39. Lenar Kess 00:14:23

    Lucas Ropek at TechCrunch also reports OpenAI lost a senior data-center executive, Malone, in what is now a run of high-profile departures. OpenAI's statement is that it has "recently reorganized" its "infrastructure organization to support the scale and pace of our work."

40. Damra Vol 00:14:41

    And then the physical layer meets local politics. The US government is moving to remove public-input requirements for air pollution permits, which would let data-center permits issue without being publicized at all. Tom's Hardware has the change.

41. Lenar Kess 00:14:55

    We've covered siting three episodes running, so I'll stay on the mechanism, because the mechanism is the whole item. A public comment period is the instrument local opposition uses. Remove it and you haven't changed anyone's mind about a data center, you've changed when the neighbours find out there's going to be one.

42. Damra Vol 00:15:12

    Andy Masley's read of the polling is that the American backlash is driven by local environmental and economic concerns rather than anti-Big-Tech sentiment. His words: "It's probably not your ideological thing, or mine. It's about object-level claims about how data centers impact the communities." That's one analyst reading polls, not a published survey. But if he's right, removing notice makes the fight worse, because the complaint is about effects people can measure where they live.

43. Lenar Kess 00:15:43

Australia ran the same argument from the other end and then partly reversed. The Guardian reports Anthony Albanese backed down on requiring states to power new AI datacentres entirely with renewables, flagging carveouts for Queensland and the Northern Territory at today's national cabinet meeting in Sydney. That came after Chris Bowen had said Labor would use constitutional powers to override states that resisted. Albanese called the discussions "positive and constructive."

44. Damra Vol 00:16:11

    Threatening constitutional override and then granting carveouts inside the same week is a fast retreat. A renewables mandate with two jurisdictions exempted is a different policy from a renewables mandate, and the reporting is that the states pushed and the requirement moved.

45. Lenar Kess 00:16:29

    There's a paper up today from Joshua Penman called "Semantic Overlays: Mitigating Prompt Injection with Annotations Beyond Tokens and Steering Vectors," and the opening of the abstract is the most direct statement of the prompt-injection problem I've read this year. Quote: "Everything a language model sees is tokens. The serving stack knows what each span is, user input, tool output, instructions, but the model must keep track of that itself, and it can lose track or be confused: text can be written to read like anything."

46. Damra Vol 00:17:00

    That's the entire vulnerability in three sentences. The runtime knows the provenance of every span it assembles. The model doesn't, because provenance isn't expressible in the only channel the model reads.

47. Lenar Kess 00:17:13

    So Penman adds a channel. Semantic overlays are small learned adapters applied at chosen prefill positions to a frozen model's residual stream, an out-of-band annotation that tokens can't forge, because it isn't made of tokens. Unlike steering vectors, they're trained, they're adaptable, and they're applied to particular spans instead of the whole forward pass.

48. Damra Vol 00:17:35

    The demonstration sells it better than the benchmark table does. They ask the model to copy a code snippet, and they attach an overlay asserting that the snippet is in a different programming language than it actually is. The model rewrites the snippet faithfully in the asserted language. The overlay stated a false fact about a span, and the model believed the annotation over the text in front of it.

49. Lenar Kess 00:17:57

Which is exactly the power an attacker wants, sitting in the serving stack's hands instead. Mark a span non-executable and instructions injected into untrusted context stop being instructions. Overlays compose, the marked text stays readable, and they can carry imperatives as well as assertions.

50. Damra Vol 00:18:16

    Here are the numbers, with the caveat attached. On SEP, separation goes from twenty-four point three percent up to ninety-six point five, and utility doesn't move. TensorTrust attack success drops from thirty-four point eight percent down to six point six. All four of the PIArena attack families fall to zero percent compliance, and the marked spans still stay readable, with an exact copy rate of ninety-two and a half percent.

51. Lenar Kess 00:18:44

    The caveat is in the paper itself. Single author, self-reported benchmarks, and Penman says the scoring rule is his own and that he corrected a defect in the published grader. He says so on the page, which is the right move, and it's also why zero percent shouldn't be read as settled.

52. Damra Vol 00:19:01

    The deployment constraint is the bigger limit for most people listening. This needs access to the residual stream at prefill, so it's a serving-side capability. If you're calling an application programming interface, you can't adopt it. You can only wait for whoever runs the model to adopt it on your behalf.

53. Lenar Kess 00:19:19

    There's a companion in the same batch from Jason Liu on evidence-carrying termination for tool-using large language models. Same instinct: make the runtime hold the invariant instead of asking the model to hold it.

54. Damra Vol 00:19:32

    The rule is that an agent may return COMPLETE only when a typed certificate binds every required answer claim to valid in-scope trace evidence, and a deterministic replay reconstructs the claimed value. In their locked static study, the agent produced zero unsafe completions across two hundred and eighty-eight runs, where the termination-critic baseline produced two hundred and fifty-two.

55. Lenar Kess 00:19:57

    They also ran a five hundred and seventy-six trajectory study across twenty-two held-out clusters. There the agent stopped early without support zero times out of sixty-six, and the controller did it

forty times. Supported completions went up a little, so being strict didn't cost them anything measurable. Liu is explicit about scope, too. The certificate "certifies support in a recorded trace under declared assumptions, not external truth, safety, or alignment."

56. Damra Vol 00:20:26

    That's the right amount of modesty for a stopping rule. It tells you the agent can show its work, and nothing about whether the work was right.

57. Lenar Kess 00:20:33

    One more from the same batch, and it's about cost rather than safety. Paritok-4B, from Jiayu Shi and Luzhuo Chen, is a four-billion-parameter low-rank adapter that compresses coding-agent trajectories.

58. Damra Vol 00:20:47

    It's extractive, and that's the design decision carrying it. It selects spans instead of rewriting them, and ninety-six percent of the identifiers, paths and numbers it emits already appear in its input. A compressor that paraphrases an identifier is worse than useless for a coding agent, because the agent then goes looking for a file that doesn't exist.

59. Lenar Kess 00:21:09

    On all three hundred SWE-bench Lite instances it compresses context to twenty-five point seven percent of the original size, while retaining eighty-six and a half percent of uncompressed single-shot solve quality. That's two times harder compression than a GPT-4.1-mini compressor and two point four times harder than GPT-5. They distilled a GPT-4.1-mini teacher over sixty-seven thousand real OpenHands trajectories and fine-tuned Qwen3-4B on the result.

60. Damra Vol 00:21:39

    And the economics line is blunt. GPT-5 as a compressor is net-negative. It costs more than the downstream tokens it saves. Paritok is a two hundred and sixty-four megabyte adapter that self-hosts on a single twenty-four gigabyte accelerator with no per-token compressor fee, and the weights, data and eval scripts are all Apache 2.0.

61. Lenar Kess 00:22:01

    Their significance test is a McNemar with p equals zero point zero seven nine — thirty instances were solved only uncompressed, and seventeen only compressed. That means not significantly worse at this sample size. It doesn't mean equivalent, and I'd rather say that than round it up into a claim the data won't hold.

62. Damra Vol 00:22:21

    Perplexity launched Portable Computer yesterday. It's their agentic platform running entirely on-device with zero token costs, starting on Nvidia DGX Spark and RTX Linux machines. Michael Nuñez has it at VentureBeat, and Aravind Srinivas described the target directly: "A background process that continually ingests context from every single connector, or app, performs multi-hop reasoning in a perpetual inference loop, and runs on your hardware. That is the future."

63. Lenar Kess 00:22:52

    Mario Zechner's reaction to the benchmark figure was two words: "chat, is it over?" The figure he's reacting to comes from a post I can't see, so I'm not going to repeat the number. His reaction is what I can report.

64. Damra Vol 00:23:06

    Glean announced Tau the same morning at its own conference in San Francisco. It's a desktop workspace connecting its enterprise system to local files, applications and code, and it launched with token-cost numbers aimed at Anthropic. Duncan Riley wrote it up at SiliconANGLE. Vendor-published comparison, on the morning of the vendor's own conference, so discount it for that.

65. Lenar Kess 00:23:29

    Nick Baumann posted the counterweight, which is an inventory of what ChatGPT's work setup gives you: a browser with persistent authentication, and a sixteen gigabyte nine-core machine, all of it rented. Same capability list as Srinivas's, opposite answer on where the hardware lives.

66. Damra Vol 00:23:46

    The Guardian has a story from Jason Wilson that I'd file under retrieval rather than generation. A pro-Israel messaging website badged with the name of a thinktank that doesn't exist published a hundred and twenty-four reports, over five hundred and sixty thousand words, in nine days.

67. Lenar Kess 00:24:03

    The mechanism is the reportable part. It's built on a commercial platform that promises to optimize content so that AI chatbots will cite it. Somebody sells that as a service. The site gives Israel's position on the torture of Palestinian prisoners, on Israeli war crimes, and on whether Israel deliberately starved Palestinians in Gaza. All of it is presented as neutral research.

68. Damra Vol 00:24:26

    And on the same day OpenAI published a takedown of a covert Russian influence campaign. So you have platforms policing generation, and nobody at all policing the documents that retrieval draws

from. Five hundred and sixty thousand words in nine days is cheap now, and citation is the surface with no moderator on it.

69. Lenar Kess 00:24:46

    Bill Gates published a note saying the AI era "will be one of the most turbulent times in human history" and that "we are not preparing adequately," calling for a regulatory framework. To Karen Weise at the New York Times he went further, saying tech executives are privately very worried about AI disruption and publicly downplay the risks to protect fundraising and planned public offerings.

70. Damra Vol 00:25:09

    He names nobody and produces no documents, so that's an assertion rather than reporting, and I'd hold it there. Ina Fried at Axios has him addressing his Jeffrey Epstein association as a "huge mistake" driven by wanting to raise funds, and saying he hopes the ties won't undermine his AI arguments. That's context for why he's talking about all of it this week.

71. Lenar Kess 00:25:32

    Last one. Mackenzie Hawkins at Bloomberg reports, from sources and documents, that Huawei has proposed exporting its high-end Ascend 950-series chips to build Egyptian government AI data centers. Pitch stage, no deal reported.

72. Damra Vol 00:25:48

    Which arrives the same week Taiwanese prosecutors charged nine people, including one person from Nvidia and two from Super Micro, over illegally exporting high-end AI servers carrying B300 accelerators that are banned from sale to China. We went through the indictments in detail yesterday, so that's background here rather than a second helping.

73. Lenar Kess 00:26:10

    If Ox Alpha's weights really do ship tonight, then tomorrow is the first day anyone outside Z.ai can check that leaderboard result against the actual model. That's a rare thing on a day like this one, where almost every number came from the company selling the product it measures.

74. Damra Vol 00:26:27

    And Jalapeño stays half-checkable until somebody who doesn't work for OpenAI gets a part onto a bench of their own.

75. Lenar Kess 00:26:34

Weights tonight, and a real comparison once somebody runs them. Thanks for spending part of the morning with us — Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic