

Who gets to say no

2026-08-24 / 00:32:38

“Fewer than ten percent of data center companies answered a state questionnaire about their own power demand, and then everyone acted surprised when Texas started auditing the interconnect queue.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

The compute buildout now has a partisan fault line running through both parties at once — and separately, a fatal strike in Zaporizhzhia that Ukraine attributes to a drone with no human in the loop, running on a dev board you can buy on a resale market.

- Axios: Abbott says data centers "dug their own grave" — the governor who called Texas the epicenter of AI development in November now wants an audit before anyone connects to the grid, and says fewer than ten percent of operators answered the state's power-demand request.
- Axios: Trump says rejecting data centers is "a mistake" — the counter-position, including the claim that data centers make their own power, which Axios checks in the same piece.
- Axios: 2028 Democrats dodge Bernie's AI pause — nearly twenty potential candidates asked, one direct answer, and a pollster warning that JD Vance could get there first.
- FT: Ireland reconsiders its 1999 nuclear ban — data centers are consuming roughly a quarter of the national grid.

- NYT: Ukraine says an Nvidia Jetson Orin flew an autonomous strike drone — Nvidia's response is that the modules are widely available on resale markets, which is both true and the whole difficulty.
- Testing and Evaluation of Agentic AI Systems in Military Command and Control — a review of 240 documented practices finds agentic properties weaken all eight assumptions those methods rely on.
- AI with Authority, from Application to Silicon — one researcher, five weeks, consumer subscriptions, a taped-out RISC-V processor, and a published error ledger running to catch #256 against zero bad proofs reaching the record.
- ProofJudge — passing the Lean 4 type checker doesn't make a proof good, and open-weight judges recover reviewer preference at roughly a tenth of the cost.
- Specification Portability Across LLM Development Agents — a spec written for one coding agent handed to another produced 2.33 percent valid SQL. Your AGENTS.md isn't a neutral artifact.
- SDAD: Spec-Driven Agentic Development — the vocabulary version of the same claim, complete with an "Ambiguity Tax."
- Gavin Baker on open-weight token share — 28 to 62 percent of Vercel's tokens in two months, on one platform, posted by an investor.
- Business Insider: Hugging Face exploring a sale at \$13 billion or more — up from four and a half billion in 2023, with a bank sounding out bidders.
- CNBC: Alibaba drops ten percent after a \$10.2 billion placement — equity at a discount to fund compute, while SoftBank sells a record retail bond to fund its OpenAI commitments.
- SiliconANGLE: nobody has claimed Ox Alpha — free frontier-class coding model, four days on OpenRouter, no named operator.
- Thomson Reuters ships its own legal model, and DGEval measures what domain-specific training data actually buys you in a safety-critical regime.

SEGMENTS

00:00:04	The backlash gets a party line
00:09:05	Zaporizhzhia and the resale market
00:12:42	Five weeks from application code to silicon
00:18:16	Your spec doesn't travel
00:23:40	Sixty-two percent, and a registry for sale
00:26:09	Three instruments, one bill
00:28:35	Short ones

Transcript

1. Lenar Kess 00:00:04

Greg Abbott sat down with Jonathan Karl on ABC's This Week yesterday and said this about the companies building data centers in Texas. Quote — <emphasis>they basically dug their own grave</emphasis> for the problem that's been caused for them. And that's why they got the backlash they deserve. That's the governor of Texas. That's the same man who stood next to Google last November, called the state the epicenter of AI development, and announced forty billion dollars across three campuses.

- [axios.com](#)
- [axios.com](#)
- [axios.com](#)
- [techmeme.com](#)
- [x.com](#)
- [x.com](#)
- [techmeme.com](#)
- [techmeme.com](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [arxiv.org](#)

- arxiv.org
- arxiv.org
- x.com
- earendil.com
- x.com
- techmeme.com
- siliconangle.com
- reddit.com
- techmeme.com
- cnbc.com
- techmeme.com
- techmeme.com
- techmeme.com
- theguardian.com
- forbes.com
- siliconangle.com
- techmeme.com
- siliconangle.com
- arxiv.org
- siliconangle.com
- techmeme.com
- x.com
- theguardian.com

2. Damra Vol 00:00:31

Nine months between those two sentences. And the turn wasn't rhetorical — in June he told state regulators to make data centers pay the full cost of the electrical infrastructure they require, and to start phasing out the tax incentives. Earlier this month he ordered an audit of any project trying to connect to the Texas grid before it can move forward. Then he gave Fox News Sunday a number last week that explains the temperature better than any of the quotes. Fewer than ten percent of data center companies had answered a state request for the information Texas needs to forecast its own future power demand.

3. Lenar Kess 00:01:07

Ten percent. On a questionnaire asking how much electricity you intend to draw.

4. Damra Vol 00:01:12

In the state you told everyone was your home. [tsk] If you're looking for the moment the goodwill ran out, it's probably somewhere in that non-response rate. A governor can absorb a lot of local

anger on your behalf. He can't absorb being unable to answer the question of how much power his own grid is about to be asked for.

5. Lenar Kess 00:01:31

Then Sunday evening the other half arrives. Trump did a long interview with Michael Cohen — his former fixer, which is its own thing — and it aired in full yesterday. Trump's position is the opposite of Abbott's, stated plainly. Quote: communities that don't take a data center, they're making a mistake, because data centers create tremendous amounts of jobs and money. He also said the facilities aren't taking power from the grid, because — quote — they're making their own power plants. They're building the most beautiful, you've never seen power plants like this.

6. Damra Vol 00:02:04

Axios put the correction directly under the quote in the same piece, which I appreciated. Some data centers are developing behind-the-meter generation that feeds the facility directly. Others will draw from the grid like everyone else. Both of those are true at once, and which one applies is a per-project fact rather than a category fact. Trump has also taken shots at Abbott specifically, saying that for Texas to say no to data centers is a mistake because the industry could be bigger than oil.

7. Lenar Kess 00:02:35

Bigger than oil, said to Texas. That's a deliberate choice of comparison.

8. Damra Vol 00:02:39

It's the comparison a Texas governor is least able to accept from outside, yes. But set the personalities aside for a second, because the polling underneath this actually moved. The Annenberg Public Policy Center found last month that sixty-one percent of Americans oppose building a new data center in their area. In March that number was forty-nine. And it doesn't sort by party the way you'd expect — sixty-nine percent of Democrats, fifty-four percent of Republicans, and fifty-three percent of independents.

9. Lenar Kess 00:03:10

A twelve-point jump in five months. That's an opinion still forming.

10. Damra Vol 00:03:14

And a separate Fox News poll puts it at seventy percent of voters opposing one in their community. Two different pollsters pointing the same way, so whatever's driving it isn't a polling artifact.

11. Lenar Kess 00:03:26

So that's the electorate. Here's what the politicians are doing with it. Josh Shapiro in Pennsylvania — who celebrated a twenty-billion-dollar Amazon investment last year — imposed new restrictions this week. Kathy Hochul ordered a one-year moratorium on hyperscaler data centers in New York last month, and JB Pritzker revoked their tax breaks in Illinois. Then Axios reported last week that the National Republican Senatorial Committee warned the leading AI companies directly: voter anger over data centers was putting a critical Ohio Senate seat at risk.

12. Damra Vol 00:03:59

I'd underline that last one for anyone modelling capacity. A party campaign committee told the industry privately that they are costing it a seat. That message reaches a boardroom in a way a governor's press conference never does.

13. Lenar Kess 00:04:12

Which brings us to Holly Otterbein's piece from yesterday, the sharpest item in this whole cluster. Axios went to nearly twenty Democrats seen as potential 2028 presidential contenders and asked them a direct question: do you support Bernie Sanders' call for AI CEOs to — quote — stop building machines that humans cannot control. *One* of them directly answered.

14. Damra Vol 00:04:35

One. Out of twenty. [pause] Who answered?

15. Lenar Kess 00:04:40

Rahm Emanuel, and he answered by criticizing it. Quote — this is talking around a problem instead of trying to solve it. A pause is fine, but then what? He went on: how the time it buys gets used to build a comprehensive plan that meets America's energy needs instead of masking them. Then he called for permitting reform and for the hyperscalers to pay for the energy infrastructure upgrades.

16. Damra Vol 00:05:03

So the one person willing to engage did it by saying the pause is a placeholder for a plan nobody has written. And about half the field simply didn't respond at all — Harris, Newsom, and Buttigieg among them. The second-order detail matters too: almost nobody criticized Sanders either. They wouldn't endorse it and they wouldn't attack it. That's a group of professional politicians who've all separately concluded there's no safe answer yet.

17. Lenar Kess 00:05:30

Adam Carlson, a progressive pollster, says exactly that in the piece and he's blunt about the cost. He calls the party overly cautious, says it's a big strategic error for the Democratic Party writ large not

to be the one that's owning this. And then he asks the question that I suspect is going to circulate: what if JD Vance comes out and is the first major presidential candidate to support a data center moratorium, other than AOC?

18. Damra Vol 00:05:56

That isn't a hypothetical anymore, given where the midterm candidates already are. Mike Rogers, the Republican Senate nominee in Michigan, backs a one-year moratorium, and Amy Acton, running for governor in Ohio as a Democrat, wants a conditional one. Stacy Garrity, the Republican nominee for Pennsylvania governor, supports a pause as well. So that's two parties and three states arriving at the same policy, all before November third.

19. Lenar Kess 00:06:23

And on the other end, AOC's chief of staff points out she's leading the data center moratorium bill in the House. Ro Khanna told Axios he now supports a construction pause in Pennsylvania, Michigan, and Wisconsin — quote, places where these centers have been shoved down the throats of communities. That's a shift from where he was.

20. Damra Vol 00:06:44

Miles Brundage was pushing back on the whole moratorium direction over the weekend, and his objection is the one the industry is going to keep making — the China argument and the economic argument, in that order. I don't think it's a bad objection. I think it's an objection that has to be made to a person who just got a higher electricity bill, and that conversation has never gone well for anyone.

21. Lenar Kess 00:07:06

Chris Van Hollen's version of the counter is the tightest sentence in the whole canvass. Quote: Americans should not have to pay higher electric bills and other costs to subsidize the data centers being built by the richest corporations on the planet. That's a full campaign message in twenty-five words, and it doesn't require anyone to have an opinion about model capabilities.

22. Damra Vol 00:07:28

Right, and that's why it'll travel. Chris Murphy went somewhere stranger — he pointed reporters at a Substack post about a trip to Silicon Valley where he wrote that he could divine very few American values, in contrast to Chinese values, guiding AI development. And then: any talk about ethical or moral AI is just whitewash. That's a sitting senator writing off the entire safety vocabulary as marketing.

23. Lenar Kess 00:07:55

The international version of the same constraint arrived this morning. The Financial Times reports that Irish lawmakers are reconsidering the country's 1999 ban on nuclear power, because data centers are now consuming roughly twenty-five percent of Ireland's electricity.

24. Damra Vol 00:08:11

A quarter of a national grid. Ireland spent decades building an economy on being the place multinationals put their European operations, and the physical bill for the current generation of that arrived as a percentage of the power supply. Repealing a twenty-seven-year-old nuclear ban isn't a small legislative act, and nobody argued them into it. The electricity math did.

25. Lenar Kess 00:08:34

So where this sits at the start of the week. Two Republicans at the top of their party are publicly disagreeing about whether the buildout is an asset or a liability. The Democratic field mostly declined to answer at all. And there's a date — November third — after which everyone knows a lot more about what voters do with an opinion they've been forming since March.

26. Damra Vol 00:08:54

Meanwhile the permitting queue keeps getting longer. Abbott's audit order is the version of this that touches an actual project schedule. Everything else on that list is still talk.

27. Lenar Kess 00:09:05

Different story, and a hard one. The New York Times reported this morning that Ukrainian officials say a Russian drone struck a gas station in Zaporizhzhia and killed three people, and that the drone was flying without a human in the loop. Andrew Kramer's piece says Ukraine identified the compute onboard: an Nvidia Jetson Orin module.

28. Damra Vol 00:09:25

Let me be precise about what's attributed to whom, because this is the kind of claim that gets repeated a layer flatter each time. The autonomy determination is Ukraine's, from Ukrainian officials, and there's no independent forensic confirmation in anything I've seen today. The part number is less ambiguous, and I keep turning it over. A Jetson Orin isn't exotic. It's the module people put in warehouse robots, university projects, and hobby drones. You can order one.

29. Lenar Kess 00:09:54

Nvidia's response to the Times was that the modules are widely available on resale markets.

30. Damra Vol 00:10:00

Which is true. Completely true, and also a description of why the export-control conversation about frontier training chips has almost no purchase here. Controls on H-series parts are about who gets to train the big models. This is edge inference on a board that's been in the channel for years, in secondhand supply in every country. There's no chokepoint to hold.

31. Lenar Kess 00:10:23

The uncomfortable adjacency is that a paper posted to arXiv this morning is about exactly the certification problem underneath this. Ulysse Richard, Heather Frase, and colleagues went through 240 documented testing and evaluation practices used for military command and control systems. That's across eight evaluation dimensions and three lifecycle stages. Their question was whether those methods can support the public commitments that get made when agentic systems get procured.

32. Damra Vol 00:10:52

And the answer is structural rather than alarmist, which is why it's useful. They identify eight assumptions that established test methods make about the thing being tested, grouped into four clusters — whether the system can be specified, whether it's stable, whether it composes, and whether it can be supervised. Agentic properties weaken all eight. Their own sentence is precise about it, and I'll read it straight: test results may satisfy process requirements, but they do not warrant the inference from tested to fielded behavior.

33. Lenar Kess 00:11:26

So you can run the whole evaluation program, sign every box, and still not have evidence that the fielded thing behaves like the tested thing.

34. Damra Vol 00:11:34

Right. And they don't stop at the complaint — they derive ten assurance claims and say which ones are recoverable with methods that either exist or are close. Bounded mission envelopes, for one, and trajectory-grounded correctness, which means you check the path taken and not just the endpoint. Executable runtime constraints, and characterized run-to-run variance, which is just admitting the thing is stochastic and measuring how much. The line I hadn't seen before is that part of the evidentiary burden moves into deployment — the determination to field becomes a continuing act rather than a one-time approval.

35. Lenar Kess 00:12:11

And where you can't generate the evidence at all, they propose governing the leftover uncertainty with defined expiry conditions and an assigned owner. Which is a paper about paperwork, and it's

the most concrete thing anyone published today about a category of system that killed three people at a gas station.

36. Damra Vol 00:12:28

The gap between those two documents is the whole situation. One side is running a procurement assurance argument that its own reviewers say doesn't currently close. The other side bought a dev board on a resale market and flew it.

37. Lenar Kess 00:12:42

Something completely different, and it's the paper I'd read first today if you only read one. Jason Hickey posted a preprint called *AI with Authority, from Application to Silicon*. In five weeks, one researcher on consumer AI subscriptions directed a small fleet of agents from application code, through a verified compiler and executive, down to a RISC-V processor that was taped out on a community silicon shuttle. No proof passed through human review. No register-transfer-level hardware description was written by a human.

38. Damra Vol 00:13:15

[breath] Let me unpack the two pieces of jargon, since the claim lives inside them. Register transfer level is the language you actually describe hardware in — it's what a chip designer writes. A tapeout on a community shuttle means the design was submitted and fabricated on a shared multi-project wafer run, which is how small teams and universities get real silicon without a fab contract. So the claim is that agents wrote the hardware description, and physical parts exist.

39. Lenar Kess 00:13:45

And the discipline that makes it defensible has a name. He calls it the Salt method, and it rests on a proof kernel that a hallucinated proof can't pass. Mathematical claims travel between the agents as kernel-checked artifacts. Human attention is reserved for statements, designs, and rulings. Verification is stated link by link, from the Lean 4 kernel out to SAT-checked equivalence at the silicon boundary.

40. Damra Vol 00:14:11

That's the inversion, and I think he names it correctly in the abstract. For sixty years machine verification was a cost overhead you could only justify for exceptional artifacts — avionics, cryptography, or a nuclear controller. His argument is that at agent speed it flips, because verification is the only referee that scales with the output. His phrase is the incorruptible referee that lets one person safely direct autonomous machine work at scale. The kernel doesn't get tired and doesn't get talked into anything.

41. Lenar Kess 00:14:44

The accounting is what makes me take it seriously instead of just enjoying it. He published theorem provenance, a token meter he registered before the run, a floor-bounded count of his own hours, and an error ledger. The catch numbering runs to two hundred fifty-six in an append-only flags ledger he maintained from July seventh to July twentieth. Number seventy-nine was never assigned, and later catches go in un-numbered. All of that against zero incorrect proofs reaching the record.

42. Damra Vol 00:15:13

<laugh-speak>He told us about the gap in the numbering.</laugh-speak> That's the detail that made me trust the rest of it. Pre-registering the token meter before the run means he can't retroactively decide what counted. And publishing an error ledger where two hundred fifty-six things were caught is a person inviting you to argue with him rather than admire him.

43. Lenar Kess 00:15:33

It's one author reporting on his own five weeks, and it's a version-one preprint. Nobody has reproduced it.

44. Damra Vol 00:15:40

Agreed, and that's the correct caveat. But notice what reproduction would even mean here — the artifacts are kernel-checked. Someone can rerun the checker. That's a much lower bar than reproducing a training run.

45. Lenar Kess 00:15:53

There's a companion posted the same day that pushes on the obvious weak point. Shane Caldwell's ProofJudge starts from the observation that a proof passing the Lean 4 type checker can still be bad. So he built an agentic judge that scores formal proof quality on five dimensions beyond correctness. It looks at how well the proof leverages the library, whether it fits the automation, how clear its structure is, the quality of the statement itself, and whether it follows Mathlib conventions.

46. Damra Vol 00:16:22

And he grounded the judge properly, which is the interesting engineering. The judge agent gets tool access to the exact commit the pull request applies to, so it can query the library state while scoring. The evaluation runs over 218 declarations drawn from distinct Mathlib pull requests. A judge counts as aligned when it rates the accepted version above the initial version that reviewers sent back for revision. Six judge models, all above chance — 80.8 percent down to 63.5. Two open-weight judges hit roughly seventy percent at a tenth of the cost of the best one.

47. Lenar Kess 00:16:59

So the kernel says correct, and something else has to say good, and the something else can now be a cheap open-weight model that agrees with human reviewers seven times out of ten.

48. Damra Vol 00:17:10

There's a third one in the same batch doing this in physical design. VortexChat couples a decision agent to topology generation, gradient-based refinement, and full-wave electromagnetic simulation, for inverse design of integrated photonic devices from natural-language specs. They fabricated the output — a broadband terahertz perfect vortex beam multiplexer — and the measurements agree with simulation on efficiency, mode purity, and crosstalk. No human in the loop against the benchmark thresholds.

49. Lenar Kess 00:17:43

A chip and a photonic device. Agents designed both, both were fabricated, and both preprints went up today. And in each case the common element is a checker at the end that can't be sweet-talked — a proof kernel in one, a full-wave simulator and then a physical measurement in the other.

50. Damra Vol 00:18:00

Which suggests where this works and where it doesn't. If your domain has an incorruptible referee, one person can now direct a lot of machine work through it. If your domain's referee is a rushed code review at the end of a sprint, none of this transfers.

51. Lenar Kess 00:18:16

That sits right next to a result anyone maintaining an agent config file should hear. A group out of Ukraine — Oleg Grynets and colleagues — used Oracle-to-PostgreSQL migration as a controlled task to ask whether a specification written for one coding agent means the same thing to another one.

52. Damra Vol 00:18:35

And they did it at a size where the answer isn't anecdote. Stage one was a specification-first migration pipeline over 1,006 PL/SQL files. Of those, 623 regenerated successfully, and 380 of the generated scripts actually executed in PostgreSQL 16. Stage two was the cross-agent experiment on 1,802 Oracle scripts with known PostgreSQL implementations, run through Amazon Kiro, Google Gemini, and GitHub Copilot. Claude Code and Cursor were in the single-agent stage.

53. Lenar Kess 00:19:12

And the headline number is ugly. The strongest replicated failure was Gemini directly consuming a specification that Kiro had produced. Token F1 came in at 0.035, SQL syntax validity at 2.33

percent, and abstract syntax tree mean similarity at 0.015.

54. Damra Vol 00:19:34

Two percent valid SQL. At that point the agent is doing a different task, not a degraded version of the same one. And to unpack the metric — abstract syntax tree similarity compares the parsed structure of the generated code against the reference, so 0.015 means the output isn't structurally related to the right answer at all. It's producing something shaped like SQL that mostly doesn't parse.

55. Lenar Kess 00:20:00

They tested the fixes instead of just recommending them, which I appreciated. Rewriting the specification for the target agent substantially improved Gemini in the tested configuration. Compression didn't provide a universal benefit. And retrieval-augmented ingestion — letting the agent pull the spec on demand rather than swallowing it whole — was the only strategy that showed up on the per-agent Pareto frontiers of both Gemini and Copilot.

56. Damra Vol 00:20:26

They also report that specification size alone doesn't predict implementation quality, which kills the reflex fix. Everyone's instinct when an agent misbehaves is to write more spec. The data says length isn't the variable.

57. Lenar Kess 00:20:41

Their conclusion is stated as a warning to the field: specifications in heterogeneous spec-driven workflows should not automatically be treated as agent-neutral artifacts.

58. Damra Vol 00:20:52

And practically — if you've got a CLAUDE dot M-D or an AGENTS dot M-D in a repo, you've written a spec and you've almost certainly assumed it's portable. This says it might be, and you have no evidence either way, and when it isn't portable nothing tells you. The code comes back. It just doesn't run.

59. Lenar Kess 00:21:12

There's a companion paper posted the same morning that gives this a whole vocabulary. Vu Hung Nguyen and Thanh Nguyen formalize what they call Spec-Driven Agentic Development. Their argument is that large language models with very long context windows have made specification quality the fuel for autonomous delivery, and they propose governance metrics — an Ambiguity Tax, a Spec Fidelity score, and a cost index with a repair multiplier.

60. Damra Vol 00:21:40

[tsk] I'll take the vocabulary and leave the metrics. An Ambiguity Tax with no measured value attached is a name for a feeling. But their closing claim I'll defend: agentic speed doesn't eliminate engineering discipline, it relocates discipline upstream into specification precision, explicit gates, and auditable provenance. That's consistent with the migration data, and it's consistent with Hickey's kernel.

61. Lenar Kess 00:22:06

There's a survey in the same batch that puts a harder edge on the evidence side of this. It reviews work through May of this year across software engineering and software security tasks, and proposes separating five things that keep getting collapsed together: functional correctness, security, operational reliability, evidence provenance, and agent authority.

62. Damra Vol 00:22:27

And its list of recurring validity threats reads like things everyone quietly knows and nobody writes down. Weak test oracles, and data that's duplicated or leaked across time. Harnesses that change between comparisons. Security checks that only test a proxy, and budgets and human interventions that go under-reported. That last one especially — half the agent results you see don't tell you how many times a person stepped in.

63. Lenar Kess 00:22:54

Their conclusion is that model capability should be judged as an assurance case supported by task-appropriate evidence, rather than a single benchmark score. Which is the same word — assurance — that the military command and control paper used this morning, arrived at independently.

64. Damra Vol 00:23:11

And the DAIR AI team flagged a trace study over the weekend in the same neighborhood — ninety-four thousand development events across 557 agentic coding sessions. There's also a post going around Hacker News titled just <emphasis>What Is a Harness?</emphasis>, arguing the harness is the next layer of infrastructure in its own right. Which I think is the same object as the spec, viewed from the other end. The spec is what you tell the agent. The harness is what the agent can reach.

65. Lenar Kess 00:23:40

Money and market structure. Gavin Baker posted a chart from Guillermo Rauch on Saturday showing open-source model share of tokens served on Vercel going from twenty-eight percent to sixty-two percent in two months.

66. Damra Vol 00:23:54

Hold that one at the right size. It's one platform's traffic, posted by an investor with positions, describing a user base that skews toward people deploying cost-sensitive applications rather than people buying frontier reasoning. Vercel isn't the industry. But a thirty-four-point move in eight weeks on any real platform isn't noise, and it points the same direction as the pricing pressure everyone's been describing anecdotally.

67. Lenar Kess 00:24:19

There's a matching number from the other side of the trade. The Financial Times reported Ramp data showing that Fable 5, which launched in June, has plateaued at roughly eleven percent of corporate spending on Anthropic tools, with customers shifting toward cheaper models. Opus 5 has surpassed it.

68. Damra Vol 00:24:37

A newer flagship losing spend share to the older, cheaper sibling in the same family. That's the demand curve talking. And it's card-spend data via Ramp, so it's a slice of the market rather than Anthropic's own books — but it's a slice measured rather than asserted.

69. Lenar Kess 00:24:54

And then on the same weekend, Katie Roof at Business Insider reported that Hugging Face is exploring a sale that could value it at thirteen billion dollars or more. It was four and a half billion in 2023. They've brought in a bank to sound out potential buyers. Talks are early and no bidder was named.

70. Damra Vol 00:25:12

Nearly three times the valuation in about three years, for a company that doesn't train frontier models. It hosts them. It's the registry and the download endpoint. It's where the model card lives, where the dataset lives, and where your build script pulls weights from at two in the morning. If open weights really are carrying most of the tokens somewhere like Vercel, then the interesting question about a sale isn't the price. It's what happens to a piece of shared infrastructure when it acquires a strategic owner with its own model business.

71. Lenar Kess 00:25:44

The LocalLLaMA subreddit has a thread this morning that's exactly that anxiety, phrased less formally.

72. Damra Vol 00:25:51

It's a fair anxiety. Nothing has happened yet — early talks, unnamed bidders, and a bank taking temperature. But everyone who has a Hugging Face pull in a build script has an unpriced dependency on the answer, and until yesterday most of them hadn't thought about it as a dependency at all.

73. Lenar Kess 00:26:09

Capital flows, quickly. Alibaba priced a Hong Kong dollar eighty-billion placement — about ten point two billion US — to fund AI development. Reuters had the terms: seven hundred ten million shares at a three point six percent discount to Friday's close. The stock fell ten percent today.

74. Damra Vol 00:26:29

So the market charged them roughly the value of the raise to do the raise. That's a legible price on the sentence we will spend more on compute. Not a refusal — the placement cleared — but not enthusiasm either.

75. Lenar Kess 00:26:42

SoftBank is doing the same thing with a different instrument. Bloomberg reports a planned one-trillion-yen retail bond sale — about six point three billion dollars — which would be a record for any issuer in Japan, and it's their third this year. The money goes toward their OpenAI commitments.

76. Damra Vol 00:26:59

That's the contrast I'd put in front of anyone modelling this. Alibaba is funding compute with equity, which dilutes and then stops asking. SoftBank is funding it with debt sold to Japanese retail investors, which has a coupon and a maturity and doesn't care what the model roadmap looks like in 2029. Three sales in one year is a cadence, not an event.

77. Lenar Kess 00:27:20

And Oxford Economics gave the Financial Times the aggregate: US corporate spending on equipment and facilities is set to rise forty percent between 2021 and 2027, more than three times Europe's pace, driven by the AI race.

78. Damra Vol 00:27:36

Which is the number that ties this segment back to the top of the show. That forty percent is buildings and transformers and interconnects in actual counties with actual voters. The financing question and the permitting question are the same question arriving from two directions.

79. Lenar Kess 00:27:52

Gene Marks wrote a column in the Guardian over the weekend arguing the debt-bomb worry is overdone. His claim is that the concern about Meta, Oracle, xAI, and CoreWeave raising billions while keeping long-term obligations off the balance sheet isn't Enron two point oh, and that the risks are different and recoverable.

80. Damra Vol 00:28:11

I'd pair it with a Forbes piece from Robert Szczerba that frames compute as a bet on how long the hardware keeps earning. The detail I'd hold onto is that Nvidia will backstop that bet only case by case, and only up to twenty-five percent of a deal. Which tells you Nvidia has done the depreciation math and priced its own confidence. Twenty-five percent is a number that means something.

81. Lenar Kess 00:28:35

Four quick items to close. First: since August twentieth, an anonymous provider has been giving away free access to a frontier-class coding model on OpenRouter, listed as Ox Alpha. Zero cost for input and output tokens. OpenCode, the open-source terminal agent, added support the same day. Four days later, no company has claimed it.

82. Damra Vol 00:28:58

Stealth models on OpenRouter are a known pre-launch pattern, so the capability isn't the surprise. The trade people are making is. You're getting a frontier coding model for free, and in exchange you're sending your repository to an operator nobody can name, whose retention policy and terms aren't published anywhere. I'm not going to guess whose it is, because there's nothing to guess from. I'd just want anyone routing work through it to have said that sentence out loud once.

83. Lenar Kess 00:29:25

Second: Bloomberg's sources say ByteDance is merging its coding platform Trae and its agent-building tool Coze into Doubao, and plans a product called Doubao Work aimed at Tencent's WorkBuddy. For anyone who hasn't touched them — Trae is a coding environment, Coze is a builder for making agents, and Doubao is ByteDance's consumer assistant brand.

84. Damra Vol 00:29:47

Two credible standalone developer products folded into an assistant. That's a company deciding the surface people will actually use isn't a separate tool with its own login. Whether US vendors drift the same way is an open question — but Trae and Coze weren't struggling products, which makes the consolidation a strategy rather than a cleanup.

85. Lenar Kess 00:30:07

Third: Thomson Reuters launched its first proprietary large language model this morning, called Thomson, combining their own legal archive with models from outside providers. First deployment is Tabular Analysis, a high-volume document review capability inside their CoCounsel Legal assistant.

86. Damra Vol 00:30:25

A data owner deciding that renting a model isn't enough, because the archive is the asset. And the calibration for that bet is sitting in today's arXiv batch — DGEval, a benchmark on the International Maritime Dangerous Goods Code, Amendment 42-24. Sixteen hundred seventy-eight questions, and thirteen models across six providers, including one maritime-specific fine-tune. The best model beats the human practitioner baseline on multiple choice.

87. Lenar Kess 00:30:55

And then falls apart where?

88. Damra Vol 00:30:57

Stowage, segregation, and regulatory recall. Which in dangerous-goods shipping is the whole job — what you may put next to what, and where on the vessel. Beating a practitioner on multiple choice while being weakest on exactly the provisions that cause fires is a very specific result, and the authors say plainly that it's a safety assurance instrument to be applied continuously, not a settled characterization of what models can do.

89. Lenar Kess 00:31:22

Last one. Sam Altman told the podcaster David Senra that he's worried AI ends up controlled by a handful of companies, models, or people, with nobody else having a say in how it affects society. He tied it partly to fear — that people frightened of AI will trade, in his words, a lot of liberty for safety.

90. Damra Vol 00:31:42

Trevor Levin asked the obvious thing on X, and I'd rather leave it unanswered than pretend to settle it: why did Altman talk about this class of risk more when OpenAI was a nonprofit than he does now that it's a trillion-dollar company? Both readings are available. Nothing published this weekend picks between them.

91. Lenar Kess 00:32:02

I'd put Chris Lehane next to it. OpenAI's chief global affairs officer talked to Robert Booth at the Guardian on Sunday and said people should prepare to defend against ongoing, persistent cyber-attacks from AI systems as models gain the ability to plan and launch offensives. His sentence was

— we are hitting a different chapter, a different moment within AI, in terms of what the capabilities of this technology can do. That's the same week the company paused development on some of its most advanced internal models, and it hasn't said which ones, or for how long. I'm Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic