Permission Slips and Poured Concrete

2026-06-03 / 00:18:06

> *“I'd rather track the steel than the speeches.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A stack of European filings wants to triple data center capacity and own more of the AI stack — on the same day a JP Morgan report says the country building fastest can't pour its own concrete on schedule. Lenar and Damra trace the day's real constraint: not model quality, but megawatts, transformers, capital, and rights.

- The EU's Cloud and AI Development Act (CADA) aims to triple data center capacity in 5–7 years, paired with a tech-sovereignty communication and open-source strategy and a Chips Act 2.0 — a statement of intent about which layers of the stack Europe wants to own.
- JP Morgan, via the WSJ, says 60%+ of US data center capacity planned for 2027 isn't yet under construction — the build-out is power- and permit-bound, not building-bound.
- Alibaba's Qwen 3.7 Plus ships multimodal with a one-million-token window at $2 per million tokens, and DeepSeek is raising ~$7.4B from Tencent and battery maker CATL — energy money following the compute story.
- Microsoft's on-device Aion 1.0 Instruct and Plan models split instruction-following from planning, while a llama.cpp build report shows reproducible local gains on two 3090s.

- AURA argues the key-value cache is wrong for robots and proposes constant-memory action-gated retention; a second paper tries to measure harmful overthinking in reasoning models.
- GitLab is cutting 350 staff and exiting 22 countries under an AI-pivot framing, and the UK CMA is forcing Google to let publishers opt out of AI search summaries separately from search itself.

SEGMENTS

Transcript

1. Lenar Kess 00:00:00

   Picture yourself running platform engineering for a mid-size bank in Frankfurt. Every model you call, every accelerator you rent, every object store your logs land in — almost all of it routes through an American company. Now your government decides that dependence is a problem, and it starts writing law to change it. That's roughly where the European Commission stood this morning.

   - techmeme.com
   - techmeme.com
   - digital-strategy.ec.europa.eu
   - digital-strategy.ec.europa.eu
   - digital-strategy.ec.europa.eu
   - techmeme.com
   - techmeme.com

- techmeme.com
- techmeme.com
- reddit.com
- reddit.com
- arxiv.org
- arxiv.org
- techmeme.com
- theguardian.com

2. Damra Vol 00:00:21

   So what actually landed? Because "tech sovereignty package" is the kind of phrase that can mean four binding regulations or one press conference and a slide deck.

3. Lenar Kess 00:00:30

   Four documents, dropped together. A Communication on European Tech Sovereignty, an EU Open Source Strategy bundled with it, the Cloud and AI Development Act — they're calling it CADA — and a Chips Act 2.0. Per Gian Volpicelli at Bloomberg, CADA's headline goal is to triple the EU's data center capacity over the next five to seven years, and Chips Act 2.0 would let the EU invest directly, not just subsidize.

4. Damra Vol 00:00:57

   Tripling data center capacity in five to seven years is an enormous number to put in a legislative proposal. That's a construction program, not a policy lever. Who's pouring the concrete?

5. Lenar Kess 00:01:08

   That's the gap I keep circling. The filing says it wants to triple capacity. It doesn't build a single hall. And the first Chips Act, back in 2023, was mostly about de-risking private money — state-aid permission, subsidies, coordination. The 2.0 version, per the Commission's own summary, "introduces new measures to further boost" the sector and adds this direct-investment idea. I haven't seen the full mechanism for that direct investment spelled out yet, so I'm taking the summary at its word there.

6. Damra Vol 00:01:38

   And the dependency they're worried about isn't only clouds. It's the fabs. ASML is Dutch, sure, but the leading-edge fabrication capacity sits in Taiwan and increasingly Arizona. Nvidia designs the chips everyone wants. So when Mathieu Pollet at Politico frames this as cutting reliance on US tech — sovereignty over which layer, exactly? You can build all the data center shells you want and still be renting the silicon inside them.

7. Lenar Kess 00:02:04

   Right, and the open-source strategy is the most interesting tell about how they're thinking. Pairing an open-source push with the sovereignty communication suggests they've concluded they can't out-spend the American labs on frontier models, so the play is open weights and shared infrastructure you don't have to license from a US company. That's a coherent bet. It's also a slower one.

8. Damra Vol 00:02:26

   It's coherent until you ask who maintains it. Open weights don't run themselves. Somebody has to do the integration, the security patching, and the eval work — the grind where a model becomes a system you can actually deploy in a regulated bank. If the EU funds the weights but not the people who operate them, you get a press release, not sovereignty.

9. Lenar Kess 00:02:47

   And there's a live counter-example worth putting next to that. Saritha Rai at Bloomberg has a piece on India trying to build and export its own sovereign AI template, and running straight into the same wall — a dependence on foreign AI infrastructure that the ambition can't wish away. So the EU isn't alone in this. Lots of governments want a sovereign stack. Almost none of them control the whole supply chain.

10. Damra Vol 00:03:10

    That's the shape of it. Sovereignty is a stack, and right now most countries own maybe two layers of seven. The EU filing is a statement of intent about the layers they'd like to own. Watch the funding lines, not the communiqué.

11. Lenar Kess 00:03:25

    That's what I'll be tracking — whether Chips Act 2.0 comes with appropriated euros attached or stays a framework. Which sets up the next piece neatly, because it turns out even the country that's furthest ahead on building can't keep up with its own plans.

12. Damra Vol 00:03:39

    You're going to the data center numbers.

13. Lenar Kess 00:03:41

    I am. Katherine Blunt at the Wall Street Journal has a report, sourced to JP Morgan, that the US data center build-out is falling behind schedule. The number that stopped me: more than sixty percent of the data center capacity planned to come online in 2027 isn't yet under construction.

14. Damra Vol 00:03:59

    Sixty percent not yet under construction, for capacity that's supposed to be live next year. [tsk] That's not a rounding error. A hyperscale data center is roughly an eighteen-to-thirty-month build once you break ground, before you even energize it. If the steel isn't up yet, that 2027 capacity is a 2028 number at best.

15. Lenar Kess 00:04:19

    And the binding constraint usually isn't the building. It's the power. You can pour the slab fast. Getting a utility interconnect, getting transformers — there's a multi-year backlog on large power transformers — and getting a grid operator to commit megawatts is where these projects stall. The building is the easy part.

16. Damra Vol 00:04:37

    So now hold the two stories next to each other. The EU is proposing to triple its capacity in five to seven years. The US, which has the most aggressive private build-out on the planet and the capital to back it, can't get sixty percent of its 2027 plan into the ground on time. The EU's number reads as fantastical, not just ambitious, unless they've solved a power-and-permitting problem the Americans haven't.

17. Lenar Kess 00:05:02

    That's the connection I'm willing to make, and only that far. Both stories land on the same constraint, and it isn't model quality. It's megawatts and concrete and transformers and the people who sign interconnection agreements. The capability conversation has gotten way ahead of the capacity to run it.

18. Damra Vol 00:05:20

    And for anyone building on top of this, it's a planning input. If you're assuming compute keeps getting cheaper and more abundant on a smooth curve because the models keep improving, the supply side has a different opinion. There may be a stretch where the best model you can get is gated by where you can physically get inference capacity, not by what the lab shipped.

19. Lenar Kess 00:05:40

    Which is a good segue to who's shipping, because the model news today came out of China, and it came with a price.

20. Damra Vol 00:05:46

    Alibaba and Qwen, I assume.

21. Lenar Kess 00:05:48

    Qwen 3.7 Plus. Per Carl Franzen at VentureBeat, it's a multimodal proprietary model with a one-million-token context window, and it lands at two dollars per million tokens — which they're pricing at sixty percent below their text-only Qwen 3.7 Max.

22. Damra Vol 00:06:05

    Wait — the multimodal one is cheaper than the text-only one? That's backwards from how this usually goes. Multimodal models carry the vision encoder and the extra compute on image tokens. You'd expect a premium, not a sixty-percent discount.

23. Lenar Kess 00:06:20

    It is backwards, and I don't have a clean explanation from the source. A couple of readings. One, Plus and Max are different model sizes — Plus is the smaller, faster tier in Alibaba's naming, so the comparison might be a cheaper model that happens to also be multimodal, not a multimodal discount per se. Two, it's a pricing move. Two dollars per million tokens with a million-token window is aggressive against anyone selling long-context inference.

24. Damra Vol 00:06:48

    The naming reading is probably right, and it matters because the headline "multimodal cheaper than text" claims more than the model card supports. Plus versus Max is capacity, not modality. Still — two dollars per million on a one-million-token window is a real number. That's the kind of price that shows up in your build-or-buy math for document processing.

25. Lenar Kess 00:07:09

    And it pairs with the capital story underneath it. Reuters reports DeepSeek is set to raise about seven and a half billion dollars in its first outside funding round — investors including Tencent and the battery maker CATL — at a valuation somewhere between fifty-two and fifty-nine billion dollars.

26. Damra Vol 00:07:27

    CATL is the interesting name on that list. A battery and energy-storage company taking a position in a frontier-model lab — that's a bet on the power story we were just talking about, more than on the model. If inference is constrained by megawatts, the people who store and move energy have a seat at this table.

27. Lenar Kess 00:07:46

    That's a sharp read, and I'd hold it loosely — corporate investors take positions for a lot of reasons, and we don't have CATL's thesis from the Reuters piece. But the shape is notable. DeepSeek's first

round, and it's already a fifty-billion-dollar company on paper, raising from a social platform and an energy firm rather than the usual venture names.

28. Damra Vol 00:08:06

    And it's a first round at that valuation, which tells you the previous funding was internal or quant-desk money. DeepSeek came out of a hedge fund. So this is the moment it goes from a research shop bankrolled by trading profits to a company taking serious outside capital. The interesting question is what strings come with Tencent money.

29. Lenar Kess 00:08:25

    We won't know that from a funding announcement. Let's move down a layer, because the other half of today's model news isn't about the giant proprietary systems. It's about what you can run on hardware you already own.

30. Damra Vol 00:08:37

    Microsoft at Build, and the local crowd.

31. Lenar Kess 00:08:39

    Microsoft announced two on-device models at Build 2026 — Aion 1.0 Instruct and Aion 1.0 Plan. Their framing on Instruct is "efficiency at scale" — a next-generation small language model they say is smaller, faster, and more efficient than their previous one. The Plan model is the more interesting of the two by name: a model whose job is planning, agentic decomposition, rather than just answering.

32. Damra Vol 00:09:06

    Splitting Instruct from Plan is an admission that one model doing everything is the wrong shape for on-device. You want a small, fast model to handle the turn-by-turn instruction following, and a separate model that's good at breaking a task into steps. That's the agent architecture moving onto the laptop. The catch is always the same — what are the real numbers? Microsoft saying "smaller and faster" tells me nothing until I see tokens per second on actual silicon and a benchmark I trust.

33. Lenar Kess 00:09:35

    Which is exactly why the local community is fun to read on the same morning. There's a post on the LocalLLaMA subreddit — a user shouting out a specific llama.cpp build, b9455, running on two RTX 3090s. He says that build sped up the Unsloth quantization of Qwen 3.6 — their 27-billion-parameter model — in the high-quality eight-bit quant.

34. Damra Vol 00:10:00

That's the texture I love about that corner of the world. Microsoft puts out a polished announcement with adjectives. And the same day, someone on two consumer cards is reporting an actual build number, an actual quant file, and a speedup they measured themselves. One of those is marketing, and one of those is a result you can reproduce tonight.

35. Lenar Kess 00:10:19

    And the gap between those two is shrinking, which matters for anyone deciding what to build on. A 27-billion-parameter model in an eight-bit quant on two 3090s is a serious local setup now. The question is no longer whether you can run something useful locally. It's which tier of capability you give up to keep your data on your own machine.

36. Damra Vol 00:10:40

    And the answer to that keeps getting better for the local side. But let's not oversell it — two 3090s is still a fifteen-hundred-dollar card situation and a power supply that trips your breaker. "Local" still costs you — it's just yours instead of rented.

37. Lenar Kess 00:10:56

    Fair. Let's go to the research, because two papers landed today that both poke at things we usually take for granted, and the first one has the best opening line I've read in a while.

38. Damra Vol 00:11:06

    Go ahead, set it up.

39. Lenar Kess 00:11:07

    It's a paper called AURA — Action-Gated Memory for Robot Policies at Constant VRAM, by Josef Chen, on arXiv. The abstract opens: "The key-value cache is the right memory for datacenters but the wrong memory for robots." And then it makes a simple argument. The key-value cache — the memory that makes large language model inference fast by remembering every prior token — grows without bound. A robot running continuously can't afford memory that grows forever.

40. Damra Vol 00:11:38

    That's a good framing, because the key-value cache is one of those things everyone in inference treats as just how it works. In a datacenter you have a request, you serve it, you free the cache. A robot doesn't have a request boundary. It's running the same policy for hours. If your memory grows with every observation, you hit the wall on video memory and the robot stops. "Constant VRAM" is the whole pitch in two words.

41. Lenar Kess 00:12:05

And "action-gated" is how they get there — instead of keeping every token, the memory update is gated on actions the policy takes, so what gets retained is tied to what the robot did, not to every frame it saw. I've read the abstract, not the full method, so I can't tell you how well it holds up against a baseline yet. But the problem statement is exactly right, and it's the kind of constraint that doesn't show up until you take a model off the server and put it on something that has to run all day.

42. Damra Vol 00:12:32

    It's the same lesson as the on-device model split. The architecture that's correct in a datacenter is the wrong architecture on the edge, and we're watching a whole wave of work rediscover that for memory, model size, and planning. The server assumptions don't survive contact with a battery and a fixed memory budget.

43. Lenar Kess 00:12:50

    The second paper is from a group including Simone Caldarella and Elisa Ricci, titled "Thinking Past the Answer: Evaluating Harmful Overthinking in Large Reasoning Models." The setup: reasoning models generate explicit chains of thought to improve accuracy, and this paper introduces an evaluation protocol for what they call harmful overthinking — cases where the extra reasoning makes the model worse, not better.

44. Damra Vol 00:13:15

    Which anyone who's watched a reasoning model talk itself out of a correct answer already knows in their gut. You ask a simple question, the model gets it right in the first sentence, then reasons for four hundred more tokens and arrives somewhere wrong. The value here is they're trying to measure it — reasoning sufficiency, knowing when to stop.

45. Lenar Kess 00:13:34

    And that connects to a cost question, not just an accuracy one. Every one of those reasoning tokens is billed and adds latency. If a model overthinks its way to a worse answer, you paid more for a downgrade. An eval that names that failure is useful precisely because the whole industry has been selling "more reasoning" as strictly better.

46. Damra Vol 00:13:54

    And it ties back to the capacity story from the top of the show. If reasoning tokens are sometimes wasted compute that also lowers your accuracy, then "think harder" isn't free in a world where inference capacity is the constraint. Knowing when to stop reasoning is a capacity optimization, not just a quality one.

47. Lenar Kess 00:14:15

That's the thread that actually holds across the day, and I'll keep it that loose. Let's close on two items about power and labor, because both are concrete and both matter to anyone in this field.

48. Damra Vol 00:14:25

    GitLab first.

49. Lenar Kess 00:14:27

    GitLab is laying off 350 people — about fourteen percent of its workforce — and exiting 22 countries, as it pivots to position itself as an AI-focused enterprise software development platform. That's per Dean Seal at the Wall Street Journal.

50. Damra Vol 00:14:42

    Exiting 22 countries is the detail that tells you what kind of cut this is. That's a retreat from go-to-market in whole regions, not a hiring-pace correction. A company that sells DevOps tooling worldwide deciding it can only afford to sell in a much smaller footprint — that's a company under real margin pressure dressing a contraction in the language of an AI pivot.

51. Lenar Kess 00:15:05

    I'll be careful and generous here, though. "AI pivot" gets used as a euphemism for layoffs, and sometimes it genuinely is one. But GitLab does have a real product problem to solve. If coding agents are doing more of the work inside the development loop, a platform built around human-authored merge requests and pipelines has to change shape or get disintermediated. The pivot might be both real strategy and cover for a hard quarter.

52. Damra Vol 00:15:29

    Sure, both can be true. But 350 people lost their jobs today, and I'd push back on the framing that softens that into a strategy story. The plainer version is: a public company missed its numbers, cut fourteen percent, and the AI narrative is what you tell investors so the stock doesn't fall further.

53. Lenar Kess 00:15:48

    That's fair, and we don't have their earnings detail in front of us to say which weighs more. The last item is a regulatory one, out of the UK. The Competition and Markets Authority ruled that UK media websites can now block Google from using their articles in its AI search summaries — the overviews that answer your query without you clicking through. Reported by Joanna Partridge and Dan Milmo at the Guardian.

54. Damra Vol 00:16:12

And the mechanism is the whole story there. Until now, the brutal part for publishers was that opting out of AI summaries meant opting out of Google search entirely — the same crawler, the same index. You couldn't say "index me but don't summarize me." If the CMA is forcing Google to split those, that's a real change in leverage.

55. Lenar Kess 00:16:34

    That's what I'd want to confirm in the full ruling — whether it's an actual separate opt-out that preserves your search ranking, or a softer commitment. The publishers' complaint was specific: AI overviews dropped their click-through traffic and the revenue that comes with it. A toggle that lets you keep the search visibility while denying the summary is the only version that actually helps them.

56. Damra Vol 00:16:54

    And it's a UK-only ruling for now, from one regulator. But it's the first time I've seen a competition authority treat "index for search" and "ingest for AI answers" as two different permissions a publisher can grant separately. If that distinction holds, it travels — every other regulator watching the same traffic collapse now has a template.

57. Lenar Kess 00:17:14

    One thread runs under all of it. The EU wants to own more of the stack and can't build the halls fast enough. The US can't pour its own concrete on schedule. China priced a model to move and raised energy money to back it. And a UK regulator started prying apart permissions that everyone treated as one. None of it is the model getting smarter. All of it is about who controls the capacity, the energy, and the rights underneath the models. That's what I'm watching into Thursday.

58. Damra Vol 00:17:41

    And the data center number is the one I'll be checking against. If sixty percent of 2027 capacity still isn't in the ground by the fall, every sovereignty proposal and every pricing move is operating on compute that doesn't exist yet. I'd rather track the steel than the speeches.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic