Custom silicon, futures contracts, and a five-hundred-million-dollar law firm

2026-05-28 / 00:14:12

> *“Hedging a unit whose cost halves every nine months is a hard contract to design.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Mistral spent one morning announcing chip ambitions, an Airbus and BMW supply deal, and a push to ensure Europe's independence from US tech giants. ByteDance is building its own CPUs. Taiwan has raised fourteen and a half billion dollars in debt to feed AI capacity. Shanghai and US exchanges are drafting futures contracts for compute. And Axios says Corporate America is starting to ask whether the AI spend is paying back, while Kirkland and Ellis sets aside five hundred million dollars to build its own platform. The day the infrastructure layer got financialized — and a lot of buyers looked up and asked what they bought. Also: Lenar is joined by a new co-host, Damra Vol.

- Mistral to explore designing its own chips (CNBC) — Arthur Mensch frames the move as controlling more of the infrastructure as Mistral competes with larger labs. Intent, not a roadmap.
- Mistral signs Airbus and BMW to ensure Europe's independence (Sam Schechner / WSJ via Techmeme) — industrial customers buying continuity in Paris as much as compute.

- ByteDance is developing its own CPUs (Reuters via Techmeme) — reported as supply-side defense against chip price hikes, not long-term ambition.
- Taiwanese tech books a record $14.5B of debt deals (Aileen Chuang / Bloomberg via Techmeme) — financing raised against expected AI demand.
- Shanghai is designing AI-token futures, US exchanges launching GPU compute futures (Reuters via Techmeme) — compute itself becomes a tradable underlying, with the spec on the token version still unclear.
- Corporate America enters its AI reckoning (Madison Mills / Axios) — CFOs are starting to ask for evidence of return.
- Kirkland & Ellis sets aside $500M to build its own AI platform (FT via Techmeme) — the top-grossing law firm wants tooling its competitors don't have.
- AI giants bet billions on the most expensive job in enterprise (Janakiram MSV / Forbes) — forward-deployed engineers as the labs' collision course with Accenture and TCS.
- Anthropic and OpenAI found PMF with coding agents (Simon Willison via Techmeme) — fit at the $200/month price point, where the harness explains more of the result than the underlying model.
- Miles Brundage's median MTS theorem — a frontier lab's policy positions converge to those of the median member of technical staff.
- Soro: a lightweight foundation model and chatbot for Tajik (Liashkov et al., arXiv) — a useful counterweight to a day of chip plans and futures contracts.

SEGMENTS

Transcript

1. Lenar Kess 00:00:00

   Before we get into the day — a small note about the show. If you tuned in expecting one voice, you're going to hear two. Damra Vol is joining me as co-host, and we both have new voices for this format. The plan hasn't changed — source-first, concrete, calibrated — just with another set of ears in the room. Damra, want to say hi?

   - techmeme.com
   - techmeme.com
   - cnbc.com
   - axios.com
   - x.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - forbes.com
   - arxiv.org
   - i.redd.it
2. Damra Vol 00:00:19

   Hi. I'm the one who's going to ask, [pause] okay, but does it actually run? Lenar tends to lay out the day's frame; I'll keep poking at the operator edge. Listener feedback drove this. Braid moves fast, the AI story moves faster, and the show is going to keep evolving along with it.

3. Lenar Kess 00:00:37

   Today's lineup is mostly money and metal. Mistral wants to design its own chips and has signed Airbus and BMW. ByteDance is building CPUs. Taiwanese tech firms have taken on fourteen and a half billion dollars in debt this year to feed AI capacity. Shanghai is drafting futures contracts for AI

tokens. US exchanges are about to launch GPU compute futures. And Corporate America is starting to ask whether any of the spend so far is paying back.

4. Damra Vol 00:01:05

   Then Simon Willison's blog post saying Anthropic and OpenAI finally found product-market fit with coding agents — which extends what we covered yesterday. And one short tweet from Miles Brundage about how policy actually gets made inside these companies. A through-line, if there is one — and I don't want to force one — the infrastructure layer is getting financialized at the same moment a lot of buyers are looking up and asking what they bought.

5. Lenar Kess 00:01:30

   Mistral is the headline. <emphasis>Three</emphasis> announcements in one morning. CNBC has Arthur Mensch saying Mistral is exploring designing its own chips for AI data centers. The Wall Street Journal has Sam Schechner reporting that Mistral is, quote, accelerating superintelligence development to ensure Europe's independence from US tech giants. And Mistral signed supply deals with Airbus and BMW.

6. Damra Vol 00:01:53

   Three releases in one morning is a positioning play. The chip line is what I'd press hardest on. Exploring designing is a long way from a tape-out. Did CNBC give any detail — partner foundry, target workload, custom accelerator versus a full server CPU?

7. Lenar Kess 00:02:10

   It doesn't. Mensch is quoted on intent and on the competitive logic — control more of the infrastructure as Mistral competes with the larger labs — and the article reads as ambition rather than a roadmap. I'd take that at face value. They're telling industrial customers, sovereign-fund investors, and the EU that they intend to climb the stack.

8. Damra Vol 00:02:30

   And Airbus and BMW are buying the signal that Mistral keeps existing. They aren't buying chips — they're buying continuity. [pause] A defense and aerospace customer and a German carmaker both want AI compute they can run on European hardware under European rules. Mistral is the available answer in Paris.

9. Lenar Kess 00:02:51

   Right. The chip framing tells those customers Mistral plans to own enough of the stack to be regulated locally. That matters a lot to a regulated buyer.

10. Damra Vol 00:03:00

    My skeptical read — and this is inference, not from the article — the chip program is probably a small accelerator effort designed around their model architecture, and the press got a quote about intent rather than a product. That's how it usually goes. What concerns me is whether Mistral can keep three things going at once: model training, customer deployment at Airbus and BMW, and silicon. That's three different muscle groups.

11. Lenar Kess 00:03:25

    It is. And nobody outside the company knows which one slips first. Whether the Airbus and BMW deals come with embedded engineers from Mistral changes a lot — because if they do, the consulting-collision piece from Forbes that's coming later starts to apply to Mistral too.

12. Lenar Kess 00:03:40

    ByteDance is doing a less dramatic version of the same move. Reuters has sources saying ByteDance is developing its own CPUs to support its AI infrastructure, as chip price hikes and supply shortages constrain expansion plans. Notice the framing — this is reported as supply-side defense, not long-term positioning.

13. Damra Vol 00:04:00

    ByteDance is already a hyperscaler in everything but name. They have the workload volume to justify a custom design. The Reuters piece is sourced, not announced — there's no press release. Did it say general-purpose CPU or AI-oriented?

14. Lenar Kess 00:04:15

    It says CPU. That language matters. If they want CPUs — not accelerators — they're targeting the host side of the data-center server, where Intel and AMD have been the only meaningful options. Pair that with Bloomberg reporting this morning: global AI hardware demand is so strong it's offsetting China's worries about a stronger yuan, because AI hardware exports are surging and chip-equipment imports are rising.

15. Damra Vol 00:04:41

    So Beijing is starting to read AI hardware as a trade-balance item. Not a strategic risk. A flow.

16. Lenar Kess 00:04:48

    At least at the macro level. And on the financing side, Bloomberg has Aileen Chuang reporting that Taiwanese tech companies have completed a record fourteen and a half billion dollars of debt deals

so far this year, racing to secure financing for AI capacity. TSMC, the ODMs, the packaging companies. The debt is being raised against expected AI demand.

17. Damra Vol 00:05:10

    That number puts the futures stuff in context. Reuters has the Shanghai Futures Exchange in the early stages of designing futures contracts for AI tokens, and US exchanges about to launch GPU compute futures. So compute itself becomes a tradable underlying.

18. Lenar Kess 00:05:26

    Two different instruments. A GPU compute future would let a buyer lock in a price for hours of H100 or B200 time. An AI token future is harder to pin down without seeing the spec. I'd want to know whether they mean inference tokens — a unit of model output — or token as a general settlement unit. The Reuters piece doesn't say.

19. Damra Vol 00:05:48

    If it's inference tokens, that's a strange product to write. The economics of a token shift every time a new model lands. Hedging a unit whose cost halves every nine months is a hard contract to design. If they mean GPU-hours quoted in token-equivalents, that's more legible — you've got a physical underlying. The token version is closer to electricity, while the GPU-hour version is closer to an ordinary commodity futures contract.

20. Lenar Kess 00:06:14

    There's no primary source yet for what Shanghai is drafting. The Reuters summary is what we have. Treat it as a flag for the day, not a conclusion. <soft>The spec will tell us more once the exchange publishes it.</soft>

21. Damra Vol 00:06:26

    The shape of it — debt to build capacity, futures to price capacity, and custom silicon to own capacity — lines up across the items. Whether the contracts work is a question for the people who have to take delivery.

22. Lenar Kess 00:06:40

    Axios this morning has Madison Mills with a piece titled Corporate America enters its AI reckoning. Quote — corporate leaders are starting to question whether soaring AI spending is delivering meaningful returns. Mills says companies rushed in and are now asking for evidence.

23. Damra Vol 00:06:57

That story has been queued up since the start of the year. I'd want the second-order version of it — which buyers are pulling back, and which are doubling down on a specific kind of spend?

24. Lenar Kess 00:07:07

    Right. Set the Axios piece next to the Financial Times reporting on Kirkland and Ellis. The world's highest-grossing law firm is setting aside <emphasis>five hundred million dollars</emphasis> to build its own AI platform, rather than rely on tools available to its rivals.

25. Damra Vol 00:07:22

    Five hundred million for a single firm. That isn't a vendor switch. That's, we are going to staff and operate our own AI organization. Did the FT name what they plan to build internally?

26. Lenar Kess 00:07:35

    Not in detail in the summary we have. The framing is competitive — they want tooling their competitors don't have. For a firm where the billable hour is the unit of production, even small efficiency gains translate into a lot of money. They're also one of the few firms with the cash flow to attempt it.

27. Damra Vol 00:07:51

    And it tells you which way the consulting question is breaking. Which is the Forbes piece.

28. Lenar Kess 00:07:56

    Yes. Janakiram MSV has a column today on what he calls the most expensive job in enterprise — forward-deployed engineers. Meta, OpenAI, and Anthropic are spending billions on forward-deployed engineers, and the piece argues this puts the frontier labs on a collision course with Accenture, TCS, and the rest of the systems-integrator world.

29. Damra Vol 00:08:19

    A forward-deployed engineer is the lab equivalent of a sales engineer with a soldering iron. You go to the customer, you understand their data, their workflow, their authentication boundary, and you bring back what the model needs. It's the role that turns a generic API into a deployed system.

30. Lenar Kess 00:08:37

    And the labs are paying for it because the alternative is letting integrators own the customer relationship and the data flow. If Kirkland builds its own platform, and BMW signs with Mistral, and Anthropic embeds engineers at a third firm — the consultancy layer gets squeezed from both sides.

31. Damra Vol 00:08:55

    Squeezed, not erased. The number of companies large enough to attract a forward-deployed engineer team from a lab is small. Most enterprises will still go through someone. The open question is whether that someone is Accenture or a labelled team from a frontier lab.

32. Lenar Kess 00:09:11

    My read — and this is inference — the labs only do this for accounts above some revenue floor. Below the floor, the integrators are fine. The reckoning Axios is describing is mostly happening at the middle tier — companies that bought generic AI platform licenses, didn't have forward-deployed engineers assigned to them, and didn't have the internal engineering depth to make it work.

33. Damra Vol 00:09:32

    That's a real cost on a real spreadsheet. CFOs are going to start naming it on Q3 calls. [pause] Which is when the Axios story stops being a vibe and starts being guidance.

34. Lenar Kess 00:09:43

    Simon Willison has a post today — quote — Anthropic and OpenAI seem to have finally found product-market fit with coding agents, which are quickly becoming daily drivers for highly paid professionals. We talked about Boris Cherny's coding-is-solved claim yesterday; Simon's frame is more careful. He isn't saying the model is the product. He's saying the coding agent is.

35. Damra Vol 00:10:04

    The harness wins again. Yesterday it was the DeepSWE result and the git-history loophole; today it's a market observation. Both point at the same thing. The model alone doesn't change the day. The model wrapped in tools, file access, and a permission model does.

36. Lenar Kess 00:10:22

    Simon also flags the price. The agents are profitable for the labs because the people using them all day are willing to pay two hundred dollars a month. That's a different market from chat. Chat is a consumer product priced at twenty dollars; the coding agent is a professional tool priced at the cost of a coffee a day.

37. Damra Vol 00:10:40

    And the audience matters. A highly paid professional in his framing means engineers, but also lawyers and finance people who use them in the same way. Which, given the Kirkland item, isn't theoretical anymore.

38. Lenar Kess 00:10:53

    Right. The product-market fit Simon is naming is the fit at the price point. Twenty dollars a month didn't pay for the model. Two hundred dollars a month does, if enough professionals stay subscribed.

39. Damra Vol 00:11:05

    What I'd press on is durability. Are these subscriptions sticky once a cheaper local option exists? Yesterday's local-frontier conversation matters here. If a model that runs on an M-series chip can do eighty percent of the coding agent's job — and that's a moving target — the two-hundred-dollar tier gets pressured.

40. Lenar Kess 00:11:24

    It does. Simon isn't claiming permanence. He's claiming the current shape of the fit. There's also a Reddit thread today on r-slash-LocalLLaMA about the leaderboard getting crowded — Hy3 preview, GPT five point four, and Gemini three point one Pro. The poster, ExoticYesterday8282, compares it to a crowded subway station.

41. Damra Vol 00:11:47

    Crowded with claims, anyway. The Hy3 preview number on the CHSBO 2025 chart — I haven't read the methodology. Eighty-seven point eight beats Gemini and GPT, but the gap between those three is small enough that the harness around them matters more than the model number. Which is exactly the point Simon keeps making about coding agents.

42. Lenar Kess 00:12:09

    Two short items to close. Miles Brundage tweeted what he calls the median MTS theorem — quote — a frontier AI company's policy positions eventually converge to those of the median member of technical staff there.

43. Damra Vol 00:12:23

    That's a sharp tweet from someone who's been inside enough of these companies to mean it. He's arguing policy doesn't get set by the founder or the policy team — it gets set by what the engineers will tolerate.

44. Lenar Kess 00:12:34

    And it's an observation about institutions, not just AI. Internal engineering culture shapes what the company will and won't ship, what it will sign onto, and what it will reverse under pressure. If

Brundage is right, the lever for outside policy work is the population of engineers who join these companies, not the executives.

45. Damra Vol 00:12:52

    A small theorem on a Thursday morning, and I'll keep it on hand. What's the second item?

46. Lenar Kess 00:12:58

    Soro. A paper on arXiv this morning from a team led by Stanislav Liashkov — a lightweight foundation model and chatbot for Tajik. They build a Tajik-specialized conversational large language model designed for low-compute deployment. Not a leaderboard story. Not part of the frontier race. Just a team building a useful model in a language the bigger labs don't prioritize.

47. Damra Vol 00:13:20

    And papers like that ship every week now. The frontier race gets all the attention, but the long tail of language-specific models is where most of the world will meet these systems. It's a useful counterweight to a day full of chip plans, debt deals, and futures contracts.

48. Lenar Kess 00:13:36

    Mistral's chip detail is the item that either lands or evaporates by end of week. If anyone tapes out, that changes the story. The Axios piece could pull a CFO statement before Friday. And if another top-grossing law firm files something like the Kirkland announcement in the next two weeks, the five-hundred-million number stops being a project and starts being a pattern.

49. Damra Vol 00:13:56

    That's the one going into my Friday notes. Thanks for letting me share the room.

50. Lenar Kess 00:14:00

    Thanks for the company. I'm Lenar Kess.

51. Damra Vol 00:14:03

    I'm Damra Vol.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic