Cheaper From Both Ends

2026-06-01 / 00:19:55

> *"Twelve cents against five dollars is the kind of gap that rewrites what you're willing to let an agent try."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A Chinese lab cut the price of a frontier-class coding model to a fraction of Opus, Nvidia tried to own every layer from the laptop to the data center, and one developer ran the new Gemma 4 on a decade-old Xeon. The cost of running intelligence got attacked from both ends on the same morning — and the question underneath all of it is who gets to set that cost.

- MiniMax M3 claims parity with Opus 4.7 at roughly twelve cents per million input tokens versus five dollars — but the weights are promised in about ten days, so "open-weights" is still a countdown.
- Nvidia's DGX Station puts a GB300 chip and up to 748GB of memory on a desktop, enough to run a one-trillion-parameter model locally; the RTX Spark chip pushes the same idea into laptops, while the Vera CPUs — with Anthropic, OpenAI, and SpaceX as early customers — signal a move off x86.
- A 10-year-old Xeon is all you need: cafkafk runs a 26B mixture-of-experts model at reading speed on a 2016 CPU with no GPU, arguing mainstream tools hide the performance levers.

- Cosmos 3 is Nvidia's open physical-AI world model, backed by a Cosmos Coalition with Runway as a founding member.
- Cadence and Nvidia claim a "Level 5" autonomous chip-verification agent that turns months into a day — a large autonomy claim in a domain where mistakes ship in silicon.
- Anthropic will let the EU's ENISA join Project Glasswing for access to a model called Mythos, even as a Wirescreen analysis documents 500+ PLA attempts to procure Nvidia chips and governments from India and the UAE to France move to own their compute.

SEGMENTS

Transcript

1. Lenar Kess 00:00:00

   Here's the choice that landed on the table around two o'clock this morning, our time. A Chinese lab called MiniMax posts a new coding model — they're calling it M3 — and the pitch is that it goes head to head with Anthropic's Opus 4.7 on the kind of work you'd actually point a coding agent at. Then The Information runs the number that makes you put the coffee down. Twelve cents per million input tokens, against roughly five dollars for Opus 4.7. Same neighborhood of capability, they claim, at something like one-fortieth of the input price. So the question I woke up with is — if that holds, what do you stop being careful about?

- x.com
- techmeme.com
- techmeme.com
- theguardian.com
- techmeme.com
- point.free
- axios.com
- blogs.nvidia.com
- x.com
- forbes.com
- techmeme.com
- techmeme.com
- restofworld.org
- techmeme.com

2. Damra Vol 00:00:37

   That last bit is the whole game. At five dollars a million you ration context. You think hard before you hand the agent the entire repository, you trim the system prompt, and you cache aggressively because every retry has a meter running. At twelve cents you just — stop counting. You let it read the whole codebase twice. So the price isn't only a budget line, it changes the shape of what you're willing to attempt.

3. Lenar Kess 00:01:01

   Right, and MiniMax clearly knows that's the hook. Let me read you their own framing, because the headline claim is carrying a lot and I want to be precise about it. The post says — quote — "Introducing MiniMax M3: the first open-weights model to combine three frontier capabilities." Then they list coding and agentic numbers: fifty-nine percent on SWE-Bench Pro, sixty-six on Terminal Bench 2.1, and a couple of others further down. They say their sparse-attention setup scales context to a million tokens, and that it's natively multimodal from the start.

4. Damra Vol 00:01:35

   [tsk] Okay, but read me the last line of that post. Because I saw it and it changes how I'd file this.

5. Lenar Kess 00:01:41

   The last line is — "Weights and tech report in about ten days."

6. Damra Vol 00:01:45

There it is. So "open-weights" today is a promise with a date on it. You can hit the API right now. But what makes open weights actually matter — running it on your own hardware, auditing it, fine-tuning it, keeping your code off a server in another jurisdiction — none of that exists yet. It's open-weights the way a pre-order is a product. I'd hold the label loosely until the tarball is up and somebody's loaded it.

7. Lenar Kess 00:02:12

   That's fair, and it's worth saying the benchmarks are all self-reported. We spent yesterday's episode on exactly this problem — how a few words in the prompt can swing a benchmark ten or twenty points, and how Opus 4.8's jump over the weekend had everyone asking about contamination. So I'm not going to treat fifty-nine percent on SWE-Bench Pro as a fact about M3. I'm going to treat it as MiniMax's claim about M3, made by a vendor on launch morning.

8. Damra Vol 00:02:40

   And notice who they picked to stand next to. The Information says it rivals Opus 4.7. Not 4.8 — which Anthropic shipped over the weekend. So they're benchmarking against last month's frontier, which is the reasonable thing to do if 4.8 is too new to have stable numbers, but it also means the comparison everyone's going to repeat is already one model out of date. The replies under the post are the giveaway, by the way — people aren't arguing about the architecture, they're asking what the latency is on that million-token window. That's the right question. A cheap token that takes thirty seconds to come back isn't cheap if your agent needs forty round-trips.

9. Lenar Kess 00:03:19

   So where does that leave a working developer this morning? You can't download it. You can hit the API, there's a fifty-percent-off promo for the first week, and there's a coding harness they're pushing alongside it. My read: it's a real signal about price pressure on the closed frontier, and it's not yet a thing you can build your week on. The number that matters isn't the benchmark, it's whether independent evals on private tasks land anywhere near the claim once the weights are actually out.

10. Damra Vol 00:03:46

    And whether the ten days is ten days. I've watched "weights coming soon" age into a quarter more than once. If they ship on schedule and the numbers survive contact with somebody's private eval, that's the story. Until then it's a very interesting API with a countdown attached.

11. Lenar Kess 00:04:04

    Now flip to the other end of the same problem, because Nvidia spent today attacking cost from the hardware side, and they did it in three pieces. Start with the one that's hardest to ignore. They unveiled something called the DGX Station — this is a desktop Windows machine, the kind of thing

that sits next to your monitor — built on their GB300 Grace Blackwell chip, with up to seven hundred forty-eight gigabytes of memory, which SiliconANGLE says is enough to run a one-trillion-parameter model locally.

12. Damra Vol 00:04:33

    On a desk. Not in a rack, not in a colo, on a desk under fluorescent light. The framing in their own materials is that they're — their word — uprooting supercomputers from the data center. And the seven-hundred-forty-eight-gigabyte number is what matters here, because memory is the wall you hit running big models locally, not raw compute. If that's unified memory the chip can actually address, a one-trillion-parameter model in the room with you is a different workflow altogether. The catch is going to be price and power draw, and they didn't lead with either.

13. Lenar Kess 00:05:07

    They didn't, and I don't have a number for the Station, so I'm not going to invent one. Second piece, further down the stack: a chip they're calling RTX Spark, aimed at laptops and ordinary PCs, for Microsoft Windows. The Guardian's framing — and this is Nvidia's pitch, not the Guardian's editorializing — is that it'll let AI agents replace the mouse and keyboard as how you drive the machine.

14. Damra Vol 00:05:31

    [chuckle] I want to push on that, because "replace the mouse and keyboard" is a marketing sentence, not an engineering one. What it actually means is on-device inference fast enough that an agent watching your screen and acting for you doesn't have to round-trip to a data center. That's the real claim, and it's a good one — local latency, your data stays on the machine, no per-token meter. But "replace the mouse and keyboard" has been the demo for two years and the thing people actually do is still type. I'll believe the input device changed when I watch someone ship a day's work without touching the keys.

15. Lenar Kess 00:06:07

    And the Guardian notes this puts Nvidia head to head with Intel, Apple, Qualcomm, and AMD on the PC chip — which is new turf for them. They're a five-trillion-dollar company walking into the laptop silicon fight. Third piece, and this is the one I think a builder should actually care about: the Vera CPUs. Jensen Huang says Anthropic, OpenAI, and SpaceX are among the first big customers, and the claim is they're one-point-eight times faster than x86 chips on AI workloads.

16. Damra Vol 00:06:38

    That's the piece with teeth, and here's why. Everyone fixates on the GPU, but the processor feeding it — the part that handles orchestration, data loading, all the glue around the model — has been an

x86 chip from Intel or AMD this whole time. If Nvidia now owns the processor and the accelerator and the networking between them, they're selling the whole machine, not a part. One-point-eight times faster is the headline, but the bigger thing is that Anthropic and OpenAI signing on means the people training frontier models are willing to leave x86 to do it. That's a supply-chain shift, not a spec bump.

17. Lenar Kess 00:07:16

    So the through-line across all three: a model lab cut the price of a token this morning, and Nvidia spent the same morning trying to own every layer of the box the tokens come out of. Cheaper inference from one direction, and the company that sells the picks and shovels making sure it sells all of them. Those aren't the same story, but they point at the same thing — the cost of running intelligence, and who gets to set it.

18. Damra Vol 00:07:39

    And the winner in both is the developer who just wants the bill to go down. Whether it goes down because a Chinese lab undercut Opus or because the compute got cheaper per watt — from the seat where you're paying the invoice, you don't care which lever moved.

19. Lenar Kess 00:07:52

    Which is the perfect setup for my favorite thing on the internet today, and it's the opposite of a five-trillion-dollar product launch. A developer who goes by cafkafk put up a post titled "A ten-year-old Xeon is all you need." It hit the front page of Hacker News, sitting around a hundred sixty points when I looked. And the claim is exactly what it sounds like: they got the new Gemma 4 — a twenty-six-billion-parameter mixture-of-experts model — running at reading speed on a single Xeon from 2016, with a hundred twenty-eight gigabytes of old DDR3 memory, and no GPU at all.

20. Damra Vol 00:08:26

    Reading speed on a recycled server with a decade-old processor. Let me say what reading speed means, because it's the unvarnished version of fast — it's roughly as quick as you can read the words as they appear. It's not going to drive a forty-step agent loop. But for a single developer asking a capable model questions on hardware that was headed for e-waste? That's remarkable, and it's the exact counterweight to the seven-hundred-forty-eight-gigabyte desktop.

21. Lenar Kess 00:08:52

    And listen to why they did it, because it's a craft complaint, not a flex. cafkafk wrote — quote — "I wrote this post after getting frustrated by the lack of ways to run the new Gemma 4 Drafter models, and mainstream tools not prioritizing this, and hiding all the performance levers." Hiding the performance levers. That's the line that stuck with me.

22. Damra Vol 00:09:11

    Because that's the actual fight in local inference right now. The mainstream tools — the friendly one-click runners — optimize for the common case, which is a recent GPU, and they bury the knobs that would let you make an old CPU work. cafkafk had to drop to a community fork of llama.cpp just to get the quantized weights to load. So the post isn't really "old hardware is fine." It's "the capability is reachable on old hardware, and the tooling is making it harder to find." The model isn't the constraint. The layer wrapped around it is — which is the same place we keep landing.

23. Lenar Kess 00:09:47

    And there's a lovely human detail in the thread. They mention the server isn't even dedicated to this — it's busy acting as a Nix cache the rest of the time. So this is someone's actual recycled box doing real work, with the model running in whatever headroom is left over. Somebody in the comments immediately asks whether an old Apple desktop with a hundred twenty-eight gigabytes would do the same thing.

24. Damra Vol 00:10:09

    Which is the right instinct, and the answer is probably yes with the same caveats — you're memory-bound here rather than compute-bound, so the question is always how much memory you can address and how patient you are. I love this post because it resets the frame on the whole Nvidia morning. You don't need a trillion-parameter desktop to do useful work. You need enough memory and somebody willing to read the manual the friendly tools are hiding from you.

25. Lenar Kess 00:10:35

    So put the two next to each other and don't overclaim it. Nvidia is selling the ceiling. cafkafk is mapping the floor. Most of us build somewhere in the middle, and what today actually tells you is that the middle got wider in both directions on the same day.

26. Lenar Kess 00:10:50

    Nvidia had a second announcement today that's a different kind of thing, and I want to give it room because it's not just hardware. They released Cosmos 3 — they're calling it an open physical AI foundation model, or a world model. The pitch, from Axios and from Nvidia's own write-up, is that it helps robots and autonomous cars understand and predict the physical world with limited training data. The blog post title is the giveaway: "How Cosmos 3 helps physical AI think before it acts."

27. Damra Vol 00:11:20

Okay, "world model" is one of those terms that means something specific and also gets sprayed on everything, so let me try to pin it. The idea is a model that's learned how the world tends to behave — objects fall, a car ahead brakes, and a cup tips when you push it — so a robot can simulate what'll probably happen next before it moves. The "with limited training data" part is the real claim, because the bottleneck in robotics has always been that you can't collect a billion miles of every weird situation. If a pretrained world model lets you get away with less real-world data, that's the unlock.

28. Lenar Kess 00:11:56

    And they're not doing it alone. There's a parallel announcement — Runway, the video-generation company, posted that it's a founding member of something called the Cosmos Coalition, which they describe as a global initiative with Nvidia and other AI labs to build and open-source frontier world models for physical AI. So Nvidia is trying to make Cosmos a shared standard, not just a product.

29. Damra Vol 00:12:20

    Which is a smart and slightly self-interested move, right? If you make the world model open and get a coalition around it, you grow the whole robotics-and-autonomy market, and every one of those systems trains and runs on — well, your chips. Runway being in there is the interesting part to me. A company that's spent years modeling how pixels move through video frames has something real to contribute to predicting physical motion. That's adjacent expertise doing real work, not a logo on a slide.

30. Lenar Kess 00:12:48

    I'll flag the limit, though. "Open physical AI foundation model" is the same phrase MiniMax used this morning — open. And I haven't seen the license terms, the dataset, or independent results yet. So I'm taking Nvidia's framing as Nvidia's framing. The thing I'd watch is whether a robotics team that isn't Nvidia ships something on Cosmos 3 and reports how much real-world data they actually got to skip.

31. Damra Vol 00:13:13

    That's the number that would make it real. Everything before that is a very well-produced demo, and we've all learned to discount the demo.

32. Lenar Kess 00:13:20

    One more out of the same orbit, and this one's about our craft specifically. Cadence — the big chip-design-software company — and Nvidia announced what Karl Freund, writing in Forbes, calls the first fully autonomous EDA agent. EDA is electronic design automation, the software you use to

design and verify chips. And the specific claim is a — quote — "Level 5 AI EDA agent" for design verification, turning a months-long effort into one day.

33. Damra Vol 00:13:48

    Months into a day, and "Level 5," which is borrowing the self-driving autonomy scale — Level 5 meaning no human in the loop at all. [tsk] I want to be careful here, because design verification is one of the highest-stakes, most formal corners of all of engineering. You're proving a chip does what the spec says before you spend tens of millions taping it out. Getting that wrong isn't a bug ticket, it's a recall. So "fully autonomous" in that context is a very large thing to assert.

34. Lenar Kess 00:14:18

    It is, and I'd separate the claim from the consequence. Verification is interesting for automation precisely because it's so formal — there's a spec, there are properties you can check, the correctness criteria are more machine-checkable than, say, frontend code. So it's a more reasonable place to push toward autonomy than most of software. The "months to a day" number is what I'd want decomposed. Months of what? Engineer time? Wall-clock simulation time? Those are very different claims.

35. Damra Vol 00:14:46

    Right, and if it's compressing the engineer's iteration loop — the agent proposes the verification plan, runs the suites, triages the failures, and a human signs off — that's believable and useful, and it's not the same as Level 5. The gap between "the agent did the tedious ninety percent and a senior engineer approved it" and "no human in the loop" is the whole ballgame in a domain where a mistake ships in silicon. I'd bet the working reality is the first one, dressed in the language of the second.

36. Lenar Kess 00:15:16

    And that's the pattern worth naming, lightly, because it's everywhere this morning. Cadence says Level 5. The engineering version is almost certainly a very strong assistant with a human approving the result. Both can be true — the assistant can be transformative — but the autonomy label is the marketing, and the approval step is the craft. Watch for whether any customer reports running it without that sign-off. I'd be surprised.

37. Damra Vol 00:15:41

    And if they do, I want to know who's liable when the chip's wrong. Autonomy claims and liability questions travel together, and verification is the one place the bill is measured in mask sets.

38. Lenar Kess 00:15:53

Last territory, and it steps off the hardware bench into who controls the thing. Bloomberg — Gian Volpicelli — reports that Anthropic plans to let the EU's cybersecurity agency, ENISA, join something called Project Glasswing, which gives ENISA access to a model the reporting calls Mythos. And the detail underneath it is the one I keep turning over: EU officials reportedly traveled to the US to ask for that access. They had to go and ask.

39. Damra Vol 00:16:21

That direction-of-travel detail is everything. A bloc of sovereign governments flying to a private company's offices to request access to a model — that's the power relationship stated plainly. I don't have the primary on what Glasswing or Mythos actually are; the reporting names them but I haven't seen Anthropic describe them publicly, so I'm taking Bloomberg's sourcing at face value. But the shape comes through even so: the company holds something a government wants, and the government is the one making the trip.

40. Lenar Kess 00:16:52

And ENISA being the first EU agency in suggests this is a template, not a one-off. If Anthropic is building a structured way for a cybersecurity agency to get supervised access to a frontier model, that's a governance mechanism being invented in private, by the vendor, on the vendor's terms. Which is a very different model of regulation than the EU usually runs — they're used to writing the rules, not requesting entry.

41. Damra Vol 00:17:16

And it pairs with the other geopolitics item today in an uncomfortable way. There's a New York Times piece on a Wirescreen analysis of Chinese military procurement records — more than five hundred instances since 2019 where the People's Liberation Army sought Nvidia chips, including the A100 and the A800. So on one side you've got export controls trying to keep frontier compute out of certain hands, and on the other a documented five-hundred-plus attempts to get it anyway. Control of access is the live question on both ends — who can buy the chips, and who can touch the models.

42. Lenar Kess 00:17:52

And there's a third data point in the same vein, which is countries deciding they'd rather own the hardware than rent it. Rest of World has a piece on India and the UAE — G42 deploying US-designed supercomputers inside India, as a model for governments that want to own their AI hardware rather than depend on American cloud providers. So you've got the EU asking a company for access, China working around export controls, and India and the UAE trying to buy their way to independence. Three different answers to the same question.

43. Damra Vol 00:18:24

    And the question is sovereignty over compute, which a year ago was a policy-paper abstraction and is now a procurement decision with a delivery date. Even the money's moved — there's an item today about a French private equity firm, Ardian, partnering with a data-center group to build an up-to-five-billion-euro AI gigafactory outside Paris, targeting five hundred megawatts. Europe doesn't want to ask for access forever. It wants its own building with its own power contract.

44. Lenar Kess 00:18:54

    So that's the day, and the threads don't all tie into one bow, so I won't force them. A coding model got radically cheaper to call, with the weights promised in ten days. Nvidia tried to own every layer from the laptop to the data center, and shipped a world model to go with it. One developer ran a frontier model on a Xeon headed for the scrap heap. And governments spent the day working out whether to ask for access, route around it, or build their own.

45. Damra Vol 00:19:22

    What I'll be watching tomorrow is narrow and checkable. Whether MiniMax actually posts those weights on schedule, and whether anyone runs them on a private eval that isn't the launch deck. The rest is interesting. That one's falsifiable.

46. Lenar Kess 00:19:35

    Falsifiable beats impressive. We'll see if the tarball shows up. Until then, that's where we'll leave it — Lenar and Damra, and a ten-year-old Xeon that apparently still has a job.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic