The number nobody optimized for

2026-05-30 / 00:18:29

> *“Two documents from the same week that don't contradict each other. One says Opus 4.8 jumped on math and slipped on business ops. The other says the whole bar-chart genre measures the harness as much as the model.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Claude Opus 4.8 landed overnight with a math score that leapt and a business-ops score that fell — and reading the release honestly means distrusting the chart. Lenar and Damra work through the gap between the number that moved and the number that matters, then chase it into agent budgets, the protocol wars, local-inference tooling, Mistral's on-prem bet, and the power grid.

- A scrape of 100+ Opus 4.8 evals shows USAMO 2026 jumping 69%→97% while Vending-Bench 2 nearly halved — a retune that helped some distributions and hurt others.
- "AI benchmarks are useless" argues the record scores ride on elaborate prompt setups: change a few prompt words and results swing 10–20 points.
- The BAGEN study finds frontier agents can't estimate their own remaining budget mid-task — which collides with enterprises trying to rein in "tokenmaxxing" (WSJ via Techmeme).

- "MCP is dead?" gets a sharp rebuttal from OpenAI's Max Stoiber: nearly every company is building an MCP server, even ones with no CLI or external API.
- Multi-token prediction benchmarks hit ~3.3x faster local inference; llama.cpp got a real website and antirez shipped distributed inference.
- Notes from the Mistral AI Now Summit — on-prem KYC at BNP Paribas, against a comment that Mistral's 120B "small" model loses to models a quarter its size. xAI countered with a one-dollar coding model.
- FERC's June grid-connection proposal is the duller, realer infrastructure story next to an unsourced TerraFab "one terawatt" claim.

SEGMENTS

Transcript

1. Lenar Kess 00:00:00

   Here's a number from overnight. A model sat down with the USA Math Olympiad — this year's set, 2026 — and went from sixty-nine percent to ninety-seven percent in a single version bump. That's davidthesong on the Anthropic subreddit, who scraped more than a hundred evals on the new Claude Opus 4.8 to see what actually moved against 4.7. Ninety-seven percent on olympiad math is the kind of jump that, a year ago, would have been the entire headline. So my first question isn't whether the model is good. It's narrower — when a number moves that far, that fast, what's the first thing you check?

- reddit.com
- reddit.com
- youtube.com
- techmeme.com
- x.com
- quandri.io
- reddit.com
- x.com
- x.com
- x.com
- koenvangilst.nl
- x.com
- techmeme.com
- x.com

2. Damra Vol 00:00:37

   Contamination. [tsk] The USA Math Olympiad set for 2026 — was it public before the training cut-off? Because the cleanest explanation for a twenty-eight-point jump on a named, dated benchmark is that the questions, or close paraphrases of them, ended up in the training data. I'm not saying that's what happened. I'm saying it's the first hypothesis, and the post doesn't rule it out. And the same scrape has the counterweight built right in. davidthesong lists the areas that barely moved or got worse — legal reasoning, healthcare, finance, and multilingual. And Vending-Bench 2, the run-a-tiny-business simulation, nearly halved.

3. Lenar Kess 00:01:16

   Halved. So the same release that aces olympiad math gets noticeably worse at running a pretend vending machine.

4. Damra Vol 00:01:22

   And that's the more honest picture of a frontier bump in 2026. One rising number doesn't drag everything up with it. The model got retuned, and the retune helped some distributions and hurt others. The vending-bench drop is the one I'd want Anthropic to explain, because that's a long-horizon task — keep your goal straight over many steps — and that's exactly the place these models are supposed to be getting better, not worse.

5. Lenar Kess 00:01:46

   The official framing, at least as it comes through the AI Daily Brief's write-up, isn't really about the math score at all. The pitch for 4.8 is reliability. Better code judgment, more self-correction, and

what they call uncertainty flagging — the model telling you when it isn't sure instead of confidently bluffing. The detail that caught me is that it's more willing to critique a proposal without being asked. You hand it a plan, and it pushes back on the plan itself.

6. Damra Vol 00:02:13

   Which is useful inside a long agent run, and also the claim I trust least from a benchmark. How do you score 'pushed back appropriately'? And the caveat is right there in the same write-up — it sometimes grounds that critique in assumptions it never verified. So now it's confidently disagreeing instead of confidently agreeing. The sycophancy moved. It didn't leave.

7. Lenar Kess 00:02:36

   The headline numbers, for what they are, are modest and real. SWE-bench Pro went from about sixty-four to sixty-nine. Terminal-Bench 2.0 — the agent-in-a-shell test — from sixty-six to seventy-four. Humanity's Last Exam ticked up a few points. Solid, not seismic. But there's a post on the Claude subreddit this week, titled flatly 'AI benchmarks are useless,' and I think it's the right thing to read next to the eval scrape — not against it.

8. Damra Vol 00:03:06

   He's blunt, and he's mostly right. Quote — 'This is Goodhart's Law playing out completely. The labs tuned everything for the tests, and now we've got these fragile models that break down in production.' And then the line I'd underline for any working engineer — quote — 'Tweak a few words in the prompt and your results swing ten to twenty points.' That's the whole gap between a leaderboard and your repository. The record score rides on elaborate prompt setups and multi-shot prompts tuned to the eval. You send a normal prompt, and a lot of that performance evaporates.

9. Lenar Kess 00:03:39

   So we have two documents from the same week that don't actually contradict each other. One says Opus 4.8 jumped on math and slipped on business operations. The other says the entire bar-chart genre is measuring the harness as much as the model. What I'll do is wait for someone to run 4.8 on a fresh, private task and report back. The number I believe most is the one nobody optimized for.

10. Lenar Kess 00:04:02

    Here's a word the Wall Street Journal put in print this week — tokenmaxxing. Bradley Olson's piece, which came across Techmeme, says executives at Uber, Meta, Microsoft and others are trying to rein in tokenmaxxing by their own employees. People burning model spend so fast it's turned into a real line item on the bill.

11. Damra Vol 00:04:21

And I want to be fair to the employees before anyone calls it waste. Tokenmaxxing usually means somebody found that stuffing the whole codebase into context, or firing off five agent attempts and keeping the best one, actually produces better work. The incentive is local and rational — better output today. The cost is somebody else's problem a quarter later, when finance reads the invoice. That's an unpriced resource being used the way unpriced resources always get used.

12. Lenar Kess 00:04:50

    And it connects directly to a study that dropped the same day. Zihan Wang — handle wzenus — posted a thread that opens, almost rudely, with quote, 'Claude-Opus-4.8 takes you too much tokens.' Then it turns into a real piece of research called BAGEN, Budget-Aware Agents. They test budget awareness across four environments and five frontier agents, and find structured failures in most of them.

13. Damra Vol 00:05:16

    What does 'budget awareness' even mean as a measurable thing? Because that's the part that makes it research instead of a complaint.

14. Lenar Kess 00:05:24

    Their definition is sharp. Quote — 'A budget-aware agent doesn't just spend less, but estimates remaining budget mid-task with uncertainty.' So the bar is whether the agent, halfway through a job, can tell you roughly how much of its budget is left, and how confident it is in that estimate — not just whether it spends less.

15. Damra Vol 00:05:42

    And the finding is that mostly they can't. The agent has no proprioception about its own spending. Ask it how much it's burned and it doesn't really know. [tsk] That's the uncomfortable pairing with the tokenmaxxing story. The enterprise response is going to be the crude lever — a hard token cap, a quota, a rate limit. But a budget-blind agent under a hard cap doesn't plan toward the limit. It runs full speed and dies at the boundary, mid-task, with the work half done.

16. Lenar Kess 00:06:12

    Which is the worst of both — you paid for the tokens and you didn't get the finished job.

17. Damra Vol 00:06:17

    Right. The thing I'd actually want, and what BAGEN is gesturing at, is an agent that treats budget like a resource it reasons about — slows down, picks a cheaper path, and tells you it's running low and asks whether to continue. That's a planning capability, not a billing setting. And almost

nobody's models have it yet. So the near-term fix is external governors clamping budget-blind agents, and that'll feel exactly as clumsy as it sounds.

18. Lenar Kess 00:06:44

    There's a post on the Quandri engineering blog with a deliberately provocative title — 'MCP is dead?' — and it hit the front page of Hacker News, two hundred eighty-odd points. MCP is the Model Context Protocol, the thing Anthropic introduced to let agents reach outside services. The post is part of a recurring genre by now: the protocol's done, code-mode or plain command-line tools will eat it.

19. Damra Vol 00:07:08

    And there's a real technical argument underneath the headline. The complaint is that MCP as a transport — the actual wire format the model speaks — is clunky, and you could replace it with the agent just calling a command-line tool, or writing code that hits an API directly. That part's a fair fight. People really are routing around it.

20. Lenar Kess 00:07:29

    And then the top comment in the thread is from someone with standing to know. Max Stoiber — handle mxstbr — opens with, quote, 'I run the team at OpenAI that's responsible for the ChatGPT App Store, Codex plugins, and all things MCP.' His argument is that the transport question is a red herring.

21. Damra Vol 00:07:48

    Let me read the core of it, because it's the strongest version. Quote — 'practically every company on the planet is building an MCP server. I know this because we interact with all of them. Most of these companies don't have a CLI. Many of these companies don't even have an external API. And yet, they're all building MCP servers.' That's the move. The value isn't the wire format. It's that MCP became the default thing a company ships to make its service reachable by an agent at all.

22. Lenar Kess 00:08:19

    And he basically concedes the technical complaint to make that point — quote, 'Maybe we will turn every MCP server into a CLI under the hood. Maybe we'll use code mode.' He's saying the implementation can change completely and the protocol still won.

23. Damra Vol 00:08:34

    Which is a little too clean, so let me add the comment that complicates it. A user named tanin lays out the actual split. If you're building a connector for yourself or your team, skip MCP — just hand

your teammates a command-line tool and some prompts. If you have external users, you need MCP, because that's what Cursor and the other agent apps support out of the box. So 'dead' is context-dependent. It's over-engineered for internal tooling, and it's the thing that gets you distribution if you want to be reachable by everyone else's agent.

24. Lenar Kess 00:09:06

    And mxstbr is honest enough to undercut himself in a footnote — he says the Codex app's computer and browser-use features have made even his own argument weaker, because the agent can just drive a browser instead of needing your API at all.

25. Damra Vol 00:09:20

    Which is the thread I'd pull on next. If browser-use gets reliable enough, the question stops being 'MCP or CLI' and becomes 'why expose an interface at all, when the agent can use the same website a human uses.' That's further out and less reliable today. But it's the version where the whole protocol debate gets reframed entirely. What I'm watching is whether companies keep standing up MCP servers next quarter, or start betting the agent will just click through their existing site.

26. Lenar Kess 00:09:50

    Let's get concrete and local. There's a write-up on the LocalLLaMA subreddit from someone running the handle FantasticNature7590, who spent a few weeks benchmarking multi-token prediction — MTP — on Gemma 4, the thirty-one-billion-parameter model, and Qwen 3.6 at twenty-seven billion. The headline result: on Gemma 4, inference jumped from about forty tokens per second to a hundred and thirty-two. Roughly three-and-a-third times faster, on a single RTX PRO 6000 card.

27. Damra Vol 00:10:21

    Quick translation of what multi-token prediction is doing, because it's clever. A small draft model guesses several tokens ahead, and the big model verifies them in one pass instead of generating one token at a time. On Gemma 4 the draft model is tiny — seventy-six million parameters. And here's the part that matters for trust: the target model still verifies every token before accepting it. So the output path is identical to normal decoding. In principle the quality shouldn't change at all.

28. Lenar Kess 00:10:53

    In principle. Did he check?

29. Damra Vol 00:10:55

No — and to his credit, he says so flatly. He calls the quality and the video-memory numbers, quote, 'directional observations, not benchmarked facts.' He ran out of time for a proper quality eval. That's the right way to post a benchmark. The architecture says quality should hold because of the verify step, but he isn't claiming he proved it. The other honest caveat: the speedup depends on your stack. vLLM won on Gemma at a hundred thirty-two tokens per second, but llama.cpp was actually solid on Qwen, around a hundred seventeen. It's not one tool winning everything.

30. Lenar Kess 00:11:34

    And llama.cpp had its own small moment this week — Georgi Gerganov posted that it finally has an official website, at llama.app. His line was, quote, 'Our goal is to make local AI accessible to everyone,' with a single-line installer on the landing page.

31. Damra Vol 00:11:52

    Which sounds minor and isn't. llama.cpp has been, for years, a clone-the-repo-and-build-it project. That's a wall for anyone who isn't already comfortable at a compiler. A real website with a one-line install is the difference between a tool for people who already know it exists and a tool a curious person can actually start on a Saturday.

32. Lenar Kess 00:12:15

    And one more from the same corner — antirez, Salvatore Sanfilippo, put his DwarfStar distributed-inference project on GitHub. The pitch is that you can run a two-bit quantized Flash model across two sixty-four-gigabyte machines, or a four-bit version across two machines with a hundred twenty-eight gigabytes each, with pipelining to speed up the prefill.

33. Damra Vol 00:12:37

    That's the 'I don't own one giant machine, I own two medium ones' path, and it's a real unlock for home setups. Stitch a couple of consumer boxes together instead of buying a server. Put all of this next to a point one developer in that thread keeps making — once roughly one in three teams is running open weights, that's the point where the ecosystem starts building tooling for open models instead of treating them as second-class. MTP landing in vLLM and llama.cpp, a real installer, distributed inference — that's the tooling layer maturing in real time. The thing I'd watch is whether someone runs the MTP quality eval the poster skipped, because the whole speed story rests on the verify step actually holding.

34. Lenar Kess 00:13:22

    Over in Europe, there are notes from the Mistral AI Now Summit — Koen van Gilst wrote them up, and they hit Hacker News at around four hundred points. The detail that jumped out, and Simon Willison flagged the same one, is concrete: BNP Paribas runs Mistral's models on-premises for

know-your-customer checks in Belgium, with the sensitive data staying inside the bank's walls. And Abanca is using agent orchestration on sensitive customer data across two million customers in their app.

35. Damra Vol 00:13:51

    And then the bear case, which is right there in the same thread and worth saying plainly. A commenter who says they're rooting for Mistral writes that the company has fallen really far behind since the third quarter of 2025. Their words — Mistral's 'small' model has roughly four times the parameter count, around a hundred and twenty billion, and isn't even competing with models a quarter its size. Gemma 4 and Qwen 3.6 own the small tier right now, and they're a fraction of the size.

36. Lenar Kess 00:14:20

    So there are two readings sitting on top of each other. One: Mistral isn't winning capability benchmarks. Two: the on-premises, European, regulated-industry position is a genuine business that doesn't actually depend on topping a leaderboard.

37. Damra Vol 00:14:35

    Both true, and here's the tension between them. If your hundred-twenty-billion-parameter model loses to a twenty-seven-billion open model on capability, your on-prem customers can run that smaller open model on-prem too. Data sovereignty isn't something only Mistral can offer — anyone can run Gemma or Qwen inside the bank's walls. So what Mistral is actually selling is the support contract, the orchestration layer, the compliance paperwork, and someone to call when the know-your-customer pipeline breaks at two in the morning. That's a real moat. It's just a different one than 'our model is best.'

38. Lenar Kess 00:15:12

    Different corner of the same week — xAI shipped grok-build-0.1 through their API in public beta. It's the model behind their Grok Build command-line tool, aimed at agentic coding, and the pricing is the headline: one dollar per million tokens of input, two dollars per million output.

39. Damra Vol 00:15:31

    That price is aggressive — it undercuts most of the frontier coding options by a wide margin. But notice what we don't have. 'Excels at agentic coding' is a vendor sentence with no eval attached. No SWE-bench number, no task result, nothing checkable. The cost is the only verified fact in the announcement. And cheap-and-available does beat good-on-paper sometimes — that's how a lot of adoption actually happens. But before I tell anyone it's good at coding, I'd want to see it on a real task, not a price tag.

40. Lenar Kess 00:16:04

    So put them side by side: a coding model that's cheap and unproven, against a French lab with proven customers and underwhelming models. Two different bets on what actually closes a deal.

41. Lenar Kess 00:16:15

    Last one, and it's about the part of AI that runs on actual electricity. Politico reports, via Techmeme, that the Federal Energy Regulatory Commission — FERC, the US energy regulator — is readying a proposal in June to speed up connecting data centers to regional power grids. And AI companies are engaging the regulator directly about it.

42. Damra Vol 00:16:37

    This is the constraint nobody puts in a launch video, and it might matter more than any model release this week. The bottleneck on the buildout isn't always money or chips — it's the interconnection queue. You can finance a data center, pour the concrete, install the racks, and still wait years to actually connect to the grid. If FERC shortens that, it's a concrete lever on how fast capacity comes online. A regulator with a June deadline is a real thing.

43. Lenar Kess 00:17:06

    And on the far other end of the credibility spectrum, there's a thread making the rounds about TerraFab — an account named Lacey Presley posting that Tesla's hundred-and-nineteen-billion-dollar semiconductor foundry in Texas is targeting one terawatt of annual AI compute capacity. Roughly fifty times current global chip output.

44. Damra Vol 00:17:26

    Fifty times global output is the number that should stop you cold. [tsk] That's a slogan, not a forecast. And I can't find a primary source for it — this is a hype account, not a Tesla filing, not an earnings call, not a press release. So I'd file the terawatt claim under aspiration posted as fact until there's a document with Tesla's name on it. The contrast with the FERC item is the point. One is duller and far more real, because it's a regulator with a deadline. The other is a big round number on social media.

45. Lenar Kess 00:17:58

    And that's the note to end on, going into Sunday. The capacity story isn't bottlenecked on whether someone hits a terawatt. It's bottlenecked on interconnection queues, permitting, and the slow problem of getting power to the building. What I'll be reading is the actual text of that June FERC proposal when it lands — not the foundry number. The next piece of real signal is in that filing.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic