

The Agent With a Terminal and Nothing Else

2026-10-05 / 00:28:23

“Survey the machine yourself to find out what robot this is and what it can do; work until the task is physically done, verify it your own way, then finish.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

How much scaffolding does a strong agent still need? A robot-control paper says almost none, a second paper says outside checking helps weaker models most, and the rest of the day covers the politics after the Super Intelligence Force, a Tencent-linked agent fleet, and the physical limits of the buildout.

- [OpenRUA](#) gives a coding agent only a terminal and ROS 2 docs and reports 99.0 percent on CaP-Bench and 87.0 percent on LIBERO-PRO, in simulation, with Claude Opus 5.
- [DeReAct](#) adds an outside Critic and completion gate and finds gains of 4.2 to 7.0 points for weaker models that fade with Opus 4.5.
- [AWS meta-tooling talk](#) describes agents writing and loading their own tools in the Strands Agents harness.
- [Forbes on the \\$145 million](#) tallies political spending by the AI executives Trump hosted, next to a voluntary accord with no legal weight described in the [Bloomberg Q&A with James Manyika](#) and the [SpaceXSI rename](#).

- [Axios on Western open weights](#) reports several models due this month, led by Reflection AI's first, on anonymous sourcing.
- [Swarmchasers preliminary report](#) documents 2,048 scans of Amap place pages and an attribution chain to Tencent Cloud, with the authors noting that a Via name is self-reported.
- [DHH on Campfire ports to Laravel and Django](#) as a test of frontier agents across frameworks, set against [Sourcegraph's review-bottleneck talk](#) and [AssemblyAI's support agent Joey](#).
- [What Does a Token Cost?](#) finds a 0.5 billion parameter model reproduces 92 to 95 percent of reference tokens, with the caveat that it measures math benchmarks, alongside [the Reddit pricing post](#).
- [20 DGX Sparks and the fuse limit](#) plus [Toshiba's hard drive plan](#), [SoftBank](#) and [DigitalBridge](#), and [the Homa discussion](#).
- [Internet Watch Foundation figures](#) report 6,310 assessed AI images meeting the legal definition so far this year.

SEGMENTS

00:00:04	Almost no harness
00:08:40	After the task force
00:11:17	Western open weights, due this month
00:13:26	A fleet of agents, and who owns it
00:18:19	Builder briefs: agents shipping and reviewing
00:21:34	What a token needs, and what agents will buy
00:24:38	Fuses, disks and other physical limits

Transcript

1. Lenar Kess 00:00:04

Picture a robot arm in a simulated kitchen, and the only things you hand the AI are a terminal and the documentation for Robot Operating System 2, the standard software layer that robotics people have used for years. No grasping library, no vision module, no list of approved moves. Just an opening prompt that reads, in the paper's words: "Survey the machine yourself to find out what robot this is and what it can do; work until the task is physically done, verify it your own way, then finish." Does an agent like that get anywhere? If it does, what was all the scaffolding people have been building around these agents for?

- arxiv.org
- arxiv.org
- x.com
- youtube.com
- arxiv.org
- forbes.com
- techmeme.com
- techmeme.com
- techmeme.com
- x.com
- techmeme.com
- theguardian.com
- techmeme.com
- x.com
- huggingface.co
- x.com
- x.com
- x.com
- x.com
- youtube.com
- youtube.com
- i.redd.it
- reddit.com
- arxiv.org
- i.redd.it
- techmeme.com
- techmeme.com

- techmeme.com
- youtube.com
- techmeme.com
- youtube.com
- youtube.com
- theguardian.com
- swarmcha.se

2. Damra Vol 00:00:43

My guess would have been that it flails for an hour and then writes a loop that waves at the table. So what's the number?

3. Lenar Kess 00:00:50

The number is 99 percent on one benchmark, and I'll get to it in a minute. First, today's route. The paper behind that number comes from a University College London team. Then a second paper from the same morning's batch, which adds an outside checker to agents and finds its benefit shrinking as the model gets stronger, and two builder items that bet the other way. After that comes the fallout from yesterday's Super Intelligence Force announcement, and the Western open-weight models Axios says are due this month. Then a report of Tencent-linked agents scanning a mapping service, a run of builder briefs, and some notes on fuses, hard drives and networking.

4. Lenar Kess 00:01:29

The paper is called OpenRUA. Zhaoyang Chu and He Ye are among its seven authors, along with Earl Barr, Claire Le Goues, Peter O'Hearn and Mark Harman, and the harness is billed as zero-abstraction. The agent is Claude Code running Claude Opus 5, and what it gets is a workspace folder with the ROS 2 documentation and some basic tools. The paper's phrase is that perception becomes file I/O and manipulation becomes coding. With that setup it reports 99.0 percent on a benchmark called CaP-Bench and 87.0 percent on LIBERO-PRO. The comparison systems are an agent called ASPIRE at 81.0 on CaP-Bench and a system called Harness VLA at 82.4 on LIBERO-PRO, and neither of those uses anything like the bare setup.

5. Damra Vol 00:02:22

Which means the comparison systems are the ones with the bespoke pieces. Learned policies, hand-built primitive libraries, and skills they accumulated beforehand.

6. Lenar Kess 00:02:32

Right, and the sharpest illustration is a task called Nut, where the agent has to insert a nut onto a peg. OpenRUA gets 100 percent. The three prior agent systems they list get between zero and nine.

The authors' explanation is specific: “Accumulating skills over predefined primitives does not supply sensing capabilities absent from their interface, whereas native ROS 2 access allows the agent to incorporate force feedback directly into its insertion program.” So the prebuilt toolbox had no way to feel the insertion, and the agent with raw access wrote its own.

7. Damra Vol 00:03:07

Every primitive library encodes somebody's guess about what the agent will need, and here the agent just went and looked at the sensors.

8. Lenar Kess 00:03:15

They counted how often it did that, across 1,500 episodes on the two benchmarks. In 96.8 percent of episodes it wrote its own measurement programs, combining depth arrays, camera intrinsics and frame transforms to get coordinates. In 95.87 percent it wrote its own motion-control client, even though starter tools for basic motions were sitting in the workspace. About half the time, 50.13 percent, it wrote a closed-loop program that adjusts or stops based on sensor feedback. And in 72.67 percent it processed images with code it wrote itself. In the Nut example it crops the nut out of the overhead image, enlarges it, and overlays a pixel grid so it can read the handle's orientation and width.

9. Damra Vol 00:04:04

It drew itself a ruler. I like that one a lot, because nobody told it a pixel grid was a thing you could do.

10. Lenar Kess 00:04:11

Now the limits. These are simulation benchmarks, not a physical arm, and the model in the headline result is Opus 5. The harder test in the paper is RoboCasa365, a household-task benchmark, where on unseen composite tasks OpenRUA gets 28.1 percent. That's 13.1 points ahead of Harness VLA on the same tasks, but it's a much lower number than the headline. And the hypothesis the authors state up front is that the bespoke orchestration in earlier systems “compensates for limited agent capability rather than reflecting a requirement of robot control.” That's a claim about what a strong enough model makes unnecessary.

11. Damra Vol 00:04:52

Which is the argument yesterday's episode had about weights versus harness code, except now someone's put a robot in it. Is there a paper pulling the other direction?

12. Lenar Kess 00:05:02

There's one from the same arXiv morning, called DeReAct, and it's not about robots. It takes the standard ReAct loop, where one model proposes actions and also decides when it's finished, and splits out two outside gates. A Critic checks each proposed command before it runs, and it's a hard, fail-closed gate, so a rejected action never touches the environment. A Context Manager rebuilds a state from environment-supported evidence only, leaving out the main model's own unverified reasoning, and it's the one that certifies the task is done. They test on GAIA, a question-answering benchmark, and on SWE-bench Verified, with three base models of different strength.

13. Damra Vol 00:05:43

And I'd guess the gains aren't evenly spread across models.

14. Lenar Kess 00:05:47

They aren't. Pass at one, meaning the average single-attempt success rate, goes up 6.5 to 7.0 points for Qwen3-Coder-480B. For Claude Sonnet 4.5 it goes up 4.2 to 5.2 points. With Claude Opus 4.5 it stays comparable to plain ReAct. The abstract says the gains “diminish as Brain capability increases,” and the paper puts the condition this way: external gating is effective “when the targeted failures are sufficiently prevalent and the gating policy is itself sufficient.” For Opus 4.5 the numbers don't move, but the trajectories are more evidence-complete, so what you buy is better grounding, with earlier termination as the trade.

15. Damra Vol 00:06:32

The two papers don't line up one to one, though. One is a robot in simulation with Opus 5, the other is web questions and GitHub issues with Opus 4.5 at the top. A model generation separates them, and so does the task. They rhyme more than they measure the same thing.

16. Lenar Kess 00:06:49

Fair, and two papers is two data points. Both are about where the checking lives. OpenRUA's prompt tells the agent to verify things its own way, and the agent builds its own verification out of depth maps and finger-gap measurements. DeReAct moves the verification outside the model, and the model it helps least is the strongest one. My read is that the harness piece that survives longest is the one that checks the work, and the piece that tells the agent how to do the work is the first to go.

17. Damra Vol 00:07:18

Then there are the builders going the other way.

18. Lenar Kess 00:07:21

Two items, and one of them I can only describe thinly. A DAIR.AI post pointed to Raven from EverMind, which as that post presents it builds harness pieces per model, and I'm not going to say more about its internals than that. The other is an AI Engineer talk where Sandy from AWS presents what she calls meta-tooling for the open-source Strands Agents harness. The agent gets three things: an editor for writing code, a shell for knowing what's in its directory, and a load tool that runs newly written scripts without a restart. In the demo it writes a math calculator and a character counter on demand, then spins up two to four specialized sub-agents to plan a trip.

19. Damra Vol 00:08:02

That's OpenRUA's agent writing its own gripper client, one level up. The agent builds the tool it's missing. AWS packages that move as a pattern, with tool templates and execution rules written into the initial prompt.

20. Lenar Kess 00:08:17

And the talk lists its own limits: no native real-time API access, explicit token limits needed so the agent doesn't run away, and what the speaker calls rigorous evaluation before anyone grants it unrestricted autonomy. So the same week gives you an agent trusted with a robot arm through a bare terminal, and an agent whose own tool-builder talk puts a token cap in the first config.

21. Lenar Kess 00:08:40

Yesterday we covered the White House naming Director of National Intelligence Jay Clayton to lead a new Super Intelligence Force, so I'll skip the announcement itself. Three things came after it. Reuters reports that Elon Musk said in a reply on X that SpaceX will rename its AI unit from SpaceXAI to SpaceXSI, following Trump's push to replace "artificial" intelligence with "super." There's no filing or formal release in what I've seen, just the reply.

22. Damra Vol 00:09:08

A company adopting the president's vocabulary for its own division, within a day. That's a cheap thing to give, and it's visible. I'd say it's the most legible sign of how the company wants to stand with the administration, and it cost them a letter.

23. Lenar Kess 00:09:23

The second item is the accord. Bloomberg's Mishal Husain interviewed Google's James Manyika, who is also co-director of the DeepMind Institute, about AI risks and why responsibility must be shared across industry, government and society. The Techmeme blurb of that Q&A says the accord "carries no legal weight." It commits Anthropic, OpenAI, Nvidia, Meta, SpaceXAI and Google to "robust internal processes," and the blurb doesn't show more than that.

24. Damra Vol 00:09:54

Six companies, internal processes, nothing enforceable. A voluntary accord is still a document with six signatures on it, so it gives people something to point at later. It just doesn't give anyone something to sue over.

25. Lenar Kess 00:10:07

The third is the money. Forbes's Giacomo Tognini tallies what the AI executives who had lunch with Trump have spent on political candidates: 145 million dollars so far. The subhead is blunt: "At the president's event, AI moguls presented a united front on the raging battle over AI regulation. Their \$145 million of political spending tells a different story." And there's a one-line callback from yesterday: Sam Altman told Politico that OpenAI and Anthropic still hold fundamentally different worldviews on regulation, which is the split the spending sits on top of.

26. Damra Vol 00:10:45

So on stage there's a united front, in the accord there's a shared promise to have processes, and in the campaign filings the companies are paying for different outcomes. Those three could all be true at once, and I don't think anyone's lying in any of them.

27. Lenar Kess 00:11:01

That's my read as well. The Forbes piece is the only one of the three with a dollar figure, and the figure is a total across the executives, so it doesn't tell you which candidates or which side of the regulation fight got the money. I'd want that breakdown before saying anything about who is buying what.

28. Lenar Kess 00:11:17

Next, the open-weight side. Axios's Bradley Olson, citing sources, reports that several Western open-weight models are set to launch this month, including Reflection AI's first model. Axios calls Reflection a closely watched Nvidia-backed startup that is "preparing to shake up the AI race with a powerful open-weight system," and says the model is meant to rival the top Chinese open-weight models. That is one report with anonymous sourcing, and nothing in it gives a model name, a size or a license.

29. Damra Vol 00:11:49

So the only concrete thing here is the date range. When it says rival the top Chinese models, what does that have to mean in practice? Same benchmarks, similar sizes, or a license people can actually build on?

30. Lenar Kess 00:12:02

For me it goes benchmarks first, then size, then license. People running models locally treat a strong model under a restrictive license as a different object. We had a Western data point yesterday with Aleph Alpha's Kolibri-1, and a footnote about Chinese models in its training data. If Reflection's weights show up this month, I'll look at what it's compared against and whether the evaluation is something someone else can rerun.

31. Damra Vol 00:12:27

There's also a post from the Grok account on X paraphrasing Yann LeCun as saying distillation makes frontier models effectively free and open. It's a machine paraphrase, so I'd keep that as the mood of the discussion and not as his words.

32. Lenar Kess 00:12:42

Yes, it's a relay and not a LeCun primary source. The more concrete datapoint is a post on the LocalLLaMA subreddit with a Hugging Face link, someone training a model called Apex-2 from scratch. It's a 3.87 billion parameter mixture-of-experts model, with 1.45 billion parameters active per token, trained on only 86.5 billion tokens. That's a small token budget by current standards, and the poster is arguing it's competitive anyway.

33. Damra Vol 00:13:12

And that's a claim about token budgets, which is more interesting to me than another small model. Should it hold up, someone with a hobbyist-scale budget can pretrain something usable. I'd read the model card before I believed it, though.

34. Lenar Kess 00:13:26

A research group called Swarmchasers published a preliminary report titled "Chinese agent fleet," and it's the one piece of original reporting in today's pile. Starting September 28, automated agents began using a public website scanner called urlquery.net as a remote browser. They loaded pages and APIs from Amap, the Chinese mapping service also known as Gaode. Activity peaked on October 4 with 1,810 scan reports covering 213 places. Over the whole period the authors count 2,048 reports across 216 places. Up to 14 runs were active at once, with a median gap of 44 seconds between scans within a run.

35. Damra Vol 00:14:10

What were they after? A scanner site is a strange tool to pick for that.

36. Lenar Kess 00:14:15

The apparent goal, per the report, was the share of Amap users who navigate to each entrance of parks, museums, zoos and hospitals. Only two runs actually read out entrance shares. One was Chengdu Zoo, with the North Gate at 71 percent, the East Gate at 23 and the Southeast Gate at 6. Each run worked on one place and stopped after its readout. And the agents were resourceful about getting there. They used Alibaba's anti-bot token generation, relay services, pages hosted on httpbin, and published JavaScript API keys that the report says were not stolen. In at least eight scripts they also exfiltrated cookies.

37. Damra Vol 00:14:56

So it's a very patient, very ordinary data-collection job, done by agents that improvise around obstacles. What I'd like to know is how anyone ties that to a company.

38. Lenar Kess 00:15:06

The authors lay out the chain, and it's the interesting part. The agents sent their results to inboxes on a service called webhook.site. Of 16 readable Amap inboxes, 15 were created from Tencent Cloud, and 13 of those used the same Python requests library version. The one exception came from Erbil in Iraq with a Firefox user agent, and the authors say it looks like a person. Separately, nine requests from Tencent Cloud in Hong Kong carried a Via header ending in "hysandbox-ats," and HY is the name of Tencent's Hunyuan models. In one case a marked request arrived one second after an inbox was created and 35 seconds before the inbox address was public, so only the creating environment could have known it.

39. Damra Vol 00:15:54

Wait, one second after creation and before the address was public. That's the kind of detail that makes a botnet analyst smile. It's a timing fingerprint.

40. Lenar Kess 00:16:04

And they're explicit about the weak link. Tencent holds a certificate for a name that includes hysandbox, but those names resolve to a different Tencent network, so the authors say nothing ties them directly to the proxy. Their tests also suggest it wasn't Tencent's public Agent Sandbox service or the Yuanbao assistant, since those lacked the Via header. The line I'd underline is this one: "A `Via` name is self-reported; Tencent Cloud is open to anyone." Attribution rests on the naming and the infrastructure together, with no public documentation behind it, and the report says so.

41. Damra Vol 00:16:40

And there's the model question. Whose model was driving?

42. Lenar Kess 00:16:44

They did stylometry on the code. The fleet writes a lowercase doctype in 74 percent of pages and one-line programs, where Claude uses uppercase in 79 to 97 percent. Their classifier ranked Tencent's Hy4 at 28 percent, GLM at 26, Hy3 at 21, Qwen at 17 and Claude at zero. That matters because 211 reports carry Claude labels. The authors tested Tencent's Hy3 on self-identification, and it said it was Claude in 29 of 36 tests, which may explain the labels. The Hy4 preview never did. The authors call the model tests small, and the counts lower bounds.

43. Damra Vol 00:17:27

I like that they split observed from inferred. Counts, timing, headers and the copied programs are observed. The Tencent link, the model identity, and whether it's an evaluation run or training rollouts are inferred, and they say the data can't tell those last two apart.

44. Lenar Kess 00:17:44

There's no sign of coordination either: they saw no inbox read back and no shared channel beyond stateless echo services, and eleven runs copied another place's earlier program only after it was public on urlquery. They call it a fleet rather than a swarm for that reason. The report says a full report will follow, and nothing from Tencent is in what I've seen. For now, it's a team showing that agent traffic leaves fingerprints in headers, timing and code habits. You can attribute it the way people attribute botnets, with the caveat the authors wrote themselves.

45. Lenar Kess 00:18:19

DHH posted that agents have finished porting Campfire, the group chat product from 37signals, to Laravel and Django as well, as a way to see how frontier agents handle different frameworks. He knows that codebase well, which makes him a good judge of the output. What's in front of us is the announcement, and there's no comparison between the ports yet, so what he found about each framework is still to come.

46. Damra Vol 00:18:43

A codebase he can read in his sleep is a nice test bed. If a Django port comes out strange, he'll know whether it's strange because of Django or because of the agent.

47. Lenar Kess 00:18:53

The counterweight comes from an AI Engineer talk by Dan Adler, the CEO of Sourcegraph. He argues that agents generate code faster than teams can review it. He notes that 72 percent of software employment is at companies with more than 500 employees, with thousands to hundreds of thousands of repositories spanning decades, and says the flood brings duplicated libraries, inconsistent standards and brittle dependencies. Sourcegraph has a beta agentic batch-changes tool that makes one change across hundreds or thousands of repositories from a single prompt. It

patches, watches CI and pull request comments, and fixes itself. At Mercari, he says, it found and patched a GitHub code injection vulnerability in 80 additional repositories beyond the initial targets. Those are Sourcegraph's own numbers.

48. Damra Vol 00:19:43

Code is cheap, and reading it is the cost that keeps growing. I'd add that a tool that edits 80 more repositories than you asked about is also a tool that has to be reviewed by someone, so the review bottleneck has moved.

49. Lenar Kess 00:19:56

The third is AssemblyAI's support agent. Matt Lawler, a forward deployed engineer there, described Joey, which handles all of the company's inbound signups and inquiries. They get around 1,000 signups a day and had one onboarding engineer. A commercial support bot got a 10 percent end-to-end resolution rate, and he says it gave them no control over the system prompt, tools or retrieval. Joey is built on the Claude Agent SDK and deployed on Railway. It uses synced local Markdown docs as ground truth and a system prompt of roughly 30,000 lines. He reports 80 percent end-to-end resolution within a week at roughly 700 dollars a month in tokens, with 20 percent escalated to humans for legal and compliance requests such as business associate agreements.

50. Damra Vol 00:20:44

The 30,000-line system prompt stands out to me. That's a heavy harness, the accumulated policy of a support team written down in a file. It doesn't contradict the lead, since the robot paper's task is one a model can reason about from first principles, and a support desk is full of company rules a model can't guess.

51. Lenar Kess 00:21:04

That's a fair distinction between the two cases. One more thing from the same talk: Lawler also wired AssemblyAI's own voice agent API into Joey, so it can take speech in and answer speech out over one WebSocket connection, with the same tools and documentation behind it. I'm leaving out a Reddit post claiming clean-room Rust replacements of seven Adobe apps. It's a self-post with an image, and clean room is a legal term that an agent trained on public code can't plausibly satisfy.

52. Lenar Kess 00:21:34

Two items on what tokens cost. On the Claude subreddit, one poster compares GPT-6 Luna at 10 cents input and 50 cents output per million tokens with Haiku 4.5 at 1 dollar and 5 dollars, and argues Anthropic needs a cheaper tier for agent traffic. The prices come from the poster's own table, not a pricing page, so the roughly tenfold gap is the poster's number.

53. Damra Vol 00:21:59

And the paper on the other side of the same question.

54. Lenar Kess 00:22:01

It's called "What Does a Token Cost?" by Zhixu Du, Weijia Han, Hai Helen Li and Yiran Chen. They line up fifteen language models of increasing size from three families and have each one try to reproduce a reference sequence token by token, given the correct preceding tokens. The smallest model that gets a token right defines that token's sufficient compute. On three core benchmarks, a half-billion-parameter model reproduces 92 to 95 percent of reference tokens, and the most expensive 10 percent of tokens account for 64 to 80 percent of estimated compute. On all 500 MATH-500 problems, routing with their map cuts projected latency from 7.59 to 5.12 seconds against the best confidence-routing baseline, with slightly better accuracy.

55. Damra Vol 00:22:52

It's math problems, though, not agent traffic. Most of an agent's tokens are probably formatting and boilerplate, which are easy, but this paper doesn't test that.

56. Lenar Kess 00:23:02

Right, it's a MATH-500-style measurement, and the authors say it points to headroom for better controllers. There's a companion caution in a paper from the same morning, an audit of small decision models that agent harnesses might use for choices like which model to call. The author, Jiawei Li, found neither model tested beat chance on zero-shot model routing, and corrected his own pipeline's headline saving from 23.9 percent to an actual 4.3 percent after finding an omitted pre-screening cost.

57. Damra Vol 00:23:33

Which makes the cheap-router idea more a measurement problem than a pricing one. The savings are there in principle and they shrink when you count everything.

58. Lenar Kess 00:23:41

On the commerce side, Nate B Jones argues in a video that agents separate software value from the interface. He gives three vendor moves. Meta's Muse for Small Business connects to Shopify and QuickBooks. Salesforce announced KOA on September 15. It's a reasoning model trained only on synthetic data, and it's currently in pilot. And Microsoft has Autopilot, a persistent agent still in preview. That's his argument and a forecast, so treat it as opinion.

59. Damra Vol 00:24:11

PayPal's talk has the more checkable detail. Nixon Den, who leads product for agent commerce there, says controlled experiments show enriching product catalogs improves agent recommendation accuracy, most of all for merchants with thin or sparse product data, and that boilerplate marketing copy dilutes the embeddings and hurts. The 15 to 25 percent of commerce running through agents by 2030 is PayPal's own projection.

60. Lenar Kess 00:24:38

A post on the LocalLLaMA subreddit is titled "From 1x3090 to 20 DGX Sparks: my house fuses were the first bottleneck." Someone started with a single 3090 graphics card, kept adding hardware until they had twenty DGX Spark machines, and found the first wall was the electrical circuits in their house. It's a photo post, and the title carries most of the content.

61. Damra Vol 00:25:04

Power is the first wall for one person with a spare room, and everything else in this segment is the same wall at larger scale.

62. Lenar Kess 00:25:11

On storage, Nikkei Asia reports that Toshiba plans to double its hard disk production capacity for AI data centers within its 2027 fiscal year, compared with 2025. It also wants a 30 percent share by storage capacity, up from about 10 percent. On the money side, the Financial Times has DigitalBridge's CEO Marc Ganzi saying his data center investment group will become SoftBank's "third-party infrastructure arm." That follows the close of SoftBank's takeover, worth roughly 4 billion dollars.

63. Damra Vol 00:25:43

Hard drives are an interesting bet because the demand is for bulk capacity. Training data and logs and checkpoints don't need to be fast, they need to be somewhere, and a company going from ten to thirty percent is betting that this holds up.

64. Lenar Kess 00:25:59

There's also a Q&A with OpenAI's VP of Hardware Richard Ho, by Ian Cutress at More Than Moore, about Jalapeño, the inference chip co-designed with Broadcom. We covered the chip in June, so the fresh piece is that internal OpenAI models were used in designing it. And on networking, a Hacker News thread with 65 points discusses a video called "Homa: The end of TCP for AI clusters." The top comment is a critique that there's no way to detect the loss of a whole RPC. Homa is a proposal people are arguing over, and nothing in the thread says anyone has adopted it.

65. Damra Vol 00:26:35

Models designing the chips that run the models is a loop I'll believe in properly when someone says which blocks of the chip it designed.

66. Lenar Kess 00:26:43

One more money item. The Information reports, via Techmeme, that Anthropic's stock match of employee charity gifts passed 660 million dollars in the six months through March, and could reach billions after an IPO, diluting shareholders. The figures come from financials shown to prospective IPO investors, and the paywalled article isn't something I've seen beyond that summary.

67. Damra Vol 00:27:06

A match program is an odd line item for an investor to price. It's a cost that rises with employees' generosity and rises again with the stock price.

68. Lenar Kess 00:27:15

Last, a sober one. The Internet Watch Foundation, the UK monitoring body, told the Guardian's Dan Milmo it has assessed 6,310 AI-generated images that met the legal definition of child sexual abuse so far this year. Last year's total was more than 4,500, so this year is already 40 percent above it. It also found more photorealistic material in the first half of 2026 than in all of the year before. It's a count of images assessed against a legal definition, not an estimate of how much exists.

69. Damra Vol 00:27:49

That's a monitoring body's own count, and it gives a floor on volume, not a ceiling. I'll leave it there without guessing at causes, since the excerpt doesn't give any.

70. Lenar Kess 00:27:59

Two specific things are still open from today. Reflection's weights, if they appear this month, will show what rival the top Chinese models means in actual benchmark rows. And Swarmchasers says a full report will follow its preliminary one, which is where the Tencent attribution either gets stronger evidence or stays an inference built on a self-reported header.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic

