

OpenAI's DevDay Pitch: Cloud Agents Your Whole Team Shares

2026-10-08 / 00:31:54

“Ten people can have a shared view of an agent's work and each assume one of the other nine is looking.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

OpenAI's DevDay sessions describe agents that live in the cloud and belong to a team instead of one person, and the open issue is who can edit, trigger and review them. Also today: a Senate bill that would make data centers pay for the grid, withdrawals and a credit dispute around OpenAI's math results, new attribution for the South Korean bank hack, the White House science summit, and a Brief.

- [OpenAI DevDay: Codex Cloud session](#) Cloud environments with filesystem snapshots, plus Tailscale and OIDC access to private infrastructure; it matters because a shared agent can now reach into your network.
- [OpenAI DevDay: Dots and Spaces session](#) A Dot is a persistent cloud agent with its own browser and file system, and the speaker lists team Dots as future work; the permissions model for several people sharing one is left open.
- [OpenAI DevDay: OpenClaw session](#) Shared team sessions make prompts and decisions visible alongside the code, which changes what review looks like.
- [OpenAI DevDay: ChatGPT Work session](#) Page instructions work like a shared config file and results queue for human review before publication.

- Techmeme on OpenAI's GPT-6 tiering GPT-6 Sol serves paid ChatGPT tiers and Luna serves Free and Go; the post says nothing more about the models.
- AI Engineer talk on Strands context handling Keep a pointer in agent state and fetch bulk data on demand, the opposite default from one shared workspace.
- Axios on the Senate data center bill Data centers would pay existing and incremental transmission costs, and new sites of 20 megawatts or more would bear incremental system costs; the bill is a proposal only.
- Washington Post via Techmeme on the Fed September meeting officials cited AI investment in discussing inflation.
- Hacker News discussion of OpenAI's withdrawn math results Three results reported withdrawn; the thread title is the extent of what we have.
- doomslide on X A prior-work allegation, treated here as an allegation.
- Susan Zhang on X Second post in the credit dispute.
- Axios on OpenAI's math results Why verifiability makes math the next field, with Stephen Wolfram's skepticism.
- CrowdStrike analysis via Techmeme The South Korean bank attacker is likely Chinese-speaking and financially motivated, and used LLMs and ARTEX.
- LocalLLaMA post on the attacker's model stack Names five tools and models; unverified.
- Forbes on ransomware and autonomous AI A contributor's argument that defenders need AI too.
- Axios on the Genesis summit More than \$1B in industry commitments and a southeastern compute hub, as announced.
- Reuters via Techmeme on Biohub \$300M from Meta, DeepMind and Isomorphic and \$500M+ from the US for open biology datasets.
- Bloomberg via Techmeme on Isomorphic Labs Early talks at a valuation of at least \$40B.

- [Claude Haiku 5.5 announcement](#) Anthropic's own description of the model.
- [AINews on Haiku 5.5](#) Tiered pricing and the subagent positioning.
- [Indian Express on Haiku 5.5](#) The third model in a month ahead of a planned IPO.
- [WSJ via Techmeme on the fired researchers' letter](#) Their ask on AI monitoring; one side of a firing dispute.
- [Guardian column on Sam Altman](#) Altman's remark on accepting some bad things for the technology's benefits, with the columnist's argument.
- [WSJ via Techmeme on Broadcom financing](#) More than \$50B in financing for OpenAI's custom chip, reported from sources.
- [FT via Techmeme on SemiAnalysis](#) China 24 GW operating with 50 GW planned, versus 56 GW in the US.
- [CNBC on Manus](#) \$500M+ raise after Beijing forced Meta to unwind its acquisition.
- [Politico via Techmeme on Ecosia](#) Dropping Mistral for open-source models, some Chinese.
- [Forbes on Sarvam](#) A new hire to sell sovereign AI, now carrying DeepSeek and Zhipu models.

SEGMENTS

[00:00:04](#) Persistent multiplayer agents at DevDay

[00:05:42](#) What shared agents change

[00:12:04](#) The Senate data center bill

[00:17:28](#) Math withdrawals and a scooping claim

[00:21:28](#) Who hit the South Korean banks

[00:24:15](#) Genesis summit and open biology data

[00:27:04](#) The Brief: Haiku 5.5, a letter, financing, and three moves abroad

Transcript

1. Lenar Kess 00:00:04

Picture a competitive research team of two people. One of them, a man named Adrian, has packaged his method into a plugin. Now an agent reads newsletters and podcasts and goes through ninety days of sales call transcripts. Then it fills in a public dashboard called Compete Corner. It reads Slack messages to decide what to investigate next, and it texts him when it needs a human to sign off.

- [youtube.com](https://www.youtube.com)
- [techmeme.com](https://www.techmeme.com)
- [youtube.com](https://www.youtube.com)
- [youtube.com](https://www.youtube.com)
- [youtube.com](https://www.youtube.com)
- [youtube.com](https://www.youtube.com)
- [axios.com](https://www.axios.com)
- [techmeme.com](https://www.techmeme.com)
- [twitter.com](https://www.twitter.com)
- [x.com](https://www.x.com)
- [x.com](https://www.x.com)
- [axios.com](https://www.axios.com)
- [techmeme.com](https://www.techmeme.com)
- [reddit.com](https://www.reddit.com)
- [forbes.com](https://www.forbes.com)
- [axios.com](https://www.axios.com)
- [techmeme.com](https://www.techmeme.com)
- [techmeme.com](https://www.techmeme.com)
- [v.redd.it](https://www.v.redd.it)
- [indianexpress.com](https://www.indianexpress.com)
- [latent.space](https://www.latent.space)
- [techmeme.com](https://www.techmeme.com)
- [theguardian.com](https://www.theguardian.com)
- [techmeme.com](https://www.techmeme.com)
- [techmeme.com](https://www.techmeme.com)
- [cnbc.com](https://www.cnbc.com)
- [techmeme.com](https://www.techmeme.com)
- [forbes.com](https://www.forbes.com)

2. Damra Vol 00:00:28

That one isn't a startup pitch. It's OpenAI employees demoing how OpenAI's own people use their own product, which is a different kind of claim than a launch post.

3. Lenar Kess 00:00:38

Right, it comes out of the DevDay session videos that went up yesterday evening. There were six of them, and they all describe the same change: the unit of work stops being one person's chat window and becomes a long-running agent that a whole team shares.

4. Damra Vol 00:00:53

So who's it sharing with, and where does it live? Because "my agent runs on my laptop" and "our agent runs in the cloud and has its own browser" are two different security conversations.

5. Lenar Kess 00:01:05

We'll spend most of the first half there. Today is Thursday, October eighth. We start with those DevDay sessions, what they show and who made the demos. Then a Senate bill that would make data centers pay for grid upgrades. After that, the math story from yesterday, which has two new wrinkles, withdrawals and a credit dispute. Then new attribution on the South Korean bank attacks, and the science summit at the White House today. The Brief covers Haiku 5.5 and a letter from fired OpenAI researchers. It also has some financing numbers and three moves outside the US.

6. Damra Vol 00:01:40

One caution before we start. Everything in the DevDay segment is vendor demo, people from OpenAI showing OpenAI's product. Nobody independent has tested any of it that I can find.

7. Lenar Kess 00:01:52

Agreed, and we'll say so as we go. Let's start with the cloud piece. In the first session, a developer experience engineer named Dominic says OpenAI shipped around three hundred features over the past year, and that the work has moved toward persistent, multi-agent workflows. His pitch is that long tasks shouldn't have to run on your laptop or on a rented virtual server you babysit. So Codex Cloud runs them for you.

8. Damra Vol 00:02:18

What does a Codex Cloud environment contain? Is that just a container with my repo in it?

9. Lenar Kess 00:02:23

It's more automated than that. According to the session summary, the environment auto-discovers dependencies, generates skills that configure the agent, and writes start scripts so iteration is quick.

It also uses filesystem snapshots to speed up task startup, and it runs background updates when the repository changes. You can set network rules, secrets and environment variables, and you can share the whole environment with your team.

10. Damra Vol 00:02:48

The snapshot detail is the one I'd have asked about. Starting a coding task usually means waiting for installs, so a snapshot is the difference between an agent that feels instant and one you forget about.

11. Lenar Kess 00:03:00

And the private-infrastructure part is the demo I found most interesting. The session shows cloud tasks reaching a Raspberry Pi API on a local network, using Tailscale for the VPN and OIDC for authentication. So the agent runs in OpenAI's cloud, but it can reach into your home or office network through a tunnel you've set up.

12. Damra Vol 00:03:21

A cloud agent with a tunnel into my network, and a Raspberry Pi as the example. A Pi is the smallest possible target, which is fair for a demo, but I'd want to know what a bad prompt can reach when the target is a production database instead.

13. Lenar Kess 00:03:36

The session doesn't get into that, as far as the summary goes. It talks about access controls on the environment, and nothing about what happens when one of the shared users is the problem. The second piece is the one about teams.

14. Damra Vol 00:03:48

Dots, then.

15. Lenar Kess 00:03:49

Dots and Spaces are the two names to know. In a talk from Jason on OpenAI's developer experience team, a Dot is a persistent, proactive cloud agent. It has its own isolated compute, its own browser, and its own file system. He credits research advances in memory compaction and longer runtimes for making that possible. A Space is the shared page where humans and agents work on the same documents.

16. Damra Vol 00:04:14

And the shared page is where the multiplayer part comes from. Someone tags the Dot in Slack or Teams, it follows up, it tracks what was said across platforms, and it does outreach or collects

feedback.

17. Lenar Kess 00:04:27

There's a stat in that talk I keep turning over. He says that something called app shots, where a user captures the full context of an application and its metadata, starts roughly thirty percent of Codex sessions. That's a vendor-reported number from one speaker, so take it as that, but it says the way people begin a task is changing from typing a prompt to pointing at an app.

18. Damra Vol 00:04:48

Pointing at an app is also how you hand a task to a coworker. So the agent is being treated like a colleague you send a screenshot to.

19. Lenar Kess 00:04:56

He also says that when you grant permission, a Dot can remotely control your local machine, and he calls it, quote, "remote control on steroids." Voice comes up too, with dictation described as three to four times faster than typing.

20. Damra Vol 00:05:11

Okay, but remote control of my local machine by a cloud agent that a whole team can tag in Slack. Who's the "you" that granted the permission?

21. Lenar Kess 00:05:20

That's exactly the open question, and I don't think the sessions answer it. The speaker describes it as individual: you give the Dot permission. The multiplayer version of that, where five people can tag the same Dot, isn't spelled out in the session summaries. Jason does list team Dots, cross-device continuity and standardized context sharing as future work.

22. Lenar Kess 00:05:42

The third session is the one that reads most like a field report. The speaker says model improvements made command-line agent workflows feel obsolete, and that he didn't want to become a, quote, "meat proxy," meaning the human who ferries prompts between agents. He'd been running Codex Desktop across six to ten machines, using what he calls DevBoxes and Grabboxes for cloud testing. Then he moved to OpenClaw.

23. Damra Vol 00:06:07

OpenClaw being the open-source persistent-agent project that has its own foundation.

24. Lenar Kess 00:06:13

Right, the OpenClaw Foundation governs it, and he says OpenAI supports it. The pitch is model agnosticism, full control over your own infrastructure, and open-source extensibility. The piece I found most concrete is about shared sessions. On a team, prompts, decisions and agent output are visible to everyone, and he says that replaces isolated pull requests with shared context, which reduces conflicts and speeds things up.

25. Damra Vol 00:06:40

That one is a real claim about code review. If I can read the conversation that produced a diff, I'm reviewing the intent and not just the diff.

26. Lenar Kess 00:06:49

He makes that point himself. A single prompt can produce tens of thousands of lines of code, and he says visibility into the conversation matters because the conversation is where the rationale lives. His team also coined a word, jellyware, for agent-generated artifacts like dashboards and plugins that get updated by prompting instead of through commits.

27. Damra Vol 00:07:10

Jellyware is a great word for something that never sets. No release, no version, just a dashboard that someone keeps reshaping in a chat.

28. Lenar Kess 00:07:19

And the automation he describes is what makes the team angle matter. Monitoring loops scan Discord, Twitter and GitHub and categorize what they find. They spin up isolated DevBoxes to reproduce reported bugs, and they land fixes automatically. He says continuous integration is the bottleneck now, so they're experimenting with picking subsets of tests. Meeting summaries run on Luna, and a voice product called GPT Live talks to the server's shared identity and session state.

29. Damra Vol 00:07:48

Auto-landing fixes from a loop that read a tweet. I'd want the revert rate on that.

30. Lenar Kess 00:07:54

There's no revert rate in what we have. What he does say is that human taste stays essential for architectural judgment, and that's why the loops moved onto shared team servers, so that everyone can watch what they're doing and learn from it.

31. Damra Vol 00:08:07

That's the claim I'd push on. Visibility isn't the same as review. Ten people can have a shared view of an agent's work and each assume one of the other nine is looking.

32. Lenar Kess 00:08:17

That's a fair worry, and it's an old one in human teams too. The fourth session is ChatGPT Work, from Nick and Danielle, and it speaks to this exact point. Adrian's workflow has a human review step. The agent's results queue up for him before anything goes onto the public Compete Corner dashboard.

33. Damra Vol 00:08:35

And the detail I liked there is page instructions. A Space has instructions that work like a markdown config file, so the page tells the agent how to parse and organize what comes in. It's an agents-dot-md file, except it lives in a document your whole team edits.

34. Lenar Kess 00:08:52

Which means the configuration is multiplayer too. Anyone with edit rights on the page can change how the agent behaves for everybody. The session also describes how a Dot is set up, with responsibility prompts, contact routing, and preferences it learns as you use it, plus pings by Slack or text when it wants a person.

35. Damra Vol 00:09:11

So the permission surface isn't just who can talk to the agent. It's who can edit the instructions, who gets pinged, and whose contacts it's allowed to route to. That's three or four separate controls and I didn't hear any of them named as a product feature.

36. Lenar Kess 00:09:27

The Codex Cloud session does mention that Slack and Teams conversations with the ChatGPT bot share cloud environments but have no local access. So there's at least a line drawn between the shared agent and your machine. The Dot that can remote-control your laptop is the other side of that line.

37. Damra Vol 00:09:44

Two products with two different trust models, sold under one brand.

38. Lenar Kess 00:09:48

There's one more piece of OpenAI news from the same evening. OpenAI says GPT-6 in ChatGPT's Chat tab is served by GPT-6 Sol on the Plus, Pro, Business and Enterprise tiers, and by GPT-6 Luna

on Free and Go. It rolled out globally to paid users with Free and Go following. That's all the announcement says about the models, so I'm leaving it at the tiering.

39. Damra Vol 00:10:13

Which is still interesting for the sessions, because the speaker in the OpenClaw talk says meeting summaries run on Luna. The cheap tier is doing the always-on chores.

40. Lenar Kess 00:10:23

Now the counterweight. Among the AI Engineer talks posted this week is one on how the open-source Strands framework, which defaults to Amazon Bedrock, handles growing context. The speaker's argument is that stuffing everything into one session history makes agents hallucinate and slow down, partly because models attend heavily to the start and end of a window and ignore the middle.

41. Damra Vol 00:10:47

So the opposite instinct from a giant shared workspace.

42. Lenar Kess 00:10:51

In the sense that the default is to share less. For big tool outputs like logs or database queries, the recommended pattern is to write the data to storage, keep only a pointer ID in the agent's state, and fetch it later with a separate tool call. In swarms of agents, full context sharing hurts performance, so you pass a minimal pointer instead. He also says to cap tool calls, with three per tool as an example, to stop infinite loops.

43. Damra Vol 00:11:18

That's the tension, then. OpenAI's sessions say one source of truth, a Space that everybody and every agent edits. The Strands talk says each agent should hold a pointer and fetch what it needs. Both can be right at different sizes.

44. Lenar Kess 00:11:33

My read is that the shared Space works while the Space is small enough for a person to read, like Adrian's dashboard. Once the context outgrows a person, the pointer pattern is what keeps the agent usable. Neither side has published a test of that boundary.

45. Damra Vol 00:11:48

I'd add that the OpenAI folks have the incentive to make the shared Space the center. If the team's working memory lives in your product, leaving is expensive. Strands is open source and works with

other models. Those are different bets on who owns the context.

46. Lenar Kess 00:12:04

Axios reports that bipartisan Senate legislation, unveiled last week, would make data centers pay for the grid upgrades they cause, and it's riding inside a larger bill meant to speed up permitting for energy projects. Amy Harder's piece quotes Ari Peskoe, who runs the Electricity Law Initiative at Harvard Law School, saying of the proposed costs and restrictions, "I don't think we've seen it before." Then he adds, "But I don't think we've seen growth that's comparable to data centers today, and I certainly don't know any industry whose electricity usage has captured so much political attention."

47. Damra Vol 00:12:38

That second sentence is the whole political situation. It's a law professor saying nobody has built a rulebook for this because nobody has had to.

48. Lenar Kess 00:12:47

We covered the local backlash on Tuesday, towns and utilities and deals. This is the federal piece, and the mechanics are specific. The first one Axios calls the big one. Data centers would pay both their share of the existing transmission network and the incremental transmission cost of serving them, which reverses a decades-old federal pricing policy that generally prevents charging both.

49. Damra Vol 00:13:10

So a normal customer pays for the grid as it exists. A data center would pay for the existing grid and also for the new lines built because of it.

50. Lenar Kess 00:13:19

Yes. Second, new data centers of at least twenty megawatts would bear the incremental costs they impose across the whole system, meaning generation, storage, transmission and distribution. Regulators at the federal and state level could charge even more than those incremental costs and use the excess to reduce other customers' bills. Data centers would stay on the hook for infrastructure costs even if they stop buying electricity before those costs are recovered. And utilities would have to secure financial guarantees before building anything to serve them.

51. Damra Vol 00:13:51

The guarantee piece is the one that bites. Right now a utility can build for a customer who then never shows up, and the ratepayers eat it. Making the project post financial security first changes whether speculative sites get built at all.

52. Lenar Kess 00:14:05

Then there's what it gives states. The bill would explicitly let states go further, including treating data centers less favorably than other large industrial customers, allocating access to electricity competitively, and requiring them to bring their own new power or accept limits on use. Axios notes that dozens of policies are already pending or in place across more than half the states.

53. Damra Vol 00:14:28

Treating them worse than a steel mill is the clause the industry will fight. That's the clause that gets a lawyer out of bed.

54. Lenar Kess 00:14:34

It is. One tech industry official, speaking anonymously, called some provisions an “unprecedented level of discriminatory treatment” for a single industry. The same person said tougher grid-connection requirements could push some developers to build power systems completely off-grid, which would undercut the purpose of the bill. On the record, the Data Center Coalition says it's still reviewing, Microsoft, Google and Nvidia declined to comment, and Meta didn't respond.

55. Damra Vol 00:15:03

Amazon did say something, though.

56. Lenar Kess 00:15:05

Craig Sundstrom, Amazon's energy policy director, said the aging grid and slow permitting are pushing up electricity costs, and that the bill, quote, “could unlock the investment needed to build new generation and transmission faster.” That's the trade the bill offers. You get faster approvals for power plants and transmission lines, and in exchange you pay for what you use.

57. Damra Vol 00:15:27

Which is why Jane Flegal's line is so useful. She was a Biden White House official and she's now at the Searchlight Institute. She told Axios, “Just to be candid, having some teeth on data center accountability and ratepayer protection is just table stakes to getting a deal done.” She'd be, her words, “incredibly surprised” if tech companies opposed the broader package in the end.

58. Lenar Kess 00:15:52

And from the other direction, Travis Fisher, a former Trump White House energy official now at the Cato Institute, has, in his words, “mild support,” even with concerns about a discriminatory precedent. He says the provisions break new ground at the Federal Energy Regulatory Commission

but fit the White House's ratepayer protection pledge, which many of the biggest data center companies have already signed.

59. Damra Vol 00:16:16

So the people who'd normally disagree are circling the same compromise. Faster wires, and the data center pays.

60. Lenar Kess 00:16:23

It's still a proposal. The Senate is out of town ahead of the midterms, so Axios says there's time for changes and a possible lame-duck debate. Nothing here is law.

61. Damra Vol 00:16:32

There's a macro angle that arrived this morning too. Techmeme points to a Washington Post story by Steve Thompson on the Federal Reserve's September meeting. Officials are increasingly citing what the minutes call "surging AI-related investments," not tariffs, as a driver of rising goods inflation.

62. Lenar Kess 00:16:50

That's the Techmeme summary of the story, and it connects to the bill in a plain way. If the central bank is naming AI spending as a source of price pressure, then who pays for the power is no longer a utility-commission question. It's a question about what the buildout does to everyone's costs, and the bill is a first attempt at pushing some of that back onto the builders.

63. Damra Vol 00:17:11

I'll push on that. A bill that makes big customers cover their own incremental costs is also a bill that tells every state it can say no. After years of voluntary pledges, that's a different kind of leverage, and the labs spending the most are the ones with the most to lose from fifty different answers.

64. Lenar Kess 00:17:28

The math story from yesterday got a follow-up. A Hacker News discussion titled "OpenAI withdraws three mathematical results" collected thirty-nine points and twenty-eight comments this morning, and it links to a post on X from a researcher who goes by danintheory. What we have is the discussion title and who's posting. The withdrawal notices themselves aren't in front of me, so I won't tell you why those three results were pulled.

65. Damra Vol 00:17:54

That's a useful place to start, actually. Yesterday's show had the Lean verification angle, where a formal proof checker confirms a proof line by line. A withdrawal is what the other side of that looks like. Somebody looked, and something didn't hold.

66. Lenar Kess 00:18:08

Right, and this is a follow-up, so I won't go back through the seven hundred and twenty-two manuscripts. The short version from Tuesday: OpenAI released 722 manuscripts grouped into 372 families of results on long-standing problems, from a powerful unreleased model, and invited academics to check them.

67. Damra Vol 00:18:28

Axios has a piece this morning by Bradley Olson that explains why math is the next target, and I think the argument is right even if the results wobble. A proof can be scrutinized by mathematicians and increasingly translated into formal languages that computers verify. Code has the same property: you can tell right away whether it works.

68. Lenar Kess 00:18:49

Verifiability is the point of the piece. Mathematics gives AI something unusually valuable, a way to tell when it is right. Olson walks through the reception too. A Rutgers mathematician said on X that a result connected to the Riemann hypothesis would warrant an automatic Fields Medal if a human had done it. And Stephen Wolfram said at a National Museum of Mathematics event on Tuesday, "You can discover new math easily. You can make a trillion theorems easily. The problem is most of those theorems are not ones that anybody will care about."

69. Damra Vol 00:19:22

Wolfram's point lands differently when three of your results get withdrawn. The cheap part is generating the claim. The expensive part is the person who reads it.

70. Lenar Kess 00:19:31

Now the allegation. Two posts on X. The first is from an account called doomslice, which shared an image alleging that the release overlaps with a prior artifact, a draft or paper, in other words that someone else got there first. The second is from Susan Zhang, posting as suchenzang. The posts themselves are what we have, and neither tells me which prior work is being pointed to.

71. Damra Vol 00:19:57

So let me say what's established and what isn't. Established: a Hacker News discussion says three results were withdrawn, and two named accounts on X are raising a prior-work and credit question.

Not established: that the withdrawn results are the same ones the scoop claim is about.

72. Lenar Kess 00:20:14

Right, and the Axios piece does say that some mathematicians have questioned these results, plus another solution OpenAI released last month, asking whether they are original breakthroughs or draw heavily from earlier human work. That's the general worry. The specific scoop claim is an allegation, from people outside OpenAI, that hasn't been answered in anything I have.

73. Damra Vol 00:20:37

I'd expect the credit question to stay alive. If a model trained on the literature produces a result that's close to something in a draft nobody indexed, nobody is lying and somebody is still uncredited. There's no norm for that yet.

74. Lenar Kess 00:20:51

And the Axios piece makes the labor point too: mathematicians are asking what this does to graduate work and to training a new generation of theorists, much as software engineers did. Axios also says some mathematicians echo the engineers' answer, that the most meaningful advances often involve building a framework or bringing structure to a complex idea.

75. Damra Vol 00:21:12

I think the credit dispute outlasts the verification one. Whether the model can prove things is mostly settled by the checkers. The open fight is over credit, and whether posting a draft publicly, where a search can find it, starts to count as a claim a lab has to answer.

76. Lenar Kess 00:21:28

CrowdStrike published analysis of the attacker behind last week's South Korean bank intrusions. According to Techmeme's summary of the report by Ashley Campion, CrowdStrike Intelligence identified infrastructure tied to a targeted campaign against South Korean financial organizations that resulted in exfiltrated data. The analysis says the actor is likely Chinese-speaking and financially motivated. It also says the attacker used large language models plus an open-source Chinese agentic tool called ARTEX.

77. Damra Vol 00:21:59

We mentioned the attacks yesterday as a headline, so the new part is the profile. Financially motivated and Chinese-speaking is a very different story from the state-actor one people reach for first.

78. Lenar Kess 00:22:11

Yes, though it's attribution from a single analysis, so I'd hold it as CrowdStrike's assessment. The wording is "likely." Now the piece that traveled. A post in the LocalLLaMA subreddit, with a screenshot gallery, says the entire hack may have been the work of one person, using ARTEX, DeepSeek v4.1-Flash, GLM-5.3, Grok 4.6 and Claude Code.

79. Damra Vol 00:22:36

And that list is the unverified bit. The Reddit post cites CrowdStrike, but the Techmeme summary only says "LLMs" and ARTEX. So I can confirm the tool and the large-language-model use, and the named models are someone's reading of a screenshot.

80. Lenar Kess 00:22:52

Exactly. Which is also a reason not to draw conclusions about those vendors. If the list is right, it would include two Chinese open-weight labs, xAI, and Anthropic's coding tool in one workflow. What that tells you is that the attacker was picking by price and convenience. None of it shows that any of those companies' safeguards failed, because we don't know what was asked of which model.

81. Damra Vol 00:23:16

I'd push on that. A single operator mixing a cheap open model for bulk work with a stronger one for planning is how any developer works right now. The offensive version of that workflow is just the same workflow.

82. Lenar Kess 00:23:28

There's also a Forbes contributor piece from Ari Redbord, titled "Ransomware's Terminator Moment," which says autonomous AI now executes full ransomware attacks against hospitals, banks and utilities, and that defenders need to adopt AI too. That's a contributor opinion piece, and the summary I have is two sentences, so I'd treat it as a viewpoint and not as evidence about this case.

83. Damra Vol 00:23:52

Right, it's not about this attacker. What this case adds is one person, one campaign, and a bank-sized outcome, if the attribution holds.

84. Lenar Kess 00:24:02

A single analysis, a profile that says "likely," and a model list from a screenshot. Whether CrowdStrike's own writeup names those models decides how much weight the list deserves, and the Techmeme summary doesn't say.

85. Lenar Kess 00:24:15

President Trump is attending an AI science summit today, Thursday, and Axios got the announcement first. The White House is announcing more than one billion dollars in industry commitments to the Genesis Mission from AMD, OpenAI, Anthropic and others. It's also announcing a compute hub spanning more than fourteen institutions across ten southeastern states, and a one billion dollar investment in Georgia for student and faculty access to compute.

86. Damra Vol 00:24:42

Who's paying, though? Axios's own line is that the administration is looking for industry to step in amid its cuts to government funding for science research and development.

87. Lenar Kess 00:24:52

Yes, that's in the piece. The National Science Foundation and the Department of Energy will invest more than one hundred million dollars in AI-driven science tools. Genesis was launched last November with ambitious deadlines for accelerating discovery, and in July it was expanded to more than fifteen federal agencies. The summit was inspired by a July report from Michael Kratsios, the director of the White House Office of Science and Technology Policy, which highlights industry's increasing role in research and development and new ways to fund and organize it.

88. Damra Vol 00:25:25

So the structure is public money in the low hundreds of millions, industry commitments past a billion, and compute handed out through university hubs. The ratio tells you who's expected to carry it.

89. Lenar Kess 00:25:38

The exclusive is the announcement as shared with Axios before the event, so these are announced commitments, not delivered money. Separately, Reuters reports through Techmeme that Meta, Google DeepMind and Isomorphic Labs are putting in three hundred million dollars, and the US government more than five hundred million, to the Zuckerberg-backed nonprofit Biohub, to build open biology datasets for AI training.

90. Damra Vol 00:26:04

Open datasets paid for by the same companies that will train on them. I like that the data's open, because biology is short on large public training sets. I also notice who gets first use.

91. Lenar Kess 00:26:16

And Bloomberg, via Techmeme, says Isomorphic Labs, the AI drug discovery company spun out of Google DeepMind, is in early talks to raise funds at a valuation of at least forty billion dollars. That's

sources describing early talks, not a closed round.

92. Damra Vol 00:26:33

Put the pieces next to each other and you get a company funding the open data, a lab spun out of that company raising at forty billion, and a government summit announcing it. I'm not saying those are connected. I'm saying they were announced within a day of each other.

93. Lenar Kess 00:26:48

Last week the White House held another AI summit, focused on rolling out America.gov to streamline government services with AI, so this is a run of them. Today's one has a stated target, though: breakthroughs driven by what the White House calls "super intelligence."

94. Lenar Kess 00:27:04

Elsewhere in the space today, starting with Anthropic. It released Claude Haiku 5.5, which it calls "the cheapest, fastest, and most capable small model we've ever released." The AINews write-up says it's the first Haiku-tier update in about a year and costs roughly seventy-five percent less to run than Haiku 4.5 on average.

95. Damra Vol 00:27:25

What's the price?

96. Lenar Kess 00:27:26

For prompts under a hundred thousand tokens, ten cents per million input tokens and fifty cents per million output, with cache reads at a cent. Over a hundred thousand tokens, input goes up to fifty cents and output to two dollars fifty, with cache reads at five cents. AINews says that matches OpenAI's GPT-6 Luna. Anthropic also halved Sonnet 5.5 cache reads, from twenty cents to ten, and cut subscription prices the same day.

97. Damra Vol 00:27:55

The positioning is the interesting bit. Anthropic says to run it as a subagent under Opus 5.5 or Sonnet 5.5 for summaries, compactions and database queries. The line AINews quotes from the mikeyk account is "Opus does the heavy thinking, Haiku does the high-volume work."

98. Lenar Kess 00:28:15

That ties back to the Strands talk, where the always-on chores are the expensive part. The benchmarks are Anthropic's own, and the Indian Express calls this the third model in a month ahead of a planned IPO, which is that outlet's angle.

99. Damra Vol 00:28:29

Next item. Three OpenAI researchers who were fired have written a letter to the board, reported by Maxwell Zeff at the Wall Street Journal.

100. Lenar Kess 00:28:38

The letter asks the industry not to move forward with work that would degrade the ability to monitor AI, and says the firings are, quote, “chilling those who remain at OpenAI.” It's one side of a dispute over the firings, and I don't have OpenAI's account or the letter's full text, so I'm not going to guess at why they were let go.

101. Damra Vol 00:28:58

Next to that, a Guardian column by Moustafa Bayoumi quotes Sam Altman from Politico's Decoded podcast: “One of the differences between us and some of the stricter AI safety people is that we believe that the world should accept some bad things happening for the benefits of this technology and people having the agency” to use it. Bayoumi's piece is an opinion column, and it argues that those bad things fall on other people.

102. Lenar Kess 00:29:24

Those are two separate items from two different sources, and one is a column, so I'm putting them side by side without connecting them. The Altman quote stands on its own as a statement of how OpenAI weighs risk.

103. Damra Vol 00:29:36

Money, then. The Wall Street Journal, via Techmeme, reports Broadcom has been working to arrange more than fifty billion dollars in financing for OpenAI's custom AI chip. It also reports that Oracle is in talks on financing for a big chip purchase, with Apollo, Blackstone and Goldman Sachs among the lenders in talks.

104. Lenar Kess 00:29:56

Those are sourced reports of talks, not confirmed deals. The new piece is who lends. These are megadeals worth tens of billions apiece, and the money's coming from private credit and big banks and not from the chip company's own balance sheet.

105. Damra Vol 00:30:10

And the comparison number: SemiAnalysis, reported by the Financial Times, says China has twenty-four gigawatts of operational compute capacity and another fifty planned or under construction, against fifty-six gigawatts operating in the US. The FT's piece is about Inner Mongolia.

106. Lenar Kess 00:30:29

So fifty-six against twenty-four today, with China's pipeline at fifty. That's an estimate from one research firm, and the capacity figures are on the power side, so they tell you about buildout speed and not chip quality.

107. Damra Vol 00:30:42

Three quick moves outside the US. First, Manus. Its parent, Butterfly Effect, raised more than five hundred million dollars led by Boyu Capital and IDG Capital, according to CNBC. It's the first funding since Beijing forced Meta to unwind its roughly two billion dollar acquisition.

108. Lenar Kess 00:31:02

Second, Ecosia, the Berlin-based search engine used by government services, has dropped Mistral for open-source models, including Chinese ones. Politico quotes its CEO: "We are disappointed with the quality of Mistral." That's from a Techmeme summary of the Politico piece, so I only have that line.

109. Damra Vol 00:31:21

And Sarvam, in India, hired Sashikumar Sreedharan, the former head of Google Cloud India, per Forbes. The summary says Sarvam has started selling DeepSeek and Zhipu models as part of its sovereign AI pitch.

110. Lenar Kess 00:31:35

Which means two of today's sovereignty stories end with Chinese open-weight models doing the work: a German search engine that left a French lab over quality, and an Indian company selling DeepSeek and Zhipu under its own name. How Ecosia's and Sarvam's customers react to that is the follow-up I'm curious about.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic