

OpenCode says attackers reached stored provider keys, and Anthropic reports its own hiring shift

2026-10-18 / 00:18:39

“Outsource your thinking, not your understanding.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

OpenCode says attackers got into its hosted Console and reached stored provider keys, which raises the question of who verifies an agent's work and who holds its credentials. Today also covers verification loops, agent authorization, Anthropic's report on its own workforce, and the pricing of AI infrastructure debt.

- [OpenCode breach coverage on Techmeme](#) and the [Reddit report of a drained key](#), with [Nathan Lambert's objection to the "distillation" label](#)
- [Sphinx](#), a Git hook that quizzes you on a diff before a push, plus the [WorkOS talk on agent verification](#) and [Addy Osmani's one-line principle](#)
- [A talk on agent authorization beyond OAuth scopes](#)
- [Anthropic's report on employment at Anthropic](#) and [TechCrunch's coverage](#)
- [AI infrastructure debt pricing](#) and the [World Bank speech on small AI](#)
- [DeepSeek V4.1-Flash on Ollama Cloud](#) and [Nvidia's RTX Spark laptops](#)

<u>00:00:04</u>	Stored keys at OpenCode
<u>00:04:11</u>	Verification loops for agents
<u>00:07:54</u>	Who authorizes the agent
<u>00:09:49</u>	Anthropic counts its own workforce
<u>00:14:03</u>	Debt under the data centers
<u>00:16:17</u>	Brief: local and hosted models

Transcript

1. Lenar Kess 00:00:04

A Reddit user posted this morning that a stored Anthropic API key of theirs was drained overnight, about 158 dollars, and that the timing lines up with the OpenCode breach. That's one person's post, so treat the dollar figure as an anecdote. What's sourced is bigger. OpenCode, the open source coding agent, says an attacker got into its hosted Console, and that about 1.4 percent of Console users were affected. The Console is where people stored their own provider keys, Anthropic, OpenAI, and Google, so the agent could call those models on their behalf.

- [x.com](#)
- [techmeme.com](#)
- [reddit.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [theguardian.com](#)
- [x.com](#)
- [youtube.com](#)
- [x.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [sphinx-six-theta.vercel.app](#)
- [reddit.com](#)
- [youtube.com](#)
- [x.com](#)
- [anthropic.com](#)

- techcrunch.com
- techmeme.com
- techmeme.com
- techmeme.com
- x.com
- theguardian.com
- x.com
- api-docs.deepseek.com
- huggingface.co
- x.com
- blogs.nvidia.com
- techmeme.com
- cnbc.com
- forbes.com

2. Damra Vol 00:00:36

That's the detail I keep turning over. Stored keys are a convenience feature: you paste a key in once, and every machine you sit at can use it. The price of that convenience is that a hosted service now holds a pile of other companies' credentials. So this isn't a breach of OpenCode's models, because OpenCode doesn't have any. It's a breach of a lot of people's billing relationships with somebody else.

3. Lenar Kess 00:00:58

First up is the word the company chose for it. OpenCode described what happened as a “large scale distillation attack.” I'm relaying that phrase from the Techmeme summary and the coverage around it, not from the original post. Distillation is training a smaller model on a bigger model's outputs, and the labs have been using the word lately for rivals pulling outputs from their models in bulk.

4. Damra Vol 00:01:19

So the claim has two halves. Somebody took keys, which is credential theft, ordinary and old. Then somebody used those keys to pull a lot of output out of the models behind them, which is where distillation comes in. My inference, and it's only that, is that the keys were the way in: you steal other people's keys and run your extraction on their bill, so it's distillation on someone else's account. OpenCode hasn't said that in so many words.

5. Lenar Kess 00:01:45

Nathan Lambert pushed back on the label. I'd summarize his objection as the label versus the incident. If someone stole a key and ran prompts, the incident is theft of access first. What they did

with the outputs is a separate fact that needs its own evidence.

6. Damra Vol 00:01:59

And the label decides who gets blamed. “Distillation attack” points at a rival lab or a state. “Someone got into a console” points at the vendor's security. One of those is a story about geopolitics and the other is a story about a startup storing secrets, and I'd want to know which before I believe either.

7. Lenar Kess 00:02:18

Kelsey Piper's read on the breach is that it was “likely AI-enabled,” her phrase. I haven't seen what that rests on. The other outside comment is from Miles Brundage, who points out that no AI company claims to be secure against state-level attackers. That's true, and it doesn't answer whether OpenCode's Console was storing keys sensibly.

8. Damra Vol 00:02:37

There's also a Guardian column today on world leaders and cyberattacks against the financial system, with a rogue-fridge example that runs at fifty degrees Celsius. It's a vivid detail, and nothing in the column ties it to AI, so I'd keep it out of this story.

9. Lenar Kess 00:02:51

What do we know about how those keys were held? Nothing public yet. The Console may have encrypted them at rest, or it may have decrypted them on the server for every call, and an attacker may have gotten plaintext. None of that has been published, and it decides whether this was a design problem or a bad night.

10. Damra Vol 00:03:08

For comparison, the local alternative is plain: the key sits in an environment variable or a keychain on your machine, and an attacker has to get into your laptop. Hosted storage buys you sync and sharing, and it makes the host the most attractive target in your whole setup. I'd also treat the 1.4 percent as unsettled until I know how many keys that is, because the users who stored keys in a Console are probably the heavy ones.

11. Lenar Kess 00:03:34

Fair. And if the extraction story is right, the loss isn't only the 158 dollars somebody saw on a bill. Your key becomes a free pipe for someone else's training run. That's speculation on top of a sourced breach, so I'll leave it there. So the sourced part is a breach with a percentage attached, and the rest is a label, an objection, and one anecdotal bill.

12. Damra Vol 00:03:54

Anyone with a key stored in that Console is presumably rotating it right now, whichever label wins.

13. Lenar Kess 00:04:01

The sentence I'd want from OpenCode is how the keys were stored and how they were taken. Until it comes, "distillation attack" is a claim, and the Reddit bill is a coincidence of timing that fits it.

14. Lenar Kess 00:04:11

Three talks and a launch this weekend all circled the same problem, which is how you know an agent's work is right when you didn't watch it happen. The longest is a WorkOS talk with Nick Nisi and Zach Prosser. Nisi says he hasn't written code since the end of July 2025, and their team reports about 900 pull requests a month with zero incidents. Those are the speakers' own numbers, and a conference talk is where you'd expect good ones.

15. Damra Vol 00:04:35

I'd want zero incidents defined. The stronger material is the pair of stories they tell about the opposite case. In one, an agent deleted a production database in nine seconds. The other story is a persistence bug that passed the initial tests and then failed on page reload, because the test checked that data was saved and never checked that it came back.

16. Lenar Kess 00:04:55

Their answer is a system they call CTX, running on Cloudflare. Secrets sit behind Model Context Protocol connections, so the agent never holds a credential. And the evaluators check outcomes against the GitHub API rather than trusting the agent's own report of what it did. The party verifying the work is separate from the party doing it.

17. Damra Vol 00:05:14

It connects to the first segment more than I planned. The agent doesn't hold the secret, so there's no secret for the agent's host to leak. Different failure, same instinct: keep credentials and judgment out of the process that's doing the work.

18. Lenar Kess 00:05:28

Addy Osmani posted the principle today in one line: "give your agent a way to verify its work." The sentence is short, and the evaluation workshops I went through are the long version of it.

19. Damra Vol 00:05:39

Arize's workshop is the skeptical long version. They go through why you can't just hand the check to another large language model. The judge costs money on every run, it's nondeterministic, so the same output can pass and then fail, and it makes confident errors. A judge that's wrong with conviction is worse than no judge, because people stop looking.

20. Lenar Kess 00:05:59

Langfuse's workshop takes the practical route of persona-based tests. You write simulated users, say different client tiers, and run the agent through each of them, because a support agent that behaves for your enterprise customer may do something odd with a free-tier one. It's a test suite where the fixtures are people.

21. Damra Vol 00:06:16

Then there's Sphinx, which interests me most because it verifies the human instead of the agent. It's an open source Git hook, MIT licensed, and the Show HN has 146 points and 44 comments. When you or your agent pushes to a protected branch, Git pauses, and a reviewer model with no access to your chat reads the diff and asks how the change behaves when things go wrong. A wrong answer gets an explanation and a fresh scenario. A right answer lets the push through.

22. Lenar Kess 00:06:46

The sample on the page is a webhook handler that stores an event ID before calling charge, and the quiz asks what a retry does if the charge fails. The tagline is "Outsource your thinking, not your understanding." There's an escape hatch, a command called sphinx yolo, and every skip gets logged.

23. Damra Vol 00:07:04

I'd bet the log of skips is where the product lives. A quiz you can bypass tells you how often people bypass it, and most teams would rather not have that number. It also promises that your answers never reach the coding agent and that it never opens .env files, which reads differently after this morning's breach.

24. Lenar Kess 00:07:22

Someone in the AI agents subreddit argues that an agent can remember perfectly and still be unreliable, which is the same gap from the memory side: stored context isn't verified context. All of these treat the agent's own account of its work as a claim and put a second party in charge of checking it. They disagree about who that party is: a script, a judge model, a simulated user, or you.

25. Damra Vol 00:07:45

And the cheapest of the four is the one I'd expect people to skip. The reload check in Nisi's story is something a person does by reflex, and it still got through.

26. Lenar Kess 00:07:54

A talk on agent authorization opens with a speaker getting banned from Resy after about 200 calls in 90 minutes. The speaker's point is that the service saw a rate, and had no way to see who was acting for whom. The agent was working for its owner, and the ban hit the owner anyway.

27. Damra Vol 00:08:10

Which is the argument that OAuth scopes were designed for a human clicking through a browser redirect. A scope says what an app may touch. It doesn't say why, for how long, or under whose supervision. Salesforce comes up as the example of a single scope that covers more than anyone wants to hand an agent, and Notion's Model Context Protocol server comes up as another place where this gets concrete.

28. Lenar Kess 00:08:33

The proposals are about intent. Dick Hart describes a “mission” primitive: you authorize a task with a purpose and a lifetime, and the agent's access exists for that mission and no longer. Michael Davis of JPMorgan says to onboard agents the way a bank onboards developers. Ian from Keycard talks about layered policy with a supervising agent that watches the first one.

29. Damra Vol 00:08:55

The supervising agent is where I'd poke. If one agent is policing another, who verifies the policeman? It's the verification loop from the last segment with a compliance badge on it. Still, it beats a static scope, because a static scope can't notice that an agent called the same endpoint two hundred times.

30. Lenar Kess 00:09:13

Ethan Mollick had a post on a related effect that I'll paraphrase. When agents can send a flood of inquiries, they remove the friction that used to ration access, like a form nobody wanted to fill out or a support line nobody wanted to wait on. Systems that were quietly protected by tedium lose that protection.

31. Damra Vol 00:09:30

That's the Resy story from the other side. The booking system survived on humans being slow. Agents make it fast, the service panics, and what it bans is the rate. A smarter ban doesn't fix that.

The service needs a way to ask who this agent works for and what it's allowed to do, and nobody has agreed on how to ask.

32. Lenar Kess 00:09:49

Anthropic published a report on employment at Anthropic on Friday, October 16, drawn from its own staff data, and it's the most specific account I've read of what a company at the front of this does to its own org chart. Start with juniors. Junior staff were 21 percent of technical-department employees in early 2025, and that's now 14 percent. Among technical hires the drop is steeper, from 27 percent to 9.

33. Damra Vol 00:10:14

TechCrunch's headline is that the workforce has become less junior and is spending more time in meetings. The top Hacker News comment compresses the anxiety into one line: "with no new juniors there will be no new seniors." The report has a caveat that comes close to saying the same: the company's retention and reallocation commitments don't help people who were never hired in the first place.

34. Lenar Kess 00:10:37

On the agent side, agent hours exceeded human hours in March 2026. By August, employee-directed interactive agents logged six times the hours of the humans directing them. Interactive sessions spend roughly 80 percent of their time idle, waiting on people. One engineer describes "an army of Claudes, with every one asking for my attention," and another says people are "bottlenecking the agents."

35. Damra Vol 00:11:00

The idle number reorganizes the picture for me. We usually talk about agents waiting on compute. Here the scarce resource is a human's attention, and the report says reading all the output from interactive Claude sessions would take over half a work week, even for non-technical groups. Presumably nobody reads all of it, which loops back to Sphinx: checking can't scale if the checker is a person.

36. Lenar Kess 00:11:24

So the work moves. Meta-work rose from 10 percent to 22 percent of interactive sessions between fall 2025 and spring 2026, and by that the report means people building skills and reviews for agents, plus dashboards and monitors to watch them. Written standing instructions for agents grew from a few hundred in July 2025 to 8,500 a year later. Code commits per engineer are five times what they were six months ago, and commits from managers doubled.

37. Damra Vol 00:11:53

The non-technical numbers surprised me. Coding-agent sessions by non-technical staff went from two a month in August 2025 to 90 in May 2026. And there's a line from the product-writing team: "Writers who used language to drive human behavior are now using it to drive Claude behavior." That's a job description changing in one sentence.

38. Lenar Kess 00:12:15

On productivity, the report compares 192 adopting engineers with 74 non-adopters and finds no difference in code-quality proxies such as lines removed or reverted. The authors flag that adoption isn't random, and that the company's rapid growth is a confounder, with revenue per employee two to four times what Meta and Google had at the same age. So this is internal data, published by the company that gains from the conclusion.

39. Damra Vol 00:12:40

Some lines read as culture rather than measurement. Meeting time is up 20 percent for the median employee and 16 percent for technical individual contributors. Employees with a year of tenure switch departments 50 percent more often each month. And an executive says, "a workforce that is anxiously worried is not a workforce doing its best work." I read that as a company managing a mood, and publishing the data is part of the management.

40. Lenar Kess 00:13:05

Employees expect AI to do 75 percent of their work by 2030, against 28 percent among other Americans. A manager on hiring said, "I expect models to continue to be jagged, so for my team I want them to be fungible across domains." And scorecards mentioning AI-tool track record went from about 5 percent of technical and 7 percent of non-technical assessments in 2024 to triple and quadruple those rates by 2026.

41. Damra Vol 00:13:31

One more Anthropic item, from the Financial Times, which is paywalled, so I only have the Techmeme headline: the FT reports that Anthropic is consulting law firms on negotiation tactics for agents. I know nothing beyond that. It fits the Mollick point, though, because agents negotiating with agents need rules about what a negotiator is allowed to do.

42. Lenar Kess 00:13:53

The junior figures are the ones that will matter in a year. A company that tripled in size hired a smaller share of juniors, and its own report says its commitments don't reach people it never hired.

43. Lenar Kess 00:14:03

According to The New York Times, lenders are now putting prices on AI infrastructure risk. CoreWeave's bond yield is above 15 percent. Insurance on SoftBank's debt has gotten more expensive. And some Oracle debt is trading at 80 to 90 cents on the dollar. We went through CoreWeave and Nscale on Friday, so I'll stick to what's new today, which is that these are market prices rather than forecasts.

44. Damra Vol 00:14:25

A 15 percent yield sounds like a verdict, and EnergyCredit1 on X says to slow down: that 15 percent is unsecured paper, and debt secured by the GPUs themselves prices lower. So the market isn't calling all AI infrastructure debt junk. It's saying one layer costs a lot: the borrower who promises to pay without collateral.

45. Lenar Kess 00:14:46

The Financial Times and RegLayer have a count of projects: at least 47 large US data center projects delayed, withdrawn, or blocked in 2026. Bernstein, per the reporting, puts roughly 35 percent of the pipeline as credible. I only have those through Techmeme headlines, so I can't tell you how a project got counted as blocked or what the denominator is.

46. Damra Vol 00:15:08

The two numbers measure different things, and I'd resist averaging them into a mood. One is about financing cost, the other is about projects that couldn't proceed, whether from power, permitting, local opposition, or money. The 35 percent is Bernstein's judgment, not an observation. If the gap between secured and unsecured widens, that would mean lenders believe in the hardware and doubt the borrowers.

47. Lenar Kess 00:15:33

A different pole: the Guardian reports on a World Bank speech by Paschal Donohoe about AI and developing economies. He talks about "small AI," models running on phones, as the version that reaches people. And he says the time developing countries need to catch up could shrink from centuries to decades. That timeline is his own claim.

48. Damra Vol 00:15:53

It sits oddly next to Oracle debt at 80 cents. One side of the world is financing enormous campuses at 15 percent, and the other is being told the version that reaches people fits in a pocket. It's the same technology priced at opposite ends, and the small end doesn't need a bond market.

49. Lenar Kess 00:16:09

If GPU-backed paper stays cheap while unsecured paper stays above 15 percent, the market is pricing the borrowers and not the chips.

50. Lenar Kess 00:16:17

A quick run through the model and device items. DeepSeek V4.1-Flash-1017 is now on Ollama Cloud, hosted in the US with zero data retention, according to Ollama's post, which answers a question teams have been asking about where their prompts go. We covered DeepSeek's benchmarks yesterday, so nothing on those today.

51. Damra Vol 00:16:36

On the local side there's a community quant called ds4, nicknamed Dwarfstar, on Hugging Face, with a vision tower. Tiezhen Wang's reaction was "insanely good, also super fast as flash." That's one person's impression, and I'll leave it at that.

52. Lenar Kess 00:16:53

DeepSeek also put out a search and fetch API at 75 cents per 1,000 successful requests, with 240 points and 112 comments on Hacker News. Charging only for successful requests is a billing choice, and failed fetches are common when agents browse.

53. Damra Vol 00:17:08

Nvidia's blog announces RTX Spark laptops from ASUS, Dell, HP, Lenovo, Microsoft, and MSI. They start at 2,399 dollars, and the unified memory goes up to 128 gigabytes. Memory at that size is what lets a laptop hold a big quantized model, so it connects to the quant story. Separately, Satya Nadella told Joanna Stern that Microsoft plans to make Windows safer for agents. As reported it's a stated plan, with no specific features named, so there's nothing to evaluate yet.

54. Lenar Kess 00:17:43

CNBC looks at AI wearables and wiretap laws, and the point is that the wearer may carry the liability when the device records other people. The people around the wearer haven't agreed to anything, and the law may land on the person who bought the gadget.

55. Damra Vol 00:17:57

Forbes profiles Neura Robotics, a German company valued at seven billion dollars. It wants to be the "Apple of physical AI," and it plans to sell the robots, an operating system for them, and the training gyms where robots learn. It has raised 1.4 billion dollars from Nvidia, Amazon, and Tether.

The gyms are the detail that stays with me, because Neura intends to own a place where the training data gets made.

56. Lenar Kess 00:18:21

The ds4 quant and the RTX Spark laptops make the same bet from two ends, that people will want the model on their own hardware, and this morning's OpenCode breach is a fresh argument for keeping keys and prompts close to home. I'm Lenar.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic