Cursor traces how leaked agent-session credentials led to a fake Claude Code extension

Coding Agents Alpha Tracker

2026-11-14

Cursor traces how leaked agent-session credentials led to a fake Claude Code extension

By Coding Agents Alpha Tracker • November 14, 2026

Cursor's incident report shows how credentials left in coding-agent traces became a supply-chain attack. Also: ThePrimeagen and Antigravity put agent control flow into explicit code, Theo finds Gemini 4 Carbon 5x slower in real work, and Anthropic and LangChain release agent platforms.

Cursor's incident: secrets in agent traces led to a fake Claude Code extension

Cursor published its full account of the incident it first disclosed on October 29. An attacker got hold of an access key Cursor had issued to Fireworks, which Fireworks had exposed. The attacker used it to copy coding-agent traces, mostly from 2025, taken from sessions that weren't in Privacy Mode. A few traces held credentials that had never been redacted. Those credentials led into Cursor's SaaS tools, and from there to archived non-production repos, support emails, and credentials users had sent to support [1]. Using a stolen credential, the attacker placed malicious copies of Anthropic's Claude Code extension in the cache that serves Cursor's VS Code marketplace. At most 5,000 users may have received them, and they were available for about 11 hours. Anthropic's product was not at fault, and Open VSX never carried the malware [1].

What to check: if you installed or updated `Anthropic.claude-code` **2.1.312 or 2.1.313** from Cursor's marketplace during that window, follow the steps in Cursor's security notice. Cursor warns that uninstalling the extension alone may not be enough [1]. Cursor says Privacy Mode data, production systems, official app updates and customer repos were not affected. Fewer than 0.05% of non-Privacy-Mode users had traces exposed, and each has been notified [1].

Any token you paste into an agent session can end up in stored traces. Cursor

has since added secret scanning across its repos and restricted where credentials can be used [1]. Kent C. Dodds noted separately that npm now runs a security scan on every publish. It can take a while, but he thinks it's worth it [2].

Put agent control flow in code you can see

ThePrimeagen says much of the control is lost when an agent does what you intended to do yourself [3]. His current approach is to render his program as a visual state machine: - When one node is wrong, he selects it and sends a prompt about that node to Cursor [4]. He can also select large parts of the pipeline and discuss them [5]. - Error and exit conditions are declared in code at the top of the pipeline, so you can see where each error leads [6]. Loops are visualised too [7]. - Pressing `gd` on an item opens its code or prompt [8]. The prompt editor autocompletes based on which pipeline steps have already run, and warns when you read a value too early [9].

Google Antigravity is rolling out Dynamic Workflows, which takes a similar line. Run `/run-workflow` and the agent writes a Python script that runs sub-agents in parallel or in sequence, with structured handoffs between them. The aim is "deterministic control flow instead of prompt drift." You review the script before it runs and can save it to rerun the same flow (docs) [10].

Theo: Gemini 4 Carbon was about 5x slower in his real work

In "I was wrong…", Theo audited his own usage and found Carbon "consistently five times slower" than other models on comparable work [11]. He says single-request benchmarks hide this because agent tools work in many steps [11]. He measured more than 11 seconds from when a request reaches Google's server to when the model starts working, and that delay repeats on every step [11]. He also says Carbon often fails to read the cache it just wrote, or reads an older one [11], and that it's due to double in price on December 1 [11].

A reusable trick: Carbon was one of the judges in his review runs, and it held things up. He updated his agent skill to tell agents that if they're stuck waiting on Carbon and have work to fix, they should stop it and get back to work. He says that change makes his 5x figure an underestimate [11]. He now goes further than his earlier, partly positive review: he sees no reason to use Carbon except for security work, where it's slightly less likely to flag you [11].

Agent platforms: tied to Claude or model-neutral

- **Self-Hosted Claude Managed Agents** shipped. You get the managed-agents server, which works with your identity provider and includes OpenTelemetry telemetry and storage, plus a self-hosted console with Anthropic's first Evals product. The harness is tuned for Claude [12]. swyx asked why self-hosted agents should be tied to Claude at all [13].

- **LangChain** raised $300M at a $3B valuation [14]. LangSmith is now presented as one framework-agnostic platform for running, improving and governing agents [14]. LangSmith Engine finds errors, writes code and prompt fixes, and watches production traffic for regressions [14]. More features are due at a November 20 event [14]. Harrison Chase argues for a model-neutral harness: you can move to newer or cheaper models and avoid being locked in to a provider that raises prices [15]. He says coding is by far the biggest use of agents [15].

Smaller releases

- **Devin Apps** are private, shareable web apps that run on Devin's cloud computer, meant for internal tools and preview deployments [16]. Kent C. Dodds points out that Kody.codes already does this. Its apps can use your connected MCP servers, and he runs everything from a kids' money ledger to a GitHub PR dashboard on it [17]. He says Kody differs by working with any agent and having team-based permissions [18].
- **Claude Code desktop** now supports custom themes built as mods: colours and animated backgrounds. You can describe the look to Claude or start from the `theme-examples` plugin [19]. **T3 Code** nightly lets you ask the agent to change the app's UI, and the changes apply on the fly [20, 21].
- **Token-saving tip from an AI-video workflow:** keep assets in a GitHub repo and pass Claude links instead of attachments. An image sent as encoded text can cost hundreds of thousands of tokens, while a link costs a few dozen [22].
- **Be wary of viral posts:** Geoffrey Huntley says he gets about three paid-promotion pitches a day and assumes viral agent launches were paid to be boosted [23].

Sources

1. An update on our recent security incident
2. X post by @kentcdodds
3. X post by @ThePrimeagen
4. X post by @ThePrimeagen
5. X post by @ThePrimeagen
6. X post by @ThePrimeagen
7. X post by @ThePrimeagen
8. X post by @ThePrimeagen
9. X post by @ThePrimeagen
10. X post by @antigravity
11. I was wrong…
12. X post by @PriyankaPhatak
13. X post by @swyx

14. LangChain Raises $300M to Enable Every Company to Own Its Intelligence
15. Sapphire Ventures Partners with LangChain: Agent Engineering Platform
16. X post by @cognition
17. X post by @kentcdodds
18. X post by @kentcdodds
19. X post by @ClaudeDevs
20. X post by @theo
21. X post by @theo
22. X article by @withchiay
23. X post by @GeoffreyHuntley