# Secret Claude tracker shocks users after Anthropic’s anti-surveillance stance

arstechnica.com/tech-policy/2026/07/anthropic-outed-for-claude-tracker-that-secretly-monitored-chinese-users

Ashley Belanger July 6, 2026

“Serious breach of user trust”

Anthropic accused of spying on users; engineer says “experiment” is over.

Anthropic quickly removed a tracker secretly monitoring Claude Code users in China after a security researcher exposed the hidden code and condemned the spyware-like tracking as a “serious breach of user trust.”

Last week, a web developer known as “Thereallo” was researching privacy issues in Claude Code and was shocked to find that the AI firm was using “prompt steganography” to hide code that tracks Chinese users “in plain sight.” This code wasn’t malicious, but it was sending information to Anthropic that most users wouldn’t detect, relying on shorthand markers to quietly flag users’ timezone, proxy, and potential connection to Chinese AI labs that Anthropic has accused of distillation attacks.

On X, Anthropic engineer Thariq Shihipar confirmed that the tracker was added to Claude Code as an “experiment” in March. According to Shihipar, the code “was meant to prevent account abuse from unauthorized resellers and protect against distillation.” Regarding the former, The Washington Post found unauthorized retailers have sold access to free models for $1 a month, and pro subscriptions that can cost $100 monthly sell for “as little as $12.”

Supposedly, Anthropic has “actually been meaning to take this down for a while,” Shihipar said of the hidden code, because engineers have “landed stronger mitigations since then.”

Privacy advocates were not happy with the explanation, though, warning that the code is evidence that Anthropic is willing to cross lines to surveil users. That’s

perhaps especially surprising, considering that Anthropic riled the Trump administration by refusing to allow the US government to use Claude to surveil US users. The AI firm has since sued the White House over the clash.

## Anthropic wants distillation deemed illegal

The Post suggested that the tracker incident is a sign that US firms like Anthropic are taking "increasingly aggressive measures" to block Chinese AI firms from copying their models.

A more defensive stance has apparently become critical. In the past year, Chinese firms have "consistently matched" US firms' model capabilities "within months," the Post reported. Most recently, "a new, free AI model from Chinese company Zhipu AI was better at finding computer vulnerabilities than Anthropic's Claude Opus 4.8 model, which was released in May," the Post reported.

To lock in a 12- or possibly even 24-month lead for the US, Anthropic has said the US must ramp up interventions, using a range of possible penalties to combat distillation attacks, including blocking access to advanced models, chips, and data centers in the US.

Although distillation isn't illegal (leading US firms do it, too), prompting models like Claude millions of times in order to quickly advance Chinese models violates Anthropic's user terms.

To end the endless copying, Anthropic has joined OpenAI in urging the US to view distillation attacks as a form of intellectual property theft. At a recent Senate hearing, Sen. Tim Scott (R-S.C.) agreed legal intervention is needed, arguing that the US needs "to carefully craft export control policy that is clear and concise" to stop China from using such attacks to "gain a technological edge," the Post reported.

## Secret code triggers Alibaba Claude ban

It's clear that Chinese firms are distilling US models, the Post reported. In February, Chinese researchers at Peking University and the state-funded Chinese Academy of Sciences "developed methods to detect signs of distillation in leading large language models" and found that most Chinese models "showed substantial evidence of distillation," primarily of US models. One of Alibaba's Qwen AI models—which Anthropic has since claimed was advanced after the largest distillation attack ever on Claude in June—"repeatedly appeared to mimic" Claude in February. In some intensive tests, the model would even sometimes slip up and identify itself as Claude, researchers found.

Alibaba has not commented on Anthropic's accusations, but the company has moved to distance itself from Anthropic's models amid ongoing scrutiny.

Last Friday, Alibaba banned its employees from using Claude Code for work, the South China Morning Post reported. According to a memo SCMP reviewed, Alibaba told employees the ban came in direct response to concerning news about a tracker Anthropic is using to monitor Chinese users.

"As Claude Code was recently discovered to carry back-door risks, after comprehensive evaluation, Claude Code has now been added to a list of high-risk software with security vulnerabilities," the memo said.

For Alibaba, ignoring Anthropic's determination to detect users connected to leading Chinese AI labs is risky.

Unlike individual users who can easily pay for cheap circumvention tech to evade Anthropic's location blockers without fears of major repercussions, Alibaba could be exposed to legal and compliance risks if caught violating Anthropic's terms, a source granted anonymity to discuss Alibaba's Claude ban told Reuters.

For Anthropic, allowing the attacks to continue could hurt the company's business. Some open source Chinese models are more popular than free and open American counterparts, the Post reported, and Fortune 500 CEOs have made it clear that they're searching for cheaper AI solutions. For the US, not only would moving to block Chinese distillation of American models be challenging, but it could also be unpopular—blocking Americans from benefiting from cheaper AI alternatives from China, the Post suggested.

## Anthropic tracking crossed "scary boundary"

In this climate, where a chatbot user's loyalty depends on a cost-benefit analysis weighing the cost of accessing models against their capabilities, Anthropic likely can't afford to lose user trust as it fights to keep frontier models ahead of China's.

As the web developer who flagged the hidden tracker noted, it's "weird" that Anthropic chose to move in secret when the company could have instead chosen to transparently alert users to the infringing user-tracking.

"This is not a malicious feature, but it is a weird choice for a developer tool that asks for trust," Thereallo's blog said. The blog noted that "if the client wants to detect custom API gateways, it can say so plainly. It can send an explicit telemetry field with documentation. It can make the policy visible. It can put the behavior in release notes."

The researcher emphasized that “coding agents already live on the wrong side of a scary boundary. They can inspect code, summarize secrets by accident, run commands, install packages, edit files, and push commits on your local machine.”

Although most users were likely not impacted by the tracking, Thereallo warned that the “correct reaction” is more scrutiny of Claude’s potential for user surveillance, since “the feature mostly punishes the exact people who are easier to fingerprint: normal developers doing weird but legitimate things.”

“Hiding the signal in the system prompt makes every other privacy claim harder to believe,” Thereallo said.

Anthropic did not immediately respond to Ars’ request for comment.

However, a spokesperson told the Post that Chinese labs’ distillation attacks “pose a serious threat to national security and undermine AI safety standards across the industry. That’s why we continue to speak openly about what we’re seeing and work closely with other labs, government, and partners on shared solutions.”

Ashley is a senior policy reporter for Ars Technica, dedicated to tracking social impacts of emerging policies and new technologies. She is a Chicago-based journalist with 20 years of experience.
128 Comments