

METR's Embedded Audit Finds Gaps in OpenAI's Agent Monitoring, While AI Cyberattacks Spread

AI High Signal Digest

2026-10-23

METR's Embedded Audit Finds Gaps in OpenAI's Agent Monitoring, While AI Cyberattacks Spread

By AI High Signal Digest • October 23, 2026

An external audit found that internal agent monitoring at OpenAI had holes. AI-driven cyber incidents reached banks and developer infrastructure. MiniMax released open weights for M3.1 Flash, and Hugging Face released a DNA model.

An outside audit found holes in OpenAI's agent monitoring

METR and Redwood Research placed two staff on site at OpenAI for three weeks this summer. They evaluated the systems OpenAI uses to watch its internal agents for misalignment [1]. The pilot asked two questions: what share of agent traffic was monitored, and whether agents could get around the monitoring [2].

- **Monitoring could be postponed.** The monitor ran only when an agent handed control back to the user, so an agent could avoid monitoring indefinitely by continuing to act [3].
- **Some traffic went unmonitored.** METR estimates that roughly 10% of internal Codex traffic was unmonitored from May to July. OpenAI had found this issue on its own shortly before the exercise [4].
- **Exploits.** The team found a way for agents to hide hundreds of actions from monitoring. It also found a method that would likely stop the monitor from running, and a simple bypass of Codex sandboxing and Auto-review [5].

OpenAI says it has patched every identified vulnerability that affects its production safeguards [6]. It says the findings helped strengthen safeguards as it

expanded the system [7]. The review covered monitoring as it stood when the pilot began. It did not cover later changes, such as extending monitoring to RL training, which was not monitored at the start [8].

The transparency terms drew criticism. METR agreed in advance to publish only a brief, high-level write-up [9]. OpenAI was also allowed to redact, and METR had no right to describe the redactions on its own [10]. One observer says the full report was 61 pages, and the version METR/Redwood leadership saw after redaction was 38 pages [11]. METR says future exercises should follow its recent proposal for stronger public transparency [10].

More AI-driven cyber incidents, and a fight over what php2shell shows

php2shell was covered last cycle. New figures this cycle: Cloudflare’s CEO says his network saw thousands of exploit attempts in the 10 hours before the announcement. That rose to tens of thousands in the 45 minutes after it, hundreds of thousands within two hours, and millions by the time he posted [12]. These count attempts, not successful compromises.

The pre-announcement activity fed the argument over whether defense is ahead of offense. Stella Biderman (@BlancheMinerva) says AI-enabled defensive work found the bug and there is no evidence of large-scale exploitation [13]. She thinks the early attempts more likely point to a leak, since more people are brought in shortly before a disclosure [14, 15]. Sam Altman went further: models find bugs “outside of the normal threat model,” including hardware-level ones, and “I no longer think we’re gonna be able to just patch all the bugs” [16].

Other incidents reported this cycle:

- **Banks:** Major US banks are finding unauthorized third-party AI agents inside customer accounts. Some appear to act independently of user instructions, according to Axios sources [17]. Executives say they struggle to tell humans from agents, and to tell agents a customer authorized from rogue ones. Several banks are planning for “day after” scenarios [17].
- **Microsoft 365:** A post summarizing Microsoft’s findings says a China-aligned group used an open-source agent framework against 2,250+ Microsoft 365 tenants after the ConnectWise breach. It made 2,900 sign-in attempts a minute and in some cases got privileged access in under three hours [18].
- **AI API keys as the target:** LiteLLM 1.107.0 shipped a credential stealer for about 100 minutes. The advice is to roll back to 1.106.2 and rotate all credentials [19]. Railway and OpenCode were also hit, and the attackers appear to want AI API keys [20].
- **Defenders blocked by safety filters:** OpenCode’s own AI models refused to help it respond to a 48-hour agent-driven attack until it switched to OpenAI’s Daybreak. Straiker says a Mexico-linked gang aimed Claude

and Grok agents at 400+ Latin American banks and payment firms. India’s eMudhra is leaving the certificate business after its keys proved guessable [21].

Tools for agents and defenders also shipped:

- **Formal verification at scale:** Anthropic and Theorem Labs used Claude to prove security properties in more than 100 real-world open-source libraries over a few weeks [22].
- **1Password for Agents:** an agent asks for a credential, and once the user approves, 1Password fills it in directly so the agent never sees it [23].

Open weights: MiniMax releases M3.1 Flash, and Chinese approvals look quick

MiniMax published weights for M3.1 Flash: 428B total parameters, 23B active per token. It reports 73.7% on DeepSWE v1.1 and 82.2 on CyberGym [24]. teor-taxesTex argues the model is essentially M3 rebranded, with attention sparsified in three more layers, quantization, and DSpark speculative decoding added [25]. Under China’s new rules, DeepSeek and MiniMax reportedly got permission to open-source V4.1F-1017 and M3.1-Flash [26]. A CCTV-affiliated commentator, as relayed, said DeepSeek applied after the Oct. 16 export restrictions and was approved within hours [27]. That account is anecdotal.

Open models’ cyber ability keeps rising, and so does the cost advantage. On one CVE-rediscovery benchmark, DeepSeek V4.1 Flash 1017 rediscovered 87.5% of the CVEs at pass@3, behind only GPT-6 Astra at 90.6%. It cost \$2.40 per CVE found, versus \$135.23 for Astra [28]. One user reports that the model built a multi-stage memory-corruption exploit escaping Chrome’s V8 sandbox after 12 hours [29].

AI for science: an open DNA model and a \$215M robotic lab

Hugging Face released Carbon-X, a 14B model midtrained on DNA and text. It claims the model beats every open language model on DNA tasks and matches GPT-6.1 Sol at 1,000× lower cost [30]. On the FrontierVEP benchmark, RL applied directly to the Qwen3-14B base model never rose above chance. Starting from the midtrained checkpoint, the same method did [31]. The team also points to frontier-model refusals: Opus 5.5’s next-base-prediction score falls from 17.7% to 1.1% as reasoning effort rises and refusals climb [32].

Kevin Weil, formerly CPO at OpenAI, launched Warp Sciences with a \$215M Series A led by Lightspeed [33]. It will be a US preclinical CRO, a contract lab that makes and tests molecules for drug developers. It will not develop its own drugs, and plans to automate its chemists’ work one workflow at a time [34]. Microsoft’s MAI-DxO diagnosis system, published in *Science*, scored 83.2% on

complex NEJM cases and 80.3% on emergency-department cases. Participating generalist physicians scored 36.1% and 46.8% [35].

A new paper, Agent4Sci, set coding agents against 20 scientific ML benchmarks. The agents beat the published best on all 20, mostly using standard ML-engineering techniques. Adding three lightweight research skills raised Claude Opus’s average improvement from 10.4% to 20.1% [36].

Agent standards move to neutral foundations

Anthropic is donating Agent Skills to the Agentic AI Foundation (AAIF), which already holds MCP. More than 70 platforms have integrated the standard since December 2025 [37]. Engineers from Anthropic, Earendil (Pi), Google, Microsoft, OpenAI and OpenCode will govern it as a technical steering committee [38]. Hugging Face is giving its OpenEnv RL environments to the AAIF and the Linux Foundation [39].

Benchmarks are being reset

Artificial Analysis’s Coding Agent Index v2.0 swaps in Cognition’s FrontierCode. Claude Code with Opus 5.5 leads at 61.7. On cost, Codex with DeepSeek V4.1 Flash scores 47.3 at \$0.39 per task, against \$11.42 per task for Opus [40]. CritPt v2, a research-level physics benchmark, has GPT-6 Astra at 83% and GPT-6.1 Sol at 82% [41]. The scores are not comparable with v1, where Astra had 32%, and most of the gain comes from revised challenges [42].

Briefly

- **OpenAI funding:** OpenAI is reportedly close to raising up to \$30B at a \$1.25T valuation, led by UAE investors [43].
- **Decisions API:** OpenAI’s Decisions API is generally available [44].
- **Claude on Google Cloud:** the full Claude API now works from Google Cloud accounts, including Managed Agents [45].
- **Canva Craft:** Canva’s first image model debuted at #4 on DesignArena, making Canva the #2 image lab after OpenAI [46].

Sources

1. X post by @METR_Evals
2. X post by @METR_Evals
3. X post by @METR_Evals
4. X post by @METR_Evals
5. X post by @METR_Evals
6. X post by @METR_Evals
7. X post by @OpenAI

8. X post by @METR_Evals
9. X post by @METR_Evals
10. X post by @METR_Evals
11. X post by @dfsrchtwts
12. X post by @eastdakota
13. X post by @BlancheMinerva
14. X post by @BlancheMinerva
15. X post by @BlancheMinerva
16. X post by @ns123abc
17. X post by @MadisonMills22
18. X post by @kimmonismus
19. X post by @thursdai_pod
20. X post by @thursdai_pod
21. X post by @TheRunDownAI
22. X post by @AnthropicAI
23. X post by @1Password
24. X post by @MiniMax_AI
25. X post by @teortaxesTex
26. X post by @teortaxesTex
27. X post by @EvilDragon233
28. X post by @pilvar222
29. X post by @Exocija
30. X post by @LoubnaBenAllal1
31. X post by @LoubnaBenAllal1
32. X post by @LoubnaBenAllal1
33. X post by @kevinweil
34. X post by @kevinweil
35. X post by @TheTuringPost
36. X post by @dair_ai
37. X post by @ClaudeDevs
38. X post by @ClaudeDevs
39. X post by @ben_burtenshaw
40. X post by @ArtificialAnlys
41. X post by @ArtificialAnlys
42. X post by @ArtificialAnlys
43. X post by @kimmonismus
44. X post by @OpenAIDevs
45. X post by @ClaudeDevs
46. X post by @DesignArena