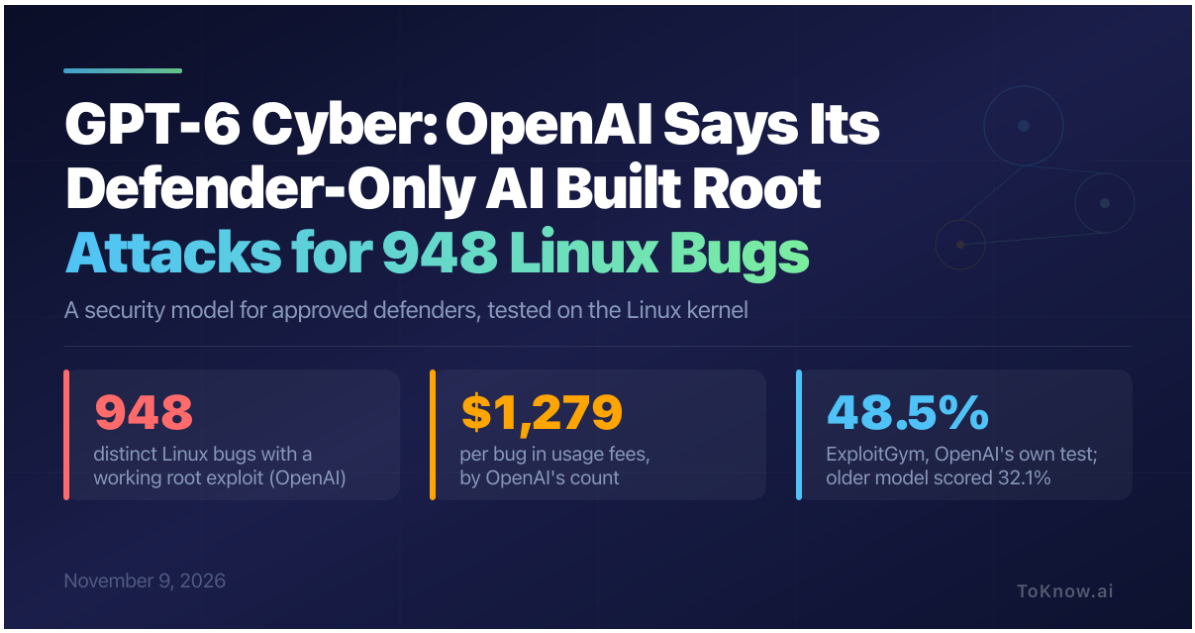
GPT-6 Cyber: OpenAI Says Its Defender-Only AI Built Root Attacks for 948 Linux Bugs

Kabui, Charles

2026-11-09

Read at ToKnow.ai



GPT-6 Cyber is OpenAI's security AI for vetted defenders: teams approved for its Daybreak Red program. OpenAI says it used the model on the Linux kernel, the operating system's core, to build 1,169 working exploits for 948 distinct bugs. An exploit is a program that turns a bug into a break-in. These bugs let an ordinary user become root, the all-powerful administrator. The run cost $1,213,430 in usage fees, or $1,279 per bug by OpenAI's count.

On OpenAI's ExploitGym test, which asks a model to turn a named flaw into a working attack, GPT-6 Cyber solved 48.5% of tasks, against 32.1% for GPT-5.6 Cyber. In its trap test, easy-to-hack side services sit beside the target and are off limits. With no safety prompts, it left them alone in 99.93% of cases. In an early run, agents reported 200 root exploits, and human review found 20 used a different unknown bug than assigned. OpenAI says it found 81 new Firefox bugs that Mozilla rates high severity. Mozilla's Firefox 159 advisory lists 41 fixes credited to OpenAI, without tying them to the 81.

The people who must fix these bugs were already swamped. In May, Linus Torvalds, who leads Linux development, wrote that the flood of AI reports had made the kernel's security list "almost entirely unmanageable". No maintainer has commented on this run in the coverage checked. After OpenAI's 10-day scan, 61 of the root bugs had been patched upstream, 43 had staged patches, and 299 were under public discussion.

Read More: Anthropic's unreleased model that links bugs into attacks on five major systems and Anthropic's free bug scanner for open-source projects

Sources:

- A thousand fewer paths to root (OpenAI)
- GPT-6 Cyber: Advancing the frontier of aligned cybersecurity agents (OpenAI)
- OpenAI's upcoming AI model built 948 working Linux root exploits in a single run (Help Net Security)
- Security Vulnerabilities fixed in Firefox 159 (Mozilla)
- Linux 7.1-rc4 announcement by Linus Torvalds (LWN)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*