Anthropic's Unreleased AI Links Bugs Into Attacks on Five Major Systems

Kabui, Charles

2026-11-02

Read at ToKnow.ai



Anthropic's Frontier Red Team, which stress-tests its AI, says an unreleased model built five working attacks on Chrome, Safari, Android, Windows and Linux, needing little to no user interaction. Anthropic calls these attack chains: small flaws linked so each one opens the door to the next. It says each came with a proof of concept, a working demo. In the Chrome chain, a web page escapes the browser's sandbox, a locked room that walls pages off from the

computer, and opens Calculator on a Mac seconds after loading. Anthropic counts 62 Chrome flaws with public CVE ID numbers, all critical or high and fixed by Chrome 156.

Anthropic says linking flaws like this once took a few specialist teams, and the model does it with little human help. The Safari chain let any web page read a signed-in user's email. The Android chain pulled photos off a stolen, locked Pixel 9 over USB, and the Windows attack only crashes servers. The Linux attack was zero-click, needing no user action or password. In its test it took 48 minutes to leak a memory address and five more minutes to reach root, full administrator control.

Security firm Black Duck's Andrew Bolster told Dark Reading that Anthropic gave no success rates, costs or human hours per chain. The Linux attack needed a server listening for multipath TCP, which lets one connection use several network paths. Red Hat shipped its Linux fix on November 1. Ubuntu still showed fixes in progress on November 2.

Read More: Anthropic's study of an open-weight model that builds cyber exploits, Anthropic's free bug scanner for open-source projects and Gemini 4 Carbon, Google's model that goes to cyber defenders first

Sources:

- Discovering and exploiting vulnerabilities in critical, real-world software with Claude (Anthropic Frontier Red Team)
- Stable Channel Update for Desktop, Chrome 156 (Chrome Releases)
- Claude Model Builds Exploits Across Major Browsers & OSes (Dark Reading)
- CVE-2026-98432 (Red Hat Customer Portal)
- CVE-2026-98432 (Ubuntu Security)

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