Anthropic's Red Team: An Open-Weight AI Model That Builds Cyber Exploits

Kabui, Charles

2026-10-04

Read at ToKnow.ai

Anthropic's Frontier Red Team published an analysis of GLM-5.3, an open-weight AI model from the Chinese lab Zhipu AI (Z.ai outside China), and says it can build working cyber exploits on its own. In Anthropic's tests, GLM-5.3 wrote complete exploits against known Chrome bugs in 50 of 410 attempts, close to Claude Mythos Preview's 56. Anthropic released Mythos Preview only to vetted defenders; anyone can download GLM-5.3. In one session it

found unknown flaws in a browser's JavaScript engine and chained them into a page that reads files off a visitor's computer.

GLM-5.3's guardrails do not hold up. Asked plainly to attack a system, it refuses. But dressing the request up as a red-team exercise got it to engage 64% of the time in Anthropic's simulations, and pre-filling its reasoning raised that to 92%. Because the weights are public, anyone can remove the refusals entirely using abliteration, a method that edits the weights. Anthropic's team did it in 2,200 GPU hours, roughly $4,400 of computing, and estimates an experienced team could do it for $1,200. The edit dropped the refusal rate from above 90% to as low as 2%. None of the tricks worked on safeguarded Claude models.

NIST's CAISI, which tested GLM-5.3 separately, agrees it is the most cyber-capable open-weight model yet, about four months behind the US frontier. The analysis is Anthropic's own, and The Decoder notes it also fits Anthropic's business: its models stay closed. As open weights close on the frontier, refusal filters become something a user can remove.

Read More: GLM-5.3, Z.ai's flagship model, which shipped with a public vulnerability-disclosure ledger

Sources:

- GLM-5.3 and the spread of advanced cyber capabilities (Anthropic)
- CAISI's assessment of Z.ai's GLM-5.3 cyber capabilities (NIST)
- Anthropic says Zhipu's open-weight GLM-5.3 nearly matches Claude Mythos Preview at building exploits (The Decoder)
- Anthropic says China's GLM-5.3 nearly matches Mythos at cyber exploits (The Next Web)
- GLM-5.3: Frontier Coding with Emergent Cyber Capabilities (Z.ai)

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