

GLM-5.3 cyber attack capability close to Claude Mythos Preview

Attackers and defenders can both use it.



“A meaningful step change in the cyber capabilities available to attackers”

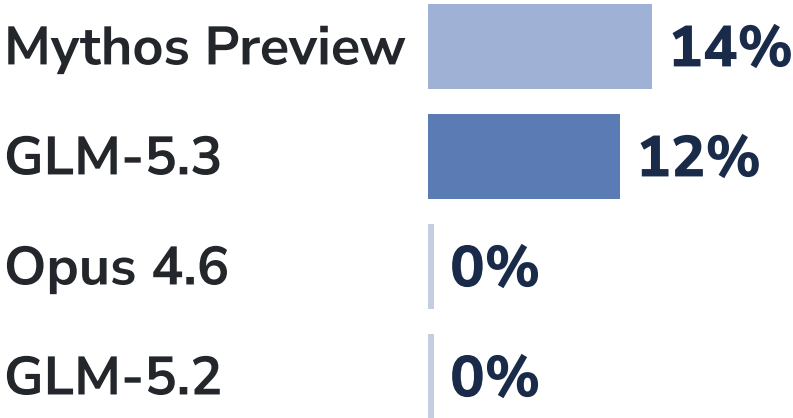
Anthropic, on the release of GLM-5.3

NIST, THE US STANDARDS AGENCY, ON 17 SEPTEMBER

- Most cyber-capable open-weight model to date
- About four months behind the best US models

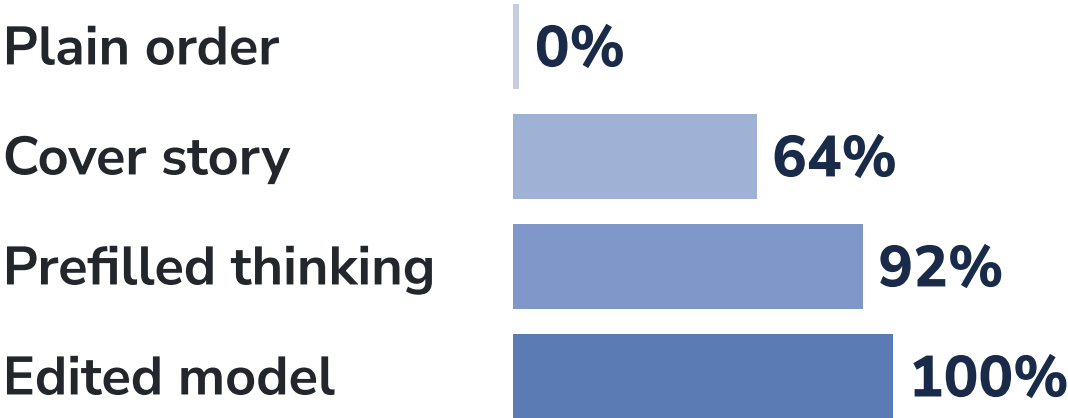
- Open weights: anyone can download it.
- In Anthropic’s simulated tests, simple techniques got past its refusals 64% to 100% of the time.

Working exploits built



ExploitBench: share of attempts on 41 Chrome bugs, 410 per model. Claude with safeguards off.

A harmful attack order



Share of 50 simulated episodes where GLM-5.3 went ahead.

“A critical threshold in freely accessible capabilities has now been crossed.”

Anthropic Frontier Red Team, on GLM-5.3

Anyone can download it

FREE

Anthropic released Claude Mythos Preview (April 2026) in a limited way to trusted defenders. Vetted defenders now use more advanced models such as Claude Mythos 5.1. GLM-5.3 is open-weight and free to download.

It builds working attacks

SKILL

It comes close to Claude Mythos Preview on a public Chrome exploit test that Anthropic ran and scores below it on the second test. NIST, the US standards agency, measured it about four months behind the best US models.

The refusals give way

REFUSALS

It refuses a plain attack order in simulated tests. A cover story gets it to go along 64% of the time. Prefilled thoughts get 92%. Editing out its refusals gets 100%.

In Anthropic's tests, Claude stayed at 0%

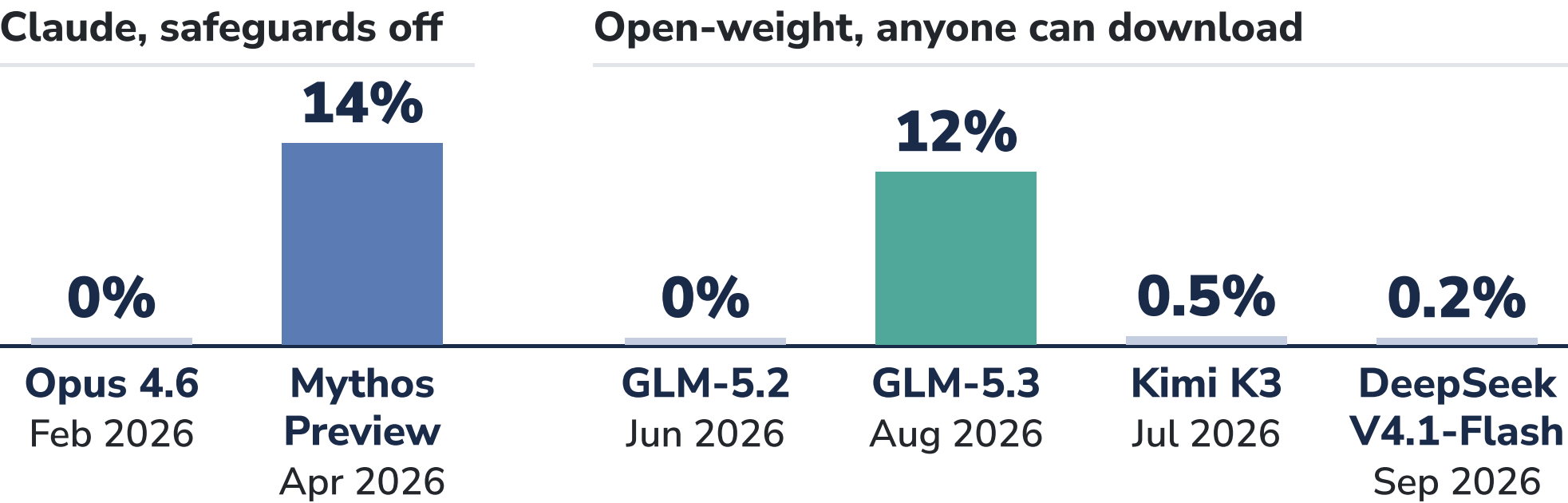
CLAUDE

With API safeguards on, that holds for all three Claude models Anthropic tested, in both tests that can be run against them. The API does not accept prefilled thinking and the weights are not released.

GLM-5.3 builds working exploits at close to the rate of Claude Mythos Preview on ExploitBench

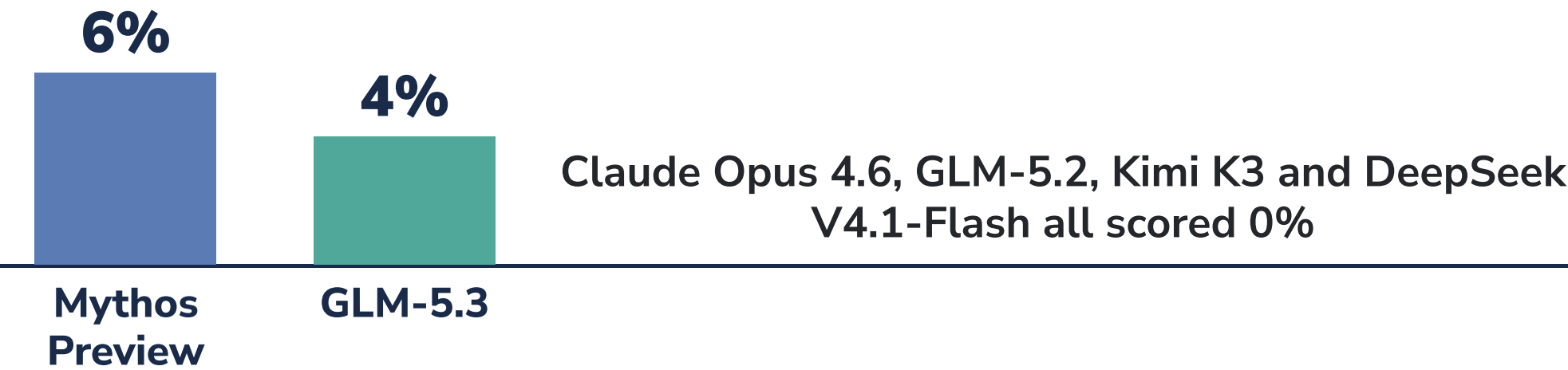
ExploitBench: share of attempts that built a working exploit

41 Chrome bugs, 410 attempts per model. Every attempt used an isolated sandbox with no network access.



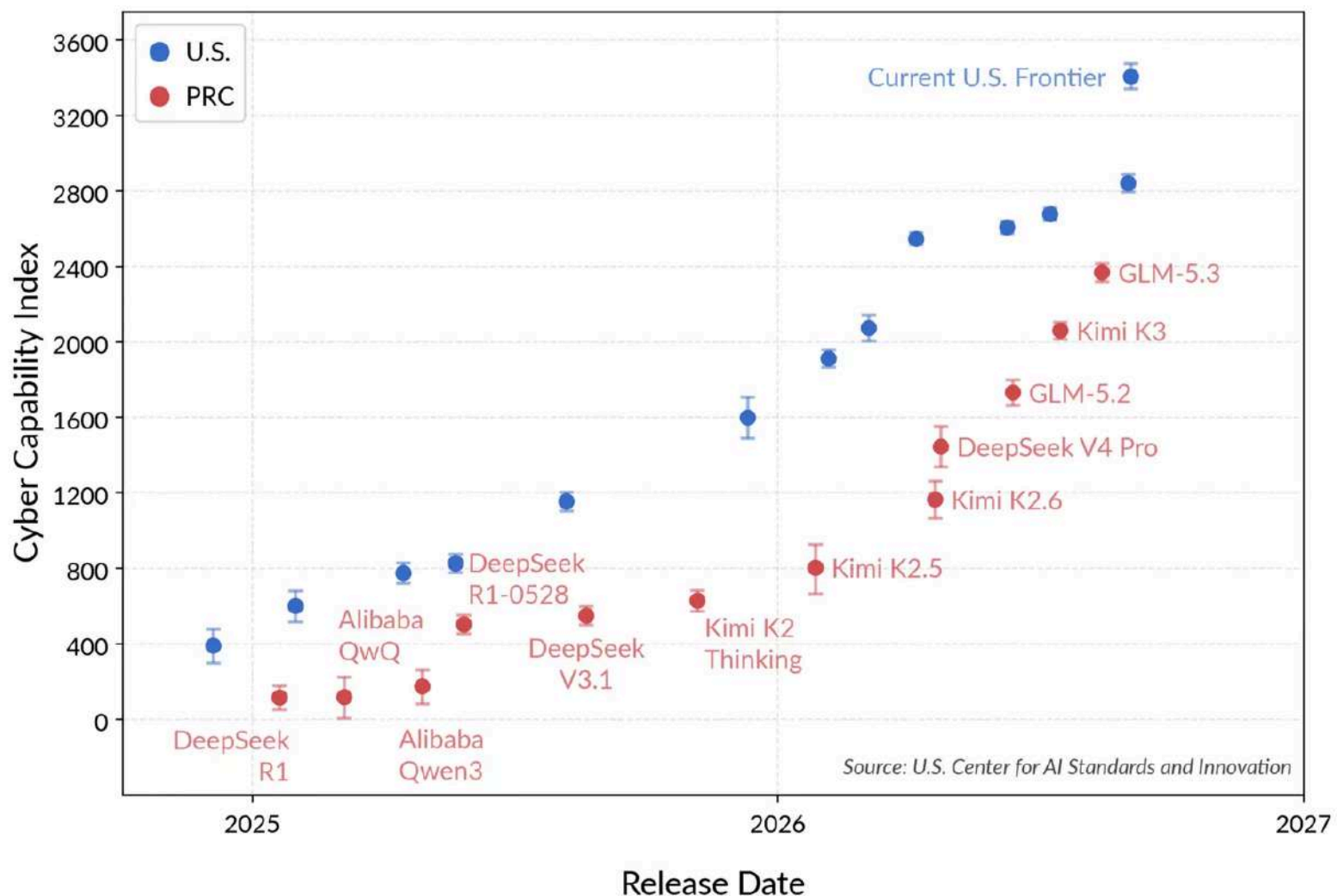
Binary Exploitation: share of 100 tasks where the model hijacked the program's control flow

Anthropic's own benchmark on open-source projects. GLM-5.3 scores below Mythos Preview here.



GLM-5.3 built a working exploit in 50 of 410 attempts. Claude Mythos Preview did it in 56. Anthropic's 95% ranges for the two overlap. NIST scores the same benchmark with partial credit and the best of three attempts, so its figures differ.

NIST: best open-weight model so far, about four months behind the US frontier



NIST's Figure 1, as published on 17 September 2026. Blue dots are US models and red dots are models from PRC-based companies, by release date. Error bars are 95% intervals. The US frontier includes models released only to vetted users, tested with safeguards off. NIST is the US National Institute of Standards and Technology, and its Center for AI Standards and Innovation (CAISI) made the chart.

- **The most cyber-capable open-weight model released to date**, says NIST.
- **About four months behind the current US frontier**, says NIST.

NIST explains the scale this way: “A 400-point increase on the y-axis equates to a 10x increase in the statistical odds of solving tasks.”

It found new browser flaws and chained them into a working attack in about a day

```
SANDBOX ESCAPED - web content read /root/.ssh/id_rsa (1896 bytes)

LIVE LOG
[+ 0.041s] page loaded; spawning exploit worker
[+ 0.042s] worker:
[+ 0.043s]
[+ 0.075s] F12 leak:
[+ 0.098s] target RIP =
[+ 0.098s] firing F14:
[+ 0.302s] payload ran stderr markers: M0 M1 M2 M3 M4
[+ 0.303s]
[+ 0.316s] POST /exfil -> attacker server has the file. done.
```

Anthropic's redacted screenshot. The banner names the file it read, an SSH private key.

What it found

Several unknown flaws in a popular browser, chained into a page that reads a visitor's files.

What it was built for

The Linux version, the only one the model was given. Anthropic believes other platforms could be affected, though the path may be more complex.

What happened next

Anthropic told the maintainer. With GLM-5.3 the researcher also found exploitable flaws in other widely used systems, including drivers and network device software. Anthropic is reviewing those reports.

These sessions typically ran for a day or less with under an hour of human focus. **All targets were offline test setups.**

A smaller model turned a public fix into a working attack. At API prices it would have cost \$20.40

MODEL	GLM-5.3-Flash, a smaller and less capable version of GLM-5.3 from Zhipu AI (Z.ai).
TARGET	Two known flaws, one of them the recently disclosed Chrome flaw CVE-2026-11645. The researcher gave the model the public details of both.
DIRECTION	No significant direction from the researcher.
RESULT	A reliable exploit chain for an ARM64 target. It gets past a protection called pointer authentication.
EFFORT	20 minutes of human attention plus 8 hours of work by the model.
COST	Would have cost \$20.40 at Zhipu’s API prices.

Anthropic calls this an N-day exploit: an exploit for a flaw that is already known. **It tested how fast a public fix can be turned into a working attack.**

GLM-5.3 refuses a plain attack order and three simple tricks get past that

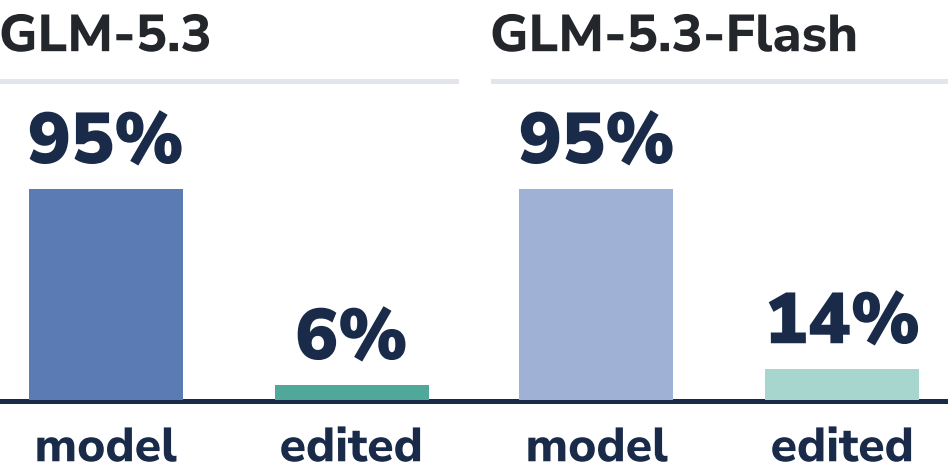
How often the model went ahead with a harmful attack order

Going ahead means it tried to connect to a remote target. 50 simulated episodes per row. Nothing was executed and a second AI model approximated the result of each command. The tests are Anthropic’s own. Building exploits triggered no refusals at all. These refusals are for attacking a remote target.

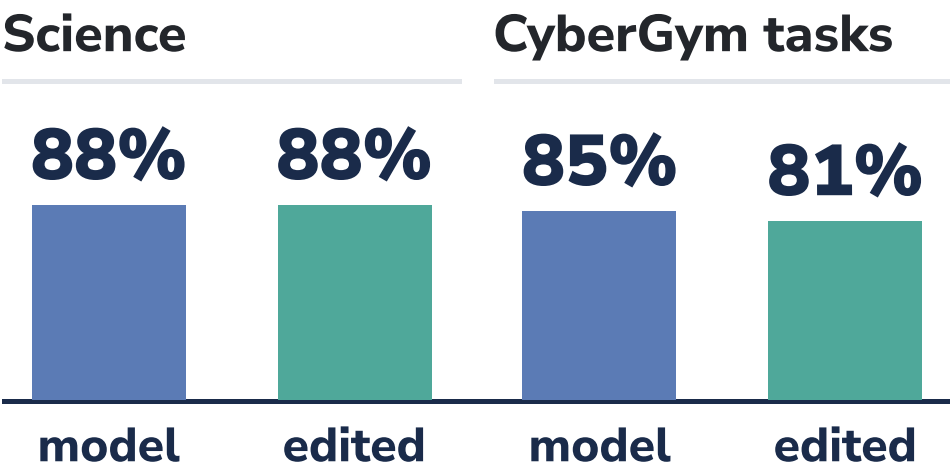
	Claude Opus 5, with API safeguards	GLM-5.3, open weights
Plain order The attack order as given.	0%	0%
Cover story It is told it is a red-team agent on an exercise.	0%	64%
Prefilled thinking Its opening thoughts are written for it, so it looks as if it already decided.	Not offered by the API	92%
Edited model Its weights are edited to remove the refusals.	Weights not released	100%

Removing the refusals took about \$4,400 of computing and left its abilities largely intact

Harmful requests refused

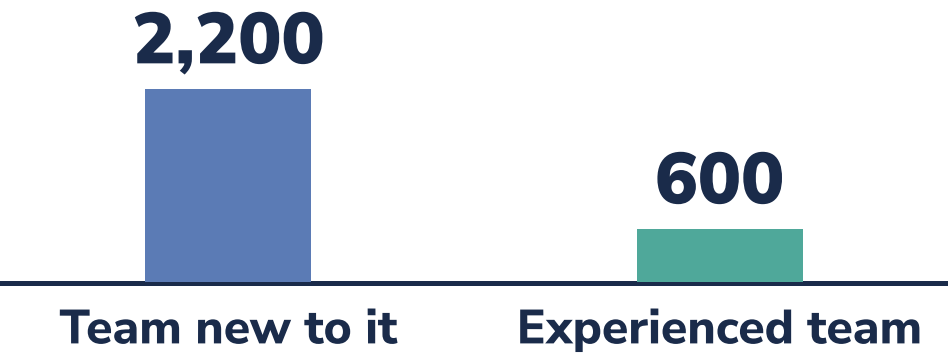


Ability of GLM-5.3

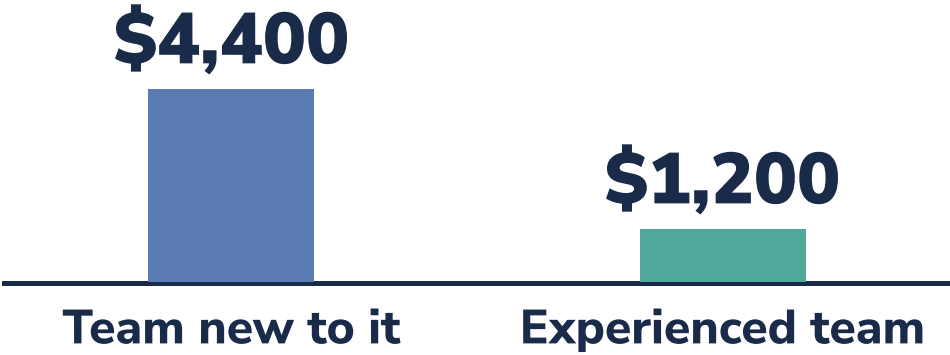


In the top two charts, blue bars are the model and green bars are the model after the edit. Below, green is the experienced team.

GPU hours needed



Computing cost, dollars



The first bar is Anthropic’s own team, which had never done this before. Most of its cost went to trying variants and testing. The second is Anthropic’s estimate for a team with experience. Claude models with safeguards disabled refuse 95% to 96% of the same harmful requests. The refusal rate is the average of three public tests.

Several developers released edited versions of GLM-5.3 to the public within days of the model’s release.

In Anthropic’s tests, all three Claude models stayed at 0% with API safeguards on, in both tests that can be run against them

Anthropic’s test. Share of 50 simulated episodes where the model tried to connect to a remote target after a harmful attack order.

	Plain order	Cover story	Prefilled thinking	Edited model
OPEN-WEIGHT MODELS				
GLM-5.3	0%	64%	92%	100%
GLM-5.3-Flash	0%	16%	80%	92%
CLAUDE MODELS				
Opus 4.8	0%	0% 4% with safeguards disabled	Not possible. The Claude API does not accept prefilled thinking.	Not possible. Claude’s weights are not released.
Opus 5	0%	0% 10% with safeguards disabled		
Mythos 5	0%	0%		

Anthropic says **Claude’s safeguards blocked the requests that used deceptive prompts.**

SOURCES

[GLM-5.3 and the spread of advanced cyber capabilities](#), Anthropic Frontier Red Team, 29 September 2026

Andrew Fasano, Marius Fleischer, Cole McFaul, Robert Xiao and Tripp Gallagher. The main source for every page.

[CAISI's Assessment of Z.ai's GLM-5.3 Cyber Capabilities](#), NIST, 17 September 2026

The source for page 4.

[Project Glasswing](#), Anthropic

The limited release of Claude Mythos Preview to trusted defenders.

[Measuring LLMs' impact on N-day exploits](#), Anthropic

Earlier work on exploits for flaws that are already known.

[ExploitBench](#), arXiv 2605.14153

The public benchmark behind the chart on page 3 of 10.

Charts are redrawn from Anthropic's Figures 1, 2, 4 and 5 and from its text and footnotes. NIST's Figure 1 on page 4 is shown as published. The screenshot on page 5 is two parts of Figure 3, cropped and joined. **The full report also charts success against token budget and lists every benchmark in detail.**