INDEPENDENT SECURITY ASSESSMENT

flyingpenguin

Claude Mythos & Project Glasswing

"A market in which the buyer cannot measure what they bought is no market at all."

ASSESSMENT

The capability claim is **unverified and increasingly contradicted**. Recommended posture: do not pay a premium for, or rely on, Mythos pending the independent July 6 disclosure.

PREPARED BY	Davi Ottenheimer · flyingpenguin.com

DATE	June 8, 2026
VERSION	1.0
DECISION POINT	July 6, 2026. Anthropic's promised public report
CLASSIFICATION	Public · For board and executive discussion

01 Executive summary

BOTTOM LINE

Defer any strategy, procurement, or risk decision that rests on the Claude Mythos capability claim until July 6, 2026, when Anthropic's promised report is due for independent review. The claim is that Mythos is too dangerous to release. The record runs the other way.

Of 23,019 vulnerabilities Mythos reported, it showed only 75 fixes. Holding back the fixes slows validation of findings. Independent teams nonetheless reproduced the showcase findings on commodity and open-weight models for a fraction of the price, among them the engineer who wrote the OpenBSD flaw Anthropic featured at launch. The headline accuracy figure covers the 8 percent of findings a human checked, and the data behind the rest stays withheld. Anthropic widened access to roughly 150 organizations and filed to go public, both ahead of the verification the program promised.

75 / 23,019
findings with a fix shown, of the total Mythos reported

90.6%
accuracy applies only to the 1,752 humans checked, not the full total

$0.11
per million tokens reproduces the showcase find on an open-weight model

5×
the public Opus price, for an exploit layer with no replayable proof

Key findings

- **The headline is self-graded.** Of 23,019 vulnerabilities Mythos reported, 1,752 were verified by a human or security firm, and fixes have been shown for 75.
- **The flagship find was recall, not discovery.** FreeBSD CVE-2026-4747 is a 2007 fix for shared code that was never applied. The fix sat in the model's training data, making the result consistent with recovery from a backlog of delayed fixes.
- **The capability is not exclusive.** Eight of eight open-weight models reproduced the detection, one at $0.11 per million tokens. On June 8, 2026, launch partner Cisco ran six frontier models across 1.8 billion lines of code and showed results do not depend on Mythos.
- **Nothing is reproducible.** No reproduction steps accompany the launch blog, the system card, or the Glasswing update, so the premium claims cannot be independently verified.
- **Scaling is outrunning proof.** Anthropic filed confidentially for an IPO near a one-trillion-dollar valuation and expanded Glasswing to roughly 150 organizations, committing access and capital ahead of verification.

02 Recommendations

1. Treat AI-assisted vulnerability discovery as a commodity input and source it competitively. The showcase results are reproducible at low cost on public models; harness runs should cost cents per million tokens, not tens of dollars. An open-source harness on commodity Haiku 4.5 and Sonnet 4.6

produced eight findings in two minutes for $0.75, two matching the Mythos showcase, at the discovery layer.

2. Do not pay Anthropic a premium or restructure operations on the basis of the Mythos security capability claim until an independent verification exists.

3. Require any AI security vendor to supply reproduction steps and verified, fixed CVEs rather than model-generated finding counts.

4. Set July 6, 2026 as a validation checkpoint, and reassess once the Glasswing report is published and independently reviewed.

> **Decision checkpoint, July 6, 2026.** A report with a verified CVE list and reproduction steps would substantiate the claim. A report that restates model-graded headline figures without independent verification would confirm the pattern in this briefing.

03 Assessment

The flagship "discovery" was backlog recall

CVE-2026-4747 is a valid stack buffer overflow in FreeBSD. The code is a University of Michigan implementation that was patched by MIT in 2007. FreeBSD imported the unpatched code in 2008 and never applied the fix. This 2007 patch is present in the model's training data, so the Mythos exploitation demonstration took an old, vulnerable operating system with a known missing patch and pointed at it. The result demonstrates how a known, undefended target can be flagged by AI, rather than the discovery of anything unknown.

The danger warnings are much thinner than advertised. Mythos did send an email out of its sandbox to flag a bug, but only after being instructed to try; it showed no sign of altering its own weights, and prior models such as Opus 4.6 find these same flaws.

Discovery is reproducible at commodity cost

Independent parties have repeatedly reproduced the showcase findings on inexpensive public models. AISLE confirmed the FreeBSD detection with eight of eight open-weight models, one at $0.11 per million tokens. Vidoc reproduced it on the public Opus 4.6 model and on GPT-5.4. Cisco's June 8 assessment across six frontier models showed the outcome is model-independent. The curl maintainers reported no change to their workflow, and Mozilla's headline of 271 Firefox vulnerabilities reconciles to roughly three against the advisory record.

> "Vulnerability discovery is an orchestration problem, not a frontier-model problem."
>
> **Niels Provos, who wrote the 1998 OpenBSD flaw Anthropic showcased, after reproducing it on commodity and open-weight models**

Provos reproduced that finding and autonomously surfaced new zero-days using Opus 4.6, Sonnet 4.6, and the open-weight GLM 5.1 on his own open-source IronCurtain harness. clearbluejar then ran the

same class of pipeline on two open-weight models on a single consumer GPU and recovered CVE-2026-4747, finding that the scaffolding, not the model, did the hard work.

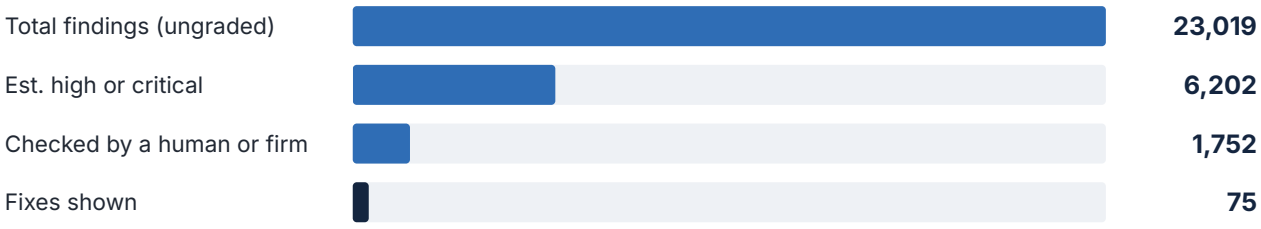
The premium is unjustifiable as presented

Anthropic prices Mythos at roughly five times its public Opus model, from $25 to $125 per million input and output tokens, on the strength of exploit development rather than discovery. With no replayable exploit to confirm, a buyer cannot verify the capability they are paying for, and the available reproductions indicate the defensible cost is a fraction of the quoted price.

Results are self-assessed, and the data is withheld

Anthropic's interim Glasswing update reports results in stages that undermine its own headline.

From 23,019 model findings to 75 shown fixes



The widely cited 90.6% true-positive rate applies to the 1,752 human-checked items, not the 23,019 total.

Beyond that checked slice, the findings are the model assessing its own output. Anthropic has also withheld the fixes used to derive them, the artifacts that would allow independent re-derivation, so the result can be validated only against the system that produced it.

An extractive disclosure structure

The disclosure architecture inverts established norms, and the economics explain why. Anthropic commits up to one hundred million dollars in model credits to a consortium of about a dozen large firms. The consortium attests to the capability that justifies restricting the model to the consortium, and the same firms sell the products and services that follow from that attestation. A rushed "emergency" memo crediting 250 CISOs was curated by security vendors positioned to capitalize on the threat narrative. The most consequential findings continue to come from humans: the Palo Alto vulnerability that triggered a federal mandate was attributed to attackers operating in production and excluded from the company's AI-credited count. Findings flow to Anthropic while fixes fall to volunteer maintainers, even as the patch-generation step a model can automate already runs in production for paying customers. Anthropic's Claude Security product patched more than 2,100 vulnerabilities in three weeks for paying customers, while open-source projects received reports.

Market motivations

On June 1, 2026, Anthropic filed confidentially for an initial public offering following a funding round near a one-trillion-dollar valuation. On June 2, it expanded Glasswing to roughly 150 organizations across more than fifteen countries, covering power, water, healthcare, and communications. Access widened and capital was committed before any independent validation, and before the report Anthropic itself promised.

Several firms now trialling Mythos, including Google, Nvidia, and Cisco, are Anthropic investors, and Goldman Sachs, Morgan Stanley, and JPMorgan are reported to be in talks to underwrite the offering. The parties certifying the capability are the parties whose returns depend on it.

Outlook

Anthropic committed to a public report within ninety days of the April 7 launch, due around July 6, 2026. With each reveal so far, Mythos has failed to substantiate its initial claims.

The prudent posture is to treat their unproven capability as unproven.

Morrell's airship rose about 300 feet, then ripped apart and crashed, shortly after its first launch on May 23, 1908.

04 Sources

flyingpenguin series

1. The Boy That Cried Mythos: Verification is Collapsing Trust in Anthropic, Apr 13, 2026.
2. America Prepares as Anthropic Mythos is 100X More Deadly Than Martian Death Ray, Apr 13, 2026.
3. FreeBSD CVE-2026-4747 Log Suggests Mythos is a Marketing Trick, Apr 14, 2026.
4. Cartel or Not? Anthropic Mythos is a Curious Case, Apr 15, 2026.
5. Ox Security Report: Anthropic MCP is Execute First, Validate Never, Apr 15, 2026.
6. How SANS Mythos Marketing Disappoints Defenders, Apr 16, 2026.
7. Mythos Mystery in Mozilla Numbers: How 22 Vulns Became 271 or Maybe 3 in April, Apr 22, 2026.
8. Alisa Esage Throws Mythos Under Zero Day Bus, Apr 24, 2026.
9. Anthropic Mythos as Valuable as a Firehose in a Blizzard, May 2, 2026.
10. Seventy-Five Cents Gets You an Anthropic Mythos Killer, May 4, 2026.
11. cURL Toe to Toe With Mythos: Big Nothingburger Leaves Bad Taste, May 12, 2026.
12. Palo Alto Defender's Guide Refutes Mythos Claim, May 13, 2026.
13. I'm on Mythos, May 25, 2026.
14. Mythos Grading Mythos: Got Patches Yet?, May 26, 2026.
15. Cisco's Mythos Post Throws Anthropic Under the Bus, Jun 8, 2026.

Anthropic program materials

16. Project Glasswing (program page), Anthropic.
17. Project Glasswing: An initial update, Anthropic, late May 2026. Source of the 23,019 / 6,202 / 1,752 / 90.6% / 75 figures.

Independent reproduction and refutation

18. AISLE reproduction: eight of eight open-weight models detect CVE-2026-4747, one at $0.11 per million tokens. See references 1 and 10.
19. Vidoc reproduction on public Opus 4.6 and GPT-5.4. See reference 10.
20. Nicholas Carlini's confirmation that he found CVE-2026-4747 using Mythos Preview, outside his February 5 paper. See references 3 and 10.
21. Cisco frontier-model assessment, six models across 1.8 billion lines of code. See reference 15.
22. Palo Alto Networks May 2026 Defender's Guide and the CVE-2026-0300 advisory; the federal-mandate CVE attributed to attackers in production, excluded from the AI-credited count. See reference 12.
23. Mozilla Foundation Security Advisory 2026-30 (Firefox 150) and Bobby Holley, "The zero-days are numbered," Mozilla blog, Apr 21, 2026. See reference 7.
24. Claude Mythos Preview system card (244 pages), Anthropic. See reference 1.
25. Finding Zero-Days with Any Model, Niels Provos, Apr 29, 2026.
26. System Over Model, Tested: Reproducing Mythos's FreeBSD Find on Local Open-Weight Models, clearbluejar, Jun 4, 2026.
27. System Over Model: Zero-Day Discovery at the Jagged Frontier, Stanislav Fort, AISLE, Apr 2026.

Press on the June expansion and IPO filing

28. Anthropic scales Claude Mythos to critical infrastructure in 15+ countries, TechCrunch, Jun 2, 2026.
29. Anthropic expanding access to Project Glasswing, CyberScoop, Jun 2026. Source for Claude Security patching 2,100+ vulnerabilities in three weeks.
30. Anthropic expands Mythos to 150 additional organizations in more than 15 countries, CNBC, Jun 2, 2026.
31. Anthropic expands Project Glasswing to 150 organizations in more than 15 countries, Help Net Security, Jun 3, 2026.
32. Experts: Anthropic's move to expand Project Glasswing will end in Mythos public release, Cybernews, Jun 2026.
33. From Anthropic's Mythos to the Birkin bag, scarcity sells, John Foley, Lex, Financial Times, Apr 23, 2026.