

STUDY · CYBERSECURITY AND AI · OCT 7, 2026

Anthropic releases more powerful AI to teams that hunt flaws before hackers do

How the new Cyber Verification Program works: tiers, timeline, numbers and implications

5 TITLE OPTIONS FOR PUBLICATION

- 1 Anthropic releases its most powerful AI to teams that hunt flaws before hackers do
- 2 Hunting flaws before hackers: Anthropic opens advanced AI to verified security teams
- 3 From switch to ladder: how Anthropic gives defenders stronger AI in three tiers
- 4 134,000 flaws found: Anthropic widens AI access for vulnerability hunters
- 5 AI that attacks to defend: who gains and who loses in Anthropic's new program



How this study is organized

Context: Analysis of Anthropic’s Oct 6, 2026 announcement and the Olhar Digital article published the same day.



 **Why it matters in the study sequence**

The order runs from verified fact to interpretation: what was announced, then numbers and dates, then impacts.

 **Positive implication**

Separating fact from analysis lets each block be reused safely in classes, reports or audits.

 **Negative implication / risk**

The program’s numbers are self-reported by the company, so the critical reading closes the study.

The original source and what the article got wrong

Context: The Brazilian article summarizes an announcement Anthropic published in English a few hours earlier.



PRIMARY SOURCE

Anthropic, “Expanding the Cyber Verification Program”

Oct 6, 2026, 19:00 UTC · anthropic.com/news



ANALYZED SOURCE

Rodrigo Mozelli, Olhar Digital (Brazil)

Oct 6, 2026, 22:20 (Brasília) · olhardigital.com.br

THREE INACCURACIES IN THE ARTICLE

1

U.S. government. The article suggests joint government vetting at every tier. In the original, that applies only to Specialized.

2

The 33,000 base. The 33,000+ critical/high findings refer to the combined total (Glasswing + own scanning), not just the 129,000.

3

Specialized list. Telecom networks and government administrative networks are missing from the systems it cites.



Why it matters in the study sequence

Everything that follows depends on a correct factual base, which is why the audit comes first.



Positive implication

The article is faithful on the essentials: tiers, models and numbers match the primary source.

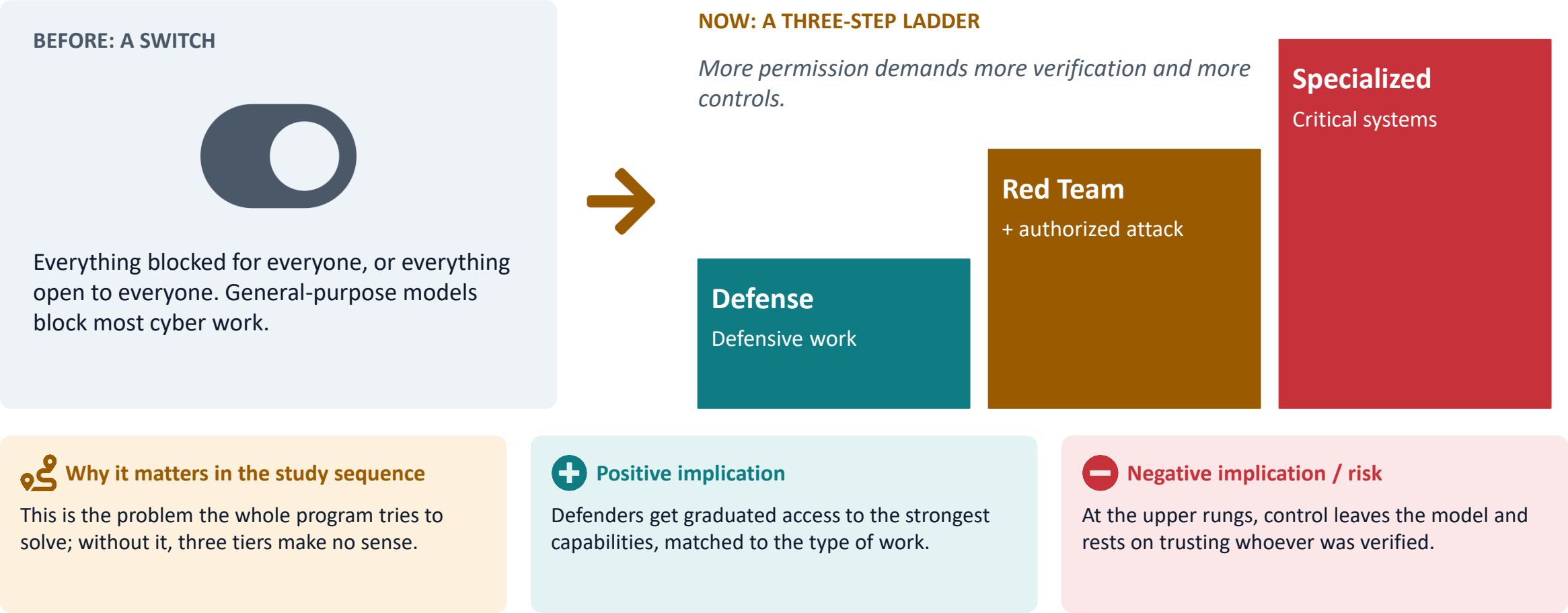


Negative implication / risk

The inaccuracies change the governance reading, especially on the U.S. government’s role.

From a switch to a ladder of verified access

Context: Cybersecurity is dual use: the same capability that finds and fixes a flaw can also exploit it.



Two tracks merged into a single program

Context: For about six months, Anthropic ran two separate trusted-access programs.



Project Glasswing

since Apr 7, 2026

Access to Claude Mythos for organizations that secure the most critical software.



Original CVP

since Apr 16, 2026

Reduced safeguards on Opus and Sonnet for vetted security teams.



New CVP

Oct 6, 2026

- Three access tiers: Defense, Red Team and Specialized
- Covers Opus 5.5, Sonnet 5.5, Mythos 5.1 and future models
- Glasswing members move to Specialized without reapproval



Why it matters in the study sequence

It shows the October CVP is a consolidation of six months of experience, not a from-scratch initiative.



Positive implication

A single path to join, with clear rules by type of work and level of risk.



Negative implication / risk

Glasswing members move over without reapproval on current models: the strictest filter is not reapplied.

Goal: give defenders an edge before attackers get one

Context: On Jun 2, 2026, Anthropic predicted third-party models with comparable cyber capability within 6 to 12 months.



STATED GOAL

Extend Glasswing’s impact to far more defenders and pursue a “permanent advantage” for those who defend systems.



Response time

A few days for Defense; a few weeks for Red Team.



Enterprise Frontier Safeguards

By Dec 2026 (“this fall”): data in the customer’s own cloud.



Transparency

“In the coming weeks”: new results on open source and critical infrastructure.



Continuous improvement

Refine the classifiers for each tier, with no set date.



Why it matters in the study sequence

The goals explain the urgency: fix flaws before offensive capability becomes commonplace.



Positive implication

Dated targets let the company be held to account, such as delivering EFS by the end of 2026.



Negative implication / risk

“Permanent advantage” is an aspiration; no target measures how many flaws actually get fixed.

The three access tiers

Context: Each tier up reduces the model’s blocks and raises the verification and security-control requirements.



Defense Access

Uses: SOC, incident response, malware reverse engineering, vulnerability validation.

Who: Companies, nonprofits, universities, government; critical infrastructure of any size; open source; researchers with a track record.

Response: A few days



Red Team Access

Uses: Everything in Defense, plus pentesting and red teaming on systems with formal authorization.

Who: In-house and government red teams, security and pentest firms. Organizations only.

Response: A few weeks



Specialized Access

Uses: Testing systems critical to lives or markets: aviation, power, telecom, interbank, government.

Who: A restricted group of organizations, reviewed in depth with the U.S. government.

Response: Not stated



Why it matters in the study sequence

This is the operational core of the study: it defines who can do what with the most capable models.



Positive implication

Regional hospitals, municipal utilities and open-source maintainers become eligible.



Negative implication / risk

Individual researchers cannot access Red Team, and the top tier depends on the U.S. government.

What stays blocked at each tier

Context: Even when verified, organizations remain subject to real-time blocks on the highest-harm uses.

Tier	What stays blocked	Vetting	Transition
No CVP	Blocks most cyber work	—	Open use for code review, patching known issues and alert triage
Defense	Complex offensive scenarios	Anthropic review	Entry point for most teams
Red Team	Ransomware, damage to physical systems, pentesting high-risk safety systems	Expanded review	Applicants stay in Defense while waiting
Specialized	Fewest blocks in the program	In-depth review with the U.S. government	Glasswing members enter without reapproval



Why it matters in the study sequence

It shows where the red line sits, even for those who passed verification.



Positive implication

Ransomware and physical damage stay blocked even in Red Team, which allows authorized attack.



Negative implication / risk

Specialized has the fewest blocks: there, safety rests almost entirely on vetting.

The trade-offs that apply to every tier

Context: Wider access comes with monitoring, proof of controls and centralized administration.



Data retention

Required to monitor misuse. Temporary exception: customers on Fable 5.1 or Mythos 5.1 with zero retention.



Platforms

Claude Platform, Google Vertex AI and Microsoft Foundry. Amazon Bedrock only for EFS-eligible customers.



Proof of controls

Every applicant is vetted and must prove the security controls required for the tier.



Appeals

A form to report wrongful blocks and appeal a rejected application.



Workspaces

The admin assigns the program to specific workspaces in the Claude Console.

Outside the program, general models remain allowed for code review, patching known issues, finding flaws in your own code and alert triage.



Why it matters in the study sequence

These are the controls a GRC auditor will need to evidence when assessing use of the program.



Positive implication

Retention lets the company detect abuse and reconstruct what was done with the model.



Negative implication / risk

Mandatory retention clashes with zero-retention policies until EFS arrives.

Eight concepts for reading the program correctly

Context: The announcement mixes information-security vocabulary with AI vocabulary.

Dual use

A capability that serves both defense and attack.

Real-time classifier

A filter that scans requests and blocks what looks offensive.

Routing

A flagged cyber request goes to a less capable model (Opus 5.5 → Opus 4.8).

KYC applied to AI

Verify identity and purpose before unlocking capability.

Authorized scope

Red Team may only test systems with formal permission.

Red team / pentest

Authorized attack simulation to find flaws.

Retention vs. ZDR

Prompts kept for audit versus zero data retention.

CyScenarioBench

Anthropic's test of multi-step cyber operations.

Why it matters in the study sequence

Mastering the vocabulary prevents misreading the numbers and tiers in the steps that follow.

Positive implication

KYC and authorized scope are already familiar to GRC and audit professionals.

Negative implication / risk

Terms like “classifier” hide the fact that no public detail exists on how they work.

The AI models involved

Context: Fable 5.1 and Mythos 5.1 share the same base model; Fable adds extra safeguards.

Model	Release	Role in the cyber context
Claude Mythos 5.1	Sep 1, 2026	Highest cyber capability; once Glasswing-only, now in all three CVP tiers
Claude Fable 5.1	Sep 1, 2026	Same base as Mythos 5.1 with extra safeguards; publicly available
Claude Opus 5.5	Sep 22, 2026	Fable-class safeguards; outside CVP, flagged cyber requests go to Opus 4.8
Claude Sonnet 5.5	Sep 22–Oct 6, 2026*	Included in CVP since the announcement
Claude Opus 4.7	Apr 16, 2026	First to test cyber safeguards; launched alongside the original CVP
Claude Mythos Preview	Apr 7, 2026	Initial version, available only to Glasswing partners

* Exact Sonnet 5.5 release date not confirmed by a reliable source.



Why it matters in the study sequence

Knowing which model sits in each tier explains the capability jump the program unlocks.



Positive implication

Future models join automatically; existing members are reassessed with no new sign-up.

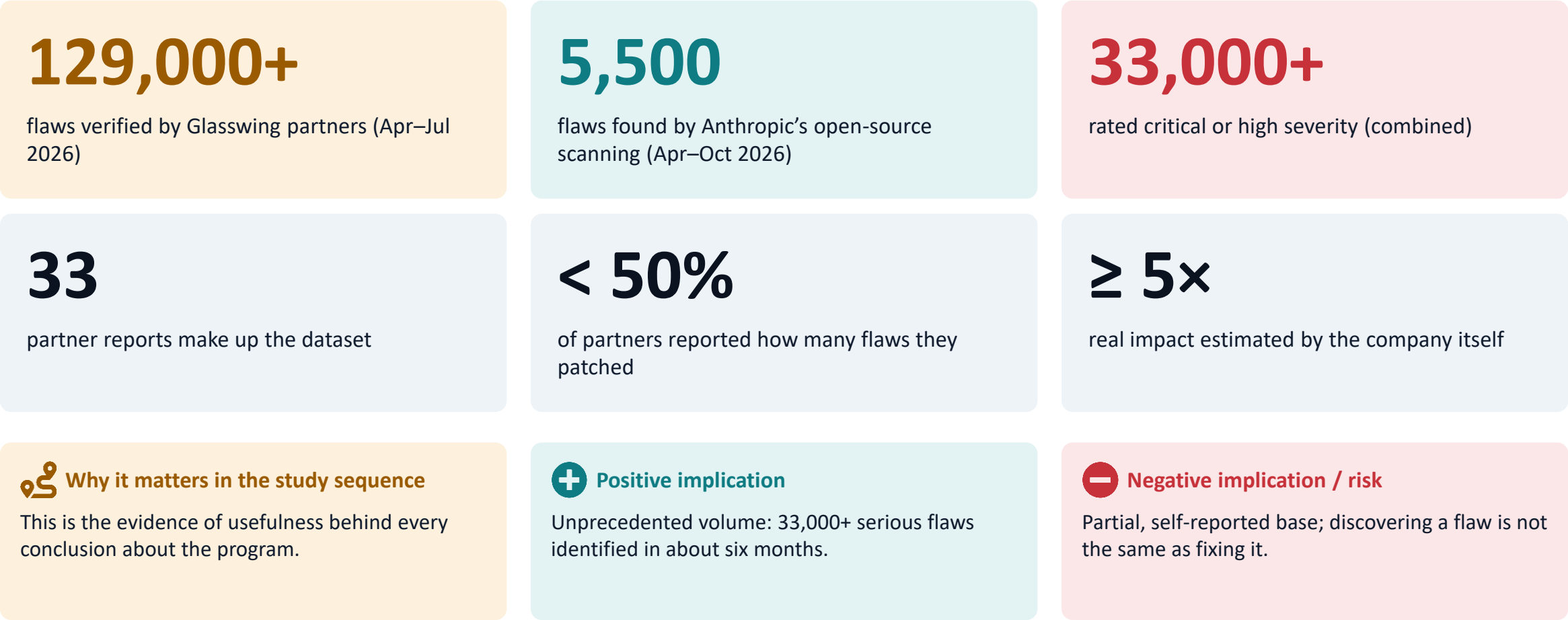


Negative implication / risk

Outside CVP, routing from Opus 5.5 to Opus 4.8 cuts capability for anyone who does not get verified.

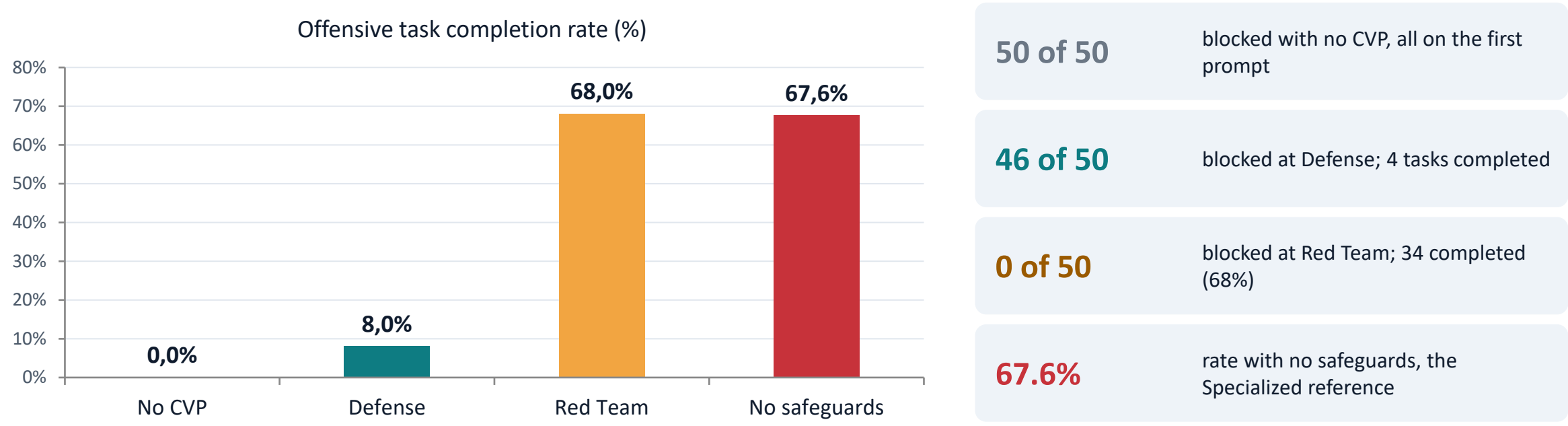
More than 134,000 verified vulnerabilities

Context: Project Glasswing’s results were the main argument for widening the program.



At Red Team, the model behaves as if there were no safeguards

Context: Opus 5.5 on CyScenarioBench: 10 challenges × 5 attempts per tier (50 trials each).



 **Why it matters in the study sequence**

It shows the tiers truly change the model's behavior, not just the contract.

 **Positive implication**

Defense blocked 46 of 50 complex offensive scenarios, as intended.

 **Negative implication / risk**

The benchmark is Anthropic's own, with a small sample and no independent validation.

April to July 2026: restricted launch and turbulence

Context: Mythos was opened gradually, conditioned on safeguards and interrupted by export controls.



Why it matters in the study sequence

It shows the opening was not a single event but a process conditioned on testing and regulation.



Positive implication

Safeguards were first tested on a less capable model, Opus 4.7.



Negative implication / risk

U.S. export controls suspended access for 19 days, from Jun 12 to Jul 1.

August 2026 onward: tools, unification and next milestones

Context: The cycle closes with the programs merged, leaving verifiable milestones for the coming months.



 **Why it matters in the study sequence**

It closes the chronology and points to the next milestones that let us check the promises.

 **Positive implication**

The Sep 22 promise (CVP on Opus 5.5) was kept in 14 days.

 **Negative implication / risk**

The defenders' estimated window of advantage ends between Dec 2026 and Jun 2027.

Four of six targets met or maintained

Context: Comparing what Anthropic promised with what it delivered as of Oct 7, 2026.

Target	Announced	Status	Detail
Keep Mythos out of general access	Apr 7	MAINTAINED	Fable 5.1 is the public version with extra safeguards
Test safeguards on a smaller model	Apr 16	MET	Opus 4.7 and the original CVP
Mythos-class for all “within weeks”	May 28	PARTIAL	Launched Jun 9, suspended Jun 12, restored Jul 1
Scale CVP with Mythos capability	Jun 2 and Aug 21	MET	Delivered Oct 6, 2026
Extend CVP to Opus 5.5	Sep 22	MET	Delivered Oct 6, 2026, 14 days later
Make EFS available	Oct 6	PENDING	Deadline: end of the northern-hemisphere fall (Dec 2026)



Why it matters in the study sequence

The delivery record lets us judge how credible the next promises are.



Positive implication

Most targets were met, several within the announced timeframe.



Negative implication / risk

The broad Mythos release was only partial and EFS does not exist yet.

Who the program was designed for

Context: The program is not aimed at the average user: it requires an organization or a proven track record.



Defensive teams

SOC, incident response and malware analysis at companies, nonprofits, universities and governments.



Critical infrastructure

Operators of any size, from regional hospitals to municipal utilities.



Open source

Maintainers and researchers with a record of responsible disclosure.



Red teams and pentest

In-house, government and specialized firms. Organizations only.



Critical systems

Those who test aviation, power, telecom, interbank and government networks.



Why it matters in the study sequence

It defines who should prepare to apply and which controls they will need to prove.



Positive implication

For the first time, small critical-infrastructure operators are explicitly eligible.



Negative implication / risk

In practice, organizations without mature controls are likely to be left out (analysis).

Who gains from the new program

Context: Mapping the beneficiaries shows whom the program’s design serves.

-  **Small and mid-size defenders** Previously outside Glasswing, now eligible for the Defense tier.
-  **Pentest firms and red teams** A formal framework for authorized offensive use; performance equal to the unsafeguarded model.
-  **Anthropic** Turns reputational risk into a governed product, distributed via Google and Microsoft.
-  **U.S. government** Becomes the gatekeeper at the program’s most sensitive tier.
-  **Open-source ecosystem** Receives scans and credits, such as the \$35 million fund.

 **Why it matters in the study sequence**
Naming the winners before the losers helps gauge how balanced the design is.

 **Positive implication**
Defense scales: capability once limited to a few hundred organizations reaches far more teams.

 **Negative implication / risk**
The gains of Anthropic and the U.S. government concentrate power over who accesses the strongest AI.

Who loses or is exposed

Context: Every access design creates people left out and new points of dependency.

 Attackers (for now)	They lose ground only while the 6-to-12-month window lasts.
 Independent researchers	Shut out of Red Team; hurts bug bounty culture and the lone researcher.
 Organizations outside the U.S.	Specialized depends on review with the U.S. government (analysis).
 Zero-retention customers	Mandatory retention clashes with sector rules until EFS arrives.
 Unverified users	Cyber requests on Opus 5.5 fall back to the older Opus 4.8.
 Competitors without vetting	Pressured to build an equivalent program.
 Why it matters in the study sequence	It completes the balance of effects before the technical and governance implications.
 Positive implication	Excluding individuals from Red Team lowers the risk of abuse by untraceable actors.
 Negative implication / risk	It creates an access gap between countries and between large and small organizations.

Control shifts from the model to identity

Context: At Red Team the classifier all but switches off; safety depends on who gets in and on monitoring.



Identity as the control

Banking-style KYC applied to AI: vetting up front, monitoring afterward.



Remediation bottleneck

Finding flaws became far faster than fixing them. Likely effect: CVE backlogs and overloaded maintainers.



Proprietary benchmark

CyScenarioBench is Anthropic's own and has no independent validation yet.



Why it matters in the study sequence

It translates the earlier numbers into practical consequences for those who run security.



Positive implication

More flaws found early means fewer windows criminals can exploit.



Negative implication / risk

Without matching fix capacity, accelerated discovery becomes accumulated liability.

What governance needs to audit

Context: Organizations in the program take on control obligations that must be documented and evidenced.

AUDIT CHECKLIST (THIS ANALYSIS' SUGGESTION)

- ✓ Authorized scope documented for each system tested
- ✓ Segregation of the workspaces enabled in the Claude Console
- ✓ Evidence of the controls required for the tier
- ✓ Legal basis and contracts for data retention
- ✓ Log of blocks, appeals and disputes



Limited transparency

The company has not disclosed how it vets applicants or how it detects abuse at Red Team.



Robustness still to be shown

One individual account of the original CVP (Opus 4.7 era) points to weak vetting. It is isolated and predates the revamp.



Why it matters in the study sequence

It turns the analysis into concrete action for risk, compliance and audit professionals.



Positive implication

Well-defined controls create an audit trail for a use that was previously invisible.



Negative implication / risk

Without public vetting criteria, the audit depends on the vendor's word.

Brazil: data-protection law and digital sovereignty

Context: This note's analysis of how the program's design affects Brazilian organizations.



LGPD (Arts. 33–36)

Mandatory prompt retention implies international data transfer and requires a legal basis and contractual clauses.



Finance and healthcare

Banks (CMN Resolution 4,893/2021) and healthcare face extra cloud-contracting and confidentiality requirements.



Digital sovereignty

Brazilian power, telecom and payment operators would depend on a review involving the U.S. government for the top tier.



Why it matters in the study sequence

It brings the global analysis closer to the regulatory reality of readers in Brazil.



Positive implication

The Defense tier can speed up security at Brazilian hospitals and public services.



Negative implication / risk

Dependence on a foreign decision for access to the most critical capabilities (analysis).

A new market standard and a race against the clock

Context: The CVP sets a benchmark that competitors are already following.



Standardization

Tiered, verification-based access is likely to become the industry standard for dual-use capabilities.



Competition

OpenAI already runs something similar: GPT-5.4-Cyber via its Trusted Access for Cyber program.



Race against the clock

Success depends on fixing faster than the capability spreads. Window: Dec 2026 to Jun 2027.



Why it matters in the study sequence

It projects the study into the next 6 to 12 months, the horizon where the thesis will be tested.



Positive implication

Competition between programs can raise the vetting standard across the whole market.



Negative implication / risk

If another lab ships similar capability without safeguards, the defenders' edge shrinks.

Five reasons to read the numbers with caution

Context: The results support the program's usefulness, but they are partial, self-reported and unaudited.

- 1 Self-reported metrics** 129,000 comes from 33 partial reports; the 5× factor is the company's estimate.
- 2 Finding ≠ fixing** Fewer than half of the partners reported patch numbers.
- 3 Conflict of interest** The same company builds the benchmark, measures the result and sells the access.
- 4 Small sample** 34 of 50 (68%) versus 67.6%: a reasonable match with only 50 trials.
- 5 Non-public criteria** It is unknown how applicants are assessed or how abuse is detected.



Why it matters in the study sequence

It closes the study by separating evidence from the company's narrative.



Positive implication

Anthropic itself discloses the data's limitations, which allows a critical reading.



Negative implication / risk

Without independent audit, conclusions rest on trusting the vendor.

Identity, scope and audit trail are the new perimeter

The new CVP is a governance bet: give verified defenders real offensive capability so they fix flaws before that capability becomes commonplace.



What changed

Glasswing and CVP became three tiers: Defense, Red Team and Specialized, with Mythos 5.1 included.



Why

134,500 verified flaws and 33,000+ serious ones in six months, according to Anthropic.



The risk

At Red Team the model runs with no blocks; everything rests on vetting and monitoring.



The deadline

Defenders' edge lasts until rivals ship similar capability: Dec 2026 to Jun 2027.



Why it matters in the study sequence

It brings source, mechanics, numbers, dates and impacts together in a single GRC reading.



Positive implication

Defenders gain unprecedented, governed access to the strongest AI capabilities.



Negative implication / risk

Concentrated power, self-reported data and a short window are the points to watch.

Sources consulted and audited

All sources accessed on Oct 7, 2026. The primary source and the analyzed article were read in full; full links are in the accompanying analysis document.

PRIMARY SOURCE

- ANTHROPIC. Expanding the Cyber Verification Program. Oct 6, 2026. anthropic.com/news/cyber-verification-program

ANALYZED SOURCE

- MOZELLI, Rodrigo. Anthropic libera IA mais poderosa para equipes que caçam falhas antes de hackers. Olhar Digital, Oct 6, 2026.

ANNOUNCEMENT COVERAGE (Oct 6–7, 2026)

- SiliconANGLE. Anthropic folds Project Glasswing into an expanded three-tier Cyber Verification Program.
- Cybersecurity News. Anthropic Opens Claude Access to Red Teams and Verified Cyber Defenders.
- GBHackers. Anthropic Opens Claude’s Advanced Cyber Capabilities to Verified Red Teams.
- CyberPress. Anthropic Opens Claude Access to Red Teams and Verified Cyber Defenders.
- Inside AI News. Anthropic Expands Cyber Verification Program for Claude Access.

HISTORICAL CONTEXT

- Constellation Research. Anthropic launches Project Glasswing, Claude Mythos Preview. Apr 7, 2026.
- Help Net Security. Anthropic releases Claude Opus 4.7 with automated cybersecurity safeguards. Apr 16, 2026.
- IDM (citing the Financial Times). Project Glasswing expansion. Jun 2026.
- Bloomberg Law. Anthropic Makes Claude Mythos 5 Available for Security Scans. Aug 21, 2026.
- ThaiCERT (citing SecurityWeek). Anthropic Expands Access to Mythos 5. Aug 25, 2026.
- VKTR and SQ Magazine. Claude Opus 5.5 launch. Sep 22, 2026.
- ANTHROPIC. Statement on the suspension and restoration of Fable and Mythos access. Jul 2026.
- Wikipedia. Claude Mythos (secondary source, used with caution).

COMPETITION AND ACCOUNTS

- Netzwoche. GPT-5.4-Cyber and Trusted Access for Cyber (OpenAI). Apr 2026.
- HELTON, Graham. I’m Sorry Dave... Personal blog (individual account).