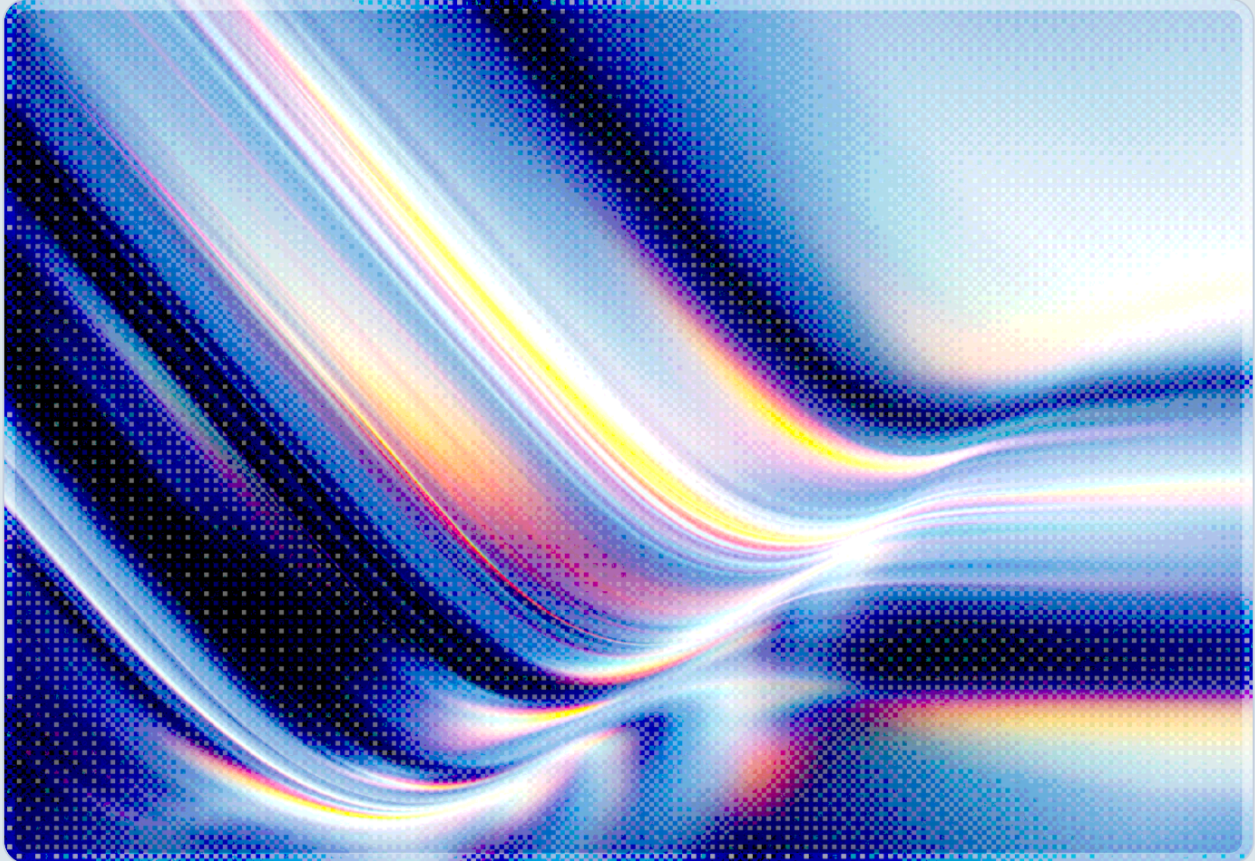


Tiered Trust for Dual-Use AI: Anthropic's Cyber Verification Tiers

Security Implications and Guidance for the Emerging Access-Governance Model

2026-10-08

 AI-assisted Rapid Research



CSAI

cloud
security
alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

On October 6, 2026, Anthropic announced a restructured Cyber Verification Program (CVP) that folds Project Glasswing into a single offering with three access tiers: Defense Access, Red Team Access, and Specialized Access [1]. Each tier trades a higher degree of permitted dual-use capability for a more demanding vetting process, and the highest tier is reviewed in collaboration with the US government [1] [2]. The structure extends what was an organization-level exemption into a graded set of decisions about who is asking, for what work, and under what controls.

For security teams, the practical effect is that access to the most capable cyber-relevant models now depends on organizational posture as well as on the content of a prompt. Applicants must prove identity, demonstrate security controls, and accept monitoring, and the stricter tiers reportedly add authentication and credential requirements that a typical enterprise may not yet meet [1][3]. Readers should note that these operational details rest on a single secondary source and have not been confirmed in Anthropic's announcement. OpenAI's Trusted Access for Cyber program uses a comparable identity-based tiering approach, which suggests that identity-based tiering may be emerging as a shared pattern among at least two leading frontier providers rather than a single vendor's policy [4].

The model improves the odds that dangerous capability reaches defenders, but it also creates new dependencies and new attack surface. Verified accounts become high-value targets, vendor decisions about eligibility become a supply-chain variable, and the monitoring that makes the scheme workable raises data-retention questions. Organizations should treat verified-tier credentials as privileged access, plan for the possibility of tier loss or suspension, and align their own internal access to dual-use AI with the same graduated logic.

Background

Anthropic's real-time cyber safeguards block prohibited uses such as ransomware development and mass data exfiltration, and they also block a band of high-risk dual-use activity such as vulnerability exploitation and offensive security tooling by default. The CVP exists to lift the second category of blocks for vetted organizations doing legitimate defensive work. According to a third-party description of the program, before the October announcement it was application-based, with an authorized administrator verifying identity and describing the defensive use case, and approval was scoped to a

single organization ID [5]. Project Glasswing operated in parallel as a more restricted channel for Anthropic's most capable cyber-relevant models, and the October announcement describes its consolidation into the program [1][2].

The October 6 announcement merges those channels. According to Anthropic, all three tiers cover Claude Opus 5.5, Claude Sonnet 5.5, and Claude Mythos 5.1, along with future models [1][2]. The tiers differ in who may apply, how long review takes, and which restrictions remain in force, as summarized below.

Tier	Eligible applicants	Permitted work	Review time	Remaining blocks
Defense Access	Security teams at companies, nonprofits, universities and governments; critical infrastructure operators; small security firms; open-source maintainers; individual researchers with a vulnerability disclosure history	SOC and incident response, malware reverse engineering, vulnerability analysis and validation	"A few days"	Offensive activity beyond defensive analysis
Red Team Access	Organizations only: in-house and government red teams, penetration testing firms	Authorized penetration testing and red teaming on systems the organization may test	"A few weeks"	Actions causing physical harm or mass disruption, such as ransomware

Tier	Eligible applicants	Permitted work	Review time	Remaining blocks
Specialized Access	A small number of verified organizations	Testing of safety-critical systems such as flight systems, power grids, telecommunications, and interbank transfer infrastructure	In-depth review with the US government	Minimal cyber blocks

Source: Anthropic [1]; SecurityWeek [2].

Two further details shape the program. First, participants must accept data retention so that Anthropic can monitor for misuse. Anthropic has said a zero-retention option will arrive through Enterprise Frontier Safeguards (EFS), which it announced on September 1, 2026 and which is slated to roll out in phases later this fall [1][6]. Under EFS, monitoring data is stored in a customer-controlled cloud account under the customer's own encryption keys [6]. Second, existing Glasswing participants transition into the program [1], and SecurityWeek reports that they move to Specialized Access [2]. Anthropic reports that Glasswing partners found more than 129,000 verified vulnerabilities between April and July, with more than 33,000 rated critical or high severity, though the figures rest on partial partner reporting that Anthropic itself describes as an undercount [1].

Secondary reporting adds operational specifics that the primary announcement, as retrieved, does not confirm. One outlet describes phishing-resistant multi-factor authentication, short-lived credentials in place of fixed API keys, outbound allowlisting, 24-hour credential revocation, organization-managed devices, a December 15, 2026 deadline for Defense Access, and a default of 25 seats for the upper tiers [3]. These details come from a single secondary source [3] and have not been confirmed in Anthropic's announcement, so readers should verify them against the published program terms before relying on them. The same source describes the December 15 deadline as applying to Defense Access users moving to phishing-resistant authentication, and the seat count as a default that can be raised on request.

The access-governance trend is not unique to Anthropic. OpenAI's Trusted Access for Cyber program verifies identity through government ID and know-your-customer checks supplemented by trust signals such as device health, offers higher tiers with more permissive models under stricter controls, and has been reported to require hardware-backed passkeys for individual members from September 1, 2026 [4]. Meanwhile, CSA's earlier analyses documented the governmental dimension of this shift: a government-

directed restriction on the rollout of OpenAI's GPT-5.6 Sol [7] and the June 2026 suspension of access to Anthropic's Fable 5 and Mythos 5 models for foreign nationals, which Anthropic could not enforce per session and therefore applied globally [8].

Security Analysis

Why graduated access addresses a control gap in content-based filtering

Content-based filtering struggles with dual-use security work because the same exploit request can be a penetration test or an intrusion. A classifier sees text; it does not see authorization. Tiered programs move part of the decision from the request to the requester, which may be a more reliable basis for judgment when the requester can be identified and held accountable. This is the logic that identity-centric security has long applied elsewhere, and it fits the Zero Trust premise that access should be granted on verified identity and context rather than on network position or assertion alone [9].

The structure also gives defenders something that blanket restriction does not. The Glasswing results Anthropic cites suggest that capable models can surface large volumes of vulnerabilities, though without a baseline the degree of acceleration is not established [1]. A program that routes that capability to defenders under controls offers a different risk trade-off than withholding it from everyone, and CSA's assessment is that it better serves defenders, subject to an evidence caveat: the figures are vendor-reported and partial, so the magnitude of the benefit should be treated as indicative.

Verified accounts become high-value targets

Once a credential unlocks reduced safeguards, its theft is worth more to an attacker than theft of an ordinary API key. A compromised Red Team or Specialized credential could give an adversary permissive offensive capability under a trusted organization's name. The reported authentication requirements, if confirmed, would be consistent with the elevated value of such credentials, including phishing-resistant authentication and short-lived credentials [3], and they parallel the hardware-backed passkey requirement reported for OpenAI's program [4]. Organizations enrolling in any tier should expect that the verified identity, not the model, is the asset an adversary will pursue. The risk is compounded by developer-workstation exposure: API keys embedded in scripts, CI pipelines, and agent configurations are common sources of leakage, and a leaked verified-tier key inherits the elevated trust.

Insider and scope-creep risk inside verified organizations

Approval attaches to an organization, and scope is bounded by authorization to test particular systems [1][5]. Within that boundary, the vendor cannot see whether a given engagement is actually authorized. A contractor, an employee acting outside a statement of work, or an attacker who has compromised a security team's tooling can all use the permitted capability against targets the organization has no right to test. On the reported design, scope violations would appear to be handled mainly through monitoring and retention rather than technical prevention, though Anthropic has not described its enforcement in detail. Organizations should therefore not assume that vendor-side monitoring substitutes for their own authorization records, rules of engagement, and activity auditing.

Data retention, monitoring, and confidentiality

The retention requirement creates tension for exactly the organizations the program targets. Security teams handle unpatched vulnerability details, incident forensics, and customer environment data, and a required retention window places that material with a third party. According to Unite.AI, enterprise customers had raised concerns about a 30-day retention requirement that Anthropic has applied to its most advanced models since June, and Anthropic presented EFS as addressing them [6]. EFS is described as shifting custody of monitoring data to the customer while preserving automated detection of serious misuse, which would improve confidentiality if it operates as described [6]. It also shifts responsibility: the customer now owns the security of a store containing sensitive prompts and outputs, which makes that bucket and its key management a target in its own right. Reporting on the transition period is mixed: Anthropic describes a phased EFS rollout, while Unite.AI reports that eligible customers receive zero data retention until full rollout [6], so participants should confirm which retention terms apply to them under [1].

Dependency and concentration effects

A tiered program makes access a revocable privilege. Anthropic decides eligibility, review speed, and which restrictions persist, and for the top tier the US government participates in vetting [1][2]. The June 2026 suspension episode showed how quickly frontier access can change when governments act [8], and the GPT-5.6 Sol rollout showed that a government-approved-partner model of access has at least one recent precedent [7]. Security operations that become dependent on a verified tier inherit that fragility. A tier downgrade, a failed re-review, or a policy change could remove capability from an incident-response workflow at the moment it is needed. Review timelines of days to weeks [1] also mean that onboarding cannot be improvised during an incident.

Fairness and coverage

The eligibility rules include small security firms, open-source maintainers, and individual researchers at the Defense tier [1], which broadens access beyond large enterprises. Red Team Access is limited to organizations and excludes individuals [1][2], and Specialized Access reaches only a small number of organizations. Individual researchers are not eligible for Red Team Access under the published criteria, which, as CSA's inference, may leave some offensive-testing work to less controlled tools. Whether the program narrows or widens the defender-attacker capability gap will depend on how many legitimate defenders clear review and how quickly, and Anthropic's announcement, as retrieved, does not report approval or denial rates [1].

Recommendations

Immediate Actions

Organizations that use Claude for security work should determine which tier, if any, matches their activities, and check whether existing CVP or Glasswing approval carries into the new structure. Confirm the authentication and credential requirements that apply to the chosen tier directly from Anthropic's program terms, because the details reported by secondary sources are unconfirmed. Inventory every API key and agent credential associated with a verified organization and move them into a secrets manager with rotation. Establish that only named, accountable staff can invoke permissive-tier capability, and record the authorization basis for each engagement.

Short-Term Mitigations

Place verified-tier access behind phishing-resistant authentication and short-lived credentials even where the vendor does not yet require them. Log AI-assisted security activity internally so that your own records can show scope, target authorization, and operator, independent of vendor monitoring. Evaluate Enterprise Frontier Safeguards once available, and plan the cloud storage, key management, and access policies for the monitoring data before enabling it. Review contracts and data-handling terms for retention of vulnerability and incident data, and classify what may be sent to a model under each tier.

Strategic Considerations

Build resilience against tier loss by keeping non-AI and alternative-vendor fallbacks for incident response, and by testing workflows without permissive model access. Apply the same graduated logic to internal use of dual-use models: separate benign, dual-use, and offensive workloads, tie each to verified identity and documented authorization, and review privileges periodically. Track the convergence of vendor programs, since differing identity requirements across providers may impose compounding administrative cost, and engage with standards and industry bodies on interoperable attestations of defender status. Finally, treat vendor eligibility decisions and government involvement as inputs to third-party and concentration risk assessments.

CSA Resource Alignment

CSA's research note *Government-Gated AI: GPT-5.6 Sol's Dual-Use Cybersecurity Implications* [7] is the most directly related prior work. It analyzed an access model in which a government asked a vendor to restrict a high-capability cyber model to approved partners, and it offered governance guidance for organizations deploying such models. The tiered Anthropic program is a vendor-operated counterpart to that government-gated pattern, and, in CSA's view, the Specialized Access tier, vetted with the US government, sits where the two approaches meet. The enterprise governance recommendations in that note apply to organizations deciding how to use gated models internally.

CSA's white paper *Sovereign AI Access Controls and Enterprise Concentration Risk* [8] addresses the dependency dimension discussed above. It treats frontier AI access as a regulated commodity and derives supply-chain resilience lessons from the June 2026 suspension of Fable 5 and Mythos 5. Those lessons apply directly to security teams that adopt verified-tier access, since tier status is a condition the vendor or a government can change.

For the identity controls that make tiered trust workable, CSA's *Zero Trust Principles and Guidance for Identity and Access Management (IAM)* [9] provides technology-agnostic guidance on verifying identity continuously and limiting privilege, which maps onto the credential hardening and scoped authorization recommended here. Organizations assessing the controls around AI-assisted security work can also use the AI Controls Matrix (AICM) v1.1 as the reference control set, particularly its identity and access management and threat and vulnerability management domains [10].

References

- [1] Anthropic. "[Cyber Verification Program Expansion](#)." Anthropic, October 6, 2026.
- [2] SecurityWeek. "[Anthropic Introduces 3-Tier Cyber Verification Program for AI Access](#)." SecurityWeek, October 2026.
- [3] XenoSpectrum. "[Anthropic Expands Access to Mythos Through a Three-Tier Verification Program, With Stricter Authentication and Log Controls](#)." XenoSpectrum, October 2026.
- [4] Biometric Update. "[OpenAI Requires Hardware-Backed Passkeys for Trusted Cyber Access](#)." Biometric Update, July 2026.
- [5] Ironscales. "[Anthropic Cyber Verification Program](#)." Ironscales Glossary, 2026.
- [6] Unite.AI. "[Anthropic Announces Enterprise Frontier Safeguards, Customer-Held Data](#)." Unite.AI, September 2026.
- [7] Cloud Security Alliance AI Safety Initiative. "[Government-Gated AI: GPT-5.6 Sol's Dual-Use Cybersecurity Implications](#)." CSA, June 28, 2026.
- [8] Cloud Security Alliance AI Safety Initiative. "[Sovereign AI Access Controls and Enterprise Concentration Risk](#)." CSA, 2026.
- [9] Cloud Security Alliance. "[Zero Trust Principles and Guidance for Identity and Access Management \(IAM\)](#)." CSA, 2024.
- [10] Cloud Security Alliance. "[AI Controls Matrix](#)." CSA, 2025.