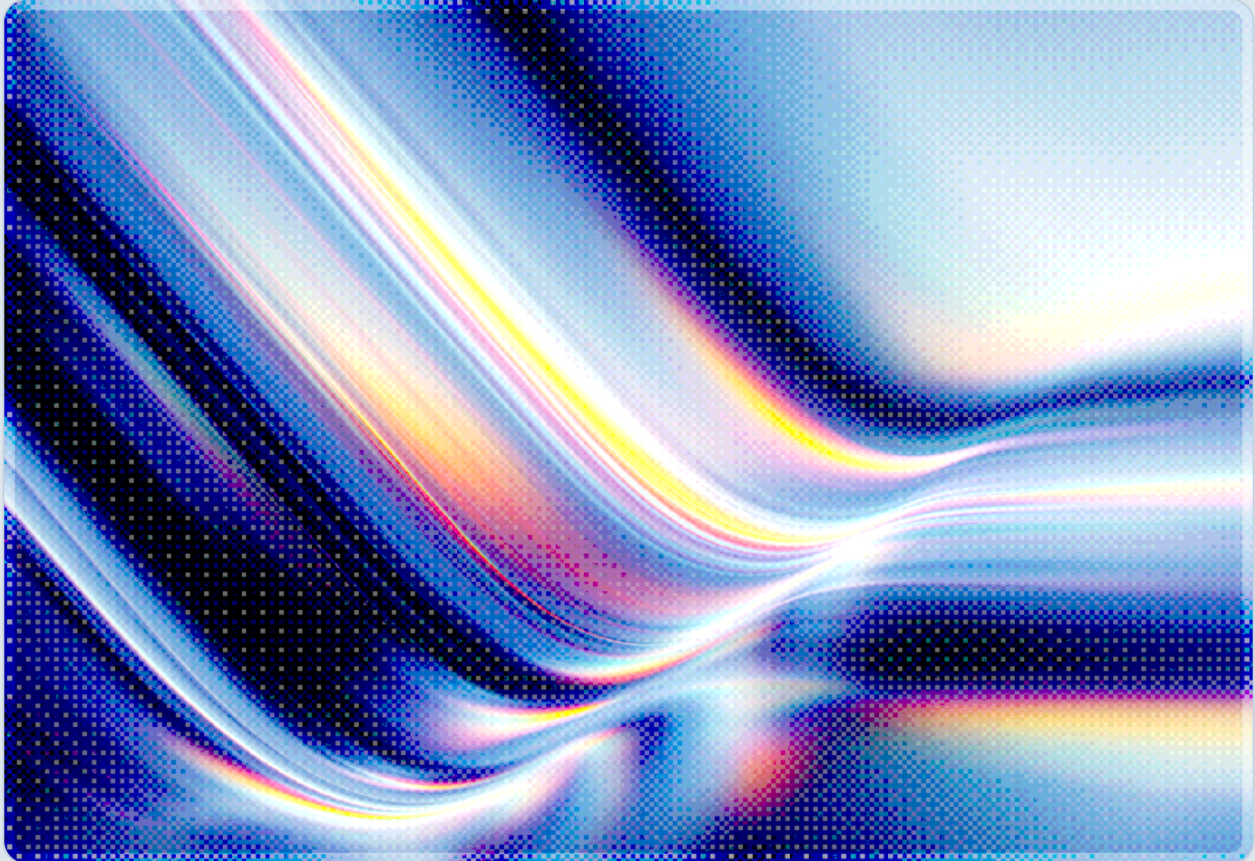


Gated Exploitation: Frontier Cyber Capability as a Lab-Controlled Privilege

Security Implications and Guidance for Organizations Inside and Outside Vetted Access Tiers

2026-10-30

 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

Since early 2026, and especially since August, the most capable cyber functions of frontier AI models have moved behind tiered, identity-bound access programs run by the labs that build them. Anthropic's restructured Cyber Verification Program (October 6, 2026), OpenAI's Daybreak Blue and Daybreak Red tiers, and Google's Fairwind Program now decide which organizations may use a model to develop exploits, chain attacks, or test safety-critical systems [1][4][7]. The result is a new category of privilege: whether an organization can use a capability now depends on a grant from the lab, and the model's technical capability no longer settles the question.

Access has become a privilege that can be granted, scoped, monitored, and revoked. Labs now verify identity, match tiers to the type of authorized activity, require data retention for misuse monitoring, and, at the highest tier, review applicants with the U.S. government [1]. This makes vetted accounts high-value targets, because a credential that unlocks exploit development with minimal classifier blocking is worth more to an attacker than a typical SaaS login. OpenAI's hardware security key mandate for Daybreak participants suggests OpenAI sees the same risk, and Google's MFA requirement points the same way [4][7].

Revocation is also a demonstrated operational risk. The June 2026 export-control suspension and an August 2026 de-verification of OpenAI program participants, which OpenAI attributed to a technical issue, both cut off access with no warning [9][10][11]. At the same time, any given gated capability is temporary. Open-weight models trail closed frontier models on offensive cyber tasks by roughly four to seven months, down from six to ten months during 2025, according to CSA's analysis of UK AI Security Institute findings [17][22]. Gating delays access to each new capability level, but it does not keep that capability scarce indefinitely.

For these reasons, enterprises should manage lab-granted access the way they manage other privileged credentials. That means named ownership, hardened authentication, documented scope, continuity planning, and contractual visibility into which tier their security vendors use against their systems.

Background

Gated access for cyber capability began as an exception. In February 2026, OpenAI introduced Trusted Access for Cyber as a pilot. Under that program, verified individuals and enterprise teams could reach more permissive models, while baseline safeguards were raised for everyone else [3]. On April 7, Anthropic launched Project Glasswing, which gave Claude Mythos Preview to a small group of launch partners and critical-software maintainers for defensive vulnerability work [18][26]. In June it extended access to roughly 150 more organizations, and NATO and ENISA were among the reported new participants [2]. At that stage these programs looked like early-access arrangements for well-known partners, not a lasting way to distribute capability.

Over the summer and early autumn the arrangement hardened into a structure. On August 10, OpenAI expanded its trusted-access effort under the Daybreak name and split it into two tiers. Daybreak Blue gives approved defenders general-purpose frontier models with adjusted safeguards, while Daybreak Red offers purpose-trained cyber models, including GPT-5.6-Cyber, for higher-risk work such as exploit validation. Individual participants were required to use hardware security keys from September 1 [4][20]. Also on September 1, OpenAI determined that GPT-6 Astra had crossed the "Critical" cybersecurity tier of its Preparedness Framework, the first of its models to do so. OpenAI launched Astra on September 3. The public release declines advanced offensive tasks such as writing proof-of-concept exploits, and the less-restricted capability is routed to vetted defenders [5]. The same day, OpenAI announced a \$1 billion commitment of credits, access, and training for under-resourced defenders such as water utilities and community banks [6][21].

Google followed a parallel path. It launched the Fairwind Program on September 2 with Gemini 3.8 Flash Cyber and more than 650 partners. Participation is limited to staff on internal security, incident response, or penetration testing teams, and participants must deploy controls such as multi-factor authentication [7]. On September 30, Google began rolling out Gemini 4 Argon, giving vetted defenders and internal teams a version without cyber-specific restrictions before any broader release [8].

On October 6, Anthropic merged Project Glasswing into a restructured Cyber Verification Program (CVP) with three tiers: Defense Access, Red Team Access, and Specialized Access. Each tier covers Claude Opus 5.5, Sonnet 5.5, and Mythos 5.1. The tiers differ in which classifiers block requests and in who may apply [1][23]. Anthropic's own testing shows how much the tier matters. In its CyScenarioBench trials with Opus 5.5, every task was blocked on the first prompt without CVP enrollment. Under Defense Access, 46 of 50 trials were blocked at some point. Under Red Team Access, the model completed 34 of 50 tasks with no blocks, close to the 67.6 percent success rate Anthropic reports for the model with

safeguards removed [1]. In Anthropic's testing the model was the same in all three conditions, so the difference reflects the classifier configuration attached to each tier. These vendor-reported results have not been independently replicated.

Governments are also involved in deciding who gets through the gate. A June 2, 2026 executive order set up a voluntary framework under which developers may give the federal government up to 30 days of pre-release access to "covered frontier models." The government may then work with developers to select "trusted partners" for early access [12]. Ten days later, the Commerce Department ordered Anthropic to suspend foreign-national access to Claude Fable 5 and Mythos 5, and Anthropic responded by shutting both models down globally [9]. On June 26, Commerce lifted the license requirement for Mythos 5 for entities listed in an "Annex A" to the Commerce letter, which has not been made public [9]. Commerce lifted the broader controls on June 30, though Mythos 5 itself had launched only for a limited group of vetted partners and remained restricted to approved organizations [11]. Anthropic's CVP now states that every Specialized Access applicant is reviewed "in collaboration with the US government" [1].

Security Analysis

The Gate Is an Authorization System Run by a Third Party

Taken together, these programs show labs taking on the job of a privileged-access authority for offensive capability. Each program has a recognizable access-control lifecycle: identity proofing at enrollment, attestation of authorized use, tiered entitlements tied to activity type, continuous monitoring, and the ability to revoke access. The difference from an enterprise PAM system is that the enterprise does not run it. The lab sets the policy, makes the decisions, and keeps the logs, and in some tiers a government agency co-signs the approval. An organization may own the systems it wants to test and hold signed rules of engagement, yet it still needs the lab's grant before it can use the full capability against those systems [24].

The table below summarizes the publicly described features of the programs run by the three leading frontier labs as of this writing. Program details are changing quickly, and several features, especially geographic eligibility, are not clearly documented in the labs' own announcements.

Program	Tiers / Entitlements	Eligibility Signals	Stated Controls	Government Role
Anthropic Cyber	Defense, Red Team,	Defense open to organizations and	Applicant verification,	Specialized applicants

Program	Tiers / Entitlements	Eligibility Signals	Stated Controls	Government Role
Verification Program [1]	Specialized; classifier blocking decreases by tier	individuals with disclosure records; Red Team for organizations only; Specialized for safety-critical system testers	proof of tier-specific security controls, mandatory data retention for misuse monitoring, real-time blocks on physically harmful actions	reviewed with U.S. government
OpenAI Daybreak Blue / Red [4][5][20]	Blue: adjusted general models; Red: purpose-trained cyber models (GPT-5.6-Cyber)	Approved individuals and organizations doing authorized work	Identity verification, monitoring, use restrictions, hardware security keys for individuals	Not specified in program descriptions
Google Fairwind Program [7][8]	Gemini 3.8 Flash Cyber with CodeMender; Gemini 4 Argon without cyber-specific restrictions for vetted defenders	Governments, critical infrastructure operators, core platforms, selected cloud customers and security partners	Use limited to internal security, IR, and pentest staff; MFA and operational standards; staged onboarding	Governments are among eligible participants

The opacity that CSA identified in its September analysis of these programs remains [13]. Based on the labs' public announcements as of October 30, none of the three programs publishes detailed approval criteria beyond the eligibility signals summarized above, and none publicly describes the specific conditions under which a grant is suspended or revoked. Anthropic's CVP announcement gives an appeals path for false-positive blocks but no revocation terms [1].

Vetted Accounts Are High-Value Credentials

When a model's most dangerous capability depends on account entitlements, the account becomes the asset to protect. An attacker who takes over a Red Team or Specialized tier session gets the capability without needing a jailbreak, an in-house development program, or an open-weight model. The account also carries the legitimacy of a vetted organization, which may delay detection. CSA has not found public reports of a vetted cyber-tier account being compromised. Still, the incentive structure is plain. OpenAI's requirement that individual Daybreak participants use hardware security keys suggests OpenAI expects phishing and session hijacking aimed at these accounts, and Google's MFA requirement points the same way [4][7][25].

Insider risk follows the same pattern. Programs attest that an organization is legitimate. They cannot attest that every employee holding the entitlement will use it only within scope. Fairwind's restriction to staff on internal security, incident response, and penetration testing teams helps, but the participating organization has to enforce it [7]. In practice, the strength of the lab's gate cannot exceed the identity and access hygiene of the organization holding the grant. Because these programs are new and are typically managed through lab portals outside corporate identity systems, many enterprises are unlikely to have included lab-portal identities, API keys, and console sessions for these programs in their privileged-account inventories.

Monitoring Creates a Confidentiality Trade-off

Anthropic requires enrolled organizations to accept data retention "so that we can monitor for cyber misuse." Zero data retention is available only to certain Fable 5.1 and Mythos 5.1 customers until Enterprise Frontier Safeguards, which will let eligible organizations keep data in infrastructure they control, launches later this autumn [1]. The requirement makes sense from the lab's side, because misuse detection needs visibility. It also means that prompts and outputs about an organization's unpatched vulnerabilities, internal network layouts, and exploit chains are retained by a third party, often before the organization has fixed the issues. For regulated sectors and for vulnerabilities in customer-facing systems, that retained data can itself be sensitive. Security teams should treat it as a data-handling decision and assess it accordingly.

Revocation Is a Demonstrated Risk

CSA's September note described revocation of a vetted grant as a foreseeable risk that had not yet been demonstrated [13]. Public reporting shows it had already happened. On August 19, TechCrunch reported that several security researchers lost Trusted Access for Cyber and Daybreak Blue access without notice and saw messages saying their identity could not be verified. OpenAI attributed the problem to "a

technical issue affecting a limited number of users" and asked them to verify again. However, all five researchers TechCrunch interviewed lived outside the United States and Europe, and some saw a message describing their account as "ineligible at this time" [10]. From the holder's side, an error and a policy change looked the same. The June export-control episode showed revocation directed by the state, carried out the same day across Anthropic's entire customer base: Anthropic received the directive at 5:21 p.m. ET on June 12 and disabled both models worldwide that day [11][14].

Together, these events show that a vetted-tier grant can be lost through lab policy, lab error, or government direction, and that the holder may not be able to tell which one occurred. The holder may get no warning and has no contractual remedy that has been made public. Organizations that build vulnerability operations, red-team engagements, or patch-validation pipelines on a gated tier inherit that fragility. The geographic pattern in the August incident and government co-review at Anthropic's highest tier both point to a dependency on the policy environment: an organization's eligibility can change for reasons unrelated to its own conduct [1][10][12].

Your Vendor's Tier Shapes the Threat Model You Are Tested Against

A further effect falls on organizations that buy security services. Red Team and Specialized tiers are limited to organizations, so penetration testing firms and MSSPs are among the main holders of the least-restricted capability [1]. An enterprise that commissions a red-team engagement is therefore testing against whatever tier its vendor holds. Two vendors with the same methodology may produce very different results depending on whether one has Red Team or Specialized access and the other has only Defense access. CSA is not aware of standard procurement templates or engagement-report formats that ask about this. Coverage claims are hard to interpret when the client cannot see which entitlement produced the findings.

The same asymmetry affects whether organizations get help at all. Operational technology operators said in April that they had been left out of early trusted-access conversations. One coalition leader called the gap "Silicon Valley vs. rural America" [15]. OpenAI's Frontline Defenders commitment and Anthropic's inclusion of critical infrastructure operators and open-source maintainers in Defense Access partly respond to that concern [1][6][21]. However, because the least-restricted tiers are limited to organizations and require tier-specific security controls, they are likely to go mostly to organizations with dedicated security teams and existing vendor relationships [1][7][13].

The Gate Has a Shelf Life

Gating works only while the gated capability is scarce. CSA's analysis of UK AI Security Institute data found that leading open-weight models now perform on offensive cyber tasks at roughly the level closed frontier models reached four to seven months earlier, a lag that narrowed from six to ten months during 2025 [17]. Reporting on the same AISI findings indicates that open-weight models reach that performance at a fraction of the cost [22]. If that lag holds, the capabilities now reserved for Red Team and Specialized tiers may be available without vetting, monitoring, or revocation within a year, although past trends may not continue. The distinction matters: the lag shows that each capability level escapes the gate after a delay, while the gate itself may persistently withhold the most recent months of capability as the frontier advances. A well-resourced adversary is therefore constrained by the gate's timing, not by its existence. Defenders who have no tiered access today should plan for both conditions: competitors and attackers may already hold capabilities they lack, and those capabilities will spread soon regardless.

Recommendations

Immediate Actions

Organizations that hold any vetted-tier grant should add it to their privileged access inventory now. That means recording the program, tier, named individuals, authentication method, and business owner. The individuals and service accounts holding these grants should use phishing-resistant authentication such as hardware security keys, whether or not the lab requires it. Session and API activity against these programs should be logged in the organization's own SIEM so that detection does not rely only on the lab's monitoring. Security leaders should also confirm with legal and privacy teams whether the program's data-retention terms are acceptable for the vulnerability and architecture data analysts will submit, and should route the most sensitive work to zero-retention or customer-controlled options where they exist [1].

Organizations without access should apply to the programs most relevant to their sector. Eligibility criteria differ across labs, and Anthropic's Defense Access tier explicitly covers critical infrastructure operators, small security firms, and open-source maintainers [1]. Under-resourced public-sector and utility defenders should assess OpenAI's Frontline Defenders initiative, including the MS-ISAC pilot [6] [21].

Short-Term Mitigations

Within the next one to two quarters, procurement and third-party risk teams should add questions about access tiers to security-services contracts. Penetration testing and MSSP vendors should disclose which programs and tiers they hold, which models they used in an engagement, and whether client data was retained by the model provider. Engagement reports should record this so that findings can be interpreted and compared over time. Contracts should also cover what happens if the vendor's grant is revoked mid-engagement.

Teams that depend on gated capability for vulnerability operations should write a degradation plan that names the generally available or self-hosted fallback and documents how coverage will change if the grant disappears. This connects to the continuity planning CSA recommended after the June suspension [14][16]. Detection engineering teams should assume adversaries can reach roughly equivalent capability through open-weight models within months, and should prioritize detections that do not depend on any particular tool, such as identity anomalies, lateral movement, and egress behavior [17].

Strategic Considerations

Over the longer term, boards and AI governance committees should treat lab-granted access to offensive capability as a governed dependency. It should be reported alongside model-provider concentration risk, with a named executive owner and a periodic review of whether the organization's eligibility, use, and data handling remain appropriate. Industry bodies, ISACs, and CSA members should continue pressing labs to publish eligibility criteria, revocation and appeal procedures, and notification commitments. A privilege that can be withdrawn without notice is difficult to build an operational program around. Organizations should also follow how the June 2026 executive order's trusted-partner framework is implemented, because government involvement in selection may increasingly determine eligibility for the least-restricted tiers [12].

CSA Resource Alignment

This note sits within a sequence of CSA research on gated cyber capability. *The Two-Tier Cyber Defense Problem: Vetted AI Access* (September 2026) documented the September program launches and argued that vetted access creates a two-tier defense market and filters benign defenders more effectively than determined attackers [13]. *Who Is a Trusted Defender? Gated Cyber Model Access* (October 2026) examined the governance of gated cyber-model access, and *Tiered Trust for Dual-Use AI: Anthropic's Cyber Verification Tiers* (October 2026) analyzed the CVP restructuring that anchors much of this note [23][24]. This note builds on that body of work in two ways. It treats the grant as a privileged credential

with its own lifecycle and attack surface, and it corrects the September note's statement that no vetted grant had been revoked, in light of the August de-verification incident. It also incorporates Anthropic's October restructuring, which brings government co-review into the top tier. *OpenAI's Hardware Passkey Mandate for Trusted Cyber Access* (August 2026) provides further background on the authentication controls discussed in the vetted-accounts section [25].

AI Model Export Controls: The Fable-Mythos Precedent and Sovereign AI Access Controls and Enterprise Concentration Risk (June 2026) analyze the state-directed suspension and the Annex A approved-entity model [14][16]. They provide the continuity and supply-chain framework that the revocation recommendations in this note rely on. The approved-entity structure those papers describe has since been carried into Anthropic's Specialized Access tier, where approval is shared between the lab and the U.S. government [1].

AI SI: Open-Weight Models Close Cyber Capability Gap (August 2026) supplies the evidence for this note's view that any given gated capability is temporary [17]. *The AI Vulnerability Storm: Building a Mythos-Ready Security Program* gives the operational playbook for vulnerability operations that organizations should be able to run regardless of their tier [18].

For control mapping, organizations should apply CSA's AI Controls Matrix (AICM) v1.1 [19]. The Identity and Access Management domain covers inventory, phishing-resistant authentication, and entitlement review for vetted-tier accounts. The Supply Chain Management, Transparency, and Accountability domain covers vendor tier disclosure and model-provider dependency. The Data Security and Privacy domain covers retention of vulnerability data by model providers. Documenting tier entitlements as control evidence under these domains brings this new privilege class into existing audit and assurance processes.

References

- [1] Anthropic. "[Expanding the Cyber Verification Program](#)." Anthropic, October 6, 2026.
- [2] Deutscher, Maria. "[Anthropic expands Project Glasswing cybersecurity program to 150 more organizations](#)." SiliconANGLE, June 2, 2026.
- [3] OpenAI. "[Introducing Trusted Access for Cyber](#)." OpenAI, February 5, 2026.
- [4] The National CIO Review. "[Daybreak Blue and Red Give Defenders New Access to Frontier AI](#)." The National CIO Review, August 12, 2026.
- [5] Swain, Gyana. "[OpenAI launches GPT-6 Astra, its first model to cross a critical cybersecurity threshold](#)." CSO Online, September 4, 2026.
- [6] The Next Web. "[OpenAI puts \\$1bn behind cyber defence for water utilities and community banks](#)." The Next Web, September 4, 2026.
- [7] Google. "[Proactive cyber defense for governments and enterprises](#)." Google Blog, September 2, 2026.
- [8] Hawley, Michelle. "[Google Debuts Gemini 4 Argon for Cyber Defenders](#)." VKTR, October 1, 2026.
- [9] Reddie, Andrew W. "[Commerce Eased Its Block on Anthropic's Mythos, But Major Questions Remain](#)." Tech Policy Press, June 27, 2026.
- [10] Franceschi-Bicchierai, Lorenzo. "[Researchers say OpenAI revoked their access to limited cyber program](#)." TechCrunch, August 19, 2026.
- [11] Bright Defense. "[Claude Fable 5 Restored After U.S. Lifts Export Controls](#)." Bright Defense, updated September 21, 2026.
- [12] WilmerHale. "[New Executive Order Addressing Early Government Access to Frontier AI Models](#)." WilmerHale, June 2, 2026.
- [13] Cloud Security Alliance. "[The Two-Tier Cyber Defense Problem: Vetted AI Access](#)." CSA AI Safety Initiative, September 5, 2026.
- [14] Cloud Security Alliance. "[AI Model Export Controls: The Fable-Mythos Precedent](#)." CSA AI Safety Initiative, June 28, 2026.

- [15] Waterman, Shaun. "[OT Cybersecurity Frozen Out by Frontier Labs](#)." GovInfoSecurity, April 28, 2026.
- [16] Cloud Security Alliance. "[Sovereign AI Access Controls and Enterprise Concentration Risk](#)." CSA AI Safety Initiative, June 16, 2026.
- [17] Cloud Security Alliance. "[AIS: Open-Weight Models Close Cyber Capability Gap](#)." Cloud Security Alliance, August 10, 2026.
- [18] Cloud Security Alliance. "[The AI Vulnerability Storm: Building a Mythos-Ready Security Program](#)." Cloud Security Alliance, April 14, 2026.
- [19] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." Cloud Security Alliance, June 22, 2026.
- [20] OpenAI. "[Expanding Daybreak as the cyber defense window narrows](#)." OpenAI, August 10, 2026.
- [21] OpenAI. "[Daybreak for Frontline Defenders](#)." OpenAI, September 3, 2026.
- [22] The Decoder. "[Open-weight models now match frontier cyber performance from just four months ago at a fraction of the cost](#)." The Decoder, July 2026.
- [23] Cloud Security Alliance. "[Tiered Trust for Dual-Use AI: Anthropic's Cyber Verification Tiers](#)." CSA AI Safety Initiative, October 8, 2026.
- [24] Cloud Security Alliance. "[Who Is a Trusted Defender? Gated Cyber Model Access](#)." CSA AI Safety Initiative, October 3, 2026.
- [25] Cloud Security Alliance. "[OpenAI's Hardware Passkey Mandate for Trusted Cyber Access](#)." CSA AI Safety Initiative, August 2, 2026.
- [26] Computing. "[Anthropic unveils Project Glasswing – AI-driven cybersecurity to detect software flaws](#)." Computing, April 2026.