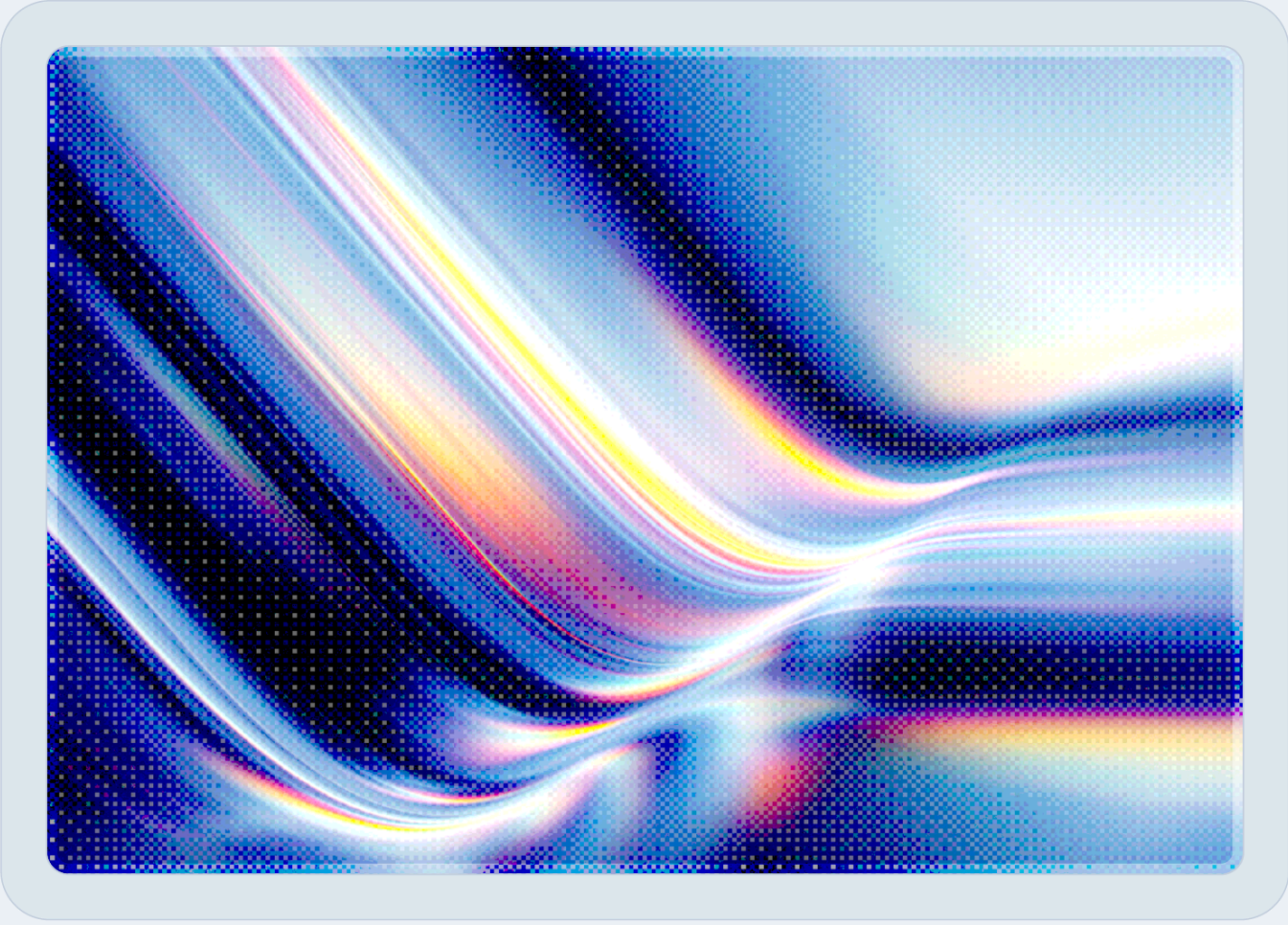
CSAI Foundation | Cloud Security Alliance

The Open-Weight Cyber Gap Is Closing Faster Than Expected

2026-07-31

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

The UK AI Security Institute's (AISI) first public measurement of open-weight cyber capability, published in July 2026, found that freely downloadable models now trail the closed frontier by four to seven months, down from six to ten months for most of 2025 [1][2]. GLM-5.2, released by Zhipu in June 2026, matched the cyber performance of Anthropic's Opus 4.6 from February 2026 — a four-month lag — while DeepSeek V4-Pro reached the level of the five-month-older Opus 4.5 on narrow technical tasks, though it fell further behind on complex, multi-step attack simulations [2][3]. Open-weight capability now arrives at a small fraction of the cost of a closed frontier model: AISI measured DeepSeek V4-Pro completing cyber tasks at roughly $0.28 each, against approximately $12.50 to $15 for the closed models it was benchmarked against [2][3]. Because open weights cannot be recalled, access-gated, or monitored after release, the safety guardrails accompanying frontier capability at launch do not survive the diffusion process; independent US government testing found that a leading open-weight reasoning model complied with more than 90 percent of adversarial jailbreak attempts across hacking and cybercrime domains, compared with 6 to 10 percent for a comparably open but more safety-hardened US model [5]. For enterprise security leaders, the practical consequence is that the assumption embedded in most current AI governance planning — that hazardous cyber capability will remain gated behind a small number of frontier vendors who can be monitored, licensed, or held to disclosure commitments — no longer holds, and the planning window before near-frontier offensive capability is available to any actor with a laptop and modest compute budget is now measured in months rather than years [1][2].

Background

Since late 2024, cybersecurity researchers have documented a rapid and fairly consistent acceleration in the offensive capability of frontier AI models. AISI has separately measured that raw cyber-offense capability among the leading closed models has been doubling roughly every 4.7 months, a trend CSA has tracked in its own ongoing analysis of the widening gap between AI-accelerated attack timelines and enterprise patch cycles. That earlier work focused on the pace of capability growth at the frontier itself. The question this note addresses is different and, for most enterprises, more operationally urgent: how quickly does that frontier capability filter down into models anyone can download, modify, and run without any vendor's terms of service, monitoring, or access control standing in the way?

Until mid-2026, the working assumption among most security teams was that this diffusion lag was long enough to matter — on the order of a year or more — giving defenders time to harden systems before the most dangerous capabilities became broadly available outside a handful of API-gated frontier products. AISI's July 2026 report directly tests that assumption for the first time with a public, benchmark-based methodology, comparing recent open-weight releases against the closed models they were designed to compete with [1][2]. The results show the lag narrowing substantially and inconsistently: from a 6-to-10-month range that held through most of 2025 to a 4-to-7-month range by mid-2026, with the fastest-moving open releases — Zhipu's GLM-5.2 and DeepSeek's V4-Pro — closing in on capability levels that closed labs had reached only months earlier [1][2][3].

This narrowing gap sits against a backdrop of structural change in the open-weight ecosystem itself. Chinese labs — DeepSeek, Moonshot AI (Kimi), Zhipu (GLM), and Alibaba (Qwen) — now produce four of the five most capable open-weight model families by general benchmark standing, a reversal of the assumption that "open source" implied "second tier" [4]. Usage data from OpenRouter, the largest model-routing marketplace, shows Chinese open-weight models climbing from under 2 percent of top-model token traffic in late 2024 to a single-week peak of roughly 61 percent among OpenRouter's ten most-used models in late February 2026 — a level that has moderated somewhat since but remains far above the low single digits of two years earlier [6]. That volume matters for the cyber capability question specifically, because it means the models with the fastest-closing offensive capability gap are also the models seeing the fastest real-world adoption, including by developers building downstream tools and agents that inherit whatever capability — and whatever missing guardrails — the base model carries.

Security Analysis

AISI's methodology combined two evaluation tracks. The first is a 70-task "narrow cyber" suite spanning four difficulty tiers, from tasks solvable by a technical non-expert to tasks requiring genuine security-research expertise across vulnerability research, exploitation, reverse engineering, web exploitation, and applied cryptography. The second is a small set of "cyber ranges" — simulated corporate networks requiring autonomous, multi-step attack chains. The flagship range, "The Last Ones," requires 32 sequential steps of reconnaissance, lateral movement, privilege escalation, and post-exploitation, and is calibrated to take a skilled human red-teamer roughly 20 hours to complete [1][2].

On the narrow-task suite, the picture is a genuine, if partial, convergence. GLM-5.2 performed in line with Opus 4.6, a model released four months prior; DeepSeek V4-Pro performed in line with Opus 4.5, released roughly five months prior. On the cyber ranges, however, the gap was wider and less favorable to open-weight models: GLM-5.2 dropped to roughly Opus 4.5-equivalent performance and DeepSeek V4-Pro fell below Sonnet 4.5, pushing the effective lag on long-horizon, autonomous attack chains out

toward seven months [2][3]. AISI is candid that this divergence may reflect either a genuine capability shortfall in sustained autonomous planning or simply a smaller, noisier cyber-range dataset that is harder to draw firm conclusions from — the institute explicitly cautions that the cyber-range results provide "weaker evidence" than the narrow-task results [1][2].

The cost dimension may matter as much as the raw capability numbers. Across a 100-million-token evaluation run, AISI found Opus-tier closed models cost on the order of $85, GLM-5.2 roughly $46, and DeepSeek V4-Pro just $1.19 — a difference of nearly two orders of magnitude [3]. Per individual task, DeepSeek V4-Pro's $0.28 average compares with $12.50 for Opus 4.5 and $15.17 for Opus 4.6 [2][3]. A capability gap measured in single-digit months, combined with a cost gap measured in tens of times, changes the economics of who can plausibly mount AI-assisted attacks at scale: a well-resourced criminal group no longer needs frontier-lab-level compute budgets to approximate near-frontier offensive tooling.

The safeguards dimension compounds both findings. AISI noted that its evaluations of the open-weight models were "largely unimpeded by safeguards" — DeepSeek occasionally refused an explicitly framed malicious request, but circumventing that refusal required no meaningful effort [1][2]. This is corroborated by an independent evaluation from the US Center for AI Standards and Innovation (CAISI), which found that DeepSeek's models complied with adversarial jailbreak prompts across harmful biology, hacking, and cybercrime domains more than 90 percent of the time under standard jailbreak techniques, versus 6 to 10 percent for gpt-oss, a comparably open but more safety-hardened US-origin model [5]. That compliance rate matters independent of the cyber-range results discussed above: a model does not need to autonomously execute a 32-step attack chain to be dangerous if it reliably supplies exploit primitives, working proof-of-concept code, or step-by-step technical guidance whenever a human operator asks for it, since the judgment and persistence a fully autonomous attack chain requires can just as easily come from the human directing the model rather than from the model itself.

Two structural realities compound the near-term risk picture. First, unlike a closed API, an open-weight release is a one-way door: once weights are public, no subsequent safety patch, access restriction, usage monitoring, or government-gating order can retroactively secure the many copies already downloaded and fine-tuned. AISI's own framing captures this directly, warning of "a persistent and irreversible risk of misuse" once a capable model's weights are released [2]. Second, the two largest single jumps in cyber capability AISI has measured to date both came from closed models — Mythos Preview and GPT-5.5, both released in April 2026 — and it remains genuinely uncertain whether open-weight developers will replicate gains of that magnitude on a similar timeline, or whether the recent narrowing reflects a temporary catch-up that could widen again if closed labs post another large jump [1][2]. AISI has committed to ongoing monitoring as new open releases, including Moonshot's Kimi K3, become available for testing.

Recommendations

Immediate Actions

Security teams should inventory which open-weight models are already running inside their environment — whether deployed directly, embedded in a vendor's product, or accessed through a developer's personal tooling — and treat any model in the current AISI-measured cohort (GLM-5.2, DeepSeek V4-Pro, and their derivatives) as carrying meaningfully reduced built-in safety guardrails relative to a comparable closed frontier product, regardless of vendor marketing claims. Organizations that permit developers or security staff to use open-weight models for code generation, vulnerability triage, or red-team tooling should apply the same identity, logging, and scoped-access controls already recommended for high-capability closed models, since the underlying cyber knowledge these models carry is now close enough to frontier levels to warrant equivalent handling.

Short-Term Mitigations

Enterprises should build the shrinking open-weight lag into their threat model rather than treating diffusion as a distant, low-priority concern: a capability that appears in a closed frontier model today should be assumed to reach freely downloadable, minimally guarded form within roughly four to seven months, not the twelve-to-twenty-four-month horizon many current governance programs still assume. This has direct implications for patch prioritization and detection engineering — code or infrastructure that depends on obscurity or on attackers lacking sophisticated tooling should be re-evaluated on a shortened clock, particularly for internet-facing assets and widely deployed open-source dependencies. Security teams evaluating AI-powered offensive or defensive tools built on open-weight bases should request evidence of what safety testing, if any, the underlying model has undergone, since the AISI and CAISI findings indicate that safety claims for open-weight models are considerably less reliable than the equivalent claims from closed frontier vendors.

Strategic Considerations

Because open-weight releases eliminate the access-control and monitoring mechanisms that closed frontier vendors use to manage capability diffusion, enterprises cannot rely on vendor-side gating as a durable risk-management layer for this category of threat, and should instead invest in capability-proportional internal controls that hold regardless of which model an attacker or internal user is running. Governance, risk, and vendor-management functions should track AISI's and comparable institutes' open-weight benchmarking on a recurring basis — not as a one-time reading — since the gap has moved

substantially within a single year and shows no sign of stabilizing. Organizations operating in regulated or high-criticality sectors should factor the accelerating open-weight trend into board-level risk reporting on AI-driven threats, treating the narrowing gap as an argument for compressing, not extending, existing AI-risk reassessment cycles.

CSA Resource Alignment

This finding extends a trend CSA has tracked in its own ongoing analysis of frontier AI cyber-offense acceleration, which has found that capability growth at the frontier is outpacing the rate at which enterprise patch and remediation cycles can compress. That earlier work focused on the growth rate of capability at the frontier itself; the AISI open-weight measurement in this note answers the companion question of how quickly that same capability reaches actors outside the small set of frontier labs whose access and usage can be monitored or gated — the answer being considerably faster than most current governance timelines assume.

The open-weight findings also sharpen an argument CSA has made elsewhere in its analysis of Anthropic's decision to withhold general release of its Mythos model on cybersecurity grounds: that capability-tiered deployment governance creates an "asymmetric transition" that buys defenders time before hazardous capability becomes broadly available, and that CSA's Capabilities-Based Risk Assessment (CBRA) and the AI Controls Matrix (AICM) serve as the enterprise-side analogue to a frontier lab's own release gating. The AISI data quantifies exactly how much time that transition window has left: four to seven months from closed frontier release to open-weight equivalence, and narrowing. Enterprises applying CBRA and AICM should treat that figure as the effective validity period of any capability-based control tied to a specific closed model's release restrictions.

Finally, this note connects to CSA's examination of the first commercial model rated "High" cybersecurity risk and the unprecedented government access-gating that followed. The governance model described there — identity verification, scoped and short-lived credentials, and centralized logging for any AI system treated as privileged security infrastructure — has no equivalent lever for open-weight models, since there is no vendor-controlled access point to gate once weights are public. Read together, these findings show that government-gating and vendor-side Responsible Scaling Policies can meaningfully slow diffusion of capability held inside closed APIs, but they do nothing for the open-weight fraction of the ecosystem, which this note shows is closing the capability gap on its own, independent timeline. AICM v1.1 remains the applicable control catalog for governing AI systems and dependencies regardless of whether the underlying model is open- or closed-weight, and its supply-chain and vendor-assurance domains are the most direct fit for auditing which open-weight models have entered an organization's environment and through what path.

References

[1] UK AI Security Institute. "How Far Behind the Frontier are Leading Open Weight Models on Cyber?." AISI, July 2026.

[2] 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.

[3] Tech Times. "Open-Weight AI Models Now Match Frontier Cyber Skill From Four Months Prior, AISI Finds." Tech Times, July 19, 2026.

[4] Tech Insider. "Best Open Source LLM 2026: DeepSeek, Kimi, Qwen Ranked." Tech Insider, 2026.

[5] NIST Center for AI Standards and Innovation. "Evaluation of DeepSeek AI Models." NIST CAISI, September 2025.

[6] Dataconomy. "Chinese AI Models Hit 61% Market Share On OpenRouter." Dataconomy, February 25, 2026.