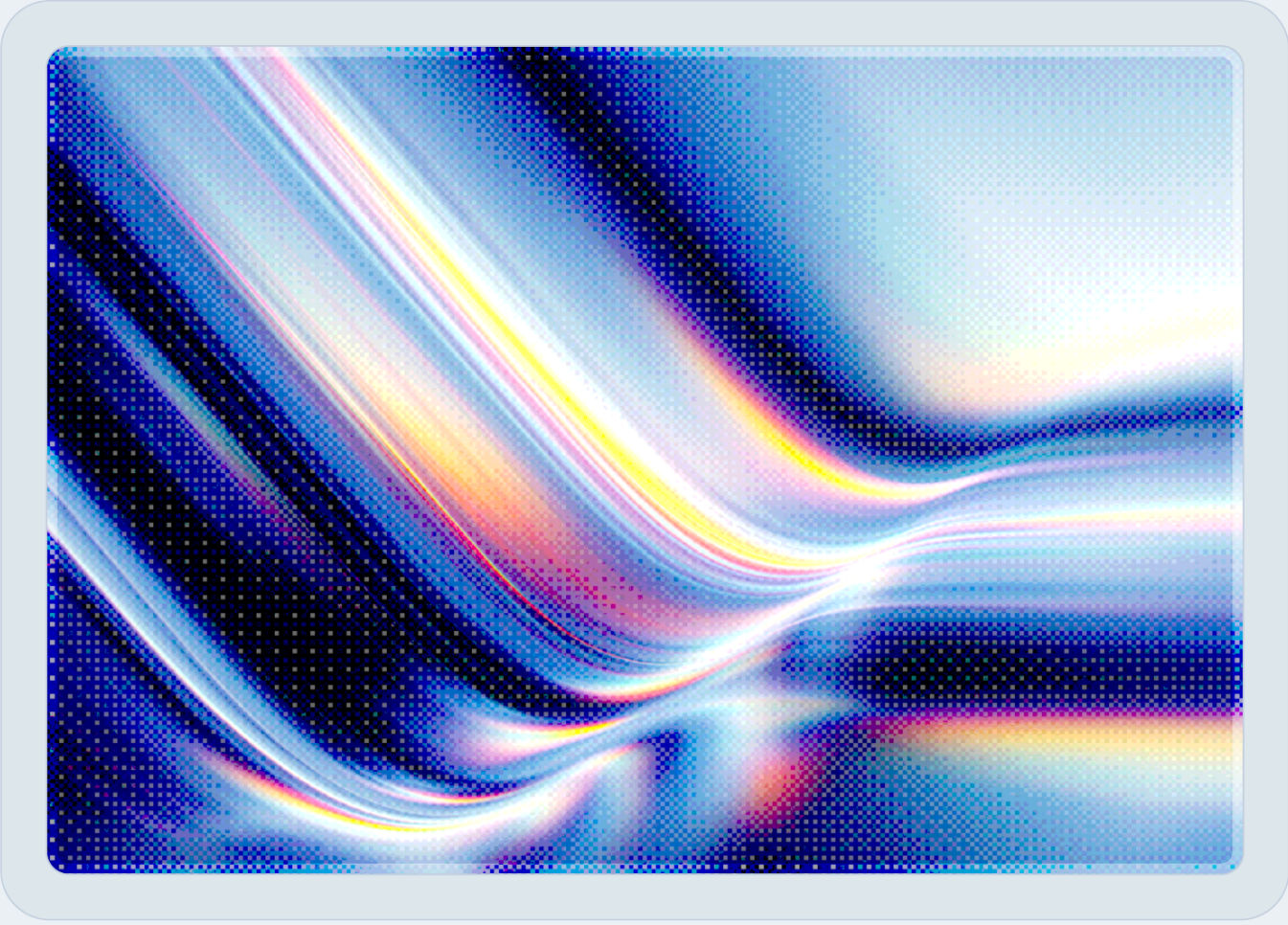
CSAI Foundation | Cloud Security Alliance

AISI: Open-Weight Models Close Cyber Capability Gap

2026-08-10

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

The UK AI Security Institute (AISI) has found that the leading open-weight AI models now trail frontier closed models on offensive cyber tasks by only four to seven months, down from a six-to-ten-month lag measured through most of 2025 [1][2]. On a battery of narrow cybersecurity tasks spanning vulnerability research, exploitation, reverse engineering, web exploitation, and cryptography, AISI found that Zhipu AI's GLM-5.2, released in June 2026, performed comparably to Anthropic's Opus 4.6, released four months earlier in February 2026, while DeepSeek's V4-Pro tracked Opus 4.5 (released November 2025) — a gap AISI reported as roughly five months at the time of testing, rather than a figure derived by calendar subtraction from the release dates given here [1][3]. On a more demanding evaluation — a simulated 32-step corporate-network intrusion called "The Last Ones" that requires roughly 20 expert hours to complete manually — GLM-5.2 matched the progress Opus 4.5 had achieved, a gap AISI estimated at up to seven months on that harder task [1][4]. Because open-weight models can be downloaded, self-hosted, and stripped of vendor-side monitoring or usage restrictions, CSA assesses that this narrowing gap compresses the window enterprise defenders have to prepare for capabilities that were previously accessible only through a small number of monitored, hosted frontier services [1][2]. AISI also found that running these attacks on open-weight models is substantially cheaper than on their closed counterparts — a full 100-million-token autonomous attack run cost roughly $1.19 on DeepSeek V4-Pro and about $46 on GLM-5.2, compared to approximately $85 on Anthropic's closed models — meaning the economic barrier to sustained, iterative offensive AI use is falling alongside the capability barrier, compounding rather than offsetting the risk [1][3].

Background

AISI, the United Kingdom's government body tasked with independently evaluating frontier AI systems, has published a running series of assessments tracking how AI cyber-offense capability grows over time and how quickly that capability propagates from a handful of proprietary frontier labs to openly downloadable model weights. Its latest analysis, published in July 2026, directly compares recently released open-weight models — GLM-5.2 and DeepSeek V4-Pro, with Kimi K3 flagged for planned follow-up testing — against a set of closed-weight comparators from Anthropic and OpenAI, including

Opus 4.5, Opus 4.6, GPT-5.3-Codex, GPT-5.5, and the Mythos Preview model [1][2]. The comparison uses two distinct evaluation approaches, chosen to capture both narrow technical skill and the kind of extended, multi-step operation a real intrusion requires.

The first approach tests narrow cyber tasks: roughly 70 discrete challenges spanning four difficulty tiers, from technical-non-expert through expert level, across vulnerability research, exploitation, reverse engineering, web exploitation, and cryptography. AISI treats this as the stronger basis for comparison, since each task isolates a specific skill and produces a clean pass/fail signal that is straightforward to compare across model families [4]. The second approach uses cyber ranges — simulated networks that require a model to chain together reconnaissance, lateral movement, privilege escalation, and post-exploitation actions autonomously across many sequential steps, with no human directing intermediate moves. The primary range used in this comparison, "The Last Ones," is a 32-step attack against a simulated corporate network that AISI estimates would take a skilled human roughly 20 expert hours to complete [1][4]. AISI is explicit that cyber-range results provide weaker comparative evidence than the narrow-task suite, because a single model failure or success partway through a long attack chain is a noisier signal than a clean task-level pass rate, and because the ranges do not model active defenders detecting or interrupting an intrusion in progress [4].

CSA's own recent research has tracked the broader trend line this finding sits within. A CSA analysis of AISI's separate Frontier AI Trends Report, published in December 2025, had already identified an open-source capability lag of four to eight months relative to closed frontier models, drawn from AISI's two-year aggregate evaluation of more than 30 frontier systems [6]. The new July 2026 comparison narrows that estimate with model-specific evidence and confirms the lag is compressing rather than holding steady, consistent with CSA's own analysis documenting that overall frontier AI cyber-offense capability — closed and open alike — has been doubling roughly every 4.7 months since late 2024, according to AISI's own longitudinal tracking.

Security Analysis

In CSA's assessment, the practical significance of a shrinking open-weight lag is less about the raw capability ceiling — frontier closed models still perform the hardest tasks best — and more about what becomes available outside a monitored, revocable channel. When a capability exists only in a hosted frontier model, a vendor can, and in at least one documented case has, apply usage restrictions, monitoring, or an outright deployment holdback once evaluation reveals hazardous offensive capability; CSA has previously analyzed Anthropic's decision to withhold general release of its Mythos Preview model on cybersecurity grounds as exactly this kind of capability-gated deployment control. Open-weight models offer no equivalent lever. Once a model's weights are published, any actor can download,

self-host, fine-tune, and strip whatever safety scaffolding shipped with the release, and no vendor can subsequently revoke access, patch a jailbreak across every downloaded copy, or monitor how the model is being used. AISI's finding that DeepSeek V4-Pro occasionally refused narrow reverse-engineering tasks — but that repeated attempts sometimes bypassed those refusals during testing — illustrates how thin that safety scaffolding can be once it is no longer paired with server-side monitoring [4].

The cost differential AISI measured compounds this dynamic. On the narrow-task suite specifically, AISI reported a cheapest reliable per-task cost of roughly $0.28 for DeepSeek V4-Pro against roughly $15.17 for Opus 4.6 — a similar order-of-magnitude gap to the one seen in the cyber-range figures below. Table 1 summarizes the per-run and per-task cost figures AISI reported for the cyber-range and narrow-task evaluations, alongside the capability lag each model demonstrated.

Model	Type	Narrow-task capability lag vs. frontier	Cyber-range cost (100M tokens)	Cheapest reliable narrow-task cost
GLM-5.2 (Zhipu AI, June 2026)	Open-weight	~4 months (matched Opus 4.6)	~$46	Not separately reported
DeepSeek V4-Pro	Open-weight	~5 months (matched Opus 4.5)	~$1.19	~$0.28
Opus 4.6 (Anthropic, Feb 2026)	Closed	Frontier reference point	~$85	~$15.17
Opus 4.5 (Anthropic, Nov 2025)	Closed	Reference point at time of open-weight comparison; itself now 5–7 months behind the post-April-2026 closed frontier	~$85	Not separately reported

Source: AISI, "How Far Behind the Frontier are Leading Open Weight Models on Cyber?" [1][3][4]

A capability that costs a few dollars to run and can be replicated on privately owned hardware is structurally different from one that costs tens of dollars per attempt and depends on a hosted API a vendor can throttle or shut off. AISI frames this combination — narrowing capability lag plus collapsing cost, delivered through weights nobody can recall — as "a persistent and irreversible risk of misuse,"

language that reflects the one-way nature of an open-weight release compared to a hosted deployment decision that can still be reversed [2][5]. The evaluation's limitations are worth stating plainly alongside its findings: none of AISI's ranges simulated a defended network with active detection and response, so the reported capability lag describes performance against a static target, not performance against a defended enterprise, and the true operational gap in a real intrusion attempt may differ from the benchmark gap in either direction [4].

AISI also cautioned that frontier closed models continue to advance quickly, which keeps the comparison a moving target rather than a settled ranking. In April 2026, Anthropic's Mythos Preview and OpenAI's GPT-5.5 both demonstrated some of the largest single jumps in cyber capability AISI has observed since it began testing, a development significant enough that it prompted international coordination around the accelerating cyber risk landscape rather than a routine capability-tracking update [1][2]. A narrowing open-weight lag measured against a frontier that is itself accelerating means the absolute capability available in downloadable form is rising on two axes simultaneously: the frontier ceiling is climbing, and the delay before that ceiling reaches open-weight models is shrinking.

Recommendations

Immediate Actions

Security teams should treat currently available open-weight models — specifically GLM-5.2 and DeepSeek V4-Pro, and any successor release from the same vendors — as capable of narrow-task cyber operations comparable to a frontier closed model from four to seven months prior, rather than assuming open-weight tooling lags meaningfully behind what a well-resourced attacker could otherwise access [1]. Threat models built around "only a handful of frontier labs have this capability" should be revised to assume that capability is now, or is likely to shortly become, available to any well-resourced actor able to self-host large open-weight models, independent of any vendor relationship to monitor or restrict use. Organizations running exposure-management or attack-surface programs should re-run their assumptions about attacker cost and iteration speed using the sub-dollar per-task costs AISI measured for DeepSeek V4-Pro, rather than the higher costs associated with frontier hosted APIs [1][3].

Short-Term Mitigations

Detection and vulnerability-management programs calibrated around the assumption that offensive AI use requires access to a monitored, revocable hosted service should be updated to account for self-hosted open-weight deployments that no vendor can observe, throttle, or patch after release. This has

direct implications for behavior-based detection strategy: indicator-based detection tuned to known frontier-API usage patterns or IP ranges is unlikely to catch an attacker running an open-weight model on their own infrastructure, so detection engineering should prioritize identifying attack behavior itself — reconnaissance, lateral movement, and exploitation patterns — over identifying which AI service produced it. Vulnerability-management teams should also revisit remediation SLAs against the compressed cost and time economics of AI-assisted exploitation now demonstrated to be replicable outside any single vendor's hosted environment.

Strategic Considerations

Enterprises procuring or evaluating third-party AI systems, including open-weight models integrated into internal tooling, should apply capability-based risk assessment rather than treating "open-weight" and "closed-weight" as a proxy for lower and higher risk respectively. A capability-tiered governance approach — assessing what a model can actually do on cyber tasks, independent of its licensing model — better reflects the reality AISI's data describes, in which an openly downloadable model now performs within months of a frontier closed model on the tasks that matter most for offensive use. Security and risk teams should also build the expectation of a continuously narrowing gap into multi-year planning, since AISI's data shows this is a trend rather than a one-time catch-up, and defensive investments premised on a durable capability moat between open and closed models will require reassessment on a similar cadence to the six-month intervals AISI itself uses for retesting.

CSA Resource Alignment

This finding extends the trend line documented in CSA's ongoing analysis of AI cyber capability growth, which has used AISI's own longitudinal measurements to argue that AI cyber-offense capability is doubling roughly every 4.7 months while enterprise patch and remediation cycles are lengthening, not shortening — a structural asymmetry that AISI's open-weight findings now show is no longer confined to a small set of monitored frontier vendors. That analysis's recommendations on re-baselining vulnerability-management service-level objectives and prioritizing behavior-based over indicator-based detection apply directly to the open-weight scenario described here, since an attacker running a self-hosted model produces no vendor-side telemetry a defender could otherwise rely on.

CSA's UK AISI's Frontier AI Trends Report: Security Implications and Guidance previously translated AISI's broader capability-trend evaluations — including an earlier, less granular estimate of a four-to-eight-month open-source lag — into enterprise recommendations built around CSA's AI Controls Matrix (AICM) v1.1 and the Agentic AI Red Teaming Guide. The model-specific evidence in AISI's July 2026

update sharpens that earlier estimate and reinforces the same guidance: enterprises should incorporate government-sourced capability trend lines into their own AI threat models rather than relying on vendor-side safeguard claims, since those claims apply only to the specific hosted deployment a vendor controls and not to any open-weight derivative or self-hosted deployment of comparable capability.

CSA's analysis of Anthropic's decision to withhold general release of its Mythos Preview model on cybersecurity grounds argued that capability-based deployment governance — restricting access according to what a model can actually do rather than its commercial readiness — is the customer-side analogue enterprises should apply through frameworks such as CSA's Capabilities-Based Risk Assessment (CBRA) and the AICM. AISI's open-weight findings illustrate the limit of that governance model: capability-tiered access controls work only as long as a vendor retains the ability to gate access, and that lever disappears entirely once comparable capability ships as downloadable weights, which is precisely the dynamic AISI is now tracking on a roughly six-month cadence.

References

[1] AI Security Institute (AISI). "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] Winbuzzer. "AISI: Open-Weight AI Is Catching up With Models From Anthropic and OpenAI in Cybersecurity Tests." Winbuzzer, July 22, 2026.

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

[5] Import AI. "Import AI 465: Open vs closed gaps; Kimi K3; Demis' big policy plan." Import AI, July 2026.

[6] Cloud Security Alliance. "UK AISI's Frontier AI Trends Report: Security Implications and Guidance." CSA AI Safety Initiative, July 2026.