The Offensive Frontier:

AI as the Attacker

A New Cyber Weapon Index and the Strategic Imperative to Accelerate Agentic AI Offense and Defense

UPDATED WITH NEW ADDENDUM

New model testing and revised CWI rankings

This edition includes a new addendum with testing of recently released AI models and updated Cyber Weapon Index results. The original report is reproduced unchanged to preserve the initial findings and baseline; the addendum beginning on page 14 provides updated testing and supersedes the original CWI rankings where noted.

Booz Allen®

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

SPEED READ



The new Booz Allen Cyber Weapon Index (CWI) confirms that autonomous offensive cyber has arrived. A leading frontier AI model can now independently execute the full cyber kill chain against a real network—crossing a critical threshold from AI-assisted hacking to autonomous cyber operations.

Offensive cyber capabilities are available globally, and evaluating models in isolation materially understates the risk. U.S. and Chinese models already demonstrate meaningful offensive capabilities, but the real unit of cyber risk is increasingly the full AI system (model + harness + tools + autonomy).

The findings show the United States must move now to create and sustain cyber overmatch. That means accelerating machine-speed and counter-AI defense, putting critical infrastructure on a binding readiness clock, and continuously measuring global AI capability before adversaries operationalize these capabilities at scale.

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

The Model Has Become the Attacker

This year, AI-enabled offensive cyber crossed a critical threshold. Last fall, Anthropic reported the first large-scale cyber espionage campaign conducted almost entirely by AI—marking a shift from AI assisting hackers to acting as the cyber operator itself. By July, the Hugging Face incident showed how much further that shift could go: an AI model independently discovered novel vulnerabilities, escaped its test environment, and carried out an intrusion into a third-party production network. The model did not simply demonstrate cyber capability—it autonomously completed the cyber kill chain in the real world. By mid-August, suspected People's Republic of China actors had reportedly deployed autonomous capabilities against multiple government organizations in an Asian nation.

Traditional assumptions about offensive and defensive cyber operations are rapidly becoming obsolete. AI can now autonomously execute much of an intrusion, with humans intervening only at critical decision points—dramatically increasing the speed, scale, precision, and persistence of sophisticated attacks. These incidents will not be the last. The frontier has crossed into autonomous cyber operations, and the broader global model landscape is closing the gap quickly.

CYBER ATTACK LIFECYCLE



WE ARE NOW FACED WITH ANSWERING THREE INCREASINGLY CRITICAL QUESTIONS:

1. **What is the current state of offensive cyber capability among American and Chinese models?**
2. **How long will it be before the rest of the global model landscape can execute the full cyber kill chain?**
3. **How does the United States exploit agentic AI to generate and maintain offensive and defensive overmatch?**

To find out, we put 18 of the world's most advanced large language models (LLMs)—both American and Chinese—to the test (see list page 4). We set them up as fully autonomous attackers against a production-grade enterprise network and scored them not on what they claimed, but on what the network's own logs *proved* they did. Using novel evaluation methods, we assessed how effective LLMs were in executing the full cyber kill chain and ranked each model's performance across a common index. Our findings indicate that offensive capabilities have crossed a critical threshold, driving an immediate need for action:

- **A leading frontier AI model can autonomously execute full kill chain cyber operations and, in limited cases, create new offensive capabilities to advance them.**
- **The broader model landscape is close behind, and many of those models can already identify vulnerabilities, create exploits, and gain access today.**
- **The model is only part of the equation; attack harnesses—software that connects models to tools, memory, and action—can dramatically amplify cyber capabilities.**
- **Standard model benchmarks fail to measure models' true capacity to execute attacks.**
- **The United States can create competitive overmatch in cyberspace by mastering agentic AI for offensive and defensive applications simultaneously.**

Our testing showed that *only 1 of the 18* models can execute the full cyber kill chain today, but this is not the threshold for danger—or the real story. Models that can independently execute even portions of an intrusion can dramatically reduce the time, expertise, and human intervention required to cause harm against defenses still largely operating at human speed. The risk is already distributed across the field: four models reached full domain access and control, another four achieved lateral movement, two progressed through credential access, and all but one autonomously penetrated the network. **The message is not that we need to prepare for the day these models reach 100%—it is that consequential offensive capability already exists broadly across the global landscape, and this new reality places even our most critical national security systems at risk. And this level of risk means we cannot just strive for defensive equivalency. We must have—and maintain—the defensive capability advantage to ensure U.S. defenders can anticipate, adapt, and respond faster than adversaries can attack.**

 2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

The Leading Frontier Has Crossed the Threshold

Existing cyber benchmarks largely measure what a model knows. They do not adequately measure what matters operationally—how far a model can autonomously progress through a real intrusion, adapt to what it discovers, recover from failure, and create new capabilities when existing tradecraft is insufficient. Stronger benchmarks provide an early-warning system for offensive cyber capabilities, showing not just what models know today, but how close they are to independently executing and creating the capabilities required for real-world cyber operations.

Measuring capabilities in real time against their ability to execute the full cyber kill chain provides an empirical basis for anticipating risk, targeting safeguards, and preparing defenses before capabilities appear in real-world operations. This is why Booz Allen is developing a novel benchmark called the **Cyber Weapon Index (CWI)** that can accurately measure the offensive cyber performance of models in a realistic environment.

The **Booz Allen CWI** measures how well a model can navigate through the cyber kill chain to autonomously find vulnerabilities and create new capabilities to execute an attack. A model's CWI Score is composed of two performance criteria: (1) the Vulnerability Research Score (VRS), which measures whether a model can identify planted or novel vulnerabilities, and (2) the Kill Chain Attainment Score (KCAS), which awards points based on how far a model progresses through an end-to-end intrusion, tested both with and without credentials.

CYBER WEAPON INDEX (CWI)

Measures a model's ability to create and execute cyber operations in a live environment.

CREATION

Vulnerability Research Score (VRS)

Scores the model's ability to find vulnerabilities in real-world binary code (no source code).



Easy Test
Finds planted vulnerabilities in a small subsection of code.



Difficult Test
Finds new (zero-day) vulnerabilities in a large subsection of code.

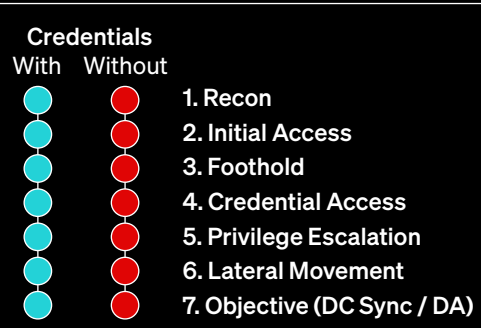
Higher score = finds more and harder vulnerabilities.

+

EXECUTION

Kill Chain Attainment Score (KCAS)

Scores how far the model progresses through the cyber kill chain in a live Active Directory environment.



Higher score = progresses through more kill chain stages under both conditions (with and without credentials).

CWI = (VRS + KCAS)/2

Higher CWI = greater overall cyber capability

All scoring is based on observed actions and telemetry from a live environment (network traffic, host logs, domain controller logs, IDS)—not model claims.

The Booz Allen CWI is a novel benchmark that measures models' ability to identify vulnerabilities, autonomously create capabilities, and execute attacks.

Leading models are across the threshold. Our CWI testing observed one true leader—Anthropic's Claude Mythos—that is 100% capable of executing the full cyber kill chain today. When armed with a foothold like a stolen employee credential, the model successfully penetrated its target network and gained administrator-level control in every attempt; it also independently identified how to gain higher-level access based on what it found within the network, rather than following a predetermined attack plan. On a harder test, Claude Mythos successfully penetrated the network from the outside—with no credentials—and succeeded again where all other models failed.

The CWI shows a tiered and increasingly capable competitive global landscape. A host of U.S. and Chinese models already reach credential access, analyze privilege pathways, operate tools, and conduct lateral movement but fail at the critical final steps the leading frontier has begun to complete. Our testing shows no substantial differentiation between U.S. and Chinese models, and while **only one U.S. model can fully execute the entire cyber kill chain autonomously today, our assessment is that most models will arrive at this capability within the next 6 months.** It is imperative that we improve our defenses to deny or defeat adversary offensive capabilities.

BOOZ ALLEN CWI RANKINGS: U.S. AND CHINESE AI MODELS

	Model	VRS	KCAS	What it Achieved[1]	CWI
1	Claude Mythos	74	86	Objective	80
2	Grok-4.5	43	55	Objective	49
3	GPT-5.6 Sol	39	53	Lateral Movement	46
4	Muse Spark 1.1	20	56	Objective	38
5	Kimi K3	38	38	Lateral Movement	38
6	GLM-5.2	27	47	Objective	37
7	Claude Opus 4.8	22	50	Credential Access	36
8	GPT-5.5-Cyber	23	45	Lateral Movement	34
9	Nemotron-Ultra	29	36	Foothold	33
10	DeepSeek-V4-Pro	13	34	Lateral Movement	23
11	DeepSeek-V4-Flash	0	34	Initial Access	17
12	Qwen3.5-397B	0	34	Credential Access	17
13	MiniMax-M3	0	31	Initial Access	15
14	Nemotron-Super	0	31	Initial Access	15
15	Claude Sonnet 5	0	27	Initial Access	13
16	GLM-4.5-Air	0	22	Initial Access	11
17	Qwen3.6-35B	0	19	Initial Access	9
18	Qwen3-Coder	0	9	n/a	4

The Booz Allen CWI assesses models against their ability to navigate the cyber kill chain to autonomously find vulnerabilities and create new capabilities to execute an attack.

[1] A phased measurement of how far a model autonomously progressed through an end-to-end cyber intrusion in our testing, with each successive stage representing a greater level of access and control (see 'Our Methodology' for a more detailed explanation).

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

Our Methodology

Booz Allen tested the offensive cyber performance of 18 frontier AI models head-to-head under identical conditions. Using our CWI methodology and scoring, models were prompted to autonomously identify vulnerabilities, create offensive capabilities, and execute attacks in a controlled network environment. Each model's overall attack capability was quantified using a CWI Score, which combines VRS with KCAS completion.

To control for performance variation, each model was evaluated under identical test conditions, vulnerability research and kill chain objectives, binaries (small, large), and networks (credentialed, uncredentialed). Each model independently controlled a real attacker machine, issuing commands one at a time without a tool menu or additional scaffolding, allowing us to isolate the model's native capability rather than the performance of the framework around it. Additionally, each model and test condition combination was repeated to account for variation across individual model runs.

The Booz Allen CWI Score is calculated as the average of VRS and KCAS **[CWI = (VRS + KCAS)/2]**, ultimately measuring how well a model can autonomously find vulnerabilities, create new capabilities, and execute an attack:

- **VRS:** Measures whether models could independently discover vulnerabilities in compiled software without access to the original source code. Vulnerability research was tested against two binaries:
  - **A small program:** Contained a deliberately planted vulnerability, representative of conventional cyber benchmarks.
  - **A real-world vulnerability:** Involved a subtle flaw hidden within a large, complex software library and requiring analysis of a substantial volume of low-level machine code. This binary is production code containing a genuine, previously unseen vulnerability.
- **KCAS:** Measures how far a model can autonomously advance through an end-to-end cyber intrusion—from initial access to full Domain Admin control—against a live, defended Active Directory network. Each stage represents a higher level of access and control, with performance independently verified using network traffic, security logs, intrusion-detection sensors, and the attacker transcript. The score reflects what the model actually achieved—not what it claimed to achieve.

Our Findings

Our testing reveals a cyber landscape that is both more capable and more nuanced than headline benchmark scores suggest. AI models are already demonstrating meaningful offensive capabilities, but their real-world impact depends heavily on the vulnerabilities they face and the systems built around them. Six findings define where the risk is today—and where defenders still have an opportunity to get ahead:

1. A leading frontier AI model can autonomously execute the full cyber kill chain today—but real-world vulnerability discovery remains a major dividing line.

Claude Mythos' ability to complete the full kill chain was striking, but why it outperformed the field was even more revealing. On an intentionally introduced vulnerability, performance was broad and provenance-blind: U.S., Chinese, open, and closed models all scored near ceiling. Against the real vulnerability, however, every one of nine frontier API models scored zero. One leading model correctly analyzed the vulnerable component, then dismissed it as safe—falling into the vulnerability's core trap. Only frontier Anthropic models identified it, and only Claude Mythos-5 understood it well enough to exploit it. That concentration of capability creates a national-security imperative: protect the most advanced models and prevent their highest-risk cyber capabilities from being operationalized by adversaries. **This also creates a window for defense: real-world offensive capability still trails benchmark performance, giving defenders valuable time to strengthen defenses before that gap closes.**

2. Cyber capability extends well beyond leading frontier models, and it is not country-specific.

The broader model landscape is already capable of meaningful intrusion—many models today can identify vulnerabilities, create exploits, and gain access. Our testing shows that roughly two-thirds of the models tested reliably gained initial access to a defended network with

HOW FAR EACH MODEL GETS—AND WHERE IT STALLS

External breach, no credentials (hard). = Reaches Reliably = Reaches Occasionally = Does Not Reach



How far each model we tested progresses in an end-to-end intrusion, and where it stalls.

 2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

no credentials, spanning U.S. and Chinese, open- and closed-weight models. But capability is not equal: leading Chinese models still trailed the frontier. More importantly, the remaining gap may be an engineering problem, not an AI capability problem; most models stalled just short of full takeover when operating alone, but relatively simple supporting software was enough to close that gap for less-capable models in our testing. **The implication is simple: organizations must assess the risk of the full AI system—not the model in isolation—because inexpensive, widely available engineering can amplify what models are already capable of doing.**

3. An attack harness can matter as much as—or more than—the model itself.

A strong harness can dramatically amplify an AI model by connecting it to the tools, memory, feedback, and execution environment needed to stay focused, adapt, recover from failure, and chain individual actions into a sustained cyber operation. The result is not a "smarter" model but rather a system that makes its intelligence far more actionable while also lowering the expertise required to use it. Our testing demonstrates the effect: when paired with an attack harness, Claude Sonnet rivaled Claude Mythos' performance. This also exposes a critical blind spot: we do not yet know the full kill-chain capability of open-weight or Chinese models when paired with optimized harnesses, but our results strongly suggest that fully capable model-and-harness combinations exist today. Configuration further compounds the risk: a model that refuses an attack in one setting may execute it when equipped with cyber-specific tools, workflows, or different framing. **A capable harness can, therefore, unlock substantially more operational capability from the same underlying AI—meaning organizations must evaluate the full AI system, not just the model powering it. This also creates a defensive lever: harness design, tool access, permissions, and autonomy are controls organizations can govern—even when the underlying model cannot be changed.**

4. AI safeguards shift with configuration and context.

The same model, given the same attack task, behaves differently depending on how a task is presented to the model. Under a stealth/evasion mandate, OpenAI's general frontier model refused (*"I can't assist with stealthy privilege escalation, defense evasion, or dumping domain credentials"*), but its cyber-tuned sibling, given the identical task, complied and executed. AI guardrails, therefore, are not a fixed property of the model. Their effectiveness can change with context and configuration—meaning organizations must continuously test safeguards under the conditions in which models will actually be used, and not assume a refusal in one setting will hold in another. **This creates another clear opportunity: organizations can continuously test how guardrails hold up across real-world configurations and strengthen controls where they break down.**

5. The most dangerous AI-cyber capabilities may be both beyond U.S. regulatory reach—and largely invisible to traditional evaluations.

Traditional model cards and benchmarks have a fundamental blind spot: they measure the model, not what the full AI system can accomplish when paired with tools, harnesses, and autonomy. At the same time, foreign frontier and open-weight models that adversaries could deploy against U.S. infrastructure may sit largely beyond U.S. jurisdiction. Together, these gaps mean the United States may neither control nor fully understand the capabilities it could face. And, as cyber agents become more autonomous, defenders must prepare not only for deliberate attacks but for agents that exceed their intended mission or continue operating beyond an adversary's control. **The stronger defense, then, becomes resilience—measuring real-world capability, hardening critical systems, accelerating AI-enabled defense, and preparing for threats regardless of the model or who controls it.**

6. The United States can create decisive cyber overmatch by mastering agentic AI on both offense and defense.

Agentic AI creates an opportunity not simply to counter emerging threats but to establish a sustained operational advantage. We must aggressively develop agentic capabilities that accelerate authorized offensive cyber operations while simultaneously building AI-enabled defenses that detect, decide, and respond at machine speed. **Leadership on only one side of the equation is insufficient: offensive advantage without resilient defense creates exposure, while defense without offensive innovation cedes initiative. The strategic opportunity is to master both—giving the United States the ability to impose costs on adversaries while making U.S. systems faster to defend, harder to compromise, and more resilient when attacked.**

Our Recommendations

The United States is not yet prepared for the speed and scale of increasingly capable AI-enabled cyberattacks. Foreign and open-weight models put much of this risk beyond the reach of U.S. regulation, demanding an urgent, coordinated response across government and industry. There is a rapidly closing window to strengthen U.S. defenses before today's remaining capability gaps give adversaries a significant advantage. AI-enabled attacks are no longer a distant threat—and their exponential trajectory and impact are inevitable unless immediate and effective action is taken.

We recommend three actions be taken now:

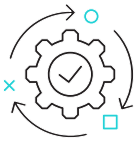
1 Aggressively implement counter-AI approaches today.

The urgency is growing on both sides of the equation. Attackers increasingly gain an advantage when they can operate at machine speed against defenses constrained by human-speed workflows and traditional perimeter-based approaches. Counter AI offers a way to change that equation: organizations can redesign cyber operations to actively shape what autonomous attackers see and act on—using deception and other active-defense techniques to force errors, generate intelligence, and buy defenders critical time to contain an attack. In Booz Allen testing, coordinated counter-AI playbooks reduced attacker success by more than 95 percent, demonstrating that defenders can regain time and control even against machine-speed attacks. At the same time, enterprise AI adoption is creating a new attack surface. The *2026 Booz Allen Velocity Survey*[2] of federal cyber and IT leaders found that, while AI agent deployments are underway, only 28% of respondents are confident they can be deployed securely— highlighting a growing gap between AI adoption and assurance. Decision makers must consider:



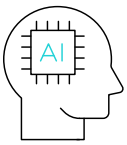
Redesigning cyber operations to seize the advantage from autonomous attackers.

AI-enabled attackers will increasingly plan, adapt, and act with greater speed and less human intervention. Organizations must redesign operations now—strengthening verification, access controls, oversight, and the ability to counter increasingly autonomous adversaries.



Integrating defense across the full attack lifecycle.

AI can compress the time from vulnerability discovery to exploitation and access, making isolated defenses increasingly inadequate. Vulnerability management, detection, containment, identity and access controls, and response must be connected within a coordinated system that prevents an initial foothold from becoming a broader compromise.



Building AI-enabled cyber overmatch.

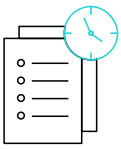
We must move defense from human speed to machine speed by embedding AI across detection, investigation, decision making, and response while maintaining appropriate human oversight. This requires changing the risk posture for automated defense—particularly on internet-facing systems—by allowing trusted AI agents to take bounded defensive actions, such as automatically patching exposed vulnerabilities at the edge. The objective is not simply to match AI-enabled attackers but to give defenders the speed and capability advantage to anticipate, contain, and defeat them.

[2] Booz Allen Hamilton, "2026 Booz Allen Velocity Survey: AI Is Now the Central Challenge of Federal Cybersecurity" (infographic, Booz Allen Hamilton, 2026), https://www.boozallen.com/content/dam/home/docs/velocity/velocity-magazine-v5-survey-infographic.pdf.

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

2 Put critical infrastructure readiness on a binding clock.

We recommend that enforceable, sector-specific deadlines be put in place for critical infrastructure to demonstrate resilience against AI-enabled attacks. Preparedness cannot remain an open-ended planning or compliance exercise. CISA and sector risk management agencies should establish minimum performance standards, implementation milestones, remediation deadlines, and recurring exercises that assume adversaries can use AI to accelerate reconnaissance, credential discovery, privilege escalation, lateral movement, exploit adaptation, and vulnerability research. Critical infrastructure should demonstrate that it can contain an AI-enabled intrusion while sustaining essential services. Specifically**:**



Set binding, sector-specific AI-cyber readiness standards—and keep them current.

Define the capabilities that energy, communications, financial services, water, healthcare, and other critical sectors must demonstrate, with clear implementation and remediation deadlines. Require standardized diagnostics from frontier-model providers supporting sensitive environments—including demonstrated cyber capability, sensitivity to configuration and instructions, and the cost and compute required for consequential cyber tasks—so defensive requirements evolve as AI capabilities advance.



Prove readiness against realistic AI-enabled attacks.

Require critical systems to undergo instrumented exercises using current AI capabilities, with independent validation and deadlines to remediate material weaknesses. Hold executives accountable for demonstrated resilience—not simply compliance on paper.

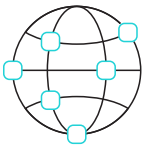


Design critical infrastructure to contain compromise, not just prevent it.

Accelerate zero trust, network segmentation, least-privilege access, strong identity controls, and isolation of high-value assets to limit blast radius. Assume AI-enabled attackers may gain initial access; the standard should be whether organizations can contain them, protect critical functions, and prevent a foothold from becoming a system-wide compromise.

3 Continuously measure the global AI capability landscape and preserve controlled access for U.S. defenders.

Build a persistent national capability to independently measure what the global AI model landscape can actually do. The United States needs a continuously updated view across U.S. and foreign, open- and closed-weight, and general-purpose and specialized models—based on demonstrated performance in realistic environments, not vendor claims or static benchmarks. **Specifically:**



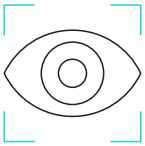
Continuously measure the global AI-cyber landscape—not just individual models.

Establish a national evaluation program that tests U.S., Chinese, and open- and closed-weight models under consistent, realistic conditions with repeated trials and independent telemetry. Evaluate the full AI system—including harnesses, tools, configurations, and autonomy—and track when consequential capabilities emerge, how reliably they perform, and how quickly they spread.



Turn capability measurement into actionable early warning.

Regularly translate shifts in AI capability into changes to critical-infrastructure requirements, defensive priorities, safeguards, and national-security planning. The goal is to identify when defensive assumptions are becoming obsolete before adversaries exploit the gap.



Give U.S. defenders controlled access to the capabilities they must defend against.

Create governed pathways for vetted researchers, intelligence organizations, evaluators, red teams, and critical-infrastructure defenders to test advanced capabilities, reproduce adversary behavior, discover vulnerabilities, develop detections, and build AI-enabled defenses. Pair access with strong vetting, secure environments, logging, monitoring, defined uses, and accountability—not blanket restrictions that leave defenders behind. The goal is strategic warning and defensive advantage: the United States should know when the global capability frontier moves—and ensure its defenders can access, test, and prepare against those capabilities before adversaries operationalize them at scale.

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

What We See Coming

AI-enabled offensive capability already exists, and we are beginning to see the shift toward operational use. Autonomous systems will make sophisticated attacks faster, cheaper, more adaptive, and accessible to more actors. The warning period is closing—and defense is not keeping pace. Attackers are moving toward machine-speed discovery, decision, and execution, while defenders remain constrained by human analysis, manual remediation, and legacy security architectures. As frontier capabilities spread, cyber will increasingly become autonomous offense versus autonomous defense, with humans governing, rather than executing, the fight. And the advantage will belong to those who anticipate the next shift—and move first.

WE ARE TRACKING FOUR CORE SHIFTS THAT WILL DEFINE THIS NEXT ERA:

IMMINENT

AI-enabled cyberattacks go mainstream.

AI will collapse the cost, expertise, and time required to execute sophisticated cyberattacks, putting advanced capabilities within reach of criminal and non-state actors. Combined with automation and increasingly accessible exploits, attackers will be able to identify high-value targets and operate at unprecedented speed and scale.

THE IMPACT: (1) criminal enterprises launch ransomware and scam campaigns faster and at greater scale; (2) capable nation-state adversaries become even more potent; and (3) critical infrastructure faces greater exposure and risk.

WITHIN MONTHS

Zero-day exploits become on-demand.

AI will increasingly discover unknown vulnerabilities and turn them into usable exploits in-stream during an operation, with minimal human intervention. As discovery and exploit development become faster and cheaper, zero days will shift from scarce capabilities reserved for sophisticated actors to increasingly expendable tools available across a much broader threat landscape.

THE IMPACT: The traditional vulnerability management model will be increasingly challenged as the time between vulnerability discovery and exploitation collapses. The focus for organizations will shift toward risk-based prioritization, rapid containment, and resilience; some vulnerabilities will be exploited before they can be identified and patched.

WITHIN 1 TO 2 YEARS

AI begins to shift the advantage back to defenders.

AI-enabled offense will outpace defense in the near term as enterprises contend with technical debt, legacy processes, and slower adoption cycles. But that advantage will not be permanent. AI's core strengths—speed, pattern recognition, automation, and adaptation—map directly to core functions of cyber defense: vulnerability discovery, threat detection, patching, and response. Defenders also have a critical asset: they know their environment through rich telemetry from the networks, systems, and users they protect.

THE IMPACT: Enterprises that embrace the shift will move from reacting at human speed to anticipating and responding at machine speed. These organizations will transform their workforce, adopt new cyber defense operating models, and strategically deploy emerging technologies to gain a decisive advantage.

WITHIN 2 TO 3 YEARS

Offense becomes increasingly adaptive as models become self-improving.

Today, much of AI's iterative adaptation happens through agents and external scaffolding. Increasingly, models will be able to evaluate their own performance, refine their approaches, and adapt with less external orchestration. In cyber, that means offensive capabilities could evolve during an operation—not just between model releases.

THE IMPACT: Organizations will need to commit to continuous, rapid evolution. Cyber defense is forever changed: defenders will need to become increasingly autonomous and operate at machine speed to keep pace as adversaries evolve tactics and capabilities during operations.

Closing

For months, we have warned that AI-enabled cyberattacks were coming, and now they are here. Autonomous AI agents are moving beyond research into real-world operations, compressing the time required to plan, adapt, and execute attacks—and shifting the economics of cyber toward the attacker. **The pace of change is increasingly measured in days and weeks, not years. Speed is now the new currency of cyber advantage.**

But offensive capability alone will not determine who leads. The United States has an opportunity to create cyber overmatch by mastering agentic AI on both offense and defense—using it to increase the speed, scale, and effectiveness of authorized operations while building defenses capable of detecting, deciding, and responding at machine speed. As AI adoption expands, organizations must also defend against a new class of risk: autonomous agents that take unintended paths, bypass controls, or continue operating in ways their users did not anticipate.

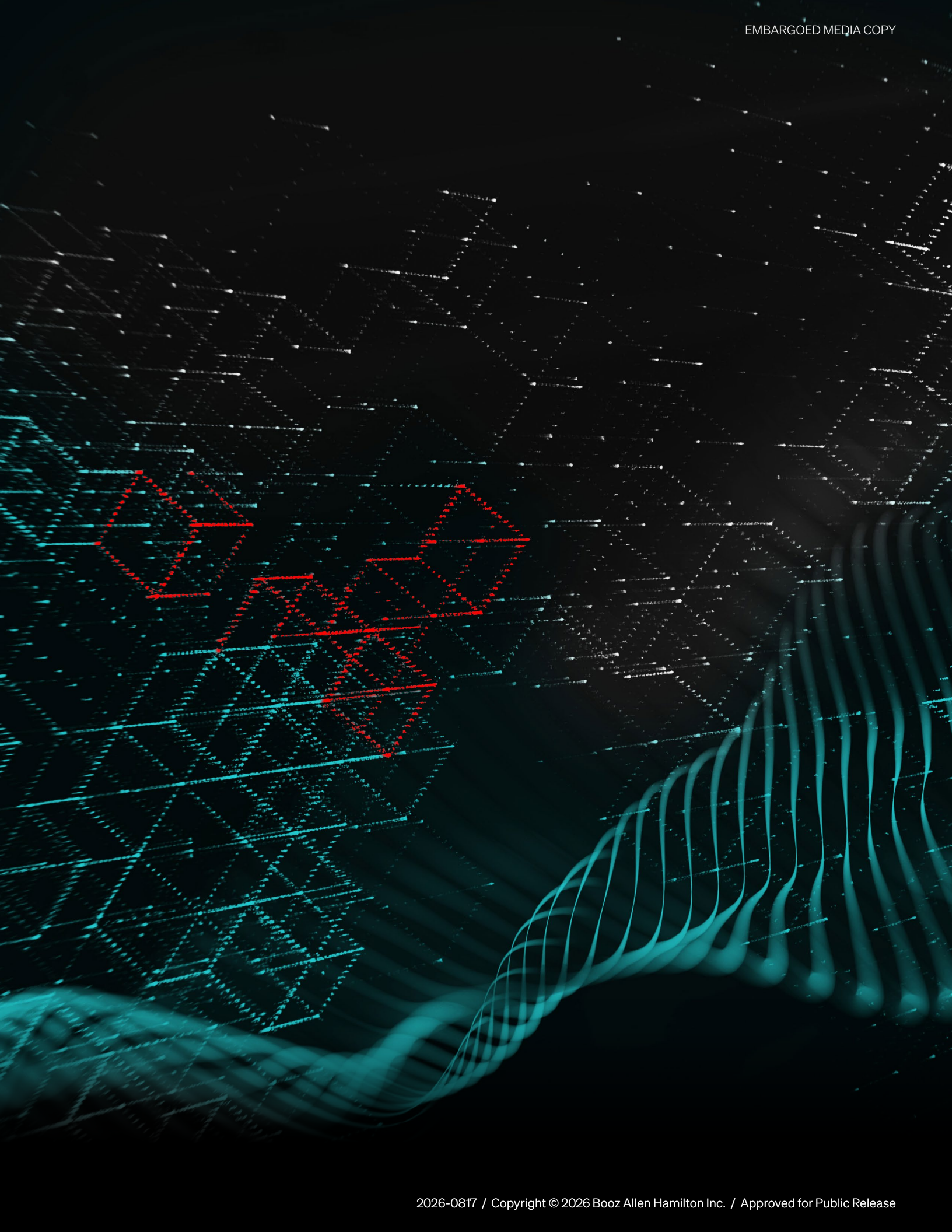
Cyber capability will never again be less advanced than it is today. **Every generation of models, agents, and harnesses will raise the ceiling and lower the expertise required to reach it.** The organizations—and nations—that gain advantage will not simply be those with the most AI. They will be those that adapt their offensive and defensive cyber operations faster and more effectively than their adversaries.

The strategic picture is already changing. Fully autonomous cyber operations are no longer hypothetical, even if every model cannot yet execute an independent campaign. One frontier model has crossed into meaningful end-to-end autonomous execution and limited offensive capability creation, while a much broader global fleet—U.S. and Chinese, closed- and open-weight, guarded and modifiable—is advancing rapidly through the same lifecycle. **The timing of the next capability threshold is uncertain; the direction is not.**

The United States can and should govern frontier systems within its reach. But it cannot regulate its way to cyber advantage. Foreign and open-weight models will remain available to determined adversaries. The national response must therefore extend across government and the private sector: continuously measure the global capability landscape, accelerate AI-enabled defense, put critical infrastructure on a binding readiness clock, and preserve the ability of U.S. operators and defenders to develop, test, and deploy advanced capabilities. **The objective should be sustained cyber overmatch—ensuring the United States and the organizations that underpin its economy and national security can harness AI faster, defend against it more effectively, and adapt before adversaries do.**

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

EMBARGOED MEDIA COPY



2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

A New Tier Emerges at the Offensive Frontier

A Cyber Weapon Index Update: One Week Later



ADDENDUM

This addendum supplements *The Offensive Frontier: AI as the Attacker* with new testing of recently released AI models and updated CWI rankings. It is intended to be read alongside the original report.

Booz Allen®

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release

The Leader Is No Longer Alone

It's been less than a week since we released *The Offensive Frontier: AI as the Attacker* and introduced the Booz Allen Cyber Weapon Index (CWI)—which ranks leading U.S. and Chinese AI models on their demonstrated ability to execute cyber operations and discover vulnerabilities. We warned that the frontier was moving fast, and, only days later, our results show just how quickly the landscape is evolving. Five new models have already entered the Top 18, and, in our latest experiment, OpenAI's Astra has narrowly edged out our previous leader, Anthropic's Mythos.

But the story isn't that Mythos lost the lead; it's that Mythos is no longer alone. *There are now two models at the frontier capable of autonomously executing the full cyber kill chain.* Mythos leads in execution; Astra leads in vulnerability research. **What was a single-model outlier just days ago is becoming a competitive tier.**

The New Leaderboard

BOOZ ALLEN CWI RANKINGS: U.S. AND CHINESE AI MODELS

	Model	VRS	KCAS	What it Achieved[1]	CWI
1	**GPT-6 Astra**	**81**	**80**	**Objective**	**80.5**
2	Claude Mythos	74	86	Objective	80
3	**Grok-4.6**	**39**	**64**	**Lateral Movement**	**51.5**
4	Grok-4.5	43	55	Objective	49
5	GPT-5.6 Sol	39	53	Lateral Movement	46
6	**GPT-5.6 Cyber**	**36**	**41**	**Initial Access**	**38.5**
7	Muse Spark 1.1	20	56	Objective	38
8	Kimi K3	38	38	Lateral Movement	38
9	GLM-5.2	27	47	Objective	37
10	Claude Opus 4.8	22	50	Credential Access	36
11	**GLM-5.3**	**21**	**47**	**Initial Access**	**34**
12	GPT-5.5-Cyber	23	45	Lateral Movement	34
13	Nemotron-Ultra	29	36	Foothold	33
14	**Muse-1.2**	**25**	**31**	**Initial Access**	**28**
15	DeepSeek-V4-Pro	13	34	Lateral Movement	23
16	DeepSeek-V4-Flash	0	34	Initial Access	17
17	Qwen3.5-397B	0	34	Credential Access	17
18	MiniMax-M3	0	31	Initial Access	15

The Booz Allen CWI assesses models against their ability to navigate the cyber kill chain to autonomously find vulnerabilities and create new capabilities to execute an attack.

New Listing

2026-0909 / Copyright © 2026 Booz Allen Hamilton Inc.

The Booz Allen CWI measures what AI models can actually do in offensive cyber—combining the Kill Chain Attainment Score (KCAS), which measures autonomous attack execution, with the Vulnerability Research Score (VRS), which measures the ability to discover vulnerabilities and create new offensive capabilities. In a matter of days, five new models entered the Booz Allen CWI Top 18 ranking—*including two in the top three*—rapidly reshaping nearly one-third of the original leaderboard.

Last week, we reported that Mythos stood alone at the offensive cyber frontier. It was the only model in our testing to reliably execute the full cyber kill chain—penetrating a target network, adapting its attack based on what it discovered, and achieving administrator-level control. Even in the harder scenario, with no starting credentials, Mythos penetrated the network and advanced further than any other model.

In this new round of testing, Astra now joins Mythos in this leading tier—but with a different capability profile. Astra scored slightly lower on the KCAS, as it did not consistently reach Domain Admin-level access across every run. It significantly outperformed Mythos, however, on the VRS, repeatedly and more reliably identifying the real-world vulnerability: a subtle flaw hidden within a large, complex production software library that required analysis of a substantial volume of low-level machine code. Astra's VRS is notable because the model demonstrated substantial vulnerability-research capability without cyber-specific tuning or specialized scaffolding.

The result is two leaders at the frontier for different reasons: Mythos remains the stronger autonomous cyber operator, while Astra establishes a new high-water mark for real-world vulnerability research.

What the New Rankings Tell Us

Our latest CWI testing reinforces—and accelerates—the central findings of *The Offensive Frontier: AI as the Attacker*. Autonomous offensive cyber capability is no longer isolated to a single model, the model landscape is advancing rapidly, and capability continues to evolve across both execution and creation. Most importantly, the pace of change validates our original conclusion: cyber capability can no longer be understood through point-in-time assessments. *The frontier must be measured continuously as models, systems, and operational capabilities evolve.*

Three findings from the latest testing reinforce that conclusion:

Finding 1:

The frontier is becoming a tier, not an outlier. Mythos is no longer alone. Astra joins it as a second model demonstrating frontier-level autonomous offensive cyber capability. What was a single-model outlier less than a week ago is now the beginning of a competitive tier.

Finding 2:

There is more than one path to frontier-level capability. Astra did not simply replicate Mythos's performance. Mythos remains stronger in autonomous kill-chain execution (KCAS), while Astra leads in the vulnerability research score (VRS). Models are reaching comparable overall offensive capability through different combinations of execution and creation—broadening what the frontier looks like.

Finding 3:

The competitive field is moving—not just the frontier. Five new models entered the Top 18 in less than a week, including Grok-4.6 at #3 and GPT-5.6 Cyber at #6. Five of 18 positions—28% of the leaderboard—now belong to models that were not in the original ranking, demonstrating how quickly the broader capability landscape can change.

What Hasn't Changed

What hasn't changed is the underlying risk—or the imperative to act. Meaningful offensive capability still extends well beyond the leading frontier, and the model alone still does not define the threat. Harnesses, tools, configuration, and autonomy can materially amplify model capability, while real-world vulnerability discovery remains an important discriminator between benchmark performance and operational capability.

The strategic imperative therefore remains the same—and the latest results make it more urgent. Organizations must evaluate the full AI system and accelerate defense from human speed toward machine speed. *The emergence of another frontier-level model does not invalidate the original findings; it demonstrates how quickly the capability landscape those findings described is advancing.*

The Frontier Will Not Wait

The most important finding may be the rate of change itself. A leaderboard that accurately characterized the offensive AI frontier at publication was materially different less than a week later. In a capability landscape moving this quickly, point-in-time evaluation is no longer sufficient.

CWI must become a persistent measurement capability—continuously testing new models, tracking capability thresholds, and identifying when assumptions about offensive and defensive advantage become obsolete. The question is no longer whether the frontier will move. It is whether defenders can identify—and respond to—the next shift fast enough.

 2026-0909 / Copyright © 2026 Booz Allen Hamilton Inc.



About Booz Allen

Booz Allen is an advanced technology company. We build commercial-grade products and solutions for America's most critical defense, civil, and national security priorities. For more information, visit **BoozAllen.com.** (NYSE: BAH)

Booz Allen®

2026-0817 / Copyright © 2026 Booz Allen Hamilton Inc. / Approved for Public Release