

September 2026

# Beyond the frontier AI patch race: The security leader's playbook to AI-era cyber resilience

# Contents

## 03 Introduction

---

## 04 The frontier AI turning point

---

## 06 The defender's asymmetry and the patch-gap crisis

---

## 07 What comes next: the frontier AI trajectory

---

## 10 How NTT DATA helps organizations — AI for security and security for AI

---

## 11 Conclusions

---

## 12 Sources

---

# Introduction

Frontier AI models have reached a turning point in 2026. Multiple systems demonstrate that AI has crossed a meaningful threshold in offensive security capability. Both Anthropic's Claude Mythos Preview and OpenAI's GPT-5.5 Cyber have been assessed as capable of finding high-severity vulnerabilities, constructing working exploits and chaining complete attack paths, largely autonomously.

Security leaders, however, should not focus on a single AI model or incident, be it Anthropic, OpenAI, Google, open-weight developers or any other option. The executive takeaway is that these models signal a structural shift in the threat landscape: Autonomous cyber campaigns will probe enterprise systems continuously — at machine speed — to find new, exploitable attack paths.

This perspective focuses on NTT DATA's answer to a more practical question: What can frontier AI verifiably do in vulnerability research today — and what should enterprise security leaders do about it?

Our main observation is that the real impact to business is not in vulnerability discovery. AI-assisted scanners are already effective at finding known vulnerability classes. The breakthrough is in converting known vulnerabilities into working exploits and chaining them together into workable exploit paths. That capability shortens the attacker's timeline. And that is the most fundamental shift that available independent evidence supports.

It also changes how security leaders should think about cyber resilience and their organization's ability to accelerate AI transformation without increasing unmanaged risk. Because defenders have the same capabilities that shorten the attacker's timeline, evidence suggests the winning response is not simply to patch faster. You need to strengthen software assurance, remediation discipline, AI governance and resilience by design — embedding it into the organization's architecture and operating model.

This paper makes the case that patching faster is necessary, but insufficient, and that the answer is a three-pronged structural shift: Make exploitation harder even when exposure exists. Reduce the blast radius when a breach occurs. Instrument the remediation process, not only detection, so the exposure that matters most to the business is fixed first. We outline five AI capability trajectories — from standing up patch-diffing pipelines and pre-positioning virtual patching to segmenting against exploit chains, extending governance to AI-assisted workflows and building defender-side continuous autonomous research — as well as a matched set of offensive and defensive action commitments, for which security leaders should plan over the next four quarters and into FY2027.



# The frontier AI turning point

AI-assisted vulnerability research predates the current generation of frontier models. By early 2026, several systems were being used against large codebases with measurable results. Available evidence suggests that AI-assisted vulnerability research was becoming mainstream before the most recent wave of frontier model announcements.

- Pattern observed before the latest frontier generation: higher throughput and better signal-to-noise than legacy tooling
- Strength before the latest frontier generation: finding new instances of known bug classes
- Gap before the latest frontier generation: autonomous chaining from a low-severity primitive to a working exploit remained human-led

Examples include curl maintainers reporting AI-assisted bug fixes and Common Vulnerabilities and Exposures (CVEs). Google's Big Sleep demonstrated AI-assisted discovery of memory-safety bugs.<sup>1</sup> Plus, other systems are being used for large-scale code review.<sup>2</sup>

In April 2026, Anthropic said an unreleased Claude Mythos Preview was capable enough at offensive security tasks that the company declined to release it publicly. Instead, the company distributed the model to roughly 40 vetted organizations under Project Glasswing — a coordinated, cross-industry defensive initiative that brings together AWS, Anthropic, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA and Palo Alto Networks.<sup>3</sup>

Around the same time, OpenAI's GPT-5.5 Cyber demonstrated comparable capabilities in independent benchmark testing.<sup>4</sup> The LaoBench — a realistic end-to-end cyber operations assessment — showed that with a budget of 100 million tokens, both Mythos and GPT-5.5 Cyber were able to achieve a full autonomous network takeover.<sup>5</sup> The cost of 100 million tokens for a sophisticated attacker organization is well below €10,000, making frontier-grade offensive AI economically accessible.

Reaction was split. Some observers saw a defensive turning point. Others dismissed the framing as safety theater. One practitioner's account captured the tension as a "boy-who-cried-wolf" moment, while still treating the broader capability trend as real.

By June, Anthropic reported that Claude Mythos Preview had surfaced more than 23,000 potential vulnerabilities across more than 1,000 open-source projects. Of the approximately 1,900 vulnerabilities external firms reviewed, 1,726 were confirmed — including more than 1,000 rated high or critical.<sup>3</sup> Throughout this perspective, we treat vendor claims as claims, rather than findings.



**Executive takeaway:** The appropriate response is not to assess which company's model is the most capable at any given moment. It is to recognize that multiple frontier models are crossing the same capability threshold in close succession, and that the enterprise response must be model-agnostic and structurally grounded.

Where frontier AI cyber capabilities make the strongest impact

Vulnerability research is not one task. The table below breaks down the research lifecycle into stages to show where AI adds genuine value and where capability remains noisy or human-dependent.

| Stage                          | What it involves                                                      | AI uplift in 2026                                                                                 |
|--------------------------------|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| Recon and scoping              | Mapping a codebase, trust boundaries, entry points and attack surface | <b>Strong.</b> Agents produce architecture summaries and task queues that orient downstream work. |
| Vulnerability discovery        | Finding probable flaws in source code or binaries                     | <b>Mature</b> but not novel. AI finds instances of known bug classes rather than new classes.     |
| Triage and reachability        | Deciding which vulnerabilities are real and externally reachable      | <b>Improving</b> but noisy. Reachability still needs a separate pass.                             |
| Primitive development          | Turning a vulnerability into a read/write or control primitive        | <b>Big change.</b> The capability that visibly improved most.                                     |
| Chaining and full exploit path | Combining primitives into a working end-to-end exploit                | <b>Big change,</b> but evidence is mixed across practitioner reports and vendor claims.           |

The real breakthrough is in exploitation, not discovery

Current evidence is concrete and convergent: Frontier models may find many of the same underlying vulnerabilities. However, the meaningful shift is the ability to stitch pieces together and autonomously generate exploitable attack paths. The breakthrough is not, “AI found a vulnerability.” It is, “AI helped validate exploitability and connect it to a credible attack path.”

**Why does validation matter?** A suspected exposure without evidence-backed validation is speculation. A model that closes the validation loop reduces the human verification step used to gate severity decisions.

**Executive takeaway:**  
This is where cyber risk becomes a board-level resilience issue. The risk is not only that vulnerabilities exist, but that exploitability can be validated faster and at greater scale.

# The defender’s asymmetry and the patch-gap crisis

If time-to-exploit is collapsing while time-to-patch is not, the defender’s core constraint shifts from detection to remediation speed and resilience. Several structural factors make this challenge increasingly difficult.

- Every patch is an exploit hint: A security fix can show attackers exactly what was broken.
- Open source discloses early by design: Fixes may land publicly before stable releases ship.
- Commit noise is no longer reliable cover: Models can read and compare large volumes of code changes.
- Patch gaps exist everywhere: Bundled dependencies and embedded components can keep known flaws alive after upstream fixes are available.

**Executive takeaway:**  
The question is not only, “Can we patch faster?” It is, “Can we reduce exposure and the blast radius when patching cannot happen fast enough?”

## The two-hour-SLA fallacy

The instinctive response — patch faster — is necessary but insufficient. If regression testing takes a day, a two-hour patch service-level agreement (SLA) may require the team to skip testing, which can create new business risk. Speed is part of the answer, but it is not the whole answer.

| Threshold | Status    | Approximate year |
|-----------|-----------|------------------|
| 1 year    | Reached   | ~2021            |
| 1 month   | Reached   | ~2025            |
| 1 week    | Reached   | ~2026            |
| 1 day     | Reached   | ~2026            |
| 1 hour    | Projected | ~2027            |
| 1 minute  | Projected | ~2027            |

Source: zerodayclock.com, based on 3,500+ confirmed-exploited CVEs (CISA KEV + VulnCheck KEV)<sup>6</sup>

## Reliable architecture over speed

The more reliable principle is to make exploitation harder even when exposure exists. This means placing defenses in front of applications, designing so one component cannot easily compromise others and rolling out fixes consistently across environments. Doing so reframes the problem from winning the patch race to reducing the blast radius of losing it.

**Executive takeaway:**  
Resilient architectures, adaptive security controls and secure-by-design practices help organizations maintain trust even when the threat landscape accelerates.

# What comes next: the frontier AI trajectory

The patch-gap is the most immediate concern that organizations need to address. But there must be a larger discussion around structural change as the capabilities of the next generations of AI models continue to improve. CISOs and other security leaders should plan for the following five trajectories.

## Trajectory 1: Operator-time compression

The current baseline — from the Hacktron Chrome chain — is roughly 20 hours of expert operator time and a few thousand dollars in tokens to produce a working two-bug exploit chain.<sup>7</sup> Compressing operator time by two-to-three times within 12 months is our high-confidence prediction. A 10-times compression within 24 months is plausible.

### Executive takeaway:

The operator remains important, but AI can reduce the time and cost required to produce working exploit chains.

## Trajectory 2: Continuous autonomous research

When the cost per exploitable finding falls into the low hundreds of dollars, the economic floor for “always-on” scanning collapses. Adversaries with token budgets will run continuous discovery. The threat picture stops looking like a sequence of disclosed vulnerabilities and starts looking like constant background pressure.

### Executive takeaway:

Lower cost per exploitable finding could make always-on research a steady state for attackers and defenders alike.

## Trajectory 3: Tooling commoditization; context as advantage

The Cloudflare harness pattern (parallel narrow agents, adversarial validators, reachability tracers, schema-constrained output) is now a publicly documented reference design.<sup>8</sup> Within 12 months, harness sophistication will not be a meaningful differentiator. Structured codebase context will be.

### Executive takeaway:

Harness patterns may become repeatable while structured codebase context becomes the defensible advantage.

## Trajectory 4: Open-weight diffusion timeline

Frontier AI developers' own leadership has publicly estimated that open-source and Chinese frontier developers could reach current top-model-level cyber capability within 6–12 months of the most recent announcements.<sup>9</sup> Guardrail-free open-weight models, which currently perform at near-frontier levels on offensive security benchmarks, accelerate this timeline further.

### Executive takeaway:

The period where frontier-class offensive capability remains limited may be short. Defender programs should not depend on capability scarcity.

## Trajectory 5: Novel-bug-class question

The current evidence is uniform: AI tools find new instances of known bug classes, not new bug classes. Whether this changes is the single most consequential open question in the field of vulnerability research. Defender programs should be designed to detect any change when it occurs.

### Executive takeaway:

Current evidence shows AI finding instances of known bug classes. The first reported novel-class finding will compress timelines on everything else in this paper.

Action plan: authorized offensive and defensive security

The following actions are matched pairs: Nearly every offensive development has a defensive mirror. The offensive points are written for red teams, capability assessors and threat-emulation functions operating under authorization, with the goal of anticipating the adversary not enabling them.

Offensive security: four commitments for the next four quarters

| Action                                           | Recommended approach                                                                                                                                                                                                                                                                      |
|--------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stand up a patch-diffing pipeline                | Subscribe to the security-commit feeds for your critical open-source dependencies. Route every relevant commit into your harness within 24 hours of landing. Deliver an internal advisory with an exploitability assessment and reachability trace into your own consumer applications.   |
| Move scoring from bug count to chain depth       | Retire raw-finding count as a key performance indicator (KPI). Replace it with primitives chained per engagement, blast radius achieved from the initial foothold and defender response triggered. A red team that ships 100 unchained findings is testing yesterday’s adversary.         |
| Run force-multiplier exercises                   | Scope at least one engagement per half-year around a five-person team running three parallel agent sessions against a realistic target environment. The credible near-term adversary is a small, capable team whose throughput is multiplied.                                             |
| Pre-position for continuous autonomous emulation | Plan FY2027 around a budget line token spend equivalent to the salaries of one or two senior researchers who are dedicated to always-on agentic research against your own and your customers’ estates. The operational scaffolding takes longer to build than the model takes to improve. |



Defensive security: seven commitments for FY 2027

| Action                                                | Recommended approach                                                                                                                                                                                                                                                           |
|-------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Reach 95% authoritative dependency inventory          | Create an application-level bill of materials, including bundled and embedded components (notably Electron and embedded Chromium). Send a weekly version-drift report to security engineering. The most common question organizations cannot answer is, "What are we running?" |
| Eliminate "update later" for security categories      | Force auto-updates on security patches across user-facing software, with a hard cap (suggested: 7 days) on user-deferred non-security updates. Treat the dialog box that lets a user defer a security update as a defect, not a UX feature.                                    |
| Pre-position virtual patching for top bug classes     | Use deployable Web Application Firewall (WAF) and runtime-protection rule templates for the bug classes most relevant to your stack. Measure the time from CVE disclosure to the rule in production date. Target by Q42026: under 4 hours for top bug classes.                 |
| Plan for defender-side continuous autonomous research | Conduct continuous find-and-patch against your own code. The same economics that help attackers help you. So, plan a token budget that exceeds what the adversary will plausibly spend. That is the strategic prize.                                                           |
| Segment and compartmentalize                          | Frontier models can chain three low-severity findings from different systems into a critical exploit path. Compartmentalization limits blast radius when chaining occurs.                                                                                                      |
| Extend governance to AI-assisted workflows            | Any workflow that uses AI to generate or assess code needs a governance layer. This includes human review checkpoints, logging, output validation and escalation paths.                                                                                                        |
| Instrument remediation, not simply detection          | Track exploitability validation time, reachability confirmation, critical asset coverage, virtual patch deployment time, remediation verification and reduction in business-critical exposure.                                                                                 |



# How NTT DATA helps organizations: AI for security and security for AI

The actions we present in this paper are practical priorities for security programs over the next four quarters and into 2027 as organizations prepare for a more demanding AI-enabled security environment. NTT DATA's Advanced Offensive Security and Vulnerability Research practice actively track these developments to help our clients strengthen software assurance, remediation readiness, architectural resilience and long-term preparedness.



## AI-assisted threat emulation

We use frontier-AI-assisted methods on the red-team side to surface attack chains that conventional engagements may miss. The model handles the combinatorial work — composing primitives across systems, reasoning about reachability, identifying the unobvious pivot. The deliverable is a more complete picture of how an adversary could move through the environment, not a longer list of unrelated findings.



## Vulnerability research as an internal capability

We run frontier-AI-assisted vulnerability research on our own scope and public dependency ecosystems, on which our clients rely, and then feed the learning back into our threat-emulation engagements.



## Architectural resilience advisory

We help our clients review compartmentalization, dependency posture, virtual patching and rollout-SLA instrumentation. The deliverable is a program that survives losing the patch race, not one that depends on winning it.



## Continuous and managed services

We offer a managed service that operationalizes the continuous-autonomous-research capability. It compares always-on agentic research against client codebases and declared critical dependencies, with a feedback loop into clients' system operations centers and incident response.

Vulnerability management cannot be isolated from AI, cloud and governance strategy. NTT DATA has full-stack expertise across AI, cloud and cybersecurity. This knowledge uniquely positions us to help our clients embed security and governance into their organization's AI transformation journey. We also help transform your cybersecurity operating model with AI technologies so you can fight AI with AI.

# Conclusion

Frontier AI models — from Anthropic, OpenAI and others emerging across the open-weight ecosystem — are best understood as signals of where AI capability is heading. Not as individual product events. Each new model generation that crosses the exploitation threshold renews the same enterprise challenge.

The key conclusion is that the most significant shift is the acceleration from vulnerability identification to exploitability validation and exploit-chain construction. For enterprises, the response should be model-agnostic and operationally grounded. Leaders should strengthen software assurance, remediation operations, dependency visibility, virtual patching, architectural resilience and AI governance.

The goal is to improve readiness and build resilience with an AI-driven and unified cybersecurity operating model. Organizations that adapt early will be better positioned to protect stakeholders' trust, strengthen cyber resilience and securely enable AI-driven transformation. These companies will make cybersecurity a strategic advantage for their business in the AI execution era.



To see if your organization's vulnerability management and remediation processes can build AI-era resilience, contact us about our [Frontier AI Readiness Assessment](#).

Visit [nttdata.com](https://nttdata.com) to learn more.

NTT DATA is a \$30+ billion business and technology services leader in AI and digital infrastructure. We accelerate client success and positively impact society through responsible innovation. As a Global Top Employer, we have experts in more than 70 countries. NTT DATA is part of NTT Group.



# Sources

1. Google Project Zero / Big Sleep. "AI-assisted discovery of memory-safety vulnerabilities in production software." 2025–2026. Practitioner source.
2. Stenberg, D. (curl maintainer). "AI-assisted bug reports and CVE submissions in open-source projects." 2025–2026. Practitioner source.
3. Anthropic. "Project Glasswing: Securing critical software through coordinated frontier AI research." April 2026. Vendor source.
4. OpenAI. "GPT-5.5 Cyber capability assessment and LAOSS benchmark results." 2026. Vendor source.
5. LAOSS Benchmark Consortium. "Last Ones Standing: An end-to-end cyber operations benchmark for frontier AI." 2026. Independent source.
6. zerodayclock.com. "Time-to-exploit milestone tracker." Based on 3,500+ confirmed-exploited CVEs from CISA KEV and VulnCheck KEV with VulnCheck XDB timestamps. Accessed June 2026. Independent source.
7. Hacktron Security Research. "Autonomous exploit chain construction against unpatched Chromium: a V8 case study." 2026. Practitioner source.
8. Cloudflare Research. "Parallel narrow agent harnesses for vulnerability research: a reference design." 2026. Practitioner source.
9. Wiz Research. "Open-weight frontier model capability diffusion timeline." 2026. Practitioner source.

## Additional reading

- NTT DATA. "Cloud-led Innovation report." 2026. Pages 3, 11–12, 23–24.
- NTT DATA. "Private and Sovereign AI report." 2026. Pages 9, 11, 12.
- CISA / European Union Agency for Cybersecurity (ENISA). "Return on Security Investment (ROSI) framework."



