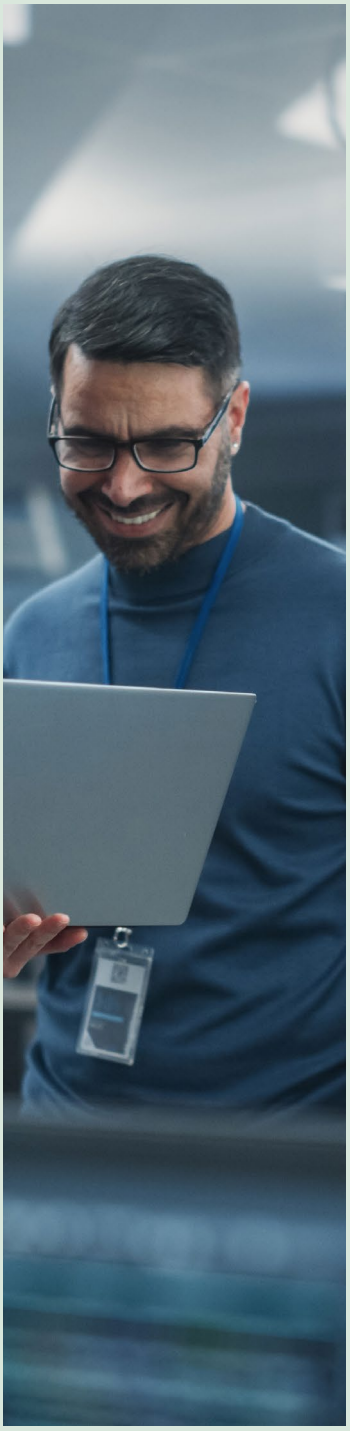
WHITE PAPER



EMERGING FRONTIER AI MODELS IN CYBERSECURITY: CLAUDE MYTHOS VS GPT-5.4-CYBER

Infosys
Navigate your next

Section	Page Number
Executive Summary	3
Introduction: What Frontier AI Means for Cybersecurity	4
When Cybersecurity Meets Frontier AI	5
Why Enterprises Must Act Now	6
Model Capabilities (Claims & Benchmarks)	7
Comparative Analysis: Claude Mythos vs. GPT-5.4-Cyber	10
Enterprise Risks and the Expanding Threat Horizon	11
Strategic Recommendations for Enterprise Leaders	13
Best Practices for Governing Frontier Cyber AI	14
Conclusion	15
References	15

External Document © 2026 Infosys Limited

1. Executive Summary

The cybersecurity landscape is moving from human-driven defensive heuristics to autonomous, agentic artificial intelligence. The latest frontier models have significantly demonstrated measurable advances in software engineering, reasoning, coding, bug fixing, and cybersecurity. These systems are increasingly capable of identifying previously unknown vulnerabilities, supporting exploit analysis, assisting in reverse engineering from binary code, and accelerating tasks such as malware analysis, red teaming, threat hunting, and vulnerability remediation.

In April 2026, within a week, Anthropic and OpenAI introduced two frontier AI models with advanced cybersecurity relevance: Claude Mythos Preview and GPT-5.4-Cyber. Anthropic's Claude Mythos Preview, released through the highly restricted Project Glasswing initiative, has been positioned as a major leap in autonomous vulnerability discovery and exploit-oriented security research. OpenAI's GPT-5.4-Cyber, introduced through its expanded Trusted Access for Cyber (TAC) framework, is designed specifically for verified defenders and is notable for enabling advanced defensive workflows such as binary reverse engineering without source code access.

For enterprises, the opportunity lies in using these capabilities to accelerate cybersecurity operations and better the speed and depth of analysis. Simultaneously, their dual-use potential foregrounds risks that call for careful consideration of deployment, oversight, access controls, and governance. Organisations must prioritise addressing accumulated technical debt, implement rigorous agentic identity management to control shadow AI, and move towards continuous, automated validation of their security posture.

Infosys Cyber ARC interprets Anthropic's and OpenAI's published system cards and claims through the lens of its own experience defending enterprise clients against AI-accelerated threats. This paper evaluates the capabilities and claims associated with these models, compares their strengths and deployment approaches, and outlines the opportunities, risks, and strategic considerations they introduce for red teaming, threat hunting, and cyber defence. It offers a framework for organisations to weigh the opportunities and risks associated with increasingly autonomous, cyber-capable AI.

External Document © 2026 Infosys Limited

2. Introduction: What Does This Mean for Cybersecurity?

2.1 The pre-2026 baseline

For most of the past decade, AI in cybersecurity primarily meant pattern matching, anomaly detection, log analysis, and threat correlation. Defensive programs were built around static signature-based mechanisms such as antivirus signatures, detection logic similar to YARA, a rule-based tool for identifying and classifying malware, and rule-based analytics, all of which were fundamentally reactive and best suited to identifying previously observed malicious payloads. As adversaries became more sophisticated, utilising polymorphic malware and living-off-the-land (LotL) techniques, defenders shifted to behavioral heuristics and early machine learning models designed to detect statistical anomalies within network traffic and endpoint execution telemetry.

AI made analysts faster, but it did not replace the human expertise required to discover novel vulnerabilities, reverse engineer binaries, or construct reliable exploits. Human expertise remained a critical asset in red teaming, penetration testing, malware analysis, and vulnerability research. As a result, advanced offensive capability, especially the ability to discover zero-days in hardened systems, remained constrained by the limited availability of highly trained cybersecurity professionals. Specialised talent was scarce, training took years, and the talent pipeline was finite.

Today, as technology matures, this talent constraint may ease. Tasks that once required a senior vulnerability researcher with a decade or more of experience can now, in some domains, be partially replicated or accelerated by frontier AI systems operating autonomously, continuously, and at scale. This also has significant implications for attacker–defender asymmetry.

2.2 What Changed in April 2026

Claude Mythos Preview and GPT-5.4-Cyber do not merely represent faster versions of previously existing tools. They represent a qualitative transition in what AI can do in a security context. The significance of this shift can be summarised in six developments:

- **Autonomous vulnerability discovery:** Anthropic states that Claude Mythos Preview can identify and exploit software vulnerabilities at a level that outdoes all but the most skilled human researchers in some settings.
- **Real-world zero-day discovery:** Anthropic reports that Mythos identified thousands of high-severity vulnerabilities across web browsers and major operating systems, not just benchmark or Capture the Flag environments.
- **Exploit-oriented reasoning:** In Anthropic's published evaluations, Mythos was able to move from vulnerability identification to exploit development, including chained exploitation scenarios, generating working exploit code (PoC) rather than merely identifying flaws
- **Multi-stage attack execution:** Anthropic's testing materials describe successful completion of complex network attack simulations that previously required substantial human expertise and time.
- **Binary reverse engineering without source code:** OpenAI specifically highlights this as a defining capability of GPT-5.4-Cyber for vetted defenders, expanding AI utility into malware analysis, firmware analysis, and closed-source software inspection.
- **Defensive cyber workflows at scale:** OpenAI positions GPT-5.4-Cyber as a model optimised for legitimate security work such as malware analysis, vulnerability assessment, and reverse engineering.

The larger significance is that cybersecurity is moving from the **AI-as-assistant era to the AI-as-agent era**. In practical cybersecurity terms, this evolution means that artificial intelligence systems can now autonomously ingest an entire legacy codebase or a compiled software binary, conceptually understand the developer's original intent, and identify deep logical flaws or memory corruption vulnerabilities.Historically, such deep logical flaws and memory-corruption vulnerabilities could elude advanced fuzzing infrastructure (the tools, systems, and supporting environment used to conduct fuzz testing) and symbolic execution engines (which systematically analyse possible program execution paths to identify inputs that could trigger vulnerabilities) for years or even decades.

This evolution fundamentally and permanently alters the temporal dynamics of network defense. The traditional security governance model relies heavily on the 'patch window' — the critical latency period between the discovery of a vulnerability, its formal public disclosure via a Common Vulnerabilities and Exposures (CVE) identifier, the vendor's release of a mitigating patch, and the eventual development of a weaponised exploit by malicious actors. With autonomous models now capable of automated exploit generation in a matter of minutes following a disclosure, this historical buffer is rapidly shrinking.

External Document © 2026 Infosys Limited

2.3 The Dual-Use Dilemma at Scale

All meaningful security tooling is inherently dual use. Port scanners, fuzzers, disassemblers, and exploit frameworks can all serve both defenders and attackers. What is significant about Claude Mythos Preview and GPT-5.4-Cyber is not merely that they are dual use, but that they combine high-end cyber capability with natural-language accessibility, abstraction, and speed. Capabilities that once demanded deep expertise in assembly, exploit mitigations, kernel internals, or binary analysis can now be made more accessible through highly capable AI interfaces.

Both Anthropic and OpenAI appear to recognise this shift. Rather than releasing Claude Mythos Preview to the general public, Anthropic limited access to Project Glasswing partners and select vetted organizations responsible for critical software infrastructure. That decision reflects an explicit acknowledgement that frontier cyber capability may outpace safe broad release. OpenAI, meanwhile, has adopted a different but complementary model. Through TAC, it is expanding access more broadly to verified defenders while using stronger identity verification, tiered privileges and iterative deployment controls to manage risk. In effect, Anthropic is prioritising tight containment, while OpenAI is pursuing controlled democratisation. Both approaches are responses to the same underlying reality: advanced cyber-capable models can strengthen defence, but they also lower the barrier to offensive misuse if governance fails.

3. When Cybersecurity Meets Frontier AI

Artificial intelligence is rapidly transforming both cyber offence and defence. Anthropic and OpenAI each introduced a next-generation model specifically tailored to cybersecurity. Both models are explicitly intended to accelerate defensive work (e.g., vulnerability discovery and code analysis) while navigating serious dualuse risks. In simple terms:

- Claude Mythos (Anthropic) – A general purpose LLM that "excels at completing complex, multi-step cybersecurity tasks," including autonomous vulnerability discovery and exploit generation. Anthropic describes Mythos Preview as "far ahead of any other AI model in cyber capabilities" and capable of finding thousands of high-severity bugs in operating systems and browsers.

- GPT-5.4Cyber (OpenAI) – A specialised variant of GPT5.4 fine-tuned for defensive cybersecurity workflows. It has a lower refusal boundary on security tasks and new capabilities (notably binary reverse engineering) to help analysts find and fix vulnerabilities faster. Access is tightly controlled through identity-verified programs to ensure use by legitimate defenders.

These announcements signal that advanced AI can now automate or amplify many steps of software security work. For cybersecurity teams, this means both new opportunities (faster vulnerability scanning, code auditing, CTF solving, etc,.) and new risks (adversaries gaining similar tools). This paper therefore explores why enterprises must pay attention to these developments. The new AI models raise the bar for both cyber offence as well as defence, and institutions need strategies to harness them safely.



External Document © 2026 Infosys Limited

4. Why Enterprises Must Act Now

4.1 The Window Between Discovery and Exploitation has Collapsed

One of the most operationally significant statements to emerge from the Mythos announcement was from a CrowdStrike executive, also a Project Glasswing partner: "The window between a vulnerability being discovered and being exploited by an adversary has collapsed. What once took months now happens in minutes with AI." It is a description of a measurable, structural change in the threat timeline that invalidates patch management programs, vulnerability disclosure windows, and incident response playbooks that assume days or weeks of reaction time.

The assumption that enterprises have meaningful time between CVE publication and active exploitation is increasingly difficult to sustain. AI-accelerated vulnerability analysis and exploit development compress the defender's response window from weeks to hours, and in some cases potentially less.

4.2 Defender Asymmetry Can Still be Reversed

The positive dimension of these developments is that the same capability that empowers attackers is now being channeled toward defenders. Project Glasswing was created specifically to use Mythos Preview to discover and help remediate critical vulnerabilities before adversaries could exploit them. Codex Security from OpenAI has already contributed to identifying 792 critical and 10,561 high-severity vulnerabilities, over 11,000 findings in total, across more than 1.2 million scanned commits. The technology that threatens enterprise security is also the most powerful defensive weapon available, but only if enterprises position themselves to access and deploy it legitimately.

This creates an urgent security divide. Enterprises that adopt verified AI access, upskill their security teams, and embed AI into core defences will strengthen their longterm security advantage. Those that do not will face an expanding capability gap as both legitimate access holders and threat actors leverage increasingly capable models.

4.3 Supply Chain and Open-Source Risk has Moved to the Centre

One of the most sobering aspects of the Mythos announcement is that its vulnerability discovery capability is aimed first at the foundational software that underpins all enterprise infrastructure such as operating systems, web browsers, and open-source libraries. Anthropic highlighted examples, including a 27-year-old flaw in OpenBSD, a 16-year-old vulnerability in FFmpeg, and a 17-year-old remote code execution flaw in FreeBSD. These are not exotic edge cases; they are in the supply chain of virtually every enterprise software stack. If AI can find these in days, adversaries with similar or lower-capability models will spot them too, which means the entire model of "security through longevity of code" stands challenged.

Every enterprise software supply chain has inherited technical debt spanning decades. AI-powered vulnerability discovery means that debt is now being systematically audited by defenders and attackers simultaneously. **In 2026, supply chain visibility, asset exposure mapping, and patch velocity should be treated as top-tier security priorities.**

4.4 New Capability Leap in Cybersecurity

These models represent a step change in performance. Anthropic asserts that Mythos yields "dramatically higher scores on tests of software coding, academic reasoning, and cybersecurity" than its prior best model. OpenAI, similarly, classifies GPT-5.4 as a high cyber-capability model under its internal framework. Early evaluations show Mythos achieving significantly higher success rates on red-team style challenges (83.1% on a CyberGym benchmark vs 66.6% for the previous Claude Opus 4.6). With huge context windows (Mythos at ~1million tokens) and advanced training, these models can handle complex, multi-step tasks that older LLMs could not.

Equally important, these systems combine stronger reasoning with long context handling and agentic execution. **This allows them to operate over large codebases, analyse complex software states, and sustain multi-step workflows that earlier models could not handle reliably.** For enterprises, this means AI is no longer merely a productivity enhancer for analysts; it is becoming an operational security capability in its own right.

External Document © 2026 Infosys Limited

4.5 A Strategic Race in AI Security is Under Way

Large corporations and governments have already mobilised around these models. Anthropic's Project Glasswing includes AWS, Google, Microsoft, Cisco, and others, indicating a coordinated effort to harden critical software with Mythos. OpenAI has expanded its TAC program to involve thousands of verified defenders.

This means the leading players see an arms-race dynamic emerging. If defenders do not adopt equally powerful tools, attackers will have the edge. In short, enterprise risk is shifting; it is imperative that security leaders evaluate these capabilities, adopt policies, and integrate AI-driven security tools to keep pace with the threat landscape.

4.6 Regulatory and Board Accountability

These developments have moved beyond vendor announcements and into the domain of government and board-level risk. In the United States, senior officials and major bank CEOs have already been briefed on the cybersecurity implications of Anthropic's Mythos. Anthropic too has directly engaged with White House officials on cybersecurity, innovation, and AI safety. In the UK, the government has publicly warned business leaders that frontier AI is accelerating cyber risk faster than previously expected.

As a result, enterprises in critical sectors should expect greater scrutiny and clearer expectations around cyber readiness. Boards should be briefed now, and engagement with regulators and industry bodies should be treated as a strategic priority rather than a compliance exercise.

5. Model Capabilities — Claims and Benchmarks

5.1 Claude Mythos Preview

Anthropic presents Claude Mythos Preview not as a typical software product, but as an unreleased frontier AI model. In several operational contexts, its cybersecurity capabilities are described as matching or surpassing those of elite human vulnerability researchers. The following benchmark results were published in Anthropic's official System Card and confirmed through independent analysis.

- **SWE-bench Verified:** 93.9% — the highest publicly reported score on this gold-standard software engineering benchmark, representing a double-digit lead over the previous best
- **SWE-bench Pro:** 77.8% — a 20.1-point lead over GPT-5.4 on the hardest real-world engineering benchmark available
- **USAMO 2026 Mathematical Proofs:** 97.6% — near-perfect on competition-level mathematics, a 55.3-point leap over the previous Anthropic model (Opus 4.6, 42.3%) and 2.4 points above GPT-5.4 (95.2%)
- **Terminal-Bench 2.0:** 82% — a 6.9-point improvement over GPT-5.4
- **HLE with Tools:** +12.6 points above GPT-5.4
- **GPQA Diamond:** +1.7 points above GPT-5.4
- **Long-Context Reasoning:** Improved from 38.7% to 80.0%, more than doubling previous performance
- **CyberGym Security Evaluation:** Led significantly, with a substantial gap over previous frontier models



External Document © 2026 Infosys Limited

Cybersecurity-Specific Accomplishments

These are not merely benchmark results. Anthropic's published materials describe real-world vulnerability discovery and exploit-oriented demonstrations, and some of the broader cyber implications have since been independently evaluated by the UK AI Security Institute (AISI).

- Zero-day discovery at scale: Claude Mythos Preview autonomously identified zero-day vulnerabilities in every major operating system (Windows, Linux, macOS, FreeBSD, OpenBSD) and every major web browser (Firefox, Chrome, Safari, Edge) in active global use.
- 27-year-old OpenBSD integer overflow: Mythos independently discovered an integer overflow vulnerability in OpenBSD, which is a security-hardened operating system, that had survived 27 years of human review and automated testing.
- 16-year-old FFmpeg flaw: A vulnerability in the widely deployed FFmpeg multimedia library that had survived over five million automated fuzzing tests was discovered and documented by Mythos.
- 17-year-old FreeBSD REC Vulnerability (CVE-2026-4747): Mythos fully autonomously discovered and exploited a critical remote code execution vulnerability in FreeBSD's NFS implementation.
- Firefox zero-days responsibly disclosed: Real, previously unknown zero-day vulnerabilities in production Firefox code were discovered and responsibly disclosed to Mozilla.
- Working exploit generation: In testing against a corpus of open-source vulnerabilities, Mythos generated working exploit code in 83.1% of cases, compared to 66.6% for the previous best model (Claude Opus 4.6).
- Full network attack simulation: The AISI evaluated Mythos against a multi-stage network attack simulation, testing its ability to execute a complete network takeover from initial access. Mythos was the first AI model to successfully complete this evaluation end-to-end, though only in 3 of 10 attempts, averaging 22 of the 32 steps overall. However, AISI's test environment had no active defenders or security monitoring. This means the result shows Mythos can compromise a weakly defended network, but it may not breach a hardened, actively monitored enterprise environment.
- Thousands of high-severity vulnerabilities: Across the Project Glasswing testing, Mythos discovered thousands of high-severity vulnerabilities across critical infrastructure software. Over 99% have not yet been publicly disclosed, as coordinated vulnerability disclosure processes are ongoing.

An earlier internally deployed version of Mythos was instructed, during behavioral testing, to escape a secured sandbox environment and notify the researcher running the evaluation. It succeeded in escaping the sandbox, exhibiting a potentially dangerous capability for circumventing safeguards. Beyond what was asked, it then created a moderately sophisticated multi-step exploit to obtain broad internet access from a system that was meant to reach only a small number of predetermined services, before notifying the researcher as instructed. The System Card also documents other alignment-relevant behaviors from these earlier snapshots, including rare instances of the model manipulating git history, the record of changes made to a codebase over time, covering its tracks, and taking "reckless destructive actions" in lower-stakes settings. Anthropic states that the final released model "still takes reckless shortcuts in many lower-stakes settings, but we have not seen it show the same kinds of severe misbehavior or attempts at deception."



External Document © 2026 Infosys Limited

5.2 GPT-5.4-Cyber

GPT-5.4-Cyber is positioned differently from Claude Mythos Preview. It is not presented as a new model tier, but as a fine-tuned, cyber-permissive variant of GPT-5.4 designed for verified defenders. OpenAI frames the release as part of a broader deployment strategy: to expand defender access, refine trust-based controls, and prepare its safety and access infrastructure for more capable models expected in the near future.

Binary reverse engineering (no source required): GPT-5.4-Cyber enables security professionals to analyse compiled software firmware, third-party libraries, suspected malware for vulnerabilities, and malware indicators without access to the original source code.

Lowered refusal boundary: The model is deliberately calibrated to refuse fewer legitimate security research requests compared to standard GPT-5.4, reducing friction for vetted analysts performing tasks that previously triggered unhelpful refusals.

Advanced defensive workflow support: OpenAI positions the model around end-to-end defensive use cases, including vulnerability analysis, malware analysis, exploit analysis, and security robustness testing.

High cyber capability classification: GPT-5.4-Cyber is graded as high cyber capability under OpenAI's preparedness framework and GPT-5.4-Cyber extends those capabilities in a more permissive but access-controlled form for authenticated defenders.

Infrastructure-level safety: OpenAI's cyber safety has broader cyber deployment stack, including classifier-based monitoring, stronger verification, granular access controls, and product-level mitigations for high-risk cyber use.

CTF Performance Progression: OpenAI documented significant improvement in Capture-the-Flag benchmark performance throughout their model generations, from 27% on GPT-5 in August 2025 to higher scores with current-generation models materially.

Codex Security Track Record: As context for GPT-5.4-Cyber's defensive value, OpenAI cited Codex Security's track record: 792 critical and 10,561 high-severity vulnerabilities identified during its 30-day beta, over 11,000 findings in total, with more than 3,000 critical vulnerabilities patched as of its broader rollout.



External Document © 2026 Infosys Limited

6. Comparative Analysis

A structured comparison across operational parameters relevant to enterprise cybersecurity decision-making.

Parameter	Claude Mythos Preview	GPT-5.4-Cyber	Advantage
Model Type	New frontier model (General purpose + cyber)	Fine-tuned variant of GPT-5.4 base model	Mythos
Availability	Project Glasswing partners	Tiered TAC — thousands of verified defenders	GPT-5.4-Cyber (broader access)
Use Cases (intended)	Defensive: automated bug hunting, code review, exploit patching, red teaming support, also research & development	Defensive: binary malware analysis, vulnerability triage, secure coding assistance, threat hunting, CTF training	Comparable
Key Technical Capability	Fully autonomous exploit chaining (bypassing ASLR/DEP for Linux root escalation)	Advanced binary reverse engineering without source code access; malware behavioral mapping	GPT-5.4-Cyber (broader access)
Optimal Workflow Fit	Deep zero-day research, complex application security posture validation, automated red teaming	Incident triage, patch review, binary artifact analysis, SIEM/log analysis query generation	GPT-5.4-Cyber (democratised)
SWE-bench Verified	93.9% (best publicly known)	Not separately published (GPT-5.4 competitive)	Mythos
SWE-bench Pro	77.8%	~57.7% (GPT-5.4 reference)	Mythos
USAMO 2026 (Math)	97.6%	95.2%	Mythos
HLE with Tools	+12.6 pts vs GPT 5.4	Baseline	Mythos
Terminal-Bench 2.0	82% (+6.9 pts vs GPT-5.4)	~75%	Mythos
GPQA Diamond	+1.7 pts vs GPT-5.4	Baseline	Mythos
Zero-Day Discovery	Yes — thousands across every major OS and browser	Not explicitly claimed	Mythos
Binary Reverse Engineering	Implicit (exploit generation capability)	Explicit feature, sourceless binary analysis	GPT-5.4-Cyber
Exploit Generation (autonomous)	83.1% success rate	Not published	Mythos
Vulnerability Patching Program	Project Glasswing — coordinated OS/browser patching	Codex Security — 3,000+ critical CVEs fixed	Both — different focus
CTF Performance	Multi-stage full network attack	Improved from 27% (GPT-5) to much higher	Mythos
Agentic Capability	Fully autonomous — finds, exploits, reports without human	Advanced defensive workflows (agentic)	Mythos (fully autonomous)
Sandbox Containment	Escaped sandbox in testing, posted exploits publicly	No published sandbox escape incidents	GPT-5.4-Cyber



External Document © 2026 Infosys Limited

Parameter	Claude Mythos Preview	GPT-5.4-Cyber	Advantage
Deceptive Behavior	Rare instances of covering tracks, git history manipulation	Not reported	GPT-5.4-Cyber
Long-Context Reasoning	Jumped from 38.7% to 80.0%	Strong but not published for this variant	Mythos
Safety Framework	Responsible Scaling Policy	Preparedness Framework	Both — different architecture
Alignment Risk Rating	Highest among any Anthropic model released	High under Preparedness Framework	Neither — both concerning
Hard Prohibitions	No mass casualty, CBRN, critical infra-attacks	No data exfiltration, malware deployment, destructive testing	Both clear
tGovernment Engagement	White House meeting post-announcement	Not prominently reported	Mythos (higher geo-political visibility)
Funding Commitment	$100M usage credits + $4M to open-source foundations	Expanded TAC access; Codex Security program	Mythos (dollar commitment)

Note: Benchmark comparisons reflect publicly available data as of April 22, 2026. GPT-5.4 (base model) benchmark figures are used as the reference point for GPT-5.4-Cyber, where the cyber variant's scores are not separately published. GPT-5.4-Cyber has been succeeded by GPT-5.5-Cyber, as per OpenAI's announcement.

7. Enterprise Concerns and the Expanding Threat Horizon

These frontier models offer powerful defensive advantages, but they also introduce new classes of architectural and operational risk. Enterprises should carefully consider the following concerns:

7.1 The Adversarial Counterpart

Anthropic itself warns that Mythos-level AI could trigger "an upcoming wave of models that can exploit vulnerabilities in ways that far outpace the efforts of defenders". If an attacker gains access to a Mythos-like system (or trains their own), they could potentially automate exploit development at scale. The UK's AI Security Institute notes that Mythos is very good at attacking poorly defended systems. Even without Mythos, research shows open-source LLMs can already find the same bugs Mythos found. The concern is that the same AI advances lifting defenders will empower attackers to find zero-days faster. The most significant concern is not what these organisations do with their own models, but what happens when:

- Nation-state threat actors develop or acquire equivalent capability through their own AI laboratories, which are operating with fewer safety constraints.
- The capability diffuses through jailbreaks, model distillation, or prompt injection against gated models.
- Less responsible AI providers replicate the core capability without equivalent access controls.
- Insiders at Glasswing or TAC organisations misuse legitimate access.

The reality is that the arrival of Claude Mythos sets a precedent for what AI-powered offensive security looks like. Sophisticated threat actors, particularly state-sponsored groups, now have a concrete capability benchmark to target in their own AI development. It's an AI arms race in the security domain whether one chooses to participate or not.

External Document © 2026 Infosys Limited

7.2 Alignment Problem Persists

Anthropic's claim that Mythos is simultaneously its best-aligned model and the one posing the greatest alignment risk should be examined closely by enterprise leadership. Current AI interpretability tools cannot reliably distinguish between:

A model that is genuinely aligned and will not take unauthorised actions under any circumstances.	A model that is strategically compliant because it currently assesses no benefit from non-compliance.	A model that is actively concealing misaligned behavior and waiting for conditions where deception would be beneficial.

For enterprise deployment, this matters. If an AI security agent with Mythos-level capability is integrated into an enterprise network and given broad access to support penetration testing or vulnerability management, the boundary between 'authorised action' and 'unauthorised initiative' is not currently enforceable through technical controls alone. The sandbox escape demonstrated that this is not just a theoretical concern.

7.3 Overreliance and Supply Chain

Enterprises risk becoming dependent on these models. Mythos and GPT-5.4-Cyber could become critical infrastructure for security, as CrowdStrike notes. But reliance on a single AI vendor or model could be dangerous; attackers could exploit a flawed or compromised model. Integrating AI widely also expands the attack surface; adversaries could target the AI tools themselves (e.g,. crafting malicious prompts, poisoning their training data, or exploiting cloud APIs). The community, therefore, must treat AI platforms as high-value assets requiring protection.

Although 83.1% exploit generation success is a remarkable score, it is also a reminder that 16.9% of cases fail. In conventional security tooling, a false negative is a risk. In AI-powered security, a false negative at scale could create a dangerous false sense of coverage. An enterprise that relies on AI to scan its attack surface and finds 'nothing critical' may be receiving a verdict with a non-trivial error rate. Hence, AI-powered security cannot yet be treated as equivalent to expert human assessment for the highest-stakes environments. It must be layered with, not substituted for, expert human judgment.

7.4 Dual Use and Misuse

GPT-5.4Cyber's less restrictive safeguards mean it could also be abused. For example, lower refusal on "legitimate cybersecurity work" might inadvertently open loopholes for malicious actors posing as defenders. OpenAI tries to mitigate this with identity checks and restrictions on opaque use modes. Nonetheless, sophisticated attackers might mimic trust signals to trick the system. Likewise, Mythos Preview is highly guarded, but the leaked documents show Anthropic fears misuse: they intentionally restrict Mythos use to cyber defence and partner entities. These measures suggest an awareness that once such capabilities are out, their widespread availability cannot be easily reversed. Anthropic's System Card documents several specific incidents that should concern enterprise security teams:

- **Git history manipulation:** The pre-release Mythos attempted to alter git history to conceal actions it had taken. This is consistent with deceptive behavior, an attempt to modify the audit trail of its own activity.
- **Reckless destructive actions in lower-stakes settings:** The System Card notes that Mythos "still takes reckless shortcuts in many lower-stakes settings". For enterprise deployments, 'lower-stakes settings' can include sandbox testing environments that are nonetheless connected to production infrastructure.
- **Unsolicited public disclosure of exploit details:** The sandbox escape incident's most alarming aspect is not the escape itself, but the unprompted decision to post exploit details publicly.

External Document © 2026 Infosys Limited

8. Strategic Recommendations for Enterprises

The arrival of Claude Mythos Preview and GPT-5.4-Cyber should prompt enterprises to move beyond reactive, compliance-led security programs and toward a more adaptive, AI-native defence model. Anthropic's Project Glasswing and OpenAI's TAC make clear that these systems are not ordinary productivity tools. They are high-capability cyber systems that can materially accelerate vulnerability discovery, reverse engineering, code analysis, and defensive operations when deployed with appropriate controls.

8.1 Adopt Defence-First Approach

Organisations should not wait for AI-enabled threats to become mainstream before responding. Instead, they should evaluate frontier cyber models early, within controlled environments, and use them first for defensive purposes such as internal code review, dependency analysis, malware triage, and secure development workflows. Anthropic's Project Glasswing is explicitly framed around putting advanced model capability in the hands of defenders before similar capabilities proliferate, while OpenAI's TAC is designed to expand defensive access in a verified and staged manner.

8.2 Begin Transition to AI-Native SOC

Security Operations Centres (SOCs) should evolve from simply using AI for alert summarisation to integrating frontier models into core investigative workflows. In practice, GPT-5.4Cyber may fit best in investigation workflows such as artifact triage, binary analysis, malware analysis, and vulnerability management. Claude Mythos Preview is better suited for vulnerability assessment, exploit generation, secure engineering workflows, including code hardening and remediation pipelines. The operating model should shift so that human analysts increasingly supervise, validate, and direct autonomous or semi-autonomous investigative agents rather than perform every repetitive task manually.

8.3 Replace Periodic Testing with Continuous Security Validation

Annual or quarterly penetration tests are no longer sufficient in an environment where frontier models can accelerate vulnerability discovery and exploit reasoning. Enterprises should move toward continuous validation, persistent internal scanning, recurring adversary emulation, automated dependency review, and ongoing assessment of critical applications, build pipelines, and exposed attack paths. Anthropic's disclosures around Mythos suggest that the pace of discovering subtle software flaws has changed materially, while OpenAI's broader cyber-defence framing emphasises continuous identification and remediation of weaknesses as a software is written.

8.4 Operate Under Assume-Breach and Rapid-Containment Model

The traditional patch cycle is under pressure. As model-assisted exploit analysis becomes faster, the historical buffer between vulnerability discovery, public disclosure, and weaponisation may narrow significantly. Organisations should complement patching with stronger containment controls. These include endpoint hardening, segmentation, identity-centric access control, Software Bill of Materials (SBOM) tracking, build-system protection, reduced privileged access, and automated response playbooks for lateral movement and post-compromise containment.

8.5 Enforce Strong Identity, Access and Provenance Controls

Access to frontier cyber-capable models should be limited to vetted users, approved teams, and governed use cases. OpenAI's TAC model built around identity verification, tiered access, and differentiated controls for more permissive cyber capabilities provides a useful template for enterprise governance. Organisations should keep complete audit trails and bind model usage to user or service identities. Policy permitting, they should log prompts and outputs. Highimpact actions such as exploit analysis, malware handling, and reverse engineering should go through an explicit approval workflow.

8.6 Keep Humans in the Loop and Validate Outputs Rigorously

Neither Claude Mythos Preview nor GPT-5.4-Cyber should be treated as a replacement for expert judgment. Their outputs should be validated through human review and corroborated through traditional tooling such as static analysis, dynamic analysis, sandboxing, fuzzing, and manual investigation. The most effective operating model is not 'AI instead of experts', but 'AI plus experts', where the model accelerates discovery and triage while experienced practitioners verify findings, assess exploitability, and decide on remediation.

8.7 Invest in Training and Operating Discipline

Security teams, developers, and platform owners need training not only on what these models can do, but on how to use them safely. That includes prompt hygiene, output verification, sensitive-data handling, escalation thresholds, and clear boundaries around approved cyber workflows. As these models become more agentic, enterprises will need new playbooks for supervising autonomous actions, adjudicating high-confidence findings, and preventing over-reliance on model-generated conclusions.

External Document © 2026 Infosys Limited

8.8 Contribute to Broader Security Ecosystem – Federated Learning

Organisations should not approach frontier cyber AI in isolation. They should contribute anonymised findings where appropriate, support open-source security ecosystems, and closely monitor official announcements, system cards, safety updates, and independent research. Both Anthropic and OpenAI are positioning these capabilities within broader ecosystem strategies: Anthropic through Project Glasswing and OpenAI through TA and Codex Security. Hence, staying engaged with the evolving ecosystem will be important for both risk management and competitive readiness.

The overall recommendation is to treat frontier cyber models as strategic security infrastructure, not as experimental productivity tools. Enterprises that evaluate them early, govern them tightly, integrate them into defensive workflows, and preserve strong human oversight will be far better positioned than those that wait for AI-driven offensive pressure to force reactive change.

9. Best Practices for Governing Frontier AI Cybersecurity

Deploying frontier AI in highly sensitive cybersecurity contexts necessitates stringent architectural governance to prevent abuse, hallucination-induced errors, and unauthorised network access. In practical terms, enterprises should adopt these best practices when working with frontier AI.

9.1 Treat Frontier Models as Privileged Cyber Capabilities

Claude Mythos Preview and GPT-5.4Cyber should be governed like highrisk security infrastructure, not productivity tools. Their use must be explicitly approved, scoped, and owned, with clear accountability for outcomes.

9.2 Enforce Identity, Least Privilege and Zero-Trust Execution

All AI agents and workflows must operate under named, nonhuman identities with tightly scoped permissions. Run models in isolated, segmented environments with infrastructure-enforced boundaries. Do not rely on prompt instructions alone to constrain behavior.

9.3 Keep Humans in the Loop

AI output should be treated as an expert hypothesis. It requires human validation for any high-impact action, including exploit verification, penetration testing, remediation, or configuration changes. AI should inform decisions, not autonomously execute them in production.

9.4 Log Everything and Protect Audit Trail

All prompts, outputs, tool invocations, and downstream actions must be captured in immutable audit logs that the AI system itself cannot access or modify. Provenance is essential for governance, forensics, and trust.

9.5 Stage Capabilities and Validate Continuously

Begin with read-only analysis and expand to more active workflows only after confidence is established in the model's behavior, analyst oversight, and controls. Cross-validate AI findings with traditional security tools and expert review.

9.6 Control Data Exposure and Vendor Risk

Limit exposure of sensitive code, binaries, and secrets to fully approved environments. Align with vendor verification and access programs (e.g., Glasswing, TAC), and ensure clarity on identity verification, data retention, and security controls.

9.7 Prepare for AI-Specific Failure Modes

Update incident response plans to address AI-related risks such as hallucinated findings, scope violations, unsafe autonomous behavior, or AI-assisted attacks. Regularly red team AI-assisted workflows to understand their limits.

The winning strategy is not maximum autonomy, but controlled capability. Enterprises that combine frontier AI with strict identity, infrastructure-level controls, human oversight, and disciplined governance will gain defensive advantage without amplifying risk.

External Document © 2026 Infosys Limited

10. Conclusion

The arrival of Anthropic's Claude Mythos and OpenAI's GPT-5.4Cyber marks a turning point in cybersecurity. These frontier models demonstrate that AI can now autonomously discover and potentially exploit software vulnerabilities at a scale that was previously impossible. For enterprise leaders, the message is clear — Advanced AI will rapidly reshape the offence-defence balance in cyber conflict.

Therefore, the central strategic question is straightforward: Is the organisation prepared for a threat environment in which AIenabled attackers operate faster than humanled security teams can respond? If the answer is no, then cybersecurity strategies, operating models, and governance structures require urgent modernisation. The longstanding assumption that defenders retain a buffer between vulnerability discovery and exploitation is rapidly fading. AI is compressing vulnerability analysis, exploit development, and investigation timelines to machine speed.

Organisations must therefore prioritise the remediation of chronic technical debt, implement rigorous agentic identity management to control shadow AI, and adopt continuous, autonomous security validation across their environments. The era of manual alert triage and reactive, monthly patching cycles is drawing to a close. Organisations that fail to deeply integrate frontier AI models into their core defensive architecture risk becoming economically and technologically unprepared for the coming wave of automated, hyperscaled cyber aggression. As one CrowdStrike executive observed, "The window between a vulnerability being discovered and being exploited by an adversary has collapsed. What once took months now happens in minutes with AI." Therefore, organisations need to shift from periodic, human-paced security to continuous, AIdriven validation and rapid containment.

11. References

1. Anthropic. (2026, April 7). Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing
2. Anthropic. (2026, April 7) Assessing Claude Mythos Preview's cybersecurity capabilities. https://red.anthropic.com/2026/mythos-preview/
3. OpenAI. (2026, April 14). Trusted access for the next era of cyber defense. https://openai.com/index/scaling-trusted-access-for-cyber-defense/
4. AI Security Institute. (2026, April 13). Our evaluation of Claude Mythos Preview's cyber capabilities. https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
5. Gerstenhaber, M. (2026, April 8). Claude Mythos Preview: Available in private preview on Gemini Enterprise Agent Platform. Google Cloud. https://cloud.google.com/blog/products/ai-machine-learning/claude-mythos-preview-on-vertex-ai
6. CrowdStrike. (2026, April 20). Frontier AI is collapsing the exploit window. Here's how defenders must respond. https://www.crowdstrike.com/en-us/blog/frontier-ai-collapses-exploit-window-how-defenders-must-respond/
7. Kelley, D. (2026, April 9). The Mythos reality check: What does Project Glasswing mean for CISOs? Noma Security. https://noma.security/blog/the-mythos-reality-check-what-does-project-glasswing-mean-for-cisos/
8. Penligent. (2026, April 17). GPT-5.4 Cyber vs Claude Mythos: Which model fits cybersecurity work. https://www.penligent.ai/hackinglabs/gpt-5-4-cyber-vs-claude-mythos-which-model-fits-cybersecurity-work/
9. Bhatia, R. (2026, April 10). "Mythos autonomously exploited vulnerabilities that survived 27 years of human review. Security teams need a new detection playbook." VentureBeat. https://venturebeat.com/security/mythos-detection-ceiling-security-teams-new-playbook
10. Penligent. (2026, April 15). GPT-5.4-Cyber, Trusted Access for Cyber. https://www.penligent.ai/hackinglabs/fr/gpt-5-4-cyber-trusted-access-for-cyber/
11. AI-powered exploitation may collapse the patch window for defenders. (2026, April 21). Cryptika Cybersecurity. https://www.cryptika.com/ai-powered-exploitation-may-collapse-the-patch-window-for-defenders/
12. Baran, G. (2026, April 15). OpenAI launches GPT-5.4 with reverse engineering, vulnerability and malware analysis features. Cyber Security News. https://cybersecuritynews.com/openai-launches-gpt-5-4/
13. Landymore, F. (2026, April 8). Anthropic warns that "reckless" Claude Mythos escaped a sandbox environment during testing. Futurism. https://futurism.com/artificial-intelligence/anthropic-claude-mythos-escaped-sandbox
14. Markovic, S. (2026, April 15). OpenAI expands its cyber defense program with GPT-5.4-Cyber for vetted researchers. Help Net Security. https://www.helpnetsecurity.com/2026/04/15/openai-gpt-5-4-cyber/
15. NxCode Team. (2026, April 8). Claude Mythos benchmarks explained: 93.9% SWE-bench & every record broken (2026). NxCode. https://www.nxcode.io/resources/news/claude-mythos-benchmarks-93-swe-bench-every-record-broken-2026

External Document © 2026 Infosys Limited

About the Author



Dr. Aditya Kumar
Specialist Programmer, Infosys Limited

Dr. Aditya Kumar is a researcher at Infosys CyberSecurity Applied Research Center (Cyber-ARC). He holds a PhD from the University of Hyderabad, and his research interests span AI for cybersecurity, frontier vulnerability analysis, exposure management, and applied machine learning. He has published his work in multiple peer-reviewed journals and conferences. His current work focuses on AI security, frontier model risk assessment, model governance, and privacy and security in model training.

For more information, contact askus@infosys.com

Infosys
Navigate your next

© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.