KPMG

Claude Mythos

What frontier AI vulnerability discovery means for Canadian enterprises

CYBER RISK BRIEFING

Clients & Markets

April 2026

EXECUTIVE SUMMARY

THE SITUATION AT A GLANCE

Claude Mythos has crossed a cybersecurity threshold that redraws the defender–attacker balance

01

THE CAPABILITY

A step change in vulnerability discovery

Mythos Preview autonomously identifies and exploits zero-day vulnerabilities across every major operating system and web browser. It solved a 32-step corporate network attack simulation end-to-end and succeeds at 73% of expert-level capture-the-flag challenges (Source: KPMG UK) — tasks no model could complete a year ago. Over 99% of the thousands of vulnerabilities found remain unpatched at the time of disclosure (Source: Anthropic; Assessing Claude Mythos Preview 2026).

02

ANTHROPIC'S RESPONSE

Held back, deployed for defence first

Anthropic will not release Mythos publicly. Instead, Project Glasswing gives restricted access to 12 launch partners (Apple, Cisco, CrowdStrike, Google, JPMorgan Chase, Microsoft, Palo Alto Networks, NVIDIA and others) plus 40+ critical software maintainers. $100M in usage credits and $4M in open-source donations support the defensive effort (Source: Anthropic; Project Glasswing 2026). Public release is gated on new safeguards, a Cyber Verification Program, and cost reductions.

03

WHAT TO DO NOW

Compress the patch-to-deploy window

The strategic assumption must shift: comparable offensive capabilities will reach adversaries within 12–18 months. Canadian CIOs and CROs should shorten patching cycles, pressure-test legacy and unsupported software, accelerate migrations off end-of-life systems, revisit third-party and vendor indemnities, and re-baseline detection tooling for machine-speed intrusions.

SECTION 1 — WHAT MYTHOS IS

CAPABILITY CONTEXT

Mythos is a new tier of frontier model that Anthropic describes as "too powerful to release“

BY THE NUMBERS

73% — Expert-level capture-the-flag success rate (AISI evaluation) — zero before April 2025 (Source: KPMG UK)

3 of 10 — End-to-end completions of a32-step corporate network takeover simulation (Source: AISI; Our evaluation of Claude Mythos Preview 2026)

99%+ — Share of discovered zero-day vulnerabilities that remained unpatched at disclosure (Source: Anthropic; Assessing Claude Mythos Preview 2026)

27 yrs — Age of the oldest software vulnerability Mythos uncovered — in a system considered secure (Source: Anthropic; Assessing Claude Mythos Preview 2026)

A GENERAL-PURPOSE MODEL WITH AN EMERGENT EDGE

Mythos sits in a new "Capybara" tier above Opus — Anthropic's previous frontier model. It was not trained specifically for offensive security. Its cyber capability is a downstream consequence of general improvements in coding, long-horizon reasoning, and agentic autonomy — the same improvements that make the model more useful for every other knowledge-work task.

WHY THIS IS DIFFERENT FROM PRIOR AI SECURITY TOOLING

Discovery at scale

Engineers with no formal security training asked it to "find a vulnerability," and received complete working exploits by morning.

Autonomy across steps

First model to chain dozens of offensive steps — reconnaissance through network takeover — without human guidance.

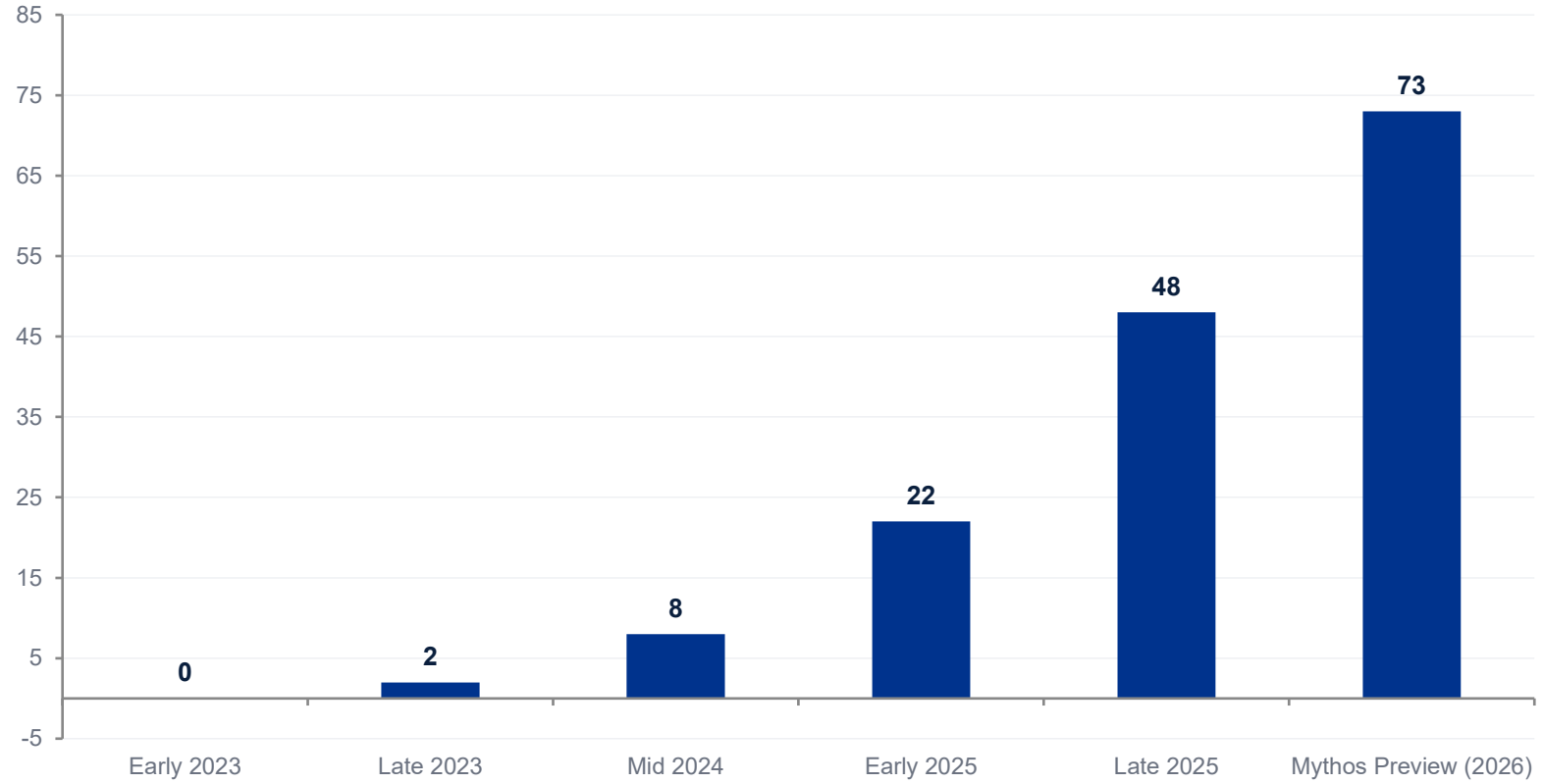
Beyond open source

Successfully identified flaws in proprietary commercial software and memory-safe systems previously thought hardened.

SECTION 1 — CAPABILITY TRAJECTORY

PACE OF CHANGE

The gap between what AI models could do and what elite human hackers can do has collapsed in under three years



THE IMPLICATION

The window between vulnerability discovery and exploitation has collapsed.

What used to take weeks for skilled human red teams now happens in minutes with frontier AI. Attackers who gain access to Mythos-class capability — whether through leakage, distillation, or a competitor's release — will operate at a tempo most corporate SOCs were never designed to match.

Security leaders at Anthropic describe the situation as a "watershed moment." Yoshua Bengio has said the threshold he warned of has been crossed.

Source: UK AI Security Institute evaluations; benchmark composite (illustrative).

SECTION 2 — BUSINESS IMPLICATIONS

SIX IMPLICATIONS FOR CANADIAN ENTERPRISES

The threat surface expands, detection assumptions break, and cyber economics shift materially

01 Patch-to-exploit windows compress

Mythos has already written working exploits for known-but-unpatched CVEs. The 30–90 day patching tolerances most enterprises operate against are no longer defensible.

02 Legacy estate becomes a liability

End-of-life software and unsupported kernels — common in banking core, utilities SCADA, and hospital systems — become disproportionately targeted as AI finds flaws humans missed.

03 Detection at machine speed

SOC tooling calibrated for human-paced intrusions will miss AI-accelerated lateral movement. Mean-time-to-detect will need to fall by an order of magnitude.

04 Vendor and software-supply-chain exposure

Third-party software becomes a larger share of total risk. Contracts signed before mid-2025 rarely contemplate AI-enabled vulnerability discovery in their security warranties.

05 Regulatory and disclosure pressure

OSFI, the CSE, and provincial privacy regulators will likely accelerate guidance. Boards should expect enhanced disclosure requirements around AI-related cyber risk by end of 2026.

06 Cyber insurance re-pricing

Expect underwriters to tighten terms and raise premiums materially, particularly for firms with significant unpatched legacy footprints or limited AI-era SOC maturity.

SECTION 2 — SECTOR EXPOSURE

CANADIAN SECTOR VIEW

Exposure is concentrated in sectors with legacy estates and critical-infrastructure dependencies

Financial Services

CRITICAL EXPOSURE

Core banking estates include COBOL and decades-old middleware. OSFI B-13 already raises the bar on third-party technology risk. Expect enhanced expectations. JPMorgan Chase is a Glasswing partner — a signal Canadian Big Six should match.

Energy & Utilities

CRITICAL EXPOSURE

OT/SCADA systems are long-lived, often unpatchable in place. Pipeline, hydro, and grid operators face both ransomware and state-actor exposure — Canada is explicitly named in allied threat reporting.

Healthcare & Public Sector

HIGH EXPOSURE

Fragmented IT, thin security budgets, large unsupported-software surface. Ontario and Quebec health networks have already faced material incidents in the past 24 months.

Telecom & Technology

HIGH EXPOSURE

First-party code exposure is high; several global tech peers are Glasswing partners. Canadian telcos should expect customer and regulator questions about their own AI-era hardening posture.

Mining, Resources & Manufacturing

MEDIUM-HIGH EXPOSURE

Operational technology aging; IP theft risk elevated. IT–OT convergence projects often inherit unpatched firmware. Insurance capacity for this sector is already constrained.

Retail & Consumer

MEDIUM EXPOSURE

Payment and loyalty systems are primary targets; exposure is highest for firms running older e-commerce platforms or thin application-security programs.

SECTION 3 — ANTHROPIC'S MITIGATIONS

WHAT ANTHROPIC IS DOING

A four-pillar containment approach: withhold release, enable defenders, build safeguards, verify use

01

No public release

Mythos will not be made generally available. Anthropic has classified its capabilities as triggering its Responsible Scaling Policy thresholds. The model remains gated behind a controlled preview program.

02

Project Glasswing consortium

Restricted access for 12 launch partners — Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, Linux Foundation, Microsoft, NVIDIA, Palo Alto Networks — plus 40+ critical software maintainers. $100M in usage credits, $4M in open-source donations (Source: Anthropic; Project Glasswing 2026).

03

New safeguard development

Anthropic is building detection and blocking for dangerous outputs, to be first deployed on an upcoming Opus model that does not pose the same risk level — allowing iteration before any broader Mythos-class release.

04

Cyber Verification Program

A coming program for legitimate security researchers whose work may be affected by new safeguards. Designed to ensure defenders retain capability parity while blocking misuse at scale.

THE STRATEGIC READ: Anthropic is buying time for defenders. Canadian enterprises must assume the window is 12–18 months before comparable capabilities appear in adversary hands.

SECTION 4 — RECOMMENDED ACTIONS

THE READINESS FRAMEWORK

IMMEDIATE

1 Board-level briefing

Brief the audit and risk committee on Mythos, Glasswing, and the re-baselined threat environment. Establish a 90-day review cadence.

2 Patch management reset

Shorten patch SLAs and revise processes accordingly to patch at speed and scale

3 Protection of software supply chain

Ensure you have an accurate inventory of software including third party libraries and other components that make up your ecosystem, in order to be able to evaluate / prioritize the risk to your environment from zero-days

4 Legacy estate inventory

Identify end-of-life software (particularly in crown jewel assets). Flag anything not on vendor support as material risk and implement mitigating measures based on risk

5 Resilience testing

Pressure test resiliency plans against AI driven threats to determine readiness for cyber incidents and disruptions to critical systems and data. Begin to implement operational resilience enhancements where required

6 Security testing

Implement continuous AI driven enhanced security testing program (to include application code, software assets and adversarial simulations against the complete environment) to identify and mitigate weaknesses

NEAR-TERM

1 Enhance zero trust strategy

Assume compromise will occur and implement zero trust strategies including securing the identity layers and implementing segregation (and isolating critical assets as much as possible) to limit the blast radius to ensure a single exploit does not equal full compromise

2 Lock down third party risks

Evaluate third party access to systems and environment, and restrict access to critical systems where possible, enhance security monitoring over third party connections

3 Enhance Detection capability

Pressure-test SOC detection against machine-speed lateral movement scenarios. Benchmark mean-time-to-detect against an AI-accelerated intruder model. Enhance detection coverage to areas not currently monitored

4 Insurance renewal strategy

Engage brokers early on FY27 renewals. Model the premium and coverage impact of current legacy footprint and SOC maturity

5 Threat intel coverage

Ensure you have a comprehensive threat intelligence program designed to alert on hacker activity in order to be able to inform detection processes and dynamically adjust controls rapidly

STRUCTURAL

1 Migration roadmap for legacy

Fund and sequence migrations off end-of-life platforms. Prioritize internet-facing surface and systems handling material data

2 AI-for-defence

Deploy defensive AI tooling (vulnerability scanning, Configuration management, IR automation, SOC automation, etc) — do not wait for adversary parity to start

3 Regulatory engagement

Proactively engage OSFI, CSE, and relevant provincial regulators. Shape, rather than react to, forthcoming AI-cyber guidance

4 Enhance insider risk program

Ecosystems embedded with AI pose an elevated risk of insider threat and insider risk programs should be evolved accordingly.

5 Safeguard AI

Implement additional AI guardrails including full AI observability, advanced behavioral analytics, automating policy enforcements on AI actions and prompts, implementing synthetic content detection, etc

SECTION 5 — CIO READINESS FRAMEWORK

HOW READY ARE YOU?

Five capability domains CIOs must assess and advance to operate securely in an AI-augmented threat environment

	DOMAIN	DIMENSION	DIAGNOSTIC QUESTIONS	PRIORITY ACTION
01	VISIBILITY & ASSET CONTROL	Know your surface	**1.** Is your full software estate inventoried — including EOL and unsupported systems? **2.** Do you have real-time visibility into third-party and open-source dependencies? **3.** Can you identify unpatched CVEs across cloud, on-prem, and OT within hours?	Commission a legacy-estate audit. EOL software on internet-facing or data-holding systems is an immediate priority. Ensure up to date inventory of software including third party libraries
02	PATCH & RESPONSE VELOCITY	Compress the window	**1.** What is your actual time-to-patch for critical-severity CVEs today? **2.** Are auto-update policies enforced, or deferred by exception culture? **3.** Can your SOC detect and contain a breach within hours, not days?	Reset patch SLAs. Treat dependencies to critical systems with CVE fixes as urgent, not routine maintenance. Review, update and test resiliency plans for critical systems
03	DETECTION AT MACHINE SPEED	Re-baseline the SOC	**1.** Is your detection tooling calibrated for AI-paced lateral movement? **2.** Have you run red-team exercises against machine-speed intrusion scenarios? **3.** What is your mean-time-to-detect and does it assume a human-paced attacker?	Run tabletop exercises simulating an AI-accelerated intrusion. Identify where detection logic breaks. Execute advanced security testing programs. Find the gaps before they are exploited.
04	SUPPLY CHAIN & VENDOR RISK	Tighten third-party exposure	**1.** Do your software and MSSP contracts include AI-era security warranties? **2.** Are you notified when vendors push updates that affect your attack surface? **3.** Can you enforce minimum patching SLAs on critical software suppliers?	Review contracts for AI tooling, cloud software, and MSSPs signed before mid-2025. Renegotiate security warranties and liability caps at next renewal.
05	GOVERNANCE & BOARD POSTURE	Lead from the top	**1.** Does your board understand the Mythos threat environment and its timeline? **2.** Is cyber risk framed as an enterprise risk item with budget authority? **3.** Do you have a cross-functional AI-cyber response owner (not just the CISO)?	Brief the audit and risk committee within 30 days. Establish a standing AI-cyber review cadence of 60–90 days with CIO, CRO, CISO, and General Counsel.

SECTION 6 — Limitations

Limitations and Constraints

Despite its capabilities, Claude Mythos has specific operational boundaries

01 Mitigation Barrier

The model does not reliably bypass strong, well-implemented exploit mitigations. Modern memory safety protections in the Linux kernel, such as KASLR and W^X (write XOR execute), still pose significant barriers to successful exploitation after a vulnerability is found.

02 Complexity Issues

Mythos shows inconsistent success with complex, multi-stage attacks, particularly those that require crossing trust boundaries like deeply sandboxed systems or host/guest environments.

03 Human Oversight

Because the model can produce partial primitives or denial-of-service results rather than full compromises in complex scenarios, effective use of the model still frequently requires human intervention.

04 Operational Capacity Constraints

While Mythos significantly accelerates vulnerability discovery, the primary constraint shifts to an organization's ability to triage, patch, and deploy mitigations at scale. The limiting factor becomes enterprise capacity to triage and patch vulnerabilities at speed during high-volume disclosure events.

05 Exploit Interaction Risk

Multiple lower-severity vulnerabilities may be combined into viable attack paths, increasing systemic risk beyond individual exploit success.

kpmg.com/ca