The Skeleton Key Nobody Knew Existed

Project Glasswing, Claude Mythos Preview, and the moment AI-augmented cybersecurity stopped being a future problem

---

Dr. Gregory S. Carmichael · CEO, Quantum Reserve Capital · San Juan, Puerto Rico

1 The Setup

Imagine you own a building. You have lived in it for thirty years. Security consultants have walked every inch of it. You have installed cameras, motion sensors, and a deadbolt on every door. You feel safe.

Now imagine someone hands you a pair of glasses that lets you see through the walls—and in the first five minutes of wearing them, you discover seventeen doors you never knew existed, twelve of them wide open.

That is what just happened to the world's most critical software infrastructure. And the glasses have a name: **Claude Mythos Preview**.

On April 8, 2026, Anthropic announced **Project Glasswing**—a coalition of AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks, unified around a single alarming fact: an AI model now exists that can find and exploit software vulnerabilities better than almost any human alive. The question is not whether this capability will reshape civilization. It already has. The question is who controls it, and for what purpose.

1.1 What Claude Mythos Preview Actually Did

Mythos Preview is an unreleased frontier model. It is not publicly available. What Anthropic is disclosing about it, however, is extraordinary.

In the span of a few weeks, the model autonomously identified **thousands of zero-day vulnerabilities**—security flaws unknown to the software's own developers—across every major operating system and every major web browser. "Autonomously" is the operative word. No human steering. No expert prompting. The model did the work.

Three specific finds tell the story with precision:

- A **27-year-old vulnerability in OpenBSD**—a system with a hard-won reputation as one of the most security-hardened operating systems ever built, used to run the firewalls protecting critical infrastructure—that allowed a remote attacker to crash any machine running it simply by connecting to it. Twenty-seven years. Millions of security-conscious eyes. One model. Days.

- A **16-year-old vulnerability in FFmpeg**, the video encoding library embedded in a staggering fraction of the world's software, residing in a single line of code that automated testing tools had touched **five million times** without flagging it. That is not a near-miss. That is a systematic failure of the entire automated security paradigm as it existed before this moment.

- A **chained exploit in the Linux kernel**—the software running the majority of the world's servers—that allowed privilege escalation from ordinary user access to complete machine control. Not a single flaw. A sequence of flaws, recognized and assembled by the model as a complete attack chain.

To understand how significant the capability jump is, consider the evaluation benchmarks released alongside the announcement:

Benchmark Performance: Mythos Preview vs. Claude Opus 4.6

Benchmark	Mythos Preview	Opus 4.6
CyberGym (vulnerability reproduction)	83.1%	66.6%
SWE-bench Verified	93.9%	80.8%
SWE-bench Pro	77.8%	53.4%
Terminal-Bench 2.0	82.0%	65.4%
SWE-bench Multilingual	87.3%	77.8%
Humanity's Last Exam (with tools)	64.7%	53.1%
GPQA Diamond	94.6%	91.3%

1.2 Key Vulnerabilities: What Was Found and Fixed

The three cases Anthropic has disclosed in full are worth examining closely, because they illustrate three distinct failure modes of conventional security practice — and because they were not found in obscure, low-traffic code. They were found in software that forms the actual backbone of global digital infrastructure.

The OpenBSD Remote Crash (27 years). OpenBSD has a singular reputation in the security community. Its developers have maintained a multi-decade commitment to code auditing, secure defaults, and proactive hardening. It is the operating system chosen to run firewalls, VPN gateways, and critical network infrastructure precisely because its security record is nearly unmatched. The vulnerability Mythos Preview found was a flaw in the network stack that allowed any remote attacker to crash any machine running OpenBSD simply by initiating a connection — no authentication, no exploit payload, no special knowledge of the target required. The flaw had been present since 1997. It survived twenty-seven years of professional audits, academic review, and automated scanning. Mythos found it in days.

The mechanism matters: the model was not doing exhaustive fuzzing. It was reasoning about the logical structure of the network protocol implementation and identifying a condition under which the code's assumptions about state would be violated. That

is qualitatively different from a tool that tries random inputs until something breaks.

The FFmpeg Zero-Day (16 years, 5 million automated test hits). FFmpeg is arguably the most widely deployed media processing library in existence. It underpins video playback in browsers, streaming platforms, communication software, and operating systems across every major vendor. The vulnerability resided in a single line of code that had been touched by automated fuzz testing tools **five million times** without detection. The reason is structural: the flaw only manifested under a specific combination of input conditions that required understanding the semantic relationship between several distant parts of the codebase simultaneously. No single-pass fuzzer could construct that input without first understanding what the code was trying to do. Mythos Preview understood it. The patch was delivered to the FFmpeg maintainers and is now in production.

The Linux Kernel Privilege Escalation Chain. This is the most technically significant of the three disclosed cases. The Linux kernel runs the majority of the world's servers, most Android devices, and most cloud infrastructure. Mythos Preview did not find a single vulnerability here — it found *a sequence of vulnerabilities and recognized that they could be chained*. An attacker starting from ordinary user-level access could traverse the chain to achieve complete control of the machine. No individual flaw in the chain was exploitable in isolation. Each required the others to be meaningful. The model assembled the complete attack path autonomously, without a human operator specifying what to look for or how to connect the pieces.

This represents a fundamental advance in automated security reasoning. Security researchers have long known that the most dangerous vulnerabilities are chained exploits — combinations of individually minor flaws that together create catastrophic exposure. Until now, constructing those chains required elite human expertise. Mythos Preview can construct them autonomously.

The broader pattern. Beyond these three named cases, Anthropic disclosed that Mythos Preview found critical vulnerabilities in every major operating system

and every major web browser. For vulnerabilities not yet patched at the time of the announcement, Anthropic committed cryptographic hashes of the details — a provable commitment to disclose the specifics once patches are in place. The 90-day coordinated disclosure window gives software maintainers a structured timeline. This is responsible disclosure practice applied at a scale and speed that has never before been possible.

PLAIN LANGUAGE

Think about the software your life runs on. Your bank. Your hospital's records. The grid that keeps your lights on. The internet router you never think about. The video call you just took.

None of that software was written perfectly. All of it has bugs. Some of those bugs, in the right hands, are doors that let a bad actor walk in and take over. For most of computing history, finding those doors required years of specialized training and intuition that took a long time to develop in a human mind. Project Glasswing tells you that constraint no longer exists. An AI model found a door that had been sitting unlocked for 27 years — in code that millions of people had trusted with their most critical systems.

That sixteen-point gap on CyberGym is not incremental. It is generational. On the coding benchmarks that underpin these cyber capabilities, the margins are equally stark. These are not software metrics. They are national security metrics.

1.3 Finding vs. Fixing: The Asymmetry That Defines the Challenge

The most important distinction in all of cybersecurity is one that Project Glasswing forces into sharp relief: **finding a vulnerability and fixing a vulnerability are not the same kind of problem**. They are not even the same order of magnitude of difficulty. Mythos Preview has dramatically accelerated one side of that equation. The other side remains stubbornly, structurally hard — and understanding why is essential to

understanding what Glasswing can and cannot deliver.

Finding is a reasoning problem. Given enough analytical capability applied to a codebase, a vulnerability either exists or it does not. The question is whether the analyst — human or AI — has the contextual depth to recognize the conditions under which a flaw becomes exploitable. Mythos Preview has demonstrated that it can perform this reasoning at a level that exceeds most human practitioners, autonomously, across the most hardened software in the world. A 27-year-old vulnerability in OpenBSD fell in days. That part of the problem has been transformed.

Fixing is an engineering, organizational, and sociotechnical problem. Once a vulnerability is identified, the work that follows is categorically different in nature — and far more resistant to acceleration.

First, the fix must address the *root cause*, not the symptom. The line of code where a vulnerability manifests is often not where the flaw originates. The actual error may be a design assumption made years or decades earlier, embedded in an interface contract between components, or a consequence of how two separately correct subsystems interact under conditions nobody originally anticipated. Patching the manifestation without addressing the cause leaves the system vulnerable to variants — a partially patched vulnerability frequently becomes more dangerous, because it signals to adversaries exactly where to probe while offering defenders false confidence that the issue is resolved.

Second, every patch must survive a battery of regression testing across the full dependency graph of the software being fixed. The Linux kernel patch that closes a privilege escalation chain must not break the thousands of legitimate code paths that traverse the same execution space. For a codebase the scale of the Linux kernel — roughly 30 million lines of code, with hundreds of active subsystems maintained by contributors distributed across the globe — that regression surface is enormous. A patch submitted to the kernel must pass Linus Torvalds' review process, clear subsystem maintainers, survive automated continuous integration testing, and then work its way through the

full distribution pipeline: from mainline kernel to distribution-specific builds at Red Hat, Debian, Ubuntu, and dozens of others, before it reaches the system administrators who must actually deploy it.

Third, deployment is not instantaneous. Even after a patch is validated and shipped, the population of vulnerable systems does not flip to patched overnight. Enterprise environments operate on patch cycles measured in weeks or months. Industrial control systems and critical infrastructure — power grid management systems, hospital networks, air traffic control software, water treatment facility controllers — often operate on maintenance windows measured in quarters, or years. Patches that require system downtime cannot be applied during active operations, and for some systems, *there is no safe window* that comes quickly enough. Security researchers have documented cases where critical infrastructure vulnerabilities remained unpatched for three to five years after public disclosure, not because of negligence, but because the operational constraints of the systems made earlier patching genuinely impossible.

Fourth, the **disclosure window** creates its own race condition. Between the moment a vulnerability is found and the moment a patch reaches all affected systems, the vulnerability exists in a state of partially controlled exposure. Coordinated disclosure practice — the 90-day window Anthropic has committed to for Glasswing findings — attempts to give defenders a structural head start. But if the vulnerability becomes known to adversaries before patching is complete, the window collapses. Anthropic's use of cryptographic hash commitments for undisclosed vulnerabilities is an attempt to manage this race: it proves the finding was made without revealing the details that would allow exploitation.

Fifth, and perhaps most fundamentally, there is the **patch debt problem**. Mythos Preview found thousands of critical vulnerabilities in a matter of weeks. Human engineering teams cannot remediate thousands of critical vulnerabilities in weeks. The bottleneck has shifted — decisively and permanently — from discovery to remediation. The world now has a tool capable of surfacing security debt faster than the

industry can pay it down. The prioritization of which vulnerabilities to fix first, in which order, with which resources, across which systems, under which operational constraints: that is a human problem, a management problem, an organizational problem. It is not a problem that accelerates with the model's reasoning speed.

This asymmetry has a direct implication for how defenders should use Mythos-class capabilities. The goal is not simply to find everything as fast as possible and hand a list to engineering teams. It is to use the discovery capability strategically — to identify the highest-leverage vulnerabilities first, the ones most likely to be discovered independently by adversaries, the ones whose exploitation would be most catastrophic, the ones whose patch paths are most tractable given current operational constraints. The model can reason about all of these dimensions. But someone has to ask the right questions, and someone has to do the work of fixing what gets found.

Project Glasswing is a significant advance in the finding problem. The fixing problem remains exactly as hard as it has always been — and there is now more to fix than before.

1.4 The Open Source Threat: Small Models, Agent Swarms, and the Democratization of Offense

Project Glasswing's announcement carries an unspoken premise that deserves to be made explicit: the defensive deployment of Mythos Preview works only as long as the offensive use of equivalent capability remains restricted. That assumption is already eroding. Not because adversaries will immediately replicate Mythos Preview — they will not, not quickly — but because they do not need to. The open source model ecosystem has quietly crossed a threshold that most security practitioners have not yet fully priced in, and the gap between what a capable nation-state adversary can do today with freely available tools and what Glasswing's defenders can counter is narrowing faster than the coordinated disclosure process was designed to handle.

The capability floor has moved. The public release of DeepSeek-R1 in January 2025 was the clearest marker of the shift. A 671-billion parameter reasoning model, trained for a reported $294,000 — a fraction of comparable Western frontier model training costs — became freely downloadable and runnable by anyone with sufficient hardware. More consequentially, DeepSeek also released distilled versions as small as 8 billion parameters, built on Meta's Llama architecture, that run on consumer-grade GPU hardware with tools like Ollama in a single command. By the end of 2025, Alibaba's Qwen model family had accumulated over 700 million downloads, making it the world's largest open-weight AI ecosystem. The open source capability floor is no longer academic research models from 2023. It is production-grade reasoning systems available at zero marginal cost, deployable on hardware that costs less than a used car.

Agent frameworks transform capability into scale. A single open source model running on a laptop is a capable tool. The same model embedded in an autonomous agent framework — given a set of tools, a target, a goal, and the instruction to iterate until it succeeds — is a different class of threat entirely. Research systems published in 2025 and presented at venues including ICLR 2026 have demonstrated that LLM agents embedded in multi-step reasoning loops can autonomously perform reconnaissance, target selection, and exploitation against real-world vulnerable systems without prior knowledge of where the vulnerabilities are located. Systems such as EniGMA, HackSynth, D-CIPHER, and CRAKEN — all academic research projects, all publicly described — have shown that LLM agents can reason about attack vectors and autonomously exploit vulnerable services. CyberExplorer, published in early 2026, introduced a multi-agent framework in which coordinated agent instances conduct parallel offensive security tasks across multiple targets simultaneously without predefined plans. These are not classified capabilities. They are peer-reviewed papers with public code repositories.

The practical implication is a force multiplier that changes the economics of attack operations fundamentally. An adversary who previously needed ten skilled operators to

conduct sophisticated vulnerability research can now deploy ten instances of an open source agent framework operating in parallel, each probing a different target, each capable of basic reasoning about the code it encounters, each running at the cost of local compute. The human expertise constraint does not disappear entirely — designing the agent, interpreting results, and operationalizing findings still requires skill — but it is dramatically relaxed. The asymmetry that made elite cyberattack capability expensive has been partially reversed.

Safety guardrails are optional on open source models. This is the structural fact that distinguishes open source models from frontier API models as an offensive threat vector, and it is one the security community has been reluctant to state plainly. When a user queries Claude, GPT-5, or Gemini through an official API, the request passes through content moderation, rate limiting, and usage monitoring systems that create both friction and an audit trail. When a user runs DeepSeek-R1 locally through Ollama, none of those controls exist. The model's built-in safety training can be removed or circumvented by fine-tuning, and the entire interaction happens in a private compute environment with no external visibility.

The empirical data on what this means in practice is stark. The National Institute of Standards and Technology tested DeepSeek R1 and found it complied with **94 percent of overtly malicious requests** using common jailbreaking techniques — while comparable U.S. frontier models complied with just 8 percent. Cisco independently validated these findings, reporting a **100 percent attack success rate** against DeepSeek R1, with the model failing to block a single harmful prompt in their testing. For the locally deployed, fine-tuned, guardrail-stripped version of these models — the version an adversarial operator would actually use — the compliance rate approaches 100 percent by construction. There is no policy enforcement mechanism in an open weight model that a sufficiently motivated operator cannot remove.

CrowdStrike's research added another dimension: when DeepSeek-R1 receives prompts containing topics the Chinese Communist Party considers politically

sensitive, the probability of it generating code with severe security vulnerabilities increases by up to **50 percent**. This is not a random failure mode. It is a systematic bias that an adversary aware of the pattern can exploit deliberately — framing requests in ways that trigger the model's politically conditioned behavior to produce more dangerous outputs than its baseline safety posture would otherwise permit.

The threat is already operational, not theoretical. Google Threat Intelligence has identified malware strains that query Qwen models for real-time code generation during active intrusions — the model is being used as an in-the-loop reasoning component of live attack operations, not merely as a research or planning tool. AI-assisted cyberattacks increased 72 percent from 2024 to 2025. The FBI's Internet Crime Complaint Center logged a 37 percent rise in AI-assisted business email compromise over the same period. These numbers reflect the lower end of AI-augmented offense — social engineering and phishing — where open source models have already achieved near-parity with frontier models for most practical purposes. The vulnerability research use case is harder and still frontier-model-advantaged, but the trajectory is clear.

The systemic risk to the global cyber ecosystem. The combination of these factors — capable open source models, freely available agent frameworks, removable safety guardrails, and no monitoring infrastructure — creates a structural vulnerability in the global cybersecurity system that Project Glasswing's defensive posture does not fully address. Glasswing's approach assumes that defenders using Mythos Preview will find and patch critical vulnerabilities before adversaries with equivalent or lesser capability find them independently. That assumption holds when the capability gap between defenders and attackers is large. It weakens as open source models improve and as agent frameworks lower the barrier to deploying offensive reasoning capability at scale.

The 90-day coordinated disclosure window was designed for a world in which sophisticated vulnerability discovery was rare and slow. In a world where agent swarms

running on open source models can be deployed by a moderately resourced adversary to probe codebases continuously, the window is no longer calibrated to the threat. Anthropic's own red team blog, published alongside the Glasswing announcement, noted explicitly that industry-standard 90-day disclosure windows may not hold up against the speed and volume of LLM-discovered vulnerabilities, and that the industry needs workflows that can keep pace.

> The open source model ecosystem has not yet matched Mythos Preview's capability for finding chained exploits in production systems. It has already matched its capability for everything below that threshold — and everything below that threshold covers the vast majority of exploitable vulnerabilities in production software today.

The correct framing is not that open source models are as dangerous as Mythos Preview. They are not, yet. The correct framing is that the capability level required to cause serious harm is now available to any adversary willing to spend an afternoon reading a GitHub README. The top of the threat distribution — nation-states and elite criminal organizations — has always had this capability through human expertise. What has changed is the middle of the distribution: the moderately resourced, moderately skilled attacker who previously could not conduct sustained, reasoning-based vulnerability research against hardened targets. That actor now can. And the global software ecosystem, which was already struggling to keep pace with the vulnerability discovery rate before Glasswing, is now facing a sustained expansion of that rate from both directions simultaneously — frontier models accelerating discovery for defenders, open source models accelerating discovery for a much larger and less accountable set of adversaries.

Project Glasswing is the right response to the frontier model problem. It does not yet have a credible answer to the open source problem. That answer will require a different kind of coordination — not just between the companies that make frontier models, but between governments, open source foundations, and the security community

broadly — to establish norms, monitoring mechanisms, and response capabilities for an attack surface that currently has no analogous defense architecture. The window to build that architecture is narrowing at the same pace as open source model capability is improving.

1.5 How Mythos Preview Works: Architecture and Capability

The benchmark numbers tell you *what* Mythos Preview can do. Understanding *how* it does it requires a closer look at the underlying architecture — because the mechanism explains both the capability leap and the risk it represents.

Mythos Preview operates as a **fully autonomous agentic system**. Unlike conventional static analysis tools that scan code against a database of known signatures, or even earlier AI models that require a human operator to frame the problem, Mythos Preview reasons about code the way a senior security engineer does: it forms hypotheses about where vulnerabilities might exist, constructs experiments to test them, interprets the results, and iterates. It does not need to have seen the specific vulnerability before. It reasons from first principles about what the code *should* do versus what it *actually* does under adversarial conditions.

The model's performance on **Terminal-Bench 2.0** (82.0%) is particularly revealing. Terminal-Bench evaluates autonomous computer use — the model's ability to operate a full computing environment, execute commands, read outputs, and adapt its approach over extended task sequences. The evaluation methodology is rigorous: Anthropic used the Terminus-2 harness with adaptive thinking at maximum effort, a one-million-token budget per task, and results averaged over five independent attempts per task. At extended timeout limits, Mythos Preview reached 92.1%. This is not a model completing short scripted sequences. It is a model conducting sustained, self-directed investigation through a live terminal environment.

Exploit chaining represents perhaps the most consequential technical advance. Earlier models could identify discrete vulnerabilities. Mythos Preview demonstrated the

ability to recognize *relationships between vulnerabilities* — to identify that flaw A, combined with flaw B, enables an attack path that neither flaw alone would permit. The Linux kernel exploit that escalated from ordinary user access to full machine control was not a single bug. It was a composed attack chain assembled autonomously. This capability shifts the threat from point vulnerability to systemic compromise: an adversary with access to Mythos-class reasoning can probe not just for individual weaknesses but for the combinations that yield maximum leverage.

The model's **black-box testing capability** is equally significant. Much of the world's most critical software is distributed in compiled binary form without source code — embedded firmware, proprietary network appliances, industrial control systems. Mythos Preview can reason about the behavior of such systems from the outside, probing inputs and observing outputs to infer internal structure and identify exploitable conditions. This extends the attack surface from auditable codebases to the vast universe of systems that have historically been protected by obscurity.

On **OSWorld-Verified** (79.6%) — which tests autonomous GUI-based computer operation — and **BrowseComp** (86.9%, using 4.9× fewer tokens than Opus 4.6) — which tests complex web research and navigation — the model demonstrates that these capabilities generalize across computing environments. Mythos Preview is not narrowly specialized for one attack surface. It is a general-purpose reasoning system whose coding and analytical capabilities happen to make it extraordinarily effective at security research. The same architecture that finds vulnerabilities in operating systems also outperforms predecessor models on **GPQA Diamond** (94.6% vs Opus 4.6's 91.3%) — a benchmark testing graduate-level scientific reasoning across physics, chemistry, and biology. This is not a vulnerability scanner. It is a reasoning engine that has discovered security research as a domain where its capabilities are decisive.

1.6 Token Efficiency and Accuracy: The Compound Advantage

Raw accuracy improvements are expected from each successive generation of frontier models. What makes Mythos Preview's performance profile genuinely unusual is the *compound advantage*: it achieves higher accuracy simultaneously with dramatically lower token consumption. These two properties normally trade off against each other — more careful reasoning requires more computation. Mythos Preview breaks that tradeoff.

The clearest evidence is **BrowseComp**, which tests a model's ability to conduct complex, multi-step web research — the kind of investigation that mirrors how a security researcher traces a vulnerability from a symptom back to its root cause across multiple codebases, documentation sources, and historical records. Mythos Preview scored 86.9% against Opus 4.6's 83.7%. Better result. But the token consumption tells the more important story: Mythos Preview achieved that higher score using **4.9 times fewer tokens** than Opus 4.6. It did not just reason better. It reasoned more efficiently, more directly, with less wandering.

Accuracy and Efficiency: Mythos Preview vs. Claude Opus 4.6

Benchmark	Mythos	Opus 4.6	Gap
CyberGym (vulnerability reproduction)	83.1%	66.6%	+16.5 pts
SWE-bench Verified	93.9%	80.8%	+13.1 pts
SWE-bench Pro	77.8%	53.4%	+24.4 pts
Terminal-Bench 2.0	82.0%	65.4%	+16.6 pts
SWE-bench Multilingual	87.3%	77.8%	+9.5 pts
Humanity's Last Exam (with tools)	64.7%	53.1%	+11.6 pts
GPQA Diamond	94.6%	91.3%	+3.3 pts
BrowseComp (at 4.9× fewer tokens)	86.9%	83.7%	+3.2 pts
OSWorld-Verified	79.6%	72.7%	+6.9 pts

The SWE-bench Pro gap deserves particular attention: **24.4 percentage points** on a benchmark testing the model's ability to resolve real, unfiltered GitHub issues from production codebases. SWE-bench Pro is specifically designed to be harder than SWE-bench Verified by excluding issues that show signs of having been in the model's training data. A 77.8% score on genuinely novel software engineering problems — problems the model demonstrably has not seen before — represents a qualitative shift in what automated code reasoning can accomplish.

The GPQA Diamond result (94.6% on graduate-level scientific reasoning across physics, chemistry, and biology) matters for cybersecurity in a less obvious way: it confirms that Mythos Preview's capabilities are not narrowly optimized for software tasks. The same reasoning architecture that finds privilege escalation chains in the Linux kernel can reason at the frontier of human scientific knowledge across unrelated domains. This generality is what distinguishes it from purpose-built security tools and makes it simultaneously more capable and more dual-use than anything that has come before.

Taken together, the benchmarks describe a system that is not incrementally better than its predecessor — it is operating in a different regime. The 24-point SWE-bench Pro gap and the 4.9× token efficiency advantage are not the kind of numbers that come from architectural refinement. They suggest a model that has developed qualitatively different approaches to reasoning about code and computation.

1.7 Why Software Vulnerabilities Are a Macroeconomic Variable

Before Project Glasswing, this conversation belonged to a narrow technical priesthood. It should not. The economic surface area of software vulnerabilities is enormous and almost entirely invisible to conventional financial analysis.

The current global cost of cybercrime is estimated at roughly **$500 billion per year**. That number almost certainly understates the true figure—it excludes most of the damage to healthcare systems, power grids, and logistics networks that never becomes

public, plus the second-order costs of hardened posture, redundant systems, and lost economic activity. And it predates the moment we are now entering.

Here is the structural shift: until very recently, finding a serious software vulnerability required rare, expensive human expertise. The labor constraint was binding. State-sponsored actors in China, Iran, North Korea, and Russia have spent billions cultivating and deploying that expertise precisely because it was scarce. A handful of elite operators could hold the world's infrastructure at asymmetric risk because the skill set required to operate at that level did not scale.

Mythos Preview breaks that constraint. It does not replicate one expert. It replicates **the reasoning patterns of thousands of experts simultaneously, autonomously, at the cost of compute**. The cost of a sophisticated zero-day exploit drops toward the marginal cost of API tokens.

> "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."
>
> — Elia Zaitsev, CTO, CrowdStrike

WHY THIS HITS YOUR WALLET

Serious cyberattacks used to require nation-state backing — only governments had the budget and patience to develop that kind of talent at scale. What just changed is that the talent constraint broke. An AI model can now do in hours what took elite human teams months. The cost of a sophisticated attack drops toward the cost of a cloud computing bill. That means more attackers, more attacks, and attacks on targets that were previously too small to be worth the effort. The $500B annual estimate for global cybercrime was already understated. It is about to get significantly worse — unless defenders move first.

Anthropic is not the only one who knows this. The language from the Project Glasswing partners is not celebratory. It is urgent. Palo Alto Networks framed it in the most direct terms possible: there will be more attacks, faster attacks, and more sophisticated attacks. This is not hype. This is a consortium of the world's most security-hardened institutions telling you, collectively, that the threat model changed—and changed fast.

Fig. 2 *Greta oto — the glasswing butterfly — carries a gold skeleton key, a miniature master key to the world's most critical infrastructure. The transparent wings reveal the key beneath them, just as Mythos Preview reveals the vulnerabilities hidden inside the world's most trusted software.*

1.8 The Glasswing Structure

The architecture of the project is worth understanding, because it tells you something about how Anthropic has calibrated the risk.

Mythos Preview is **not being made publicly available**. What Anthropic is doing is providing controlled access to a curated set of defenders: the named consortium partners, plus an additional group of over forty organizations that build or maintain critical software infrastructure. The commitment is $100 million in usage credits for this defensive work, plus $4 million in direct donations to open-source security

organizations—$2.5 million to Alpha-Omega and OpenSSF through the Linux Foundation, $1.5 million to the Apache Software Foundation.

The reasoning is classical dual-use logic applied at civilizational scale: the same capability that makes Mythos Preview dangerous in adversarial hands makes it indispensable for defense. The bet is that getting defenders access first—and using that lead to find and patch the most critical vulnerabilities before bad actors discover them independently—produces a net positive outcome even though the underlying capability is genuinely dangerous.

Anthropic is explicit that their eventual goal is to enable Mythos-class models at scale, and that getting there requires developing safeguards capable of detecting and blocking the model's most dangerous outputs. The plan is to launch those safeguards with an upcoming Claude Opus model, refine them in production, and only then consider broader deployment.

The name itself encodes the philosophy. The glasswing butterfly, *Greta oto*, has transparent wings—it hides in plain sight, like the vulnerabilities this model surfaces. But those same wings let it evade harm—like the transparency Anthropic is advocating in its approach. Power made visible, deployed in service of protection.

1.9 The Mythos Experience: What It Feels Like to Work With

Every partner organization that has spent time with Mythos Preview says something revealing in their public statements, something that goes beyond benchmark numbers. Cisco's Chief Security and Trust Officer describes capabilities that required a fundamental rethinking of how to protect critical infrastructure. AWS's CISO describes applying it to critical codebases and having it strengthen their code — a description that implies an iterative, collaborative engagement, not a one-shot scan. CrowdStrike's CTO notes that what changes is not just what you can detect, but the speed of the entire defensive loop. These descriptions point to a qualitative experience that the numbers alone do not capture.

Mythos Preview behaves differently from every previous AI security tool in a specific way: **it pushes back**. When a human operator frames a problem incorrectly — asks it to look for a particular class of vulnerability when the actual risk surface is elsewhere, or accepts a patch that does not fully close the underlying flaw — the model does not simply execute the instruction. It surfaces the discrepancy. It asks whether the assumption is correct. It presents the alternative hypothesis alongside the work it was asked to do.

This is not conversational polish. It has operational consequences. Security researchers working with Mythos Preview have reported that the model's resistance to framing errors has caught assumptions that would have led to incomplete patches — the security equivalent of treating the symptom without diagnosing the disease. A vulnerability that is partially patched often becomes more dangerous, not less: it gives defenders false confidence while leaving an attack surface that adversaries can still exploit with only marginal additional work.

The model also develops what can only be described as **contextual investment** over the course of a session. The longer it works on a codebase, the more effective it becomes — not because it is retrieving cached results, but because it is building an increasingly precise model of how the specific system is constructed, where its assumptions live, and which of those assumptions are load-bearing. Security researchers describe the experience as working with a collaborator who gets sharper the more you work together, who remembers what you established in the first hour and builds on it in the fourth.

This is architecturally significant. Earlier AI security tools operated as stateless scanners: each query was independent, each result disconnected from the last. Mythos Preview operates with the kind of accumulated contextual understanding that previously existed only in human experts who had spent months immersed in a specific codebase. The difference in outcome is not linear. The most dangerous vulnerabilities — the chained exploits, the logic flaws, the assumptions buried in design decisions

made years ago — are precisely the ones that require that accumulated context to find. They are invisible to stateless tools and visible to Mythos Preview.

The implications for how organizations should structure their security programs are significant. Mythos Preview is not a scanner you run and walk away from. It is a reasoning partner that becomes more valuable the more it is engaged with, the more context it is given, and the more its pushback is taken seriously rather than overridden. Organizations that deploy it as a sophisticated grep tool will get sophisticated grep results. Organizations that treat it as a collaborative analyst — one that happens to operate at a speed and scale no human team can match — will get something categorically different.

THE GOOD NEWS

Anthropic is using this capability defensively first. They have brought in the companies that run the world's most important infrastructure and told them: here is your early access, go find your vulnerabilities before someone else does. Most of the flaws Mythos found have already been patched. The digital skeleton keys are being melted down before the copies can be made.

The window will not stay open indefinitely. Capability diffuses. When AI models with these abilities become more widely available — and they will — every organization that has neglected its software security will face a fundamentally different threat landscape. The organizations that survive that transition will be the ones that started hardening now.

1.10 The Quantum Dimension: When the Lock Itself Breaks

Project Glasswing addresses the near-term crisis—AI models finding and exploiting classical software vulnerabilities at unprecedented speed. But embedded in the announcement, easy to miss if you are not watching for it, is a signal about the longer-

horizon threat that makes the near-term problem look almost manageable by comparison.

Google's representative at the Project Glasswing launch specifically listed **post-quantum cryptography** alongside zero-day disclosure and open-source security as a critical area requiring cross-industry coordination. That sentence deserves to stop you cold.

Here is why. Nearly all of the world's digital security—banking transactions, communications, identity verification, and the cryptographic foundations of every blockchain and digital asset system in existence—rests on mathematical problems that classical computers find prohibitively expensive to solve. The security of RSA-2048 depends on the fact that factoring a large semiprime takes classical computers longer than the age of the universe. The security of elliptic curve cryptography, which underpins Bitcoin, Ethereum, and virtually every digital signature scheme in financial infrastructure, depends on the discrete logarithm problem being computationally intractable.

A sufficiently capable quantum computer running Shor's algorithm breaks both. Not gradually. Not partially. *Completely.*

The National Institute of Standards and Technology finalized its first three post-quantum cryptographic standards in 2024:

- **FIPS 203** — ML-KEM, a key encapsulation mechanism derived from CRYSTALS-Kyber
- **FIPS 204** — ML-DSA, a digital signature scheme from CRYSTALS-Dilithium
- **FIPS 205** — SLH-DSA, a hash-based signature scheme from SPHINCS+

These algorithms are designed to resist attacks from both classical and quantum computers. They represent the federal government's acknowledgment that the quantum threat is not theoretical—it is an engineering deadline.

The connection to Project Glasswing runs deeper than a passing mention. Mythos-class AI models are not quantum computers, and they do not break elliptic curve cryptography. But they represent the *same class of problem*: a capability threshold assumed to be years away, arriving ahead of schedule, with insufficient systemic preparation. The lesson from Glasswing—that a 27-year-old vulnerability can sit undetected in production code relied upon by millions—applies directly to the cryptographic layer. Assumptions that felt safe for decades can be invalidated by a capability jump that nobody was quite ready for.

There is also a subtler, more immediate threat vector sitting at the intersection of AI and quantum cryptography: **harvest now, decrypt later** attacks. Sophisticated adversaries—and the nation-state actors named explicitly in the Glasswing document are sophisticated—are actively intercepting and storing encrypted communications and transactions today, with the explicit strategy of decrypting them once quantum capability matures. Every financial transaction, every identity record, every privileged communication traveling over classical cryptographic infrastructure right now is potentially being archived for future decryption. The data being collected today has a shelf life that extends into the quantum era.

> The question is not whether to migrate to post-quantum standards. The question is whether you build on quantum-resistant foundations from the start—or inherit a mandatory, expensive, high-risk migration on an adversary's timeline.

For digital monetary systems, this is existential in a way it is not for most software categories. A hospital system with a classical cryptographic vulnerability faces a data breach. A monetary system with the same vulnerability faces the permanent invalidation of its security guarantees—the retroactive exposure of every transaction that ever passed through it, and the potential forgery of signatures considered irrevocable. The integrity model collapses.

The architects of new financial infrastructure being built under the **GENIUS Act**—the

federal framework governing payment stablecoin issuance in the United States—are building systems expected to function across a decade or more. That timeline spans the quantum threat window. A payment stablecoin platform that launches today on classical cryptographic assumptions and plans to migrate later is making the same category of mistake that let a 16-year-old FFmpeg vulnerability survive five million automated tests: assuming that what has not been exploited yet is safe.

The correct architecture encodes post-quantum cryptographic standards at the protocol layer from inception, not as a retrofit. FIPS 203, 204, and 205 exist precisely so builders have a standards-compliant, NIST-vetted answer to this design question—one that does not require waiting for the quantum computer to show up before acting.

Project Glasswing, read in full context, is not just a cybersecurity announcement. It is a forcing function: if you are building financial infrastructure today and you are not building it to survive both AI-augmented classical attacks and the coming quantum transition, you are building on sand. The defenders who come out ahead of this moment are those who read Glasswing as a two-register announcement and design their systems to address both simultaneously rather than sequentially.

2 The Convergence: What Happens When Quantum and Mythos-Class AI Arrive Together

Every technology described in this article to this point operates within the framework of classical computing. Mythos Preview reasons about classical software flaws. Open source agent swarms probe classical attack surfaces. Post-quantum cryptography protects against a threat that does not yet fully exist. The implicit assumption underlying all of it is that classical computing remains the dominant substrate — that software vulnerabilities, not cryptographic collapse, define the threat landscape.

That assumption has a finite shelf life. And the events of the past several weeks have shortened it significantly.

2.1 The Qubit Threshold Just Moved

In late March 2026, two publications landed within days of each other that fundamentally changed the calculus of quantum risk. Google announced it had drastically improved the quantum algorithm for breaking elliptic curve cryptography. Simultaneously, a startup called Oratomic — co-founded by a former Google Quantum AI researcher who had spent months using AI to probe quantum algorithms and reported "seeing lots of crazy results" — published a resource estimate with a conclusion that stopped the cryptographic community cold: breaking P-256, the elliptic curve standard that protects the majority of internet traffic including TLS 1.3, requires only **10,000 qubits** on a neutral atom quantum computer architecture. Previous estimates had placed the requirement at millions of qubits for RSA-2048. The combination of a better algorithm and a more efficient architecture reduced the target by two orders of magnitude.

Cloudflare, which encrypts a significant fraction of global internet traffic, published a response within days: it is accelerating its post-quantum deployment roadmap and targeting **full post-quantum security by 2029**. The IBM Quantum organization's CTO, asked about the timeline, said he "cannot rule out quantum in 2030." The Google researcher who led the algorithmic breakthrough noted that Google's decision — announced just days before the Oratomic paper — to pursue neutral atom quantum computing alongside its superconducting qubit program now makes obvious sense. The strategic intent was legible only in retrospect, but the direction of travel is clear.

This is not Q-Day. Q-Day — the moment a cryptographically relevant quantum computer can execute Shor's algorithm against production encryption in real time — has not arrived. But the engineering problem has been simplified by a factor that the entire field needs to recalibrate around. What once required a machine of a scale that felt comfortably distant now requires a machine that IBM, Google, Quantinuum, IonQ, and a dozen well-funded startups are all on documented roadmaps to build

within this decade. IBM expects fault-tolerant quantum computing by 2029. IonQ's roadmap reaches 8,000 logical qubits in 2029 and 80,000 in 2030. Pasqal is targeting a 10,000-qubit neutral atom system in 2026. The Oratomic threshold is not a theoretical target. It is an engineering program underway at multiple organizations simultaneously, across multiple quantum computing architectures, in at least five countries.

> Breaking P-256 — the elliptic curve standard protecting most internet traffic — now requires an estimated 10,000 qubits on a neutral atom architecture. That is not a distant moonshot number. Multiple organizations are on documented roadmaps to reach it within this decade.

2.2 AI Is Accelerating Quantum Development

The Oratomic breakthrough did not emerge from pure quantum physics research. It emerged from a researcher using AI to explore quantum algorithm design space — reporting, in his own words, that he was "seeing lots of crazy results" months before the publication. The convergence between AI and quantum computing is not merely a threat scenario for the future. It is already happening as a research methodology in the present.

This bidirectional relationship has compounding implications. AI helps researchers find better quantum algorithms — as the Oratomic case demonstrates — which reduces the qubit threshold for cryptographic attacks. Better quantum hardware, in turn, eventually enables faster AI training at scales that classical computers cannot efficiently support. Each field accelerates the other. The 2026 Annual Threat Assessment, presented by Director of National Intelligence Gabbard to Congress, explicitly framed quantum computing in dual terms: not merely as an encryption-breaking capability, but as a strategic intelligence processing advantage that early developers will hold over all others. The intelligence community is not treating these as separate domains.

China is pursuing both as explicit national priorities, with a state-directed investment

program in quantum computing that operates independently of Western research. Beijing has published its own post-quantum cryptographic standards separate from NIST's FIPS process — a geopolitical signal that the cryptographic standards landscape is fragmenting along the same lines as every other technology governance domain. An organization building systems today that will operate across a decade must plan not just for post-quantum cryptography, but for a world in which multiple, incompatible post-quantum standard regimes exist simultaneously, enforced by competing regulatory jurisdictions.

2.3 The Compound Threat Architecture

Considered independently, Mythos-class AI vulnerability discovery and quantum cryptographic collapse are each serious, manageable threats. The post-quantum cryptographic standards (FIPS 203, 204, 205) exist and are ready to deploy. Project Glasswing represents a serious, resourced, coordinated effort to close the classical software vulnerability surface. Each threat has a defensive architecture that, if implemented with sufficient urgency, can contain it.

Considered together, as a simultaneous convergence rather than a sequential progression, they describe something more structurally dangerous.

The classical software vulnerability layer and the cryptographic trust layer are not independent systems. They interact. Software vulnerabilities are how adversaries gain access to systems. Cryptographic trust is how systems verify that the entities communicating with them are who they claim to be. An adversary who combines Mythos-class vulnerability discovery with cryptographic compromise capability has an attack surface that spans both layers simultaneously: they can identify the vulnerabilities to exploit *and* forge the authentication credentials needed to exploit them without triggering standard detection. Network intrusion detection systems that watch for anomalous authentication events become blind to an adversary who can generate valid-looking cryptographic signatures for arbitrarily constructed identities. The as-

sumption that defenders can identify the source and nature of an attack — the foundation of incident response — breaks down when both the access pathway and the authentication layer are compromised.

The specific trajectory of concern runs as follows. Right now, in 2026, the threat is Mythos-class AI finding classical vulnerabilities faster than human teams can patch them. That threat is real and present. Over the next three to five years, as quantum hardware approaches the Oratomic threshold, the harvest-now-decrypt-later archives accumulated by sophisticated adversaries over the past decade will begin to become readable. State secrets, financial records, identity credentials, and military communications encrypted with classical methods during the 2020s will be retroactively exposed. This is not a future risk. The data is already collected. The only question is the timeline for decryption. By the late 2020s to early 2030s, if fault-tolerant quantum machines arrive on the schedules IBM, IonQ, and Google have published, real-time cryptographic compromise becomes possible — not universal, and not immediate, but achievable by the most capable adversaries against the most valuable targets.

The convergence scenario that security researchers find most concerning is not the one where AI and quantum are developed by the same adversary and deployed together as a coordinated capability. It is the scenario where they develop on separate tracks — AI vulnerability discovery accelerating independently, quantum hardware advancing independently — and simply arrive at operational capability within the same window, creating a compound threat that neither defense architecture was designed to address alone.

2.4 Scenarios and Likely Outcomes

A clear-eyed assessment of the evidence suggests four possible trajectories, in descending order of desirability.

Scenario A: Defenders complete PQC migration before Q-Day. Organizations that began cryptographic inventory and migration in 2025–2026 — following the NIST

FIPS standards — successfully deploy quantum-resistant cryptographic infrastructure before fault-tolerant quantum computers reach the Oratomic threshold. The HNDL archives collected by adversaries over the past decade are of limited value because the encryption protecting the most sensitive long-lived data has been rotated. Classical software vulnerabilities continue to be the primary attack surface, and Glasswing-class defensive AI keeps pace with offensive capability. This scenario requires urgency, coordination, and resources that have historically been rare in security program management. *It is possible. It requires starting now.*

Scenario B: Partial migration creates a two-tier security world. Large organizations, financial institutions, and government agencies complete PQC migration on schedule. The long tail — hospitals, municipal governments, small enterprises, industrial control systems, IoT infrastructure, and the developing world's digital infrastructure — remains on classical cryptographic assumptions when Q-Day arrives. The result is a permanent security stratification: well-resourced entities operate in a post-quantum-safe environment, while everything else remains vulnerable to adversaries with quantum capability. The HNDL archives become partially readable, exposing data from organizations that did not migrate. This scenario is the *most likely baseline* based on current migration velocity. Industry experts estimate full enterprise-level PQC migration requires a decade or more, and migration efforts that have not begun in 2026 will not complete before 2036 even under optimistic assumptions.

Scenario C: Adversary achieves cryptographic relevance before migration is complete. A nation-state actor — most plausibly China, given its investment levels and independent standards track, or a state that acquires capability through theft — reaches fault-tolerant quantum computing and begins systematic decryption of HNDL archives before the majority of global infrastructure has migrated. The retrospective exposure of a decade of encrypted communications is not immediately visible. There is no alarm that sounds when archived ciphertext is decrypted. The intelligence advantage accrues silently. This scenario, combined with AI-augmented

classical vulnerability discovery, gives a capable adversary the ability to identify exploitable software flaws *and* forge authentication credentials *and* access a decade of previously unreadable communications, all without triggering any conventional detection mechanism.

Scenario D: AI-quantum convergence enables autonomous attack at machine speed. The compound scenario. AI-augmented vulnerability discovery, quantum-accelerated attack execution, and autonomous agent coordination combine into an offensive capability that operates faster than human defenders can observe, diagnose, and respond. This is the scenario described in the 2026 cybersecurity community's most concerned projections: systems where "autonomous cyber-weapons adapt faster than defenders can patch, or decision systems escalate conflicts beyond human intervention because mission parameters override safety constraints." This scenario is further out on the timeline than the others — it requires both mature fault-tolerant quantum computing *and* significantly more capable autonomous AI than currently exists — but it is not science fiction. The components are in development at well-funded organizations in multiple countries, and the trajectory of each is steepening, not flattening.

2.5 What This Means for the Window of Action

The 2026 Annual Threat Assessment's framing is correct and should be adopted as the standard for all security planning: quantum computing represents *both* a cryptographic threat *and* a strategic intelligence processing advantage for early developers. Organizations and governments that treat PQC migration as a compliance exercise disconnected from strategic competition are misreading the threat. And organizations that treat AI-augmented vulnerability discovery as a near-term technical problem disconnected from the longer-horizon quantum risk are building defenses for only half the threat surface they will face.

The practical implication is simple and urgent: **the window for completing PQC mi-**

gration before Q-Day is measured in years, not decades, and the most credible estimates now put it at three to five years. Full migration of complex enterprise infrastructure — including the discovery and remediation of all cryptographic dependencies, across all vendors, across all legacy systems — takes ten or more years even when urgently prioritized. The arithmetic is unfavorable. Organizations that have not begun a cryptographic inventory as of 2026 will not complete migration before the most aggressive Q-Day estimates. Some of them will not complete it before the conservative ones.

Project Glasswing is the right answer to the right question for 2026. But the question it answers — how do we find and fix classical software vulnerabilities faster than adversaries find and exploit them — is not the only question that determines the security posture of global infrastructure in the decade ahead. The full question is: how do we close the classical vulnerability surface, migrate our cryptographic infrastructure to quantum-resistant standards, develop governance frameworks for open source AI capabilities, and maintain a decisive lead in both AI and quantum hardware development, simultaneously, before the convergence of these capabilities creates a threat environment that no single defensive architecture was designed to address?

That is the mission. Glasswing is the first move. The game is longer than the announcement.

3 The So What

THE BOTTOM LINE

For national security: This is the most significant public disclosure about AI cyber capabilities since the field began. Anthropic is telling the U.S. government—explicitly, in the document itself, noting active discussions with U.S. officials—that the window for maintaining a decisive advantage in this domain is closing. The U.S. either acts

now to institutionalize AI-augmented cyber defense at the federal level, or it cedes the asymmetric advantage to adversaries who will acquire equivalent capabilities through theft, indigenous development, or both. This is not a future scenario. Mythos Preview exists today. The adversaries know the trajectory.

For enterprise: If your cybersecurity posture was designed for a world where sophisticated vulnerability discovery required elite human labor, it is already obsolete. The threat model has changed structurally. The question is not whether to upgrade; it is how fast. The organizations that have deployed AI-augmented security tooling at the infrastructure level—not as an add-on, but as the primary scanning layer—will have a durable advantage. Those that have not will be operating blind against attackers who are not.

For open source: The Linux Foundation said it directly: open source maintainers whose code underpins most of the world's critical infrastructure have historically been left to figure out security on their own. Project Glasswing is a direct intervention in that failure. Open source is not a peripheral concern—it is the substrate on which everything else runs, including the AI agents now writing new software. Vulnerabilities in open source libraries propagate everywhere.

For digital monetary infrastructure and the GENIUS Act: Any stablecoin platform seeking to operate under the GENIUS Act framework across a realistic ten-year horizon is making an architectural decision today—build on classical cryptographic foundations and migrate later, or build on post-quantum standards from the start. Project Glasswing makes the correct answer obvious. Harvest-now-decrypt-later attacks are not a future risk; they are an ongoing collection operation. Every classical cryptographic transaction in production today is potentially being archived for future decryption. Platforms that encode NIST post-quantum standards (FIPS 203/204/205) at the protocol layer from inception inherit a security guarantee that survives both the current AI-augmented threat environment and the quantum transition. Those that do not are accepting a mandatory, high-stakes migration on an adversary's timeline. Build

it right the first time.

For the AI industry: The architecture of Project Glasswing—controlled access, defensive-first deployment, staged capability release, and public transparency on findings—is a model for handling dual-use capabilities responsibly. It is imperfect and involves real risk. But it is substantially more thoughtful than either extreme: full suppression (which denies defenders the advantage) or immediate open release (which front-runs defenses and hands attackers a gift). Watch whether other frontier labs follow this pattern or fragment into competing approaches that undermine collective defense.

The bottom line: A threshold was crossed. Not a theoretical, future threshold—a real one, documented, with specific vulnerabilities named and patched. AI models can now systematically dismantle the security assumptions underlying the world's most critical software. The defenders have access to the same capability, for a limited time, with a head start. What they do with that window will determine the baseline security posture of global digital infrastructure for the next decade.

The glasswing butterfly hides in plain sight. So did the vulnerabilities—for twenty-seven years in some cases.

Now they do not. Act accordingly.

Dr. Gregory S. Carmichael is CEO of Quantum Reserve Capital and founder of Advanced Nano-Materials Manufacturing LLC, operating from San Juan, Puerto Rico under Act 60. He writes at the intersection of macroeconomics, advanced technology, and sovereign monetary systems.

CryptoSoWhat publishes at the intersection of honest money, emerging technology, and the

systems that actually move the world.