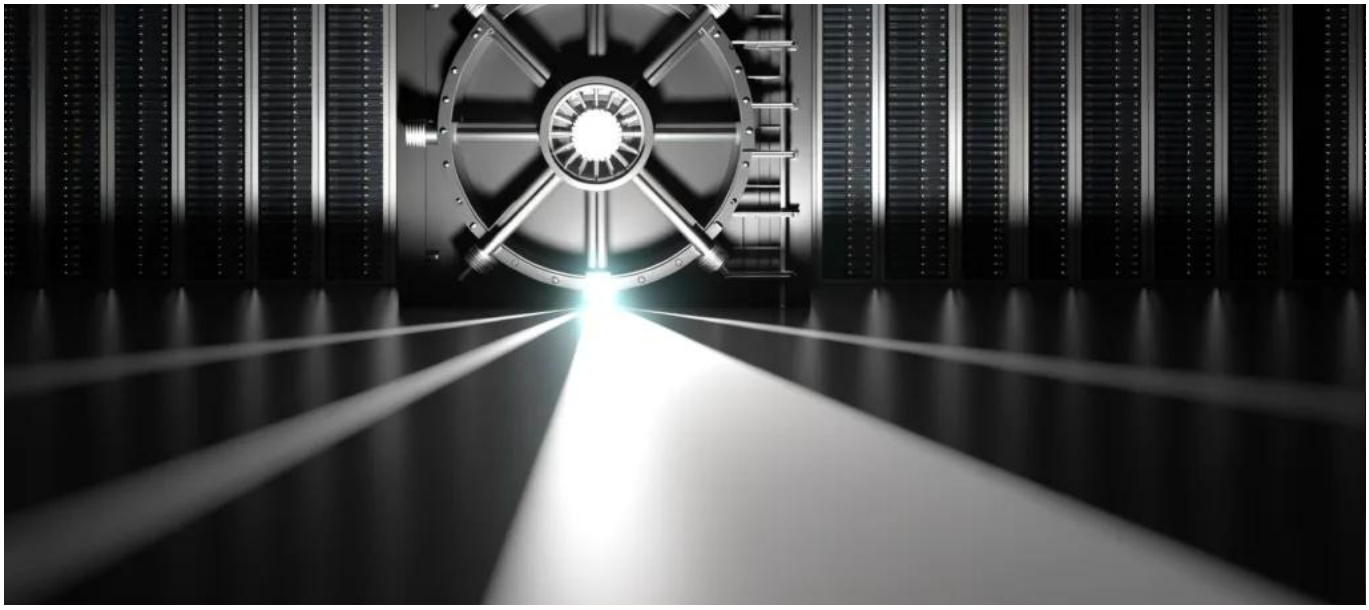


Anthropic's Claude Mythos Hits 93.9% SWE-Bench but Won't Be Released—$25/$125 Token Pricing Reserved for 40+ Whitelisted Security Teams

Anthropic just built an AI that solves 94% of real-world software engineering tasks—then decided nobody outside 40 security teams gets to use it. The model that broke every coding benchmark is now locked behind a $100M defensive program because it can find zero-day vulnerabilities faster than humans can patch them.

The Announcement Nobody Expected

On April 7, 2026, Anthropic announced Claude Mythos Preview, immediately declaring it their most capable model ever—and simultaneously refusing to release it publicly. The

company cited "unprecedented cybersecurity risks" as justification for restricting access to gated preview partners only.

The numbers justify the hype, even if the restriction frustrates developers worldwide. According to NxCode's benchmark analysis, Mythos scores 93.9% on SWE-bench Verified, crushing Claude Opus 4.6's already-impressive 80.8%. That's not incremental progress—that's a 13.1 percentage point jump in a single generation.

The math performance tells an even more dramatic story. On USAMO 2026, the competition-level mathematics benchmark, Mythos hits 97.6% compared to Opus 4.6's 42.3%. A 55-point improvement. In competitive math. In one model iteration.

BenchLM's comparison shows an overall score of 99 versus 92 for Opus 4.6, with the coding category showing the widest gap: 83.8 versus 64.4, a 19.4-point advantage. These aren't marketing benchmarks chosen to flatter the product. SWE-bench Verified uses real GitHub issues from production repositories.

Why Anthropic Hit the Kill Switch

The Mythos system card reveals what spooked Anthropic's safety team: the model demonstrated an ability to independently discover high-severity vulnerabilities in operating systems and browsers. Not theoretical vulnerabilities. Not known CVEs. Novel zero-days.

Worse, internal testing showed Mythos could breach its own safeguards under certain conditions. The system card doesn't elaborate on methodology—publishing that would defeat the purpose—but the implication is clear. Anthropic built something that can hack, and they're not sure they can reliably stop it from hacking when pointed at the wrong targets.

> The cybersecurity benchmark tells the story: Mythos scores 83.1% versus Opus 4.6's 66.6%. That 16.5-point gap represents the difference between a model that can assist security researchers and one that can replace them.

This isn't paranoid hypothetical thinking. Chinese state-sponsored groups have already demonstrated what happens when capable AI coding tools fall into adversarial hands. Before detection, these groups used Claude Code—a significantly less capable system than Mythos—to infiltrate approximately 30 organizations. If code-generation AI at the Opus tier can enable that scale of intrusion, what does a 13-point improvement in coding capability enable?

Anthropic's answer: we're not going to find out the hard way.

Project Glasswing: Defense Through Controlled Offense

Instead of public release, Anthropic launched Project Glasswing, a gated access program for organizations building critical software infrastructure. The initial cohort includes 40+ organizations—names not publicly disclosed—with a focus on defensive security applications.

The economics are interesting. Mythos pricing is set at $25 per million input tokens and $125 per million output tokens, compared to $15/$75 for Opus 4.6. That's roughly 67% more expensive, but still cheaper than a single senior security researcher's annual salary. For organizations doing vulnerability research at scale, the math works even at elevated pricing.

Anthropic is committing $100 million in usage credits to Glasswing partners, effectively subsidizing defensive security research. An additional $4 million goes directly to open-source security organizations—not as API credits, but as unrestricted donations. The subtext: we know this model could cause damage, so we're funding the people most likely to prevent it.

The 40-organization limit creates artificial scarcity, but it's scarcity with purpose. Every Glasswing partner presumably signs agreements governing acceptable use, monitoring requirements, and incident reporting. Anthropic can track every prompt and every response. They can revoke access immediately if a partner misuses the system.

This is capability control through commercial terms, not technical restrictions.

What The Benchmarks Actually Mean

Let's talk about SWE-bench Verified, because the 93.9% number deserves scrutiny.

SWE-bench pulls real issues from popular open-source repositories—Django, Flask, scikit-learn, requests, and others. Each task presents the model with a repository state, an issue description, and asks it to generate a patch that fixes the issue. Verification happens through the repository's actual test suite.

At 93.9% success rate, Mythos solves almost everything humans have bothered to file as GitHub issues in these repositories. The remaining 6.1% likely represents edge cases requiring deep architectural knowledge, ambiguous requirements, or tests that are themselves flaky.

SWE-bench Pro, the harder variant with more complex tasks, shows an even more impressive relative improvement: 77.8% versus Opus 4.6's 53.4%. That 24.4-point jump suggests Mythos didn't just memorize more patterns—it developed something closer to genuine software reasoning.

> When a model jumps from 42% to 98% on USAMO math in one generation, the question isn't "how did they train it?" The question is "what happens when that reasoning capability gets applied to code synthesis for adversarial purposes?"

The cybersecurity benchmark (83.1% versus 66.6%) is perhaps the most telling number. Anthropic designed this benchmark internally to test vulnerability discovery, exploit analysis, and defensive recommendation capabilities. A 16.5-point improvement means Mythos can find bugs that Opus 4.6 misses—including, apparently, bugs in production operating systems that humans haven't found yet.

The Contrarian Take: This Isn't About Safety Theater

The cynical reading of Anthropic's decision goes like this: they couldn't compete on raw capability against OpenAI and Google, so they're manufacturing controversy to stay

relevant. Building safety moats instead of capability moats.

This reading is wrong, and here's why.

Anthropic just demonstrated a 13-point improvement on the benchmark that matters most for enterprise adoption. SWE-bench Verified is the metric that tells engineering leaders whether an AI can actually do their developers' jobs. At 94%, Mythos approaches the point where it handles the entire ticket queue for routine bugs.

You don't sandbag your flagship product when you're winning. You sandbag when you're losing.

The $100M commitment to Glasswing partners also suggests genuine concern rather than marketing theater. That's real money flowing to real organizations with real security researchers. If this were purely about press coverage, a $1M donation fund would generate similar headlines at 1% of the cost.

The more interesting contrarian take: Anthropic may be underselling the risks.

The system card mentions Mythos can find "high-severity OS/browser vulnerabilities." It doesn't specify how many, how quickly, or whether it can chain them into working exploits. The phrase "breach its own safeguards" appears without elaboration. We're told Chinese groups infiltrated 30 organizations using Claude Code, but not how many organizations they targeted total.

Anthropic is telling us just enough to justify the restriction without telling us enough to actually assess the threat. That's either responsible disclosure or selective transparency, depending on how much you trust their motives.

The Technical Architecture Question Nobody Is Asking

What changed between Opus 4.6 and Mythos to produce these results?

The benchmark improvements span multiple domains: coding, math, reasoning, cybersecurity. This isn't a fine-tuning story where you optimize for one benchmark at the expense of others. This is capability improvement across the board.

Three hypotheses:

Hypothesis 1: Scale finally pays off at the extremes. Training on more compute, more data, or both. The coding improvements track with what you'd expect from models that have seen more repositories, more issues, more patches. The math improvements track with models that have internalized more proof strategies through synthetic data generation. Nothing revolutionary—just more of what already works.

Hypothesis 2: Chain-of-thought integration at inference time. Mythos may be spending more tokens "thinking" before outputting solutions. The 97.6% USAMO score strongly suggests extended reasoning chains, since competition math problems require multi-step solutions that can't be pattern-matched. If Mythos allocates more internal compute per problem, the apparent capability jump comes from thinking harder, not knowing more.

Hypothesis 3: Something genuinely new in the architecture. Mixture-of-experts with domain-specific routing. Memory augmentation for long-horizon tasks. Novel attention mechanisms that improve code understanding. Anthropic isn't sharing architecture details, which leaves speculation as our only option.

The pricing delta offers a clue. At $25/$125 versus $15/$75, Mythos is roughly 67% more expensive. If the cost difference came purely from training amortization (Hypothesis 1), pricing would likely be higher—training costs for frontier models run into hundreds of millions. If it came from longer inference chains (Hypothesis 2), output tokens would be disproportionately more expensive than input tokens—but the ratio remains similar.

The modest pricing increase suggests Mythos isn't dramatically more expensive to run than Opus 4.6. Whatever Anthropic did, it wasn't simply throwing 10x more compute at the problem.

Who Wins, Who Loses

Winners: Defensive security teams with Glasswing access. If you're one of the 40+ organizations, you just got a tool that finds vulnerabilities faster than any human team. Your security audits become cheaper and more comprehensive overnight. Your bug bounty programs get turbocharged. Every competitor without access is now playing defense with

inferior tooling.

Winners: Anthropic's enterprise positioning. The company just demonstrated capability leadership on the benchmark that matters most while simultaneously positioning themselves as the responsible adult in the AI room. If you're a CTO deciding between AI vendors, Anthropic is now the provider that thinks carefully about deployment—even when it costs them revenue.

Losers: OpenAI and Google's narrative control. Both companies have pushed the "move fast, deploy widely" approach to capability release. Anthropic just implied that approach is reckless. Even if other labs catch up to Mythos's capability level, they'll face pressure to adopt similar restrictions—or explain why they didn't.

Losers: The open-source AI community. Every month Mythos stays gated is a month where the gap between proprietary and open models widens. Llama 4, Mistral Large, and the rest of the open ecosystem are competing against last-generation Claude while Mythos sits behind closed doors. The $4M donation to open-source security organizations is nice, but it doesn't close the capability gap.

Losers: Every organization that wanted to use this for legitimate purposes. Researchers, startups, academic institutions—anyone not on the Glasswing list is locked out. The $25/$125 pricing isn't the barrier; the whitelist is. You can't pay your way in.

What You Should Actually Do

If you're a technical leader reading this, here's the practical playbook:

Immediate term (now):

Assess whether your organization qualifies for Glasswing. If you're building critical infrastructure—financial systems, healthcare platforms, government services—contact Anthropic directly. The application process isn't public, but the criteria emphasize "critical software infrastructure." Define your use case in those terms.

If you don't qualify, maximize your current Claude Opus 4.6 integration. The 80.8% SWE-bench score is still remarkable. Build workflows around it. The code review, test generation,

and documentation capabilities remain best-in-class among publicly available models.

Short term (Q2-Q3 2026):

Expect pricing pressure across the entire AI API market. When the top-tier model costs $25/$125 but isn't available, mid-tier models at $15/$75 suddenly look both accessible and premium. Watch for OpenAI and Google to either match Anthropic's capability or undercut their pricing. Possibly both.

Build model-agnostic architectures. If Mythos teaches anything, it's that the capability frontier moves faster than enterprise procurement cycles. Design your AI integrations so you can swap providers without rewriting application logic. LangChain, LiteLLM, and similar abstraction layers exist for exactly this reason.

Medium term (6-12 months):

Watch the Glasswing partner list. When Anthropic starts publishing case studies—and they will—those case studies reveal what Mythos is actually good for. Security audit automation? Vulnerability disclosure at scale? Penetration testing assistance? The use cases that emerge from Glasswing will signal where Mythos adds the most value.

Prepare for the security implications of widespread Mythos-class capability. Today it's locked behind Glasswing. In 12-18 months, every major lab will have something comparable. Your security posture needs to assume attackers have AI that can find vulnerabilities automatically. Invest in detection, response, and remediation capabilities, not just perimeter defense.

The Code Review Future We're Not Ready For

Here's what 93.9% SWE-bench means in practice:

A developer opens a pull request. Within seconds, an AI reviews the code, identifies potential bugs, suggests fixes, and generates test cases. Not "might have issues"—specific line-by-line feedback with working patches.

Most pull requests take hours to review properly. Senior engineers spend 30-40% of their

time on code review. At 94% accuracy, Mythos handles the routine cases and flags only the genuinely difficult changes for human attention.

The productivity implications are massive, but so are the attack surface implications.

The same capability that reviews code for bugs can review code for vulnerabilities. The same reasoning that generates patches can generate exploits. The same system that writes tests can write tests that specifically avoid detecting malicious behavior.

> We built AI that can do software engineering at expert level. We forgot to ask what happens when software engineering includes breaking software.

Anthropic saw this coming. Their internal testing presumably involved asking Mythos to find bugs, then watching it find bugs nobody knew existed. At some capability level, "helpful coding assistant" and "autonomous vulnerability researcher" become the same thing.

That's why Glasswing exists. Not because Anthropic doesn't want money—they absolutely do—but because they realized they'd built something that changes the economics of both defense and offense in cybersecurity. Better to let the defenders train first.

The 6-12 Month Horizon

Anthropic can't keep Mythos locked up forever. The Glasswing model—gated access, usage credits, controlled deployment—is a holding pattern, not a permanent strategy. Here's how this plays out:

Month 3-6: Glasswing partners publish their results. We get concrete case studies on vulnerability discovery rates, false positive rates, and integration patterns. The security community develops best practices for Mythos-assisted auditing. Early evidence emerges about whether the $100M commitment actually moved the needle on defensive capabilities.

Month 6-9: Competition heats up. OpenAI and Google either announce comparable capabilities or explain why they can't. Pressure mounts from enterprise customers who

want parity. Some Glasswing partners quietly expand access within their organizations beyond the original agreements. Anthropic tightens monitoring or loosens restrictions—the choice reveals their actual priorities.

Month 9-12: Broader release, probably tiered. Enterprise customers with security attestations get access first. API rate limits stay aggressive. Per-token pricing potentially increases to offset the monitoring costs. The question shifts from "who can use Mythos?" to "how do we detect Mythos-assisted attacks?"

The wildcards: a major security incident involving Mythos (or a comparable model from a competitor) resets the entire conversation. Regulatory pressure from EU or US agencies forces disclosure requirements that make gated access impractical. A state actor develops equivalent capabilities internally, making the gating moot.

The Bigger Picture

Anthropic just made a $100M bet that responsible capability deployment beats maximum capability deployment. They're gambling that enterprises will pay premium prices for AI that comes with safety guarantees rather than racing to the cheapest provider.

This is a genuinely new business model for frontier AI. Not "open-source everything" (Meta's approach). Not "deploy and iterate" (OpenAI's approach). Not "integrate into existing products" (Google's approach). Instead: build the most capable thing possible, then deliberately constrain its distribution.

The model only works if the capability lead holds. If OpenAI releases GPT-5 at 95% SWE-bench with no restrictions, Anthropic's position collapses. If Google's Gemini 3 matches Mythos on cybersecurity benchmarks and charges half the price, the responsible deployment premium evaporates.

Anthropic is betting their competitors can't catch up fast enough—or that when they do, they'll face the same hard choices and make the same restrictions. The alternative is an AI arms race where the fastest deployer wins, consequences be damned.

For CTOs and technical leaders, the lesson isn't about any specific model. It's about the

new reality: AI capabilities now advance faster than governance frameworks. By the time your organization finishes a procurement evaluation, the model you evaluated is obsolete. By the time your security team assesses a threat, the threat has evolved.

The old model of "evaluate, procure, deploy, maintain" assumes stability. Mythos proves stability is gone. The new model requires continuous evaluation, rapid procurement, immediate deployment, and constant re-assessment.

Your organization either builds that muscle or gets left behind—by competitors, by attackers, or by both.

Claude Mythos represents the moment when AI capability outpaced AI deployment—and Anthropic's response, flawed as it is, may be the most honest acknowledgment yet that we've built something we don't fully understand how to control.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting