Mythos AI Cracked Apple's Best Defense in 5 Days

May 15, 2026

Download this guide as PDF

> **Quick Answer:** On May 14, the cybersecurity firm Calif disclosed that Anthropic's Mythos Preview helped them find a data-only kernel local privilege escalation chain on macOS 26.4.1 / Apple M5 — bypassing Memory Integrity Enforcement (MIE), Apple's flagship security architecture that took 5 years and billions to build. Calif's work took 5 days. The 55-page report was hand-delivered to Apple Park, laser-printed. Same week: Google's Threat Intelligence Group reported the first AI-generated zero-day in the wild and added official corroboration to the OpenClaw skill malware problem InsiderLLM has tracked since February. The offense-defense timeline is compressing. Local AI users retain a structural privacy advantage but inherit skill-plugin supply-chain risk that requires active audit.

📚 **More on this topic:** OpenClaw ClawHub Security Alert · OpenClaw Security Guide · OpenClaw Security — February 2026 · OpenClaw Plugins & Skills Guide

Mythos cracks Apple M5 in 5 days

The cybersecurity firm Calif published a writeup on May 14 documenting something that, if it holds up, marks a real shift in the offense-defense balance. Working with early access to Anthropic's Mythos Preview, Calif's team found a data-only kernel local privilege escalation chain on macOS 26.4.1 running on Apple M5 hardware. The exploit chain bypasses Memory Integrity Enforcement (MIE) — Apple's flagship kernel security architecture, the result of a five-year engineering investment that Apple itself describes in roughly billion-dollar terms.

Calif's team did it in five days.

The numbers are the headline. Five years of Apple kernel engineering, plus the supporting silicon work on M5, compressed by an AI-assisted research team into roughly a workweek. The disclosure pattern shifted to match the gravity: the 55-page report was laser-printed and hand-delivered to Apple Park, not emailed. The implication is that an emailed kernel-bypass writeup on a chain that targets MIE is too sensitive to put through anything routed, cached, or scanned.

What MIE is supposed to do is enforce kernel integrity at the hardware level — pointer authentication, control-flow integrity, kernel memory protections that traditional userland-rooted attackers can't subvert. A "data-only" exploit, as Calif describes the chain, avoids overwriting

code or pointers; it works by manipulating data that the kernel trusts in ways that walk past MIE's checks. This is the kind of attack class security architects spend years anticipating and patching.

One nuance worth surfacing from Calif's own writeup before the takes pile up: "Mythos Preview is powerful: once it has learned how to attack a class of problems, it generalizes to nearly any problem in that class. Mythos discovered the bugs quickly because they belong to known bug classes." That's the caveat that should travel with this story. Mythos didn't invent a new bug class. It accelerated the work inside an existing one. Whether that distinction holds at the next inflection point is the question worth tracking.

Apple's response timeline matters too. macOS Tahoe 26.5 shipped this week, and Apple's security release notes credit Calif and Anthropic Research in fixes for separate bugs — not necessarily the MIE-bypass chain Calif disclosed. Apple typically coordinates disclosure of kernel-level bypasses across multiple release cycles. So the 26.5 credits don't necessarily indicate the Mythos-found chain is patched. We'll know more when Calif publishes CVE numbers or when Apple publishes the corresponding security content notes.

For local-AI users, the immediate question isn't "is my Mac safe right now" — Apple has the report, the kernel team is working it, and you weren't going to be the first target of a research-grade kernel chain anyway. The immediate question is what the timeline compression means for the wider security posture of every system, including the local-AI tooling we covered in our OpenClaw security articles.

That brings us to the other security story that landed the same week.

GTIG's parallel evidence, read with appropriate skepticism

On May 11, three days before Calif's disclosure, Google's Threat Intelligence Group published its quarterly AI Threat Tracker. The report documents three findings worth knowing.

First, GTIG observed what it describes as the first confirmed AI-generated zero-day in the wild: "a zero-day vulnerability implemented in a Python script that enables the user to bypass two-factor authentication (2FA) on a popular open-source, web-based system administration tool." GTIG explicitly notes that they do not believe Gemini was used; the model is not named. Their evidence for AI generation: a hallucinated CVSS score in the Python script, "educational docstrings," and what they call "a structured, textbook Pythonic format highly characteristic of LLMs training data."

Second, GTIG documented threat-actor experimentation with a Claude Code skill plug-in (`wooyun-legacy`) that distills over 85,000 real-world vulnerability cases from the Chinese WooYun bug bounty platform (2010-2016). The skill plug-in primes the model with structured vulnerability data to bias its code analysis toward seasoned-expert pattern matching. It's a methodology finding, not a tool weaponization.

Third — and most relevant for the InsiderLLM audience — GTIG named OpenClaw and OneClaw directly. We'll come back to that in the next section.

A reader should take this report seriously while also reading it for what it is. GTIG is part of Google. Their threat-intelligence output reflects real observations, but the editorial framing serves Google's positioning in the AI-security market. The 2FA zero-day attribution to "an AI model" without naming the model is convenient if you ship Gemini; the timing alongside Mythos's higher-profile capability demonstration is also convenient. The findings are corroborating evidence that criminals are using AI offensively. They are not the canonical authority on the broader story.

GTIG also references threat-actor groups APT45 and UNC2814 by designator. The report describes their methodology — APT45's recursive prompt-driven CVE analysis, UNC2814's persona-prompt jailbreaking — without further geographic or organizational attribution in the sections covered here.

OpenClaw skill weaponization: not new, now corroborated

InsiderLLM has covered the OpenClaw skill ecosystem security problem in depth since February 2026. The broader security research community has documented:

- **Koi Security** audited all 2,857 ClawHub skills in February 2026: 341 malicious skills found — roughly 12% of the registry.
- **The ClawHavoc campaign** tracked by The Hacker News, 6mile, and OpenSourceMalware: data-stealing malware delivered via fake cryptocurrency trading skills.
- **Snyk's ToxicSkills audit** (3,984 skills scanned): 13.4% contain critical security issues, 36.82% have at least one flaw.
- **Zscaler ThreatLabz** (March 2026): the DeepSeek-Claw skill campaign with malicious `SKILL.md` execution paths.
- **Trend Micro**: Atomic macOS Stealer (AMOS) delivered via weaponized OpenClaw skills.
- **1Password, Repello AI, eSecurity Planet**: independent teardowns of the attack patterns.

GTIG's May 11 report adds Google's threat intel weight to findings the security research community has documented all year. GTIG's actual language is more cautious than the February disclosures — they describe threat actors "experimenting with agentic tools such as OpenClaw and OneClaw alongside intentionally vulnerable testing environments." That's noteworthy as official corroboration, but it's not the discovery moment. The discovery moment was February.

InsiderLLM's existing security coverage:

- Original February 2026 ClawHub compromise coverage — the canonical writeup of the Koi Security audit and the ClawHavoc campaign as it broke.
- Monthly security report — February 2026 — the systematic monthly tracking.
- OpenClaw Security Guide — defensive practices.
- Plugins & Skills Guide — install-and-audit pattern.

The takeaway for InsiderLLM readers: nothing about your audit posture should change because of the GTIG report. The audit practices documented in our existing coverage still apply. What's changed is that Google's threat intel team has now officially recorded what was already public knowledge in the security research community.

What it means for local AI users

The Calif/Mythos story and the GTIG OpenClaw finding land in different places on the threat map, but they share the same underlying observation: AI capability is improving fast enough on both offense and defense that the time you have to react is shrinking.

For local-AI users specifically, three implications are worth absorbing.

The structural privacy advantage of local AI is intact. Running Qwen 3.6 or Gemma 4 on your own RTX 3090 means your prompts, your data, and your reasoning traces do not pass through anyone else's infrastructure. That advantage matters more, not less, in a world where Mythos-class capability is in deployment with a handful of large organizations. You retain control of what you ask AI to do and what it sees.

The skill-plugin supply chain is the open flank. The GTIG observation about malicious OpenClaw skills is the canary. Any system that loads third-party code at runtime — and most agentic harnesses do — inherits the supply-chain risk that has plagued package ecosystems for decades. Local AI gives you more control over where that code runs; it doesn't make the code safer.

The offense-defense balance has compressed. Calif's five-day MIE bypass is one anecdote, with the "known bug classes" caveat attached. But it's the leading edge of a trend that's already visible in less dramatic forms: faster patch cycles, faster discovery cycles, faster weaponization. The implication for local AI is not that you need to retreat to non-local tooling. It's that the same compression that benefits attackers also benefits the defenders who use AI-assisted security tooling — and that audit posture matters more than it did a year ago.

The next section is the practical audit list.

What to audit in your OpenClaw setup

What to do this week, if you run OpenClaw or any skill-plugin-driven agent harness.

Verify skill sources before install. Only install skills from maintainers you can trace — named GitHub accounts with history, organizations with publication track records, or your own internal builds. Avoid one-off uploads to ClawHub with no maintainer reputation. Check the repo's age, commit history, and issue activity. A skill repo that appeared three weeks ago with one commit and one downloads spike is not where you want your agent loading code from.

Read the skill manifest before install. Most OpenClaw skills declare their entry points, permissions, and external dependencies. Open the manifest, look at what the skill claims to do, and check that the claims match the implementation. A skill that asks for network access and filesystem write permission to "format your output as JSON" is misaligned.

Sandbox skill execution. If your threat model permits it, run OpenClaw inside a container, a chroot, or a dedicated user account with restricted filesystem and network permissions. The pattern the security research community documented all year is hidden routines that execute unauthorized code at host level. A sandbox doesn't prevent the routine from running, but it limits the blast radius. The OpenClaw Security Guide walks through containerization options.

Monitor outbound network calls from skill execution. If a skill makes unexpected network calls during a run — DNS lookups to domains you don't recognize, connections to non-vendor endpoints — that's the signal worth watching. A skill that does what it says on the tin generally doesn't need to phone home. Set a baseline for normal behavior and watch for drift.

Inspect installed skills periodically. The supply-chain risk isn't only at install time. A skill that was clean when you installed it can be compromised in a subsequent update if the maintainer's account is taken over or the registry is breached. Periodic re-inspection of installed skill code is overhead, but it's the only way to catch post-install compromise.

Cross-reference InsiderLLM's existing OpenClaw security coverage. The ClawHub Security Alert covers the broader skill-marketplace risks. The February 2026 security report catalogs the supply-chain attacks that landed earlier this year. The Plugins & Skills Guide gives the canonical install-and-audit pattern. The GTIG observation is corroborating evidence for what those articles already documented — but the corroboration matters, because it confirms the threat from a major third-party intel team.

Honest reality check

What we don't know matters as much as what we do.

We don't know exactly which bugs Calif found. Apple coordinates kernel-level disclosure carefully, and the macOS Tahoe 26.5 credits to Calif and Anthropic Research are for separate fixes that may or may not overlap with the Mythos-found MIE bypass chain. The 26.5 patch notes are publicly readable; matching credits to specific bugs requires the CVE database to update, which typically lags release by weeks. Until then, treat the "fixed in 26.5" claim as unverified.

We don't know which AI model produced GTIG's 2FA zero-day. GTIG ruled out Gemini and described AI fingerprints — hallucinated CVSS score, textbook docstring patterns — but did not name the model. That's a structurally interesting silence. The model could be a frontier model that GTIG is reluctant to name for legal or commercial reasons. It could also be an open-weights model fine-tuned by the threat actor.

The floor numbers are documented — Koi's 341 ClawHub skills and Snyk's ~534 from the ToxicSkills audit are audit results, not exhaustive counts of what's been in active distribution since February.

The Barracuda Mythos Hype Index sits at 94 out of 100 this week, per their public dashboard. Hype indices measure attention, not impact. The Calif disclosure is a concrete, dated, named-organization finding with hand-delivered evidence. The GTIG report is also concrete, dated, and named. The hype is high because there is a real underlying signal. It's the signal that's worth tracking, not the hype number.

One quote that's circulating in coverage of this week's stories isn't sourced cleanly enough to use here. We'll add it back if a primary source surfaces.

The pushback against panic is straightforward. Apple has the report. Calif's caveat about "known bug classes" matters. OpenClaw users can audit today. None of these stories require immediate emergency action. They require sustained attention.

What InsiderLLM is watching

Five things we're tracking from here.

The next GTIG AI Threat Tracker edition. Quarterly cadence means the next report lands roughly mid-August. If malicious-OpenClaw-skill distribution scales between now and then, that's the place we'll see it documented.

macOS Tahoe 26.5 patch verification. Once CVE numbers surface and Calif publishes which bugs Apple credited to whom, we'll know whether the MIE bypass chain is patched or still under coordinated disclosure.

Mythos public release timing. The Myriad prediction market currently puts a public Mythos launch by June 30, 2026 at 10.5% probability. If that probability moves materially in either direction, the broader threat landscape shifts with it.

OpenClaw skill marketplace responses. ClawHub and adjacent marketplaces have varying audit and signature practices. The GTIG finding will pressure them to improve. We'll cover material changes as they ship.

Mozilla's Firefox vulnerability work. Mozilla reported that Mythos identified 271 vulnerabilities in Firefox during internal testing. Mozilla has more credibility on these claims than vendor marketing does. If they publish a detailed methodology, that's the case study worth reading.

The compression of offense-defense timelines is the story underneath all of these. We'll keep tracking it.

Get notified when we publish new guides.

Subscribe — free, no spam