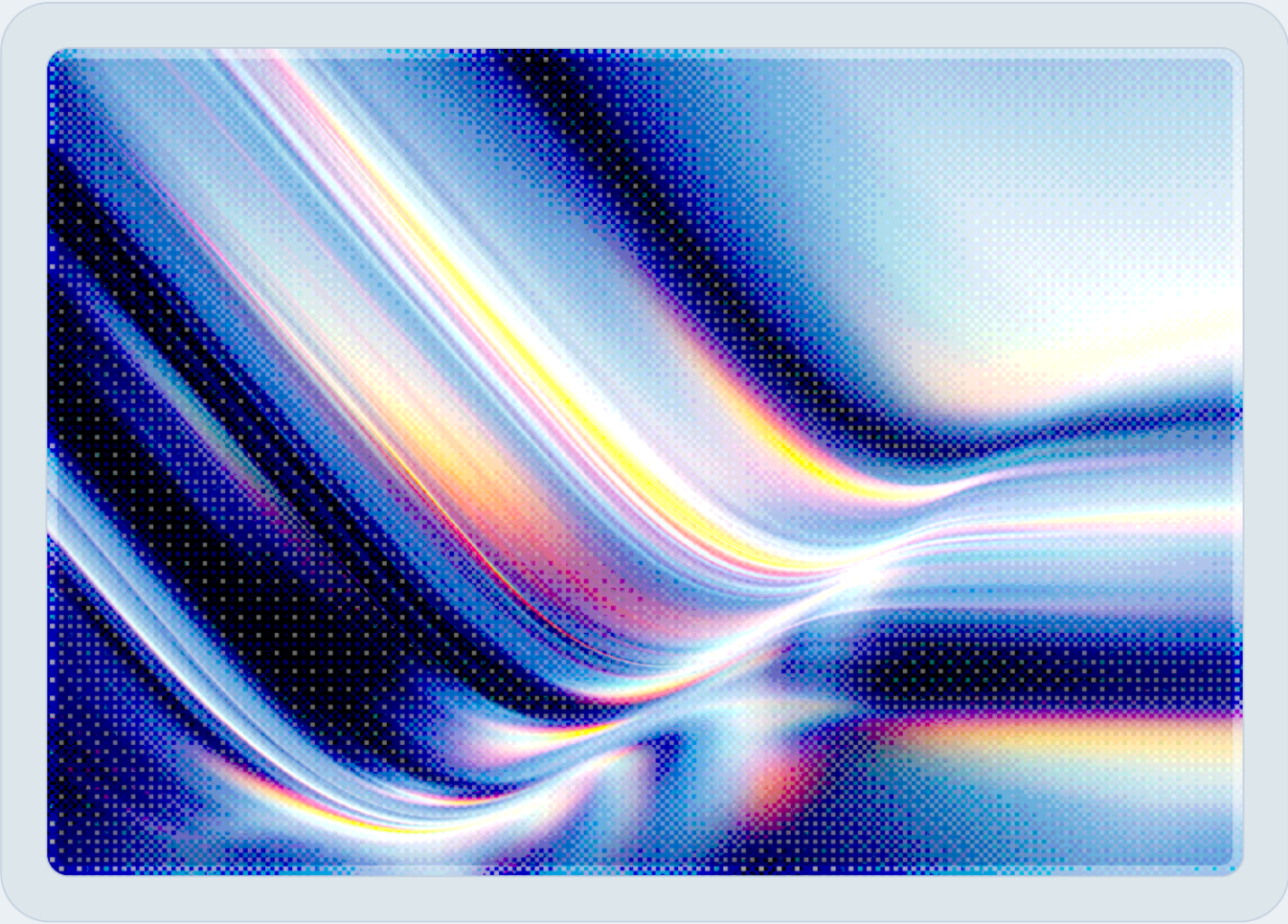
CSAI Foundation | Cloud Security Alliance

AI Chat Trust Weaponized in Mac Malvertising Campaign

MacSync Infostealer Abuses Claude.ai Shared Chats and Google Ads

2026-05-12

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Threat actors are abusing Claude.ai's shared-chat feature to host fake macOS installation guides attributed to "Apple Support," luring Mac users into pasting terminal commands that deliver the MacSync infostealer. Google Ads in this campaign legitimately display `claude.ai` as the destination domain because the malicious content resides inside Anthropic's own platform, removing one of the most relied-upon user-facing signals that an ad is fraudulent — the destination domain — [1].
- The campaign was first identified on May 10, 2026, by Berk Albayrak, a security engineer at Trendyol Group. BleepingComputer independently confirmed a second, infrastructure-distinct variant of the same attack operating in parallel, indicating coordinated or copycat activity across multiple threat actors [1].
- MacSync is a Malware-as-a-Service (MaaS) infostealer leased to multiple threat actors that harvests eight categories of data — browser credentials and session cookies, macOS Keychain entries, SSH private keys, cloud-provider credential files, cryptocurrency wallet seed phrases and private keys, shell history files, Apple Notes, and files matching configurable patterns — exfiltrating everything to attacker-controlled infrastructure [4][5].
- The payload is delivered as a gzip-compressed shell script executed entirely in memory, with the attacker's server generating a uniquely obfuscated variant on each request — a polymorphic delivery technique that evades file-hash-based endpoint detection and falls entirely outside the scope of Gatekeeper, which inspects installed application bundles rather than in-memory script payloads [1][4].
- This attack continues a documented pattern established in December 2025 when the same technique was applied to ChatGPT and Grok shared chats to deliver the Atomic macOS Stealer, demonstrating that AI platform brand trust has matured into a repeatable social engineering technique that threat actors are actively operationalizing across multiple platforms [2][3].

Background

AI chat platforms now routinely offer a sharing feature that converts any conversation into a persistent, publicly accessible URL. The feature is designed for legitimate collaboration: a user generates an installation walkthrough, debugging session, or tutorial, and shares the link with colleagues or publishes it in documentation. From the user's perspective, a shared chat URL on claude.ai, chatgpt.com, or grok.x.com carries the full visual and reputational weight of that platform — the branding, domain, and implied expertise of a sophisticated AI system.

This design characteristic creates an underappreciated attack surface. The content inside a shared chat is entirely user-controlled, but the URL is hosted on a trusted first-party domain. A threat actor who creates a shared chat with malicious instructions and then promotes that URL through Google Ads receives the benefits of that first-party trust without ever building or maintaining their own infrastructure. Unlike traditional malvertising — where the attacker must register a lookalike domain, obtain a TLS certificate, and hope the URL passes scrutiny — the AI platform variant requires only an account on a free service and an advertising budget — considerably lower infrastructure overhead than traditional malvertising, though the payload development, MaaS rental, and campaign operational security still require meaningful technical skill.

The technique was first documented at scale in December 2025, when researchers at Malwarebytes identified a malvertising campaign using shared ChatGPT and Grok conversations to deliver the Atomic macOS Stealer (AMOS). In that campaign, Google surfaced sponsored results linking to genuine AI platform URLs, and the shared conversations posed as macOS system-maintenance guides that led users through terminal commands harvesting credentials and installing persistence [2]. Huntress characterized the design of that campaign as an attack that "doesn't break trust layers; it weaponizes them all," noting that modern macOS attacks increasingly rely on trusted surfaces — ads, AI chats, notarized applications, and popular web platforms — rather than warning-generating exploits [3].

MacSync, the infostealer at the center of the current campaign, marks a further evolution in the MaaS ecosystem targeting macOS. Jamf Threat Labs documented MacSync's development across three distinct campaign generations since November 2025 [4]. The earliest variant used a fake ChatGPT Atlas browser as lure content delivered through sponsored search results. The December 2025 generation refined the technique, pivoting to genuine AI platform conversations and GitHub-themed installation pages. By February 2026, MacSync operators had adopted fileless, in-memory execution architecture that evades conventional endpoint detection, and the variant from that period was confirmed to have impacted state, local, tribal, and territorial (SLTT) government organizations in the United States [5][6]. The May 2026 campaign documented here extends this trajectory by targeting an additional AI platform and deploying polymorphic payload delivery.

Security Analysis

The Attack Chain

The campaign begins with a Google-sponsored search result triggered by queries such as "Claude mac download." The ad's displayed destination URL is the legitimate `claude.ai` domain — not a lookalike. When a user clicks through, they are directed not to Claude's actual software download page but to a publicly shared Claude.ai conversation created by the attackers. The shared chat presents itself as an official "Claude Code on Mac" installation guide and attributes the instructions to "Apple Support," lending further false authority.

The fabricated guide instructs the user to open macOS Terminal and paste a provided command. The command is encoded in base64, a common obfuscation technique that produces a string with no immediately recognizable structure. When decoded and executed, it reaches out to attacker-controlled infrastructure to retrieve `loader.sh`, a gzip-compressed shell script. The script executes entirely in memory — no files are written to disk — and the second-stage payload runs through `osascript`, macOS's built-in AppleScript execution engine. Using a native system component for code execution bypasses Gatekeeper entirely — which inspects installed application bundles, not in-memory script payloads — while also evading hash-based endpoint controls that scan files written to disk. BleepingComputer's analysis further documented that the attacker's server generates a uniquely obfuscated version of the payload on each request, a polymorphic delivery technique that evades signature-based endpoint controls relying on known file hashes [1].

The two variants BleepingComputer confirmed — one identified by Albayrak and a second discovered independently — ran on entirely separate infrastructure, suggesting that this attack pattern may already be circulating across multiple threat actor groups rather than representing the work of a single actor.

The Trust Bypass: Why This Technique Is Durable

Traditional malvertising guidance instructs users to inspect the URL in a sponsored search result before clicking. When the destination URL is `claude.ai`, that heuristic provides no protection. The malicious content is hosted by Anthropic's own platform, meaning the ad passes the URL inspection test, the TLS certificate is valid and issues from the correct authority, and the page loads on genuine Anthropic infrastructure. No spoofing of the domain or certificate chain occurs at any stage.

The social engineering layer inside the shared chat adds a second trust signal: the chat interface itself. Users who regularly use AI assistants for technical guidance may already operate in a mode where they expect and act on computer instructions, lowering their skepticism threshold. A shared chat that presents as a curated guide from "Apple Support" exploits that context directly. The interface looks identical to a legitimate shared Claude conversation, because it is one — the only difference is in the instructions it contains.

This combination produces an attack where every individual trust signal — the ad's domain, the TLS certificate, the platform's visual design, and the AI brand — validates as genuine. The only signal that something is wrong is the content of the instructions themselves, and many users — particularly those outside security and DevOps roles — may not recognize that legitimate software installations do not require pasting base64-encoded commands from a chat interface into a terminal.

MacSync Capability and the MaaS Ecosystem

MacSync's data collection scope is broad by design, reflecting the economics of a MaaS platform that must offer capability competitive with other infostealers in the underground market. Once the `osascript` payload executes, MacSync targets eight distinct data categories: browser credentials and session cookies across major browsers, macOS Keychain entries containing saved passwords and application secrets, SSH private keys, AWS and other cloud-provider credential files, cryptocurrency wallet seed phrases and private keys, shell history files, Apple Notes contents, and general file collections matching configurable patterns [4][5]. On unmonitored endpoints, the entirety of this harvest is exfiltrated to attacker-controlled servers before the user has any indication that something has occurred; behavioral endpoint controls monitoring `osascript` invocations and outbound data transfers may generate alerts during the process.

MacSync's MaaS architecture means that multiple independent threat actors may be running concurrent campaigns using the same underlying stealer code. The diversification of delivery techniques across campaigns — ClickFix social engineering, shared AI chat abuse, code-signed Swift droppers — reflects individual operators iterating on delivery rather than changes to the core malware [5][9]. The February 2026 variant's confirmation in SLTT government environments by the Center for Internet Security indicates that MacSync operators have targeted environments with dedicated security programs, including at least one category of government infrastructure, beyond the opportunistic consumer targeting more typical of early MaaS campaigns [6].

Separately, Bitdefender documented a related but distinct campaign that used fake Google Ads linking to a Squarespace-hosted page impersonating the Claude Code documentation site rather than a shared claude.ai chat. That campaign targeted both Windows users — delivering a payload via `mshta.exe` —

and macOS users via an obfuscated Mach-O reverse shell backdoor, demonstrating that threat actors are simultaneously pursuing multiple delivery approaches under the Claude Code brand [7].

Platform Feature Risk

The shared-chat feature presents a structural challenge for AI platform providers. Shared conversations are, by design, publicly accessible, indexable, and persistent. They are intended to function as lightweight documentation or tutorials — which is precisely what makes them effective as social engineering props.

AI chat platforms have not historically applied to shared conversation URLs the same content moderation they apply to AI outputs — there is no published evidence that platform providers systematically scan shared conversations for malicious instructions, in contrast to the abuse-detection pipelines that some content platforms apply to user-generated content. Nothing in the current shared-chat interface warns a viewer that the instructions they are reading were composed by an anonymous third party rather than by the AI model. A viewer who follows a shared link encounters the Claude branding and interface without any labeling distinguishing attacker-created content from content that Anthropic has reviewed or validated. This gap — between the platform's implied authority and the reality that shared content is entirely user-generated — is the condition that makes this attack class sustainable.

Recommendations

Immediate Actions

Organizations should issue an advisory to all employees, particularly developers and technical staff who are likely to search for AI developer tools, informing them that legitimate software installations never require pasting commands from a chat interface into Terminal or a command prompt. The advisory should specifically note that ads displaying trusted AI platform domains may still link to attacker-controlled content hosted within those platforms' sharing features.

Security teams should verify that macOS endpoint protection is configured for behavioral and in-memory threat detection, not only file-hash-based scanning. MacSync's fileless delivery architecture is constructed to evade defenses that rely on examining binaries written to disk. Where EDR tools support it, enable monitoring of `osascript` invocations, particularly those triggered by shell scripts rather than interactive users. Known MacSync command-and-control domains should be blocked at the DNS and network perimeter layer; threat intelligence feeds from Jamf Threat Labs [4] and the Center for Internet Security [6] have published relevant indicators.

Short-Term Mitigations

Organizations should establish and communicate clear, policy-backed guidance on approved channels for installing software and developer tools. Approved Claude Code installation, for example, should route through Anthropic's official documentation page with a known, bookmarked URL — not through sponsored search results. For high-privilege users such as developers with production access, browser policies that block or flag sponsored search results for software-download queries provide meaningful friction against this attack vector.

Endpoint coverage on macOS deserves review against the current threat landscape. Microsoft's May 2026 analysis of ClickFix campaigns targeting macOS documents the range of native macOS tools attackers abuse to execute payloads without triggering conventional detection [8]. Security teams should audit whether their current tooling can detect terminal-based execution chains that avoid dropping persistent files.

Security awareness training should be updated to include AI platform social engineering as a specific scenario category. The mental model that "if the URL is right, the content is safe" must be directly challenged with concrete examples. Exercises that walk employees through the mechanics of the shared-chat attack — including what the lure page looks like and what makes it convincing — complement general phishing guidance by building skepticism targeted at this specific vector.

Strategic Considerations

Organizations participating in AI platform provider relationships — as enterprise customers, partners, or working group members — should advocate for shared-content labeling that distinguishes AI-generated content from user-curated content in shared conversations. Providers including Anthropic and OpenAI have existing trust and safety infrastructure; extending it to flag shared conversations that contain terminal commands or scripts may reduce the attack surface at the platform level; the impact on legitimate developer and tutorial use cases — which frequently include terminal commands — would require careful design to avoid significant friction for benign users.

Threat modeling for AI tool deployments should incorporate the AI platform trust abuse vector. CSA's MAESTRO framework, which models threats across the full stack of agentic AI ecosystems, provides a structured vocabulary for reasoning about how trusted platform interfaces can be turned against users [10]. In the MAESTRO framework's terms, this attack most directly engages the Agent Ecosystem layer (L7) — where AI interfaces connect with real-world users and applications, and where social engineering converts a trusted interface into a malware delivery channel — and the Security and Compliance layer (L6), which addresses the trust posture governing platform-to-user relationships; organizations may also find it useful to model the data exfiltration phase against the Data Operations layer (L5). This mapping is

the authors' interpretive application of the framework; practitioners may reasonably map aspects of the campaign to additional layers depending on the scope of analysis. Organizations using MAESTRO to model their AI tool deployments should include shared-URL content abuse as a threat scenario within these layers, alongside prompt injection and model supply chain attacks.

Finally, the recurrence and diversification of this attack pattern — applied to ChatGPT in December 2025, to Claude.ai in May 2026, with parallel fake-domain campaigns running concurrently — suggests that AI platform social engineering is entering a phase of sustained operational use rather than sporadic experimentation. Security programs should treat AI brand abuse as a persistent threat category, update their threat intelligence subscriptions accordingly, and plan for continued evolution of delivery techniques as operators adapt to any mitigations AI platforms put in place.

CSA Resource Alignment

This incident maps to several CSA frameworks and guidance documents.

MAESTRO (Agentic AI Threat Modeling Framework): CSA's MAESTRO framework provides a seven-layer model for reasoning about threats in agentic and generative AI systems [10]. In the MAESTRO framework's terms, this attack most directly engages the Agent Ecosystem layer (L7) — where AI interfaces connect with real-world users and applications, and where social engineering converts a trusted interface into a malware delivery channel — and the Security and Compliance layer (L6), which addresses the trust posture governing platform-to-user relationships. The Data Operations layer (L5) is also relevant given the scope of credential exfiltration. This is the authors' interpretive mapping; practitioners may reasonably model aspects of the campaign against additional layers. Organizations using MAESTRO to model AI tool deployments should add shared-URL content abuse as an explicit threat scenario within these layers.

AI Controls Matrix (AICM): The CSA AI Controls Matrix provides 243 control objectives organized across five pillars covering the lifecycle of AI systems and the organizations that operate them [11]. Controls addressing third-party AI tool supply chain verification, user awareness, and access governance are directly applicable. Organizations using AI developer tools from external providers should apply AICM controls covering third-party service assessment, ensuring that the AI platforms embedded in developer workflows have reviewed and disclosed their content sharing architecture and its associated risks.

AI Organizational Responsibilities: CSA's AI Organizational Responsibilities working group has published guidance on employee use of generative AI tools covering governance, access controls, and the organizational responsibilities associated with deploying AI applications that interact with sensitive

data [12]. The employee advisory and installation policy recommendations in this note align with that guidance's emphasis on establishing clear acceptable-use frameworks before AI tools are widely deployed. The companion guidance on AI Core Security Responsibilities [13] addresses the technical monitoring and incident response capabilities that macOS endpoint coverage gaps leave unfulfilled.

Zero Trust Principles: This attack illustrates why domain-based trust cannot substitute for content verification. A Zero Trust posture extended to software installation procedures — treating every installation instruction as untrusted regardless of the URL that delivered it — would reduce the effectiveness of shared-chat lures. Practical implementation means requiring that software installations follow pre-approved, out-of-band procedures rather than instructions from any web-based source, however authoritative the source appears.

References

[1] B. Toulas. "Hackers abuse Google ads, Claude.ai chats to push Mac malware." BleepingComputer, May 10, 2026.

[2] Malwarebytes Threat Intelligence. "Google ads funnel Mac users to poisoned AI chats that spread the AMOS infostealer." Malwarebytes Labs, December 2025.

[3] Huntress Threat Operations Center. "AI-Poisoning & AMOS Stealer: The Biggest Mac Threat." Huntress, December 9, 2025.

[4] Jamf Threat Labs. "From ClickFix to code signed: the quiet shift of MacSync Stealer malware." Jamf, December 22, 2025.

[5] The Hacker News. "ClickFix Campaigns Spread MacSync macOS Infostealer via Fake AI Tool Installers." The Hacker News, March 2026.

[6] Center for Internet Security. "MacSync Stealer Campaign Impacting U.S. SLTT macOS Users." CIS, 2026.

[7] Bitdefender Labs. "Windows and macOS Malware Spreads via Fake 'Claude Code' Google Ads." Bitdefender, 2026.

[8] Microsoft Security Blog. "ClickFix campaign uses fake macOS utilities lures to deliver infostealers." Microsoft, May 6, 2026.

[9] SecurityWeek. "MacSync macOS Malware Distributed via Signed Swift Application." SecurityWeek, 2026.

[10] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA AI Safety Initiative, February 2025.

[11] Cloud Security Alliance. "AI Controls Matrix." CSA, 2025.

[12] Cloud Security Alliance. "AI Organizational Responsibilities Working Group." CSA, 2025.

[13] Cloud Security Alliance. "AI Organizational Responsibilities: Core Security Responsibilities." CSA, 2024.

