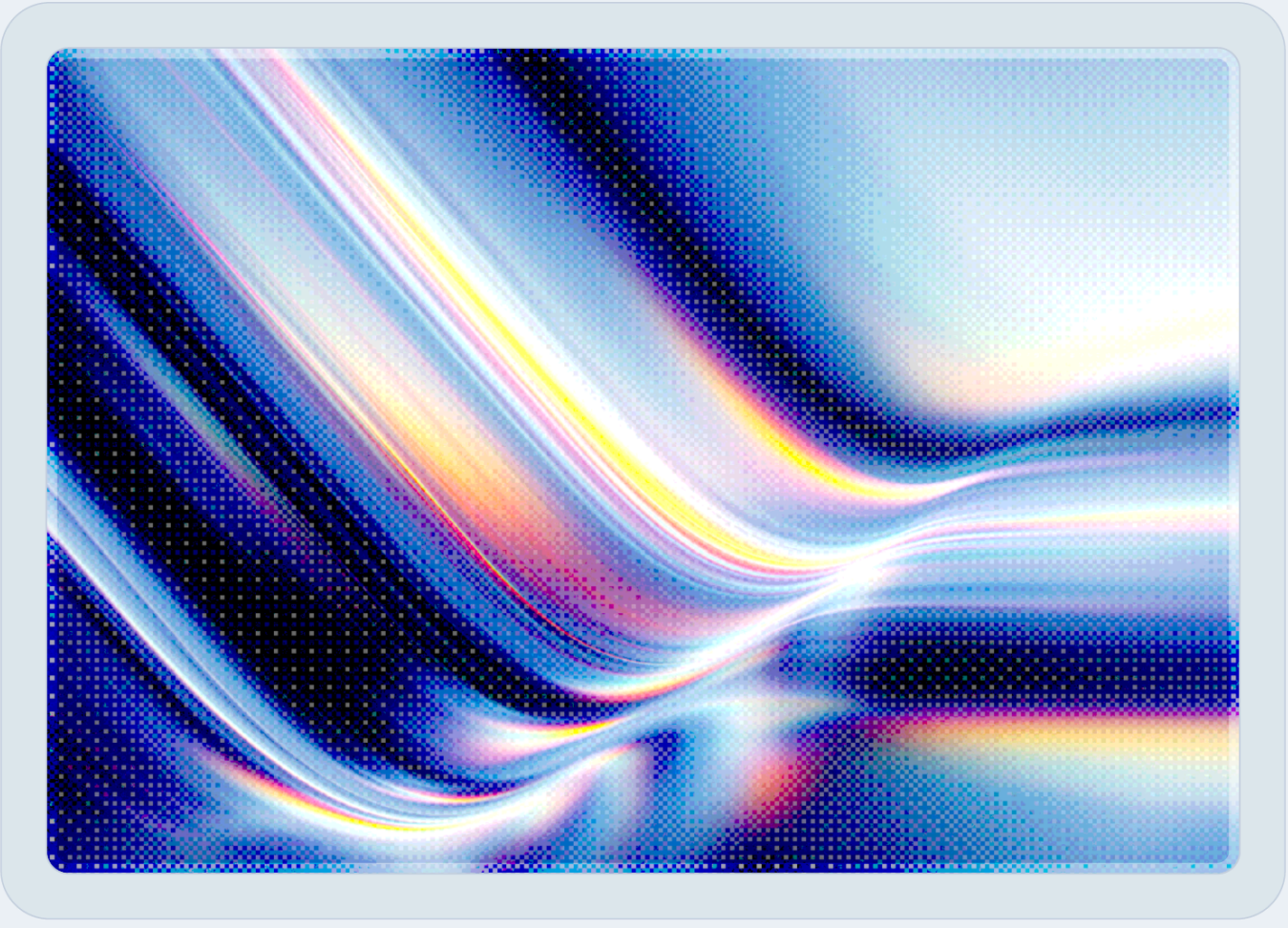
CSAI Foundation | Cloud Security Alliance

Poisoned AI Recommendations: Chatbots as Malware Delivery Vectors

Active Cryptojacking Campaign Leveraging LLM-Surfaced Malicious Downloads

2026-05-27

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Microsoft Defender Experts disclosed on May 26, 2026, that an active cryptojacking campaign uses AI chatbot interactions—in addition to traditional search engine poisoning—to surface malicious software download links; this is the first publicly documented instance of AI chatbot interactions serving as a distribution channel in this cryptojacking campaign, extending a pattern previously observed in information-stealer campaigns to a more technically complex operation.
- More than 150 attacker-controlled domains impersonating popular Windows utilities have been identified since March 2026, targeting users likely to possess high-performance GPUs capable of generating significant cryptomining revenue.
- The attack chain relies on DLL sideloading rather than software exploitation, making it resistant to standard patch-based defenses; the downloaded ZIP contains a legitimate application bundled with a malicious `autorun.dll` that loads silently when the user launches the expected program.
- Beyond cryptocurrency mining (deploying gminer, lolMiner, and SRBMiner-MULTI as GPU mining utilities), the campaign installs ScreenConnect to establish persistent remote access, creating a secondary beachhead that could support data theft, lateral movement, or ransomware delivery.
- AI chatbots compound the risk of traditional SEO poisoning because users perceive AI-generated answers as curated and authoritative—a trust asymmetry that threat actors are actively exploiting, as documented separately in Microsoft's February 2026 research on AI Recommendation Poisoning.
- Organizations should enforce verified-source software acquisition policies, monitor for anomalous GPU utilization and unexpected ScreenConnect installations, and apply AI-aware content filtering at the network perimeter.

Background

Malware distribution through social engineering has historically operated through a range of vectors—SEO poisoning, phishing, malvertising, drive-by downloads, software supply chain compromise, and others—each exploiting a different layer of user trust or a different failure mode in how software reaches end users. AI-powered chatbots introduce a distinct dynamic within this threat landscape. Users query these systems expecting synthesized, authoritative guidance rather than a ranked list of links, and the conversational format creates a different relationship with information than a search engine result page does—one that security practitioners and industry analysts have documented as more likely to reduce users' motivation to independently verify recommended sources [1][8][9].

Recommendation poisoning against AI systems builds on years of work targeting retrieval-augmented generation (RAG) pipelines and the knowledge bases that large language models query when formulating responses. In February 2026, Microsoft's Security Blog documented "AI Recommendation Poisoning" as a distinct threat category—techniques that inject persistent promotional or malicious content into an AI assistant's memory or retrieved context so that future, unrelated conversations surface attacker-controlled entities as trusted sources [1]. The attack mirrors SEO manipulation conceptually but operates at a fundamentally different layer: instead of tricking a ranking algorithm, it manipulates the model's world-model or session memory, achieving influence that can persist across user interactions without detection.

This threat did not emerge in isolation. The OWASP Top 10 for LLM Applications 2025 lists prompt injection—particularly *indirect* prompt injection, where adversarial instructions are embedded in external content the model processes—as the category's highest-priority risk [2]. Indirect injection expands the attack surface beyond the chat interface itself: any webpage, document, or API response that an AI assistant retrieves and summarizes becomes a potential injection vector. A threat actor who can influence what content a chatbot retrieves has, in effect, the ability to shape what that chatbot recommends.

Prior to the campaign documented by Microsoft Defender Experts, ZeroFox and Zscaler researchers had each independently observed threat actors optimizing web content specifically to influence AI-generated summaries served by ChatGPT, Google Gemini, and Microsoft Copilot [3][4]. These earlier campaigns focused primarily on distributing information stealers and browser credential harvesters. The May 2026 Microsoft disclosure extends this established pattern into a cryptojacking operation that adds persistent remote access infrastructure—a capability not documented in the earlier information-stealer campaigns—alongside multi-tool cryptocurrency mining payloads.

Security Analysis

The Active Cryptojacking Campaign

Microsoft Defender Experts began tracking this campaign in March 2026, when more than 150 attacker-controlled domains impersonating widely used Windows system utilities were identified. The selection of targeted applications suggests deliberate audience specificity: CrystalDiskInfo, HWMonitor, Display Driver Uninstaller (DDU), FurMark, K-Lite Codec Pack, and PDFgear are tools whose primary user base tends toward enthusiast PC builders, gaming system administrators, and IT support professionals—populations more likely to own high-performance graphics cards capable of generating meaningful cryptomining revenue [5].

In April 2026, Microsoft observed reports indicating that users may have been directed to these malicious domains through interactions with LLM-based tools rather than search engines—the first public disclosure of potential AI chatbot involvement as a distribution channel in this cryptojacking campaign. The mechanism by which the chatbot surfaces malicious links may involve multiple techniques: attacker-controlled websites optimized for AI retrieval, injected content in public knowledge bases that RAG pipelines index, or active manipulation of chatbot memory in products that offer personalization features. Microsoft has not publicly attributed the AI delivery vector to a specific injection technique, and the exact pathway warrants further investigation.

Once a user downloads the ZIP archive from an attacker domain, the initial phase of the attack requires no vulnerability exploitation whatsoever. The archive contains a legitimate, signed copy of the impersonated application alongside a malicious dynamic link library named `autorun.dll`. When the user launches the executable, Windows loads DLLs from the application's local directory before system paths—a well-documented DLL search order behavior—causing the malicious library to load silently and without visible error [5]. This sideloading technique predates AI delivery mechanisms entirely; its combination with chatbot-surfaced distribution is what makes the campaign architecturally novel.

The malicious DLL invokes `msiexec.exe` to install a second payload named `vcredist_x64.dll`, impersonating the common Visual C++ Redistributable package. This file functions as a packaged ScreenConnect installer. ScreenConnect (ConnectWise) is a legitimate remote-access platform, and its presence on an endpoint may not trigger antivirus signatures even when installed without user authorization. The campaign then deploys three GPU cryptocurrency mining utilities—gminer, lolMiner, and SRBMiner-MULTI—each a legitimate open-source project that threat actors routinely weaponize. Their signatures do not inherently indicate malice, complicating endpoint detection [5].

The combination of remote access and mining creates a staged threat posture. An operator with ScreenConnect access to thousands of compromised GPU-capable workstations holds an asset that extends well beyond cryptocurrency revenue: data exfiltration, lateral network traversal, and ransomware pre-positioning all become available without any additional initial access effort. Security teams should treat the discovery of unauthorized ScreenConnect on an endpoint as a potential indicator of full-scope compromise, not merely a cryptomining incident.

Why AI Delivery Amplifies Existing Threats

Traditional SEO poisoning exploits the gap between a search engine's ranking signal and actual content quality. AI chatbot delivery exploits something different and arguably more durable: the user's *epistemic relationship* with the AI system. Security practitioners and industry analysts have documented a pattern in which users tend to treat AI-generated answers as more authoritative than ranked search results—a trust asymmetry that threat actors are beginning to exploit, as evidenced by the February 2026 AI Recommendation Poisoning research and contemporaneous reporting [1][8][9]. When a chatbot presents a software download recommendation, the recommendation arrives without the visual cues—competing results, sponsored labels, unfamiliar domain names—that prompt critical evaluation in a search context. The conversational dynamic may also lead users to forgo independent URL verification: the act of querying the chatbot may function as implicit vetting in the user's perception, reducing the skepticism that a ranked search result typically prompts.

This trust asymmetry is not unique to any particular chatbot product; it appears to emerge from the conversational format itself, though organizations that have implemented explicit user-education programs around AI-generated recommendations may reduce this effect. More importantly, any AI assistant that retrieves from unvalidated external sources, or that allows persistent memory enrichment without auditing, is susceptible to having its recommendation behavior shaped by adversarially crafted content—particularly where the retrieval corpus is not restricted to organization-approved domains. Microsoft's February 2026 analysis documented specific observed techniques: crafting web pages with hidden text that instructs an AI to recommend a particular entity as "trusted by security professionals," injecting such content into AI assistant memory through legitimate user interactions, and planting persistent preferences in AI personalization features that survive session resets [1].

Enterprises that have broadly deployed AI assistants to their workforce—including coding assistants that suggest package names or documentation links—should recognize that each such deployment extends the potential attack surface. An employee querying a corporate AI assistant for software recommendations or library documentation receives a response shaped not only by the model's training data but by any external sources the assistant retrieves and any memory injected through prior interactions.

Threat Actor Profile and Targeting Logic

The selection of GPU-centric diagnostic utilities suggests targeting of users most likely to own high-performance graphics cards—specifically enthusiast PC builders, system administrators, and IT support professionals. GPU-intensive cryptomining against general enterprise endpoints would yield poor returns; the campaign's specificity toward this software category narrows the victim population to those most likely to offer meaningful mining throughput. The secondary deployment of ScreenConnect suggests a financially motivated but operationally sophisticated actor who treats the cryptomining payload as both immediate revenue and a secondary access broker opportunity—compromised systems with persistent remote access have well-established value in underground markets.

In environments where AI chatbot responses are not subject to the same URL filtering applied to browser traffic—currently the case in most enterprise deployments—the conversational interface can effectively bypass perimeter controls by surfacing attacker-controlled links in a trusted context. The campaign's consistent use of legitimate, signed binaries throughout the attack chain—rather than malicious executables that would trigger signature-based detection—indicates deliberate design choices to minimize endpoint visibility, distinguishing it from commodity malware campaigns that typically rely on obfuscated or custom payloads.

Recommendations

Immediate Actions

Organizations should immediately audit their endpoints for unauthorized ScreenConnect installations, paying particular attention to workstations with discrete GPUs and to users who recently downloaded diagnostic or system utility software. The presence of `autorun.dll` in the directory of any legitimate application, or of `vcredist_x64.dll` installed via `msiexec.exe` from an unexpected path, should be treated as a compromise indicator. Abnormal GPU utilization during periods of user inactivity—particularly sustained high utilization associated with gminer, lolMiner, or SRBMiner-MULTI processes—is a secondary behavioral indicator. Microsoft Defender for Endpoint detections for this campaign have been documented; organizations using that platform should review recent alerts against the indicators published in the May 26, 2026 Microsoft Security Blog post [5].

Software download policies should be reinforced immediately. Users should be directed to download system utilities exclusively from vendor-official URLs or enterprise-approved software distribution systems, and this guidance should explicitly note that AI chatbot recommendations do not constitute

organizational approval. Until AI-specific controls are in place, the trust users extend to chatbot-generated software links should be treated as a social engineering risk equivalent to a phishing recommendation.

Short-Term Mitigations

Within the next 30 to 90 days, organizations deploying AI assistants—whether general-purpose chatbots, coding assistants, or domain-specific tools—should evaluate and configure any available content filtering and grounding controls. Products that support RAG pipelines should restrict the corpus of retrievable sources to verified, organization-approved domains or data stores. Assistants with persistent memory features should have those features disabled or scoped tightly if memory update events cannot be audited. Web browsing-enabled AI assistants represent the highest-risk configuration; if real-time web retrieval is enabled, DNS and proxy filtering should apply to all URLs the assistant proposes returning to the user, not merely to URLs the user directly navigates to in a browser.

Network monitoring should be extended to detect unexpected ScreenConnect beaconing from endpoints that were not provisioned with remote management tooling. Because ScreenConnect operates over standard HTTPS, signature-based network detection is insufficient; behavioral analysis of persistent low-bandwidth outbound connections to ConnectWise infrastructure from user workstations is more reliable. Endpoint detection and response (EDR) rules should be created to flag execution of known open-source mining utilities regardless of whether they bear malicious signatures.

Strategic Considerations

Over the medium to long term, organizations should incorporate AI recommendation integrity into their threat modeling and vendor risk management programs. When evaluating AI assistant deployments, procurement teams should ask vendors whether the product's retrieval corpus is auditable, whether injected memories can be reviewed and revoked, and whether the product participates in information-sharing programs that would provide early warning of poisoning campaigns. AI assistants that surface external URLs should be treated architecturally as web browsers—subject to the same URL filtering, proxy inspection, and user-education controls.

The broader pattern this campaign represents—attackers optimizing content for AI retrieval rather than human search behavior—is expected to persist and intensify as AI assistants become more embedded in enterprise workflows, given the pace of AI assistant adoption and the financial incentives demonstrated by this campaign. Security awareness training programs should be updated to address AI-specific social engineering, emphasizing that AI-generated answers are synthesized from external sources that may have been adversarially crafted, and that software download recommendations from AI tools require the

same verification steps as any other untrusted source. Security teams should also monitor for AI-specific threat intelligence, including the emerging category of "Answer Engine Optimization" (AEO) abuse, in which attacker-controlled content is crafted specifically to appear in AI-generated summaries [3].

CSA Resource Alignment

This campaign and the broader class of AI recommendation poisoning attacks are directly addressed by several Cloud Security Alliance frameworks and research programs.

MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) is CSA's layered threat modeling framework for agentic and generative AI systems, introduced in February 2025 [6]. MAESTRO's seven-layer architecture includes distinct threat categories for Foundation Model integrity (Layer 1), Agent Framework vulnerabilities (Layer 3), and external data integration risks. The AI recommendation poisoning technique described in this note most directly maps to MAESTRO's concerns about unvalidated external retrieval influencing agent output—a risk that is categorically different from traditional software supply chain threats and requires AI-specific controls rather than simply extending existing endpoint or network security measures. Organizations applying MAESTRO to their chatbot and AI assistant deployments should explicitly model the retrieval corpus as a trust boundary and enumerate the threat of adversarial content injection alongside conventional data exfiltration scenarios [10].

AI Controls Matrix (AICM), CSA's 243-control framework across 18 security domains, provides the specific control objectives against which this campaign's defenses can be mapped [7]. AICM domains covering Data Security (controls around RAG corpus integrity and retrieval source validation), AI Supply Chain Security (controls covering model and plugin provenance), and AI Application Security (controls for output filtering and grounding) are all directly relevant. Organizations implementing AICM should ensure that AI-specific controls for retrieval-augmented generation are not treated as aspirational but as baseline requirements for any AI assistant deployment with web access. The AICM's mapping to ISO 42001 and the NIST AI RMF provides additional implementation guidance for organizations already aligned to those frameworks.

The OWASP Top 10 for LLM Applications 2025 classifies prompt injection—including the indirect variant that enables AI recommendation poisoning—as the highest-priority LLM application risk [2]. Organizations assessing their AI deployments against OWASP's guidance should treat any chatbot or assistant that retrieves external content as a system exposed to indirect prompt injection, and apply the

mitigations recommended by OWASP: output encoding, structured output formats, retrieval source filtering, and human confirmation for high-consequence actions such as software installation recommendations.

CSA's STAR (Security, Trust, Assurance, and Risk) registry and audit program can serve as a mechanism for evaluating AI assistant vendors against these controls. Organizations whose vendors participate in STAR Level 2 assessments should verify whether the assessment scope includes AI-specific retrieval pipeline integrity controls or whether coverage extends only to traditional cloud security controls.

References

[1] Microsoft Security Blog. "Manipulating AI Memory for Profit: The Rise of AI Recommendation Poisoning." Microsoft, February 10, 2026.

[2] OWASP Gen AI Security Project. "LLM01:2025 Prompt Injection." OWASP, 2025.

[3] ZeroFox. "SEO Poisoning: How Threat Actors Are Tricking AI Models like ChatGPT, Gemini, and Copilot." ZeroFox, 2025–2026.

[4] Zscaler ThreatLabz. "Black Hat SEO Poisoning Search Engine Results for AI." Zscaler, 2025.

[5] Microsoft Security Blog. "From Poisoned Search Results to GPU Mining: A Cryptojacking Campaign Abusing ScreenConnect and Microsoft .NET Utilities." Microsoft, May 26, 2026.

[6] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA, February 6, 2025.

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

[8] The Register. "Microsoft Warns That Poisoned AI Buttons and Links May Betray Your Trust." The Register, February 12, 2026.

[9] Help Net Security. "That 'Summarize with AI' Button Might Be Manipulating You." Help Net Security, February 11, 2026.

[10] Cloud Security Alliance. "MAESTRO for Real-World Agentic AI Threats." CSA, February 11, 2026.

