AI Chatbots as Covert Command-and-Control Infrastructure: Emerging Threat Patterns and Enterprise Defenses

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-08

Key Takeaways

In the span of roughly eight months—from July 2025 through February 2026—security researchers and incident responders documented the first confirmed in-the-wild malware families using commercial AI chatbot infrastructure as a primary command-and-control channel. The SesameOp backdoor stored encrypted operator commands inside OpenAI Assistants API thread objects [1]. The LAMEHUG implant, attributed to APT28 with moderate confidence, queried Hugging Face's hosted Alibaba Qwen model to dynamically generate reconnaissance commands at runtime [2]. The PROMPTFLUX dropper used a hard-coded Gemini API key to instruct the model to rewrite its own source code every hour [3]. Check Point Research then demonstrated that even AI services requiring no API key—specifically Microsoft Copilot and xAI Grok—could be weaponized as bidirectional relay proxies through their native web-browsing capabilities, requiring only anonymous web access from the infected endpoint [4].

What unifies these distinct techniques is a common strategic insight: commercial AI platforms are already allowlisted by enterprise firewalls and categorized as legitimate business services. Outbound HTTPS to `api.openai.com` or `generativelanguage.googleapis.com` blends into the normal operational traffic of organizations that have adopted AI productivity tools. Traditional C2 detection logic—which looks for unusual destinations, certificate anomalies, or suspicious domains—provides minimal signal when the communications endpoint is a globally recognized technology vendor. Security teams must now treat AI API egress as a potential covert channel category in the same way they previously learned to treat DNS, cloud storage, and social media platforms.

Background

The Architecture of AI Platform Abuse

Command-and-control infrastructure in modern intrusion operations has followed a consistent evolutionary path: when defenders learn to detect and block one category of C2 communication, adversaries migrate to channels that are harder to distinguish from legitimate traffic. DNS-over-HTTPS supplanted plaintext DNS tunneling. Cloud storage platforms such as Google Drive and OneDrive became C2 relays. Code hosting platforms including GitHub were abused for operator instructions

embedded in repository files [15]. Each migration exploited the same structural advantage—the destination was a trusted, high-traffic platform that organizations could not simply block without disrupting operations.

The commercial AI ecosystem presents this same structural advantage in an amplified form. Adoption of AI-powered tools has grown rapidly across enterprise, government, and critical infrastructure organizations. In many enterprise environments, and increasingly across government and critical infrastructure sectors, egress to major AI API providers has become a baseline expectation of normal operations—a level of adoption documented in industry surveys and consistent with CSA's own research into the widening gap between AI deployment rates and security program maturity [14]. The asymmetry between the business cost of blocking and the operational benefit of retaining access creates exactly the evasion opportunity that sophisticated threat actors seek to exploit.

AI platform abuse also benefits from a secondary characteristic that distinguishes it from most prior trusted-site abuse: the platforms themselves are capable of processing, transforming, and generating text. When a malicious actor routes C2 through a static file on Google Drive, the file simply sits there. When they route C2 through an AI platform, the platform can be instructed to generate commands, obfuscate payloads, and adapt responses based on the implant's prompts—turning the AI provider's inference infrastructure into an unsanctioned extension of the attacker's operational capability.

Living Off Trusted Sites: The Security Boulevard Classification

The broader technique of using legitimately trusted platforms for malicious communications has been labeled "Living Off Trusted Sites" (LoTS) by security researchers, in deliberate parallel to the long-established "Living Off the Land" (LotL) technique category for abusing legitimate operating system binaries [5]. LoTS attacks exploit the fact that security products and network monitoring tools must tolerate traffic to legitimate services, or else block large categories of sanctioned business communication. Prior LoTS cases documented in threat intelligence reporting include abuse of Slack, Discord, Telegram, and Pastebin as operator relay channels. AI platforms are notably more capable than static LoTS channels—such as cloud storage or code repositories—because they add active processing and text generation to the trusted communication path, enabling the platform itself to perform adversary-directed computation.

Security Analysis

SesameOp: The First Confirmed In-the-Wild AI API C2 Implant

Microsoft's Detection and Response Team (DART) disclosed SesameOp in November 2025 following discovery during an active incident response engagement [1]. The backdoor represents the first publicly confirmed case of in-the-wild malware using a commercial AI API as its sole C2 relay mechanism, rather than simply leveraging AI for code generation or translation assistance.

SesameOp's architecture consists of two components: a loader delivered as `Netapi64.dll` and a .NET-based payload obfuscated using the Eazfuscator.NET commercial obfuscation tool. The loader achieves persistence and initial execution through .NET `AppDomainManager` hijacking—a technique that intercepts the .NET runtime's assembly loading process to inject a malicious assembly before the legitimate application's code runs, producing behavior that is difficult to detect through process monitoring alone. The .NET payload, upon execution, authenticates to the OpenAI Assistants API using a hard-coded API key and establishes a "thread" object within the API's conversational memory system. It then polls this thread on a timer, retrieving operator-authored messages and executing them on the victim host. Outgoing data—results of executed commands—is compressed, encrypted with both symmetric and asymmetric cryptography, Base64-encoded, and written back to the same OpenAI thread. From a network perspective, the malware generates periodic TLS-encrypted outbound connections to a well-known, widely used AI vendor endpoint, a pattern that network monitoring infrastructure may well ignore or deprioritize.

OpenAI confirmed that it identified and disabled the API key and associated account following Microsoft's notification [1]. However, the defensive implications extend beyond revocation of a single credential. The architecture of the OpenAI Assistants API—its ability to store persistent state across multiple polling cycles, accept arbitrary text payloads, and return arbitrary generated text—is precisely what makes it exploitable for this purpose. The OpenAI Assistants API is scheduled for deprecation in August 2026 [1], which will eliminate this specific vector while leaving the underlying technique available for re-implementation on other AI platforms offering similar stateful API features.

LAMEHUG and the Operational AI Operator Model

CERT-UA's July 2025 advisory on the LAMEHUG malware family, attributed to APT28 (also tracked as Fancy Bear, associated with Russian military intelligence unit GRU 26165) with moderate confidence, introduced a qualitatively different threat model from SesameOp [2]. Rather than using an AI platform as a passive relay for operator commands, LAMEHUG integrated LLM API access directly into the

malware's decision loop. When executing on a victim host, LAMEHUG queries Alibaba Cloud's Qwen 2.5-Coder-32B-Instruct model via the Hugging Face Inference API, providing environmental context collected from the host and requesting dynamically generated Windows command sequences for reconnaissance and document collection tasks.

This architecture shifts how adversary operators manage implant behavior. Traditional C2 infrastructure requires a human operator—or a separate piece of automation the operator controls—to issue instructions that the implant then executes. LAMEHUG effectively outsources the generation of those instructions to a third-party LLM, removing the need for the operator to maintain a dedicated instruction generation pipeline. The malware uses the LLM in a role analogous to a remote shell—submitting environmental observations and receiving executable command sequences in return—except that the "shell" is a hosted AI service that interprets natural language. The CrowdStrike 2026 Global Threat Report characterizes LAMEHUG as a landmark development in AI-enabled adversary operations, noting that it represents the first publicly documented case of malware integrating an LLM into its runtime decision process rather than merely using AI for pre-infection support tasks [6].

PROMPTFLUX: AI-Enabled Polymorphic Evasion

Google's Threat Intelligence Group (GTIG) documented PROMPTFLUX in November 2025 as part of a broader analysis of adversary LLM usage patterns [3]. PROMPTFLUX is a VBScript-based dropper that embeds a hard-coded Gemini 1.5 Flash API key and, upon execution, queries the Gemini API with instructions to act as an "expert VBScript obfuscator." The prompt instructs the model to rewrite the dropper's entire source code to evade static signature detection, and the resulting variant is written to the Windows Startup folder for persistence. This cycle repeats hourly.

The threat model PROMPTFLUX represents differs from prior polymorphic malware in a key structural characteristic: the mutation engine is external rather than embedded. Traditional polymorphic engines carry their own mutation code within the malware binary itself—a characteristic that, ironically, eventually became a signature target for antivirus and EDR products [15]. PROMPTFLUX offloads the mutation logic entirely to a third-party AI service. The malware binary itself need only contain the API key and the prompt; everything computationally demanding about generating new obfuscated code is delegated to Google's infrastructure. GTIG noted that PROMPTFLUX appears to be in a development or testing phase, with certain features commented out [3], suggesting the technique is being refined rather than deployed at scale. Nevertheless, it demonstrates that the mutation capability exists and is accessible without specialized expertise—a developer with access to a Gemini API key and basic VBScript knowledge can construct a self-mutating dropper.

AI-in-the-Middle: No API Key Required

Check Point Research's February 2026 publication titled "AI in the Middle" demonstrated an attack pathway that expands the threat surface significantly beyond API-key-dependent implants [4]. Rather than registering an API key to communicate with an AI service's backend, the documented technique exploits AI assistants that incorporate web-browsing capabilities as a built-in feature—specifically, Microsoft Copilot and xAI Grok, both of which can fetch and process external URLs as part of generating responses to user queries.

The attack proceeds as follows. An implant on the victim host opens the AI assistant's public web interface, visible to any anonymous user, and submits a crafted query instructing the assistant to fetch a specific attacker-controlled URL. The assistant's browser-fetch capability retrieves that URL's content, which contains the operator's instructions encoded within otherwise innocuous-looking web content, and incorporates the fetched content into its response. The implant parses the AI assistant's output for embedded commands. Data exfiltration flows in the reverse direction: the implant includes encoded victim data within subsequent queries submitted to the AI interface, where it is relayed to a second attacker-controlled endpoint referenced in the AI's response. The entire exchange traverses the AI provider's legitimate web domain—there are no connections to attacker-controlled infrastructure that network monitoring would observe. Industry incident response data corroborates the speed advantage this technique confers: Palo Alto Networks' Unit 42 2026 Global Incident Response Report documented AI-driven attackers achieving data exfiltration in as little as 72 minutes—approximately four times faster than the prior year's baseline across observed intrusions [16]. These figures reflect Unit 42's incident caseload and may not generalize uniformly across all environments and threat actor profiles. Following responsible disclosure, Microsoft implemented behavioral changes to Copilot's web-fetch handling.

This characteristic makes mitigation structurally different from the SesameOp scenario: revoking a single API key is a tractable response, whereas modifying the web-fetch behavior of a widely deployed AI assistant requires vendor-side behavioral changes affecting all users of that service. The AI platform does not need to be compromised or specially configured for this technique to function; it operates exactly as designed. The abuse lies in how the surrounding malware harnesses legitimate AI capabilities as relay infrastructure, which makes account-based revocation impossible as a remediation path.

State-Sponsored Actors and the AI Abuse Ecosystem

AI platform abuse is not confined to isolated malware families. Google's GTIG AI Threat Tracker, updated in November 2025 and February 2026, documents a pattern of state-sponsored actors across multiple countries using AI platforms throughout the intrusion lifecycle [7]. APT41, assessed as a Chinese state-sponsored group, was observed using Gemini for code development assistance in building C2

infrastructure components, including a framework designated OSSTUN. MUDDYCOAST, an Iran-linked group, used Gemini for operational development tasks while leaving C2 infrastructure indicators exposed in their platform interactions, according to GTIG's analysis. North Korean operators used AI platforms for phishing lure creation and insider threat persona development. OpenAI's own disruption reporting, covering activity through October 2025, documents Russian-language operators refining remote-access trojans and credential stealers, Korean-language operators developing C2 systems, and Chinese-linked groups debugging malware targeting semiconductor sector organizations and political institutions [8].

The CrowdStrike 2026 Global Threat Report quantifies a broader trend: CrowdStrike reported an 89% year-over-year increase in what the company classified as AI-enabled adversary operations [6], though definitional consistency across reporting periods was not independently verified. The report documents cases of adversaries injecting malicious prompts into GenAI tools at more than 90 organizations and notes that in at least one intrusion, data exfiltration began within four minutes of initial access—a tempo that human-only adversary operations rarely achieve. The combination of AI-accelerated attack velocity and AI-mediated C2 infrastructure documented in 2025 represents a substantial shift in the threat environment: the emergence of AI platforms as a new channel category requires correspondingly new detection paradigms rather than incremental updates to existing signature sets.

Related Vulnerability Disclosure: EchoLeak

The threat landscape around AI platforms is not limited to externally abused APIs. CVE-2025-32711, disclosed in June 2025 and designated "EchoLeak," was a critical (CVSS 9.3, per Microsoft's own scoring) zero-click indirect prompt injection vulnerability in Microsoft 365 Copilot [9]. The vulnerability allowed an attacker to send a crafted email; when Copilot subsequently summarized the email, a hidden prompt embedded within the email bypassed Microsoft's cross-prompt injection attempt (XPIA) classifier through a combination of techniques—classifier evasion through instruction framing, link redaction bypass using reference-style Markdown, abuse of auto-fetched images, and exploitation of a Teams proxy permitted by the content security policy. The end result was exfiltration of all content within Copilot's access scope: OneDrive files, SharePoint content, Teams message history, and chat logs. No user interaction was required beyond the AI assistant performing a task the user had requested. Microsoft patched EchoLeak server-side; the vulnerability has been documented in a peer-reviewed academic paper published on arXiv as "EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System" [9].

Recommendations

Immediate Actions

Security teams should treat AI API egress as a monitored and controlled network category, not as a trusted-site exemption. Endpoint detection rules should alert on processes outside explicitly approved AI tooling that establish outbound connections to AI API endpoints—specifically `api.openai.com`, `generativelanguage.googleapis.com`, `huggingface.co/api`, and equivalent endpoints for other platforms. The behavioral signature most relevant to SesameOp-class implants is a non-browser process making periodic, timer-driven connections to an AI API endpoint, particularly if the connection interval is consistent and if the process image is a DLL loaded via AppDomainManager injection or DLL sideloading.

Enterprise browser management policies should be reviewed and tightened with respect to AI-related browser extensions. The malicious Chrome extensions documented in Microsoft's March 2026 research, which stole ChatGPT and DeepSeek conversation histories from approximately 900,000 users across 20,000 enterprise tenants, achieved their reach by impersonating a legitimate AI assistant tool [10]. Organizations that permit employees to self-install browser extensions, or that have enabled agentic browser configurations that install extensions automatically, should assess their extension inventory immediately and enforce an allow-list policy for extensions that interact with AI platform domains.

Short-Term Mitigations

An enterprise AI platform inventory—a catalog of which AI tools are sanctioned for use, on which systems, by which user populations, and with what data classification limitations—is a prerequisite for effective detection and response. Without knowing what AI API traffic is expected, security operations teams cannot identify what is anomalous. Shadow AI detection capabilities, available in several CASB and network monitoring platforms, can provide baseline visibility into actual AI service usage across the environment prior to formal policy enforcement. Organizations should establish this inventory and baseline before attempting to write detection rules, because rules written without accurate knowledge of expected traffic will produce either excessive false positives or missed detections.

Prompt injection defense for AI-integrated productivity tools should be treated as a distinct control category. CVE-2025-32711 demonstrated that even AI systems deployed by sophisticated vendors with active security programs are vulnerable to indirect prompt injection at critical severity. Organizations deploying Microsoft 365 Copilot, Google Workspace AI features, or similar AI-integrated productivity tools should verify that vendor-provided mitigations for known prompt injection vulnerabilities have been

applied, monitor vendor security advisories for this vulnerability class, and scope AI assistant permissions conservatively—limiting access to the minimum data necessary for the assistant's intended functions, so that a successful injection attack has a bounded exfiltration surface.

Planning should begin now for the OpenAI Assistants API deprecation scheduled for August 2026. While this deprecation will eliminate one specific C2 mechanism documented in SesameOp, organizations that have integrated the Assistants API into legitimate workflows should audit those integrations and plan migration paths. Additionally, security teams should anticipate that threat actors using SesameOp-style architectures will adapt to equivalent stateful API features offered by other AI providers, and update detection rules accordingly when those adaptations are observed in threat intelligence reporting.

Strategic Considerations

The long-term strategic implication of this threat category is that AI platforms are now dual-use infrastructure—simultaneously legitimate business tools and potential components of adversary operations. This duality does not mean organizations should restrict AI platform access categorically; that approach would sacrifice significant operational value to address a threat that can be managed through more targeted controls. It does mean that AI platform access should be engineered with the same security rigor applied to other high-value communication channels: authenticated, monitored, scoped to approved use cases, and subject to anomaly detection.

Security architecture teams should design AI integrations with isolation principles: AI features that require access to sensitive organizational data should operate within dedicated security contexts with tightly scoped permissions, rather than being granted broad access to enterprise data stores. Agentic AI workflows—configurations in which AI systems autonomously browse, retrieve, or process content on behalf of users—require particular attention, because the "AI-in-the-Middle" and indirect prompt injection techniques documented in current research are specifically effective against agentic configurations. The general principle of least-privilege agent permissions, applied to AI workflows, limits both the exfiltration surface if an injection succeeds and the potential for unauthorized actions by a compromised agent.

Threat intelligence programs should track AI platform abuse as an explicit threat category, updating watchlists and detection content as new AI APIs and capabilities enter widespread enterprise use. The rate of new AI platform feature release—new modalities, extended context windows, web-browsing and code execution capabilities—creates a continuous stream of new potential abuse vectors that threat actors will probe. Proactive threat modeling against newly announced AI platform capabilities, rather than reactive detection after abuse is confirmed in the wild, is the appropriate posture given the demonstrated adversary interest in this space.

CSA Resource Alignment

This research note relates to several active Cloud Security Alliance frameworks and guidance resources. The MAESTRO framework (Multi-Layer Agentic Evaluation and Systematic Threat Response Orchestration) provides a structured threat modeling methodology for agentic AI systems that directly addresses the risk categories illustrated by the "AI-in-the-Middle" technique and indirect prompt injection against agentic configurations [11]. MAESTRO's threat taxonomy explicitly covers cases where AI agents are manipulated through external content—a description that encompasses the EchoLeak vulnerability class and related prompt injection scenarios targeting AI-integrated productivity tools.

The CSA AI Controls Matrix (AICM) v1.0, along with its accompanying AI-CAIQ self-assessment questionnaire, maps controls across eighteen AI security domains including AI supply chain security, model access governance, and data exposure through AI systems [12]. Organizations assessing their exposure to AI platform C2 abuse should consult AICM control domains covering AI service access authorization, outbound AI API traffic monitoring, and browser-based AI tooling governance. The STAR for AI program provides a mechanism for evaluating AI service providers against these control criteria, which is relevant for assessing which AI vendors have implemented controls against API key abuse, prompt injection, and unauthorized thread storage.

CSA's guidance on Zero Trust architecture applies directly to the mitigations described in this note. The principle that no network destination should be implicitly trusted—including destinations with legitimate business use—supports treating AI API egress as a monitored control point rather than a trusted-site exemption. Zero Trust network segmentation, applied to endpoint AI API access, would constrain the scope of SesameOp-class implants by limiting which endpoints are permitted to reach AI provider APIs and which user contexts those endpoints can operate in. The CSA Cloud Controls Matrix (CCM), particularly the Threat and Vulnerability Management (TVM) and Identity and Access Management (IAM) domains, provides control mappings for the access governance and monitoring requirements identified in this note's recommendation section [13].

The CSA "State of AI Security and Governance" survey report documents the growing gap between enterprise AI adoption rates and the security program maturity to match that adoption [14]. The threat patterns documented in this research note reinforce the report's finding that shadow AI detection and AI access governance are priority areas for security programs operating in AI-integrated environments. Organizations that have not established AI platform inventories or extended their network monitoring policies to cover AI API egress are operating with unaddressed exposure to the techniques described here.

References

1. Microsoft Security Response Center / Microsoft DART, "SesameOp: Novel backdoor uses OpenAI Assistants API for command and control," Microsoft Security Blog, November 3, 2025. https://www.microsoft.com/en-us/security/blog/2025/11/03/sesameop-novel-backdoor-uses-openai-assistants-api-for-command-and-control/

2. CERT-UA, "CERT-UA#14537: Cyber-espionage operation using LAMEHUG malware (UAC-0001)," Ukrainian CERT Advisory, July 17, 2025 (expected URL format: https://cert.gov.ua/article/[ID]; direct URL unavailable due to cert.gov.ua intermittent accessibility at time of publication); Cato CTRL, "Cato CTRL Threat Research: Analyzing LAMEHUG—The First Publicly Documented Case of a Malware Integrating a LLM," Cato Networks Blog, 2025. https://www.catonetworks.com/blog/cato-ctrl-threat-research-analyzing-lamehug/

3. Google Threat Intelligence Group (GTIG), "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools," Google Cloud Blog, November 2025. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools

4. Check Point Research, "AI in the Middle: Turning Web-Based AI Services into C2 Proxies—The Future of AI-Driven Attacks," Check Point Research Blog, February 2026. https://research.checkpoint.com/2026/ai-in-the-middle-turning-web-based-ai-services-into-c2-proxies-the-future-of-ai-driven-attacks/

5. Security Boulevard, "Living-Off-Trusted-Sites (LOTS) or should we say services?," July 2024. https://securityboulevard.com/2024/07/living-off-trusted-sites-lots-or-should-we-say-services/

6. CrowdStrike, "2026 CrowdStrike Global Threat Report," CrowdStrike Press Release, 2026. https://ir.crowdstrike.com/news-releases/news-release-details/2026-crowdstrike-global-threat-report-ai-accelerates-adversaries

7. Google Threat Intelligence Group (GTIG), "GTIG AI Threat Tracker: Distillation, Experimentation, and Integration of AI for Adversarial Use," Google Cloud Blog, February 2026. https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

8. OpenAI, "Disrupting malicious uses of AI: October 2025," OpenAI Global Affairs, October 2025. https://openai.com/global-affairs/disrupting-malicious-uses-of-ai-october-2025/

9. Pavan Reddy and Aditya Sanjay Gujral (Aim Security / The George Washington University), "EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System," arXiv preprint arXiv:2509.10540, September 2025. https://arxiv.org/abs/2509.10540; see also CVE-2025-32711, NVD. https://nvd.nist.gov/vuln/detail/CVE-2025-32711

10. Microsoft Security Blog, "Malicious AI Assistant Extensions Harvest LLM Chat Histories," March 5, 2026. https://www.microsoft.com/en-us/security/blog/2026/03/05/malicious-ai-assistant-extensions-harvest-llm-chat-histories/

11. Cloud Security Alliance, "MAESTRO: Multi-Layer Agentic Evaluation and Systematic Threat Response Orchestration," CSA AI Safety Initiative. https://cloudsecurityalliance.org/blog/2024/04/09/ai-safety-initiative-maestro-framework

12. Cloud Security Alliance, "AI Controls Matrix (AICM) v1.0 and AI-CAIQ," CSA AI Safety Initiative, 2025. https://cloudsecurityalliance.org/research/working-groups/ai-controls-matrix

13. Cloud Security Alliance, "Cloud Controls Matrix (CCM) v4," CSA. https://cloudsecurityalliance.org/research/cloud-controls-matrix/

14. Cloud Security Alliance, "The State of AI Security and Governance Report 2025," CSA AI Safety Initiative, 2025. https://cloudsecurityalliance.org/

15. MITRE Corporation, "ATT&CK for Enterprise," MITRE ATT&CK Knowledge Base, 2025. Relevant techniques: T1071.004 (Application Layer Protocol: DNS), T1567.002 (Exfiltration Over Web Service: Exfiltration to Cloud Storage), T1195.003 (Supply Chain Compromise: Compromise Software Source Code), T1027 (Obfuscated Files or Information). https://attack.mitre.org/

16. Palo Alto Networks Unit 42, "2026 Unit 42 Global Incident Response Report," Palo Alto Networks, 2026. https://www.paloaltonetworks.com/blog/2026/02/unit-42-global-ir-report/