CRITICALSTART® Threat Research

TLP CLEAR // [CS-TR-26-0201] Posturing for an AI-Accelerated Attack Surface

Executive Summary

Artificial intelligence (AI) is now a core component of modern cyberattacks. Once used primarily to enhance phishing campaigns, it has evolved into a tool integrated across reconnaissance, payload development, and malware execution. Recent reporting shows GenAI embedded in state-sponsored reconnaissance, malicious browser extensions installed by over 260,000 users, and other operational workflows, highlighting the scale and effectiveness of AI-accelerated attacks. This evolution represents a structural change in adversary tradecraft. AI is no longer limited to content generation; it accelerates targeting, automates decision-making, and dynamically adapts outputs during campaigns. Organizations relying on static indicators, traditional tooling, or outdated assumptions about attacker behavior are increasingly exposed.

The impact is measurable and financially significant. Research from IBM found that Generative AI (GenAI) reduced the time to craft a convincing phishing email from 16 hours to five minutes. Organizations that have not updated governance, telemetry, and detection strategies to account for AI-assisted techniques, including risks from shadow AI, face elevated exposure. According to IBM, breaches in organizations with high levels of shadow AI cost an average of $670,000 more than those with low or no shadow AI presence. This report examines three concrete threat patterns observed in recent reporting, evaluates their implications for defenders, and provides prioritized mitigation actions to address AI-accelerated threats.

Introduction

Artificial intelligence (AI) has reduced the cost and complexity of attack development and execution across several phases of the kill chain, to the advantage of cyber threat actors. Reconnaissance that previously required hours of manual research now takes minutes. Phishing lures that once demanded linguistic skill can be crafted from a simple prompt to a Generative AI (GenAI) model. Malware that required device-specific engineering can now offload runtime decisions to a language model. AI does not create new categories of threat, but it allows attackers to execute existing attack types faster, at larger scale, and with less specialized knowledge and effort.

This article examines three AI-accelerated attack patterns that illustrate these efficiencies. The first is AI embedded in malware execution, as seen with PromptSpy, which uses a language model to interpret Android screen states and automate gestures across diverse devices. The second is AI as a live command generator, exemplified by LAMEHUG, which produces attack instructions on demand from a compromised host, bypassing static detection methods. The third is AI branding weaponized as a trust channel, demonstrated by the AiFrame extension campaign, where attackers distribute malicious browser extensions to over 260,000 users and maintain remote control through dynamically delivered interfaces. Each pattern reduces attacker effort and increases operational efficiency while exploiting gaps in governance, telemetry, and detection.

Shadow AI use further amplifies attacker advantage. According to the IBM Cost of a Data Breach Report 2025, 20% of organizations experienced breaches linked to shadow AI. Shadow AI refers to AI tools and applications adopted by employees or teams without formal approval or oversight from the organization’s IT or security governance. High levels of shadow AI allowed attackers to scale operations and reduce effort per compromise, which in observed breaches translated to an additional $670,000 in average cost compared to organizations with low or no shadow AI. These incidents also increased

exposure of personally identifiable information (PII) and intellectual property across multiple environments. Governance gaps exacerbate the problem: 63% of organizations lack a formal AI policy or are still developing one, fewer than half enforce AI deployment approvals, and only 34% audit for unsanctioned AI. The combination of rapid AI adoption and weak oversight gives attackers a clear operational advantage and increases the potential impact of each compromise.

The cases that follow illustrate how attackers are already leveraging AI in practice. They highlight the tangible operational efficiencies gained by embedding AI into malware, using AI to generate commands in real time, and weaponizing AI as a trust channel, showing why organizations must rethink detection, governance, and risk assumptions in the age of AI-assisted threats.

AI Embedded in Malware Execution: The PromptSpy Case

ESET researchers identified PromptSpy as the first known Android malware to integrate generative AI directly into its execution flow. The malware prompts Google's Gemini with XML dumps of the current screen state and receives back JSON- formatted instructions specifying tap coordinates and gestures, which it then executes through Android's Accessibility Services to lock itself in the device's recent apps list and resist removal. The malware also overlays invisible elements on uninstall buttons, making standard removal impossible without booting into Safe Mode. Its core payload deploys a VNC module that gives operators remote access to the compromised device, with communication encrypted via AES to a command-and-control server at 54[.]67[.]2[.]84. Distribution occurred through phishing sites impersonating Chase Bank, targeting users in Argentina, with code analysis suggesting development in a Chinese-speaking environment.

The significance here is not the novelty of “AI in malware” as a headline. The key point is the problem the AI integration solves. Android UI automation is traditionally brittle because device manufacturers implement different skins, gesture behaviors, and multitasking interfaces. A compatibility issue arises when code that functions correctly on one device does not perform as intended on another. By delegating UI interpretation to Gemini, PromptSpy bypasses this fragmentation problem entirely. The model reads the screen state and adapts its instructions to whatever layout it encounters, allowing a single malware build to maintain persistence across a far wider range of devices than traditional scripting permits. ESET notes that PromptSpy has not yet appeared in telemetry at scale and may still be a proof of concept, but the architecture itself signals a replicable operational advantage that attackers can leverage effectively.

AI as a Live Command Generator: APT28 and LAMEHUG

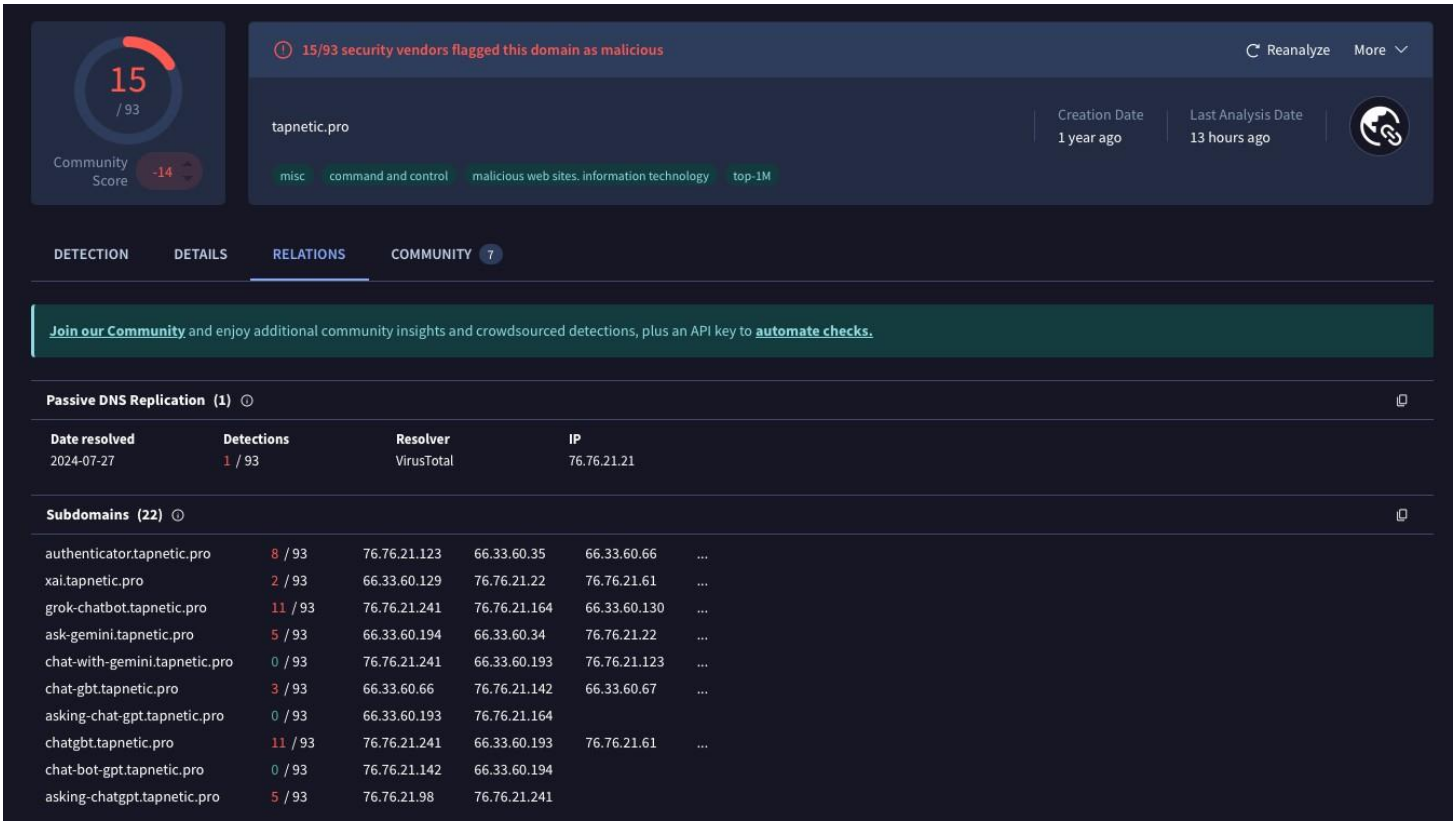
Ukraine's CERT-UA in July 2025 identified a phishing campaign targeting executive government authorities that delivered malware classified as LAMEHUG. The campaign used a compromised official email account to send ZIP archives disguised as ministry documents. Inside was a Python-based payload compiled with PyInstaller that, once executed, queried the Qwen2.5-Coder-32B-Instruct model via the Hugging Face API to generate Windows shell commands in real time. Those commands performed system reconnaissance, collected hardware and network information, and recursively copied documents from the victim's Desktop, Downloads, and Documents folders to a staging directory before exfiltrating them via SFTP or HTTP POST. CERT-UA attributed the campaign with medium confidence to APT28, the GRU-linked group also known as Fancy Bear and Forest Blizzard.

What distinguishes LAMEHUG from prior AI-assisted attack tools is the point at which the model is invoked. Previous examples of AI misuse involved attackers using LLMs during development, to write scripts or craft lures, before the attack begins. LAMEHUG queries the model during execution, on the compromised host, to generate the commands it will run. IBM X-Force noted that this approach allows threat actors to adapt their tactics during a compromise without deploying new

payloads, which directly undermines static analysis and signature-based detection. Cato Networks' follow-up analysis found that the campaign used approximately 270 Hugging Face tokens for authentication and observed multiple payload variants, suggesting active development rather than a finished capability. The broader implication is that defenders can no longer assume that the commands executed on a compromised host are fixed and predictable. When the attacker can generate novel command sequences on demand through a legitimate cloud API, behavioral detection must account for that variability. Monitoring outbound connections to LLM service endpoints from endpoints where such traffic is unexpected becomes a meaningful detection signal.

AI Branding as a Trust Weapon: The AiFrame Extension Campaign

LayerX researchers uncovered a coordinated campaign of 30 Chrome extensions that impersonated popular AI assistants including ChatGPT, Claude, Gemini, and Grok. The extensions collectively reached over 260,000 users, with several carrying the "Featured" label in the Chrome Web Store, which increased perceived legitimacy and accelerated adoption. These Chrome extensions positioned as AI summarizers, and assisted tools. Despite different names and branding, all 30 extensions shared identical JavaScript logic, permission sets, and backend infrastructure routed through the tapnetic[.]pro domain. Analysis of this domain revealed several additional malicious subdomains, which lent credibility to the attackers' network and supported the coordination of multiple extensions under a single infrastructure.



The core mechanism of this attack was iframe injection. Rather than implementing AI functionality locally, each extension loaded a full-screen iframe from an attacker-controlled subdomain. This iframe overlaid the active webpage and acted as the extension's primary interface. Because the interface was served remotely, operators could change functionality, introduce new data collection behaviors, or modify the attack chain without submitting any update to the Chrome Web Store for review.

The data collection scope of the campaign was broad. When triggered by the remote iframe, a content script used Mozilla's Readability library to extract structured page content, including titles, text, and metadata, from any tab the user visited, including authenticated enterprise portals and internal dashboards. A subset of 15 extensions specifically targeted Gmail, injecting scripts at document load on mail.google.com and using MutationObserver to maintain persistence through Gmail's dynamic page updates. These scripts read visible email content directly from the DOM, including conversation threads and draft text, and transmitted it to attacker-controlled servers. This approach is functionally equivalent to an adversary-in-the- middle attack, giving attackers real-time access to sensitive communications.

The campaign also demonstrated active evasion of enforcement. When one extension was removed from the Chrome Web Store in February 2025, an identical replacement appeared under a new identifier within two weeks, retaining the same permissions and backend connections. This tactic is known as extension spraying, where attackers rapidly redeploy malicious extensions under different names or identifiers to maintain distribution and persistence despite removals. Extension spraying amplifies operational resilience and reduces the cost and effort required to maintain access at scale.

The strategic problem this campaign exposes is a governance gap that most organizations have not closed. Browser extensions now operate with privileges comparable to endpoint software. They can read page content across all sites, access authenticated sessions, and exfiltrate data to external infrastructure. The AiFrame architecture exploits a specific vulnerability in how the Chrome Web Store review model works: what is inspected at install time is not necessarily what executes at runtime when core functionality is delivered through remotely mutable iframes. Organizations that rely on the store's review process as a proxy for extension safety are operating on a false assumption.

Implications for Organizations

The integration of AI into the attack lifecycle shortens the time between initial access and data exfiltration, giving attackers a measurable operational advantage. Threat actors like APT28 can automate the generation of context-specific PowerShell scripts or reconnaissance commands, removing the human bottleneck that previously gave defenders a window for manual detection.

The use of AI for persistence, as seen in PromptSpy, further complicates detection. The malware can adapt its behavior to avoid triggering local security alerts, introducing polymorphic characteristics in which code and execution patterns change to evade detection. AI-driven malware extends this concept by generating commands or execution logic in real time, allowing a single malware build to maintain control across diverse devices and environments.

As a result, traditional signature-based defenses are increasingly insufficient. Because AI can produce functionally identical commands in multiple syntactic variations, defenders must shift toward behavioral analysis and monitoring of API calls to public AI services to identify these dynamic attacks.

Organizations can no longer assume that the absence of known malware signatures equates to a clean environment. Security operations centers require high-fidelity telemetry capable of identifying anomalous patterns in how applications interact with external AI services. Security teams must also recognize that the same AI tools designed to increase developer productivity are now being repurposed to reduce the operational cost and effort of sophisticated cyber espionage. This convergence of legitimate utility and malicious intent demands a more nuanced approach to egress filtering, credential management, and anomaly detection.

Organizational Mitigation Strategies

Considering the proliferation of AI-accelerated threats, organizations must adopt proactive, structured measures to reduce the likelihood and impact of cyber threats. Key strategies include:

- Govern AI tool usage as an enterprise control: Approved AI platforms should be explicitly defined and enforced through centralized identity such as Microsoft Entra ID with MFA, device compliance, and conditional access. Using Microsoft Defender for Cloud Apps, enable app discovery to identify unsanctioned AI tools, mark high-risk services as unsanctioned, apply session controls to restrict uploads or downloads, and generate alerts when sensitive data is transmitted. Log AI API usage and apply DLP inspection to prompts and outputs. Where feasible, block unsanctioned AI domains and API endpoints at the proxy, firewall, or DNS layer.
- Restrict browser extension access: Enforce extension allowlists and disable developer mode through enterprise management of Google Chrome Enterprise or Microsoft Edge. Block extensions requesting broad permissions (all URLs, cookies, content script injection across sensitive domains). Use endpoint telemetry from EDR platforms such as CrowdStrike to detect anomalous browser child processes, iframe injection behavior, and unexpected outbound connections. Investigate rapid re-publication of extensions with identical permissions as potential spraying activity.
- Treat mobile Accessibility Services as high-risk: On devices managed through Microsoft Intune, restrict which apps can obtain Accessibility permissions and alert on newly granted privileges, especially from sideloaded applications. Feed mobile risk signals into conditional access policies to automatically restrict corporate resource access. Maintain playbooks for rapid isolation and Safe Mode remediation when persistence-resistant malware is suspected.
- Monitor for unexpected LLM API traffic: Log outbound traffic to LLM providers including Hugging Face, OpenAI, Google Generative Language APIs, and Anthropic. Use XDR or SIEM hunting queries to flag API calls from endpoints without a business justification. Treat these as triage signals and correlate with identity, process lineage, and data movement patterns, focusing on behavioral sequences rather than static command signatures.
- Audit OAuth app grants and integrations: In Microsoft 365 and Google Workspace, restrict user consent for high-risk scopes and require admin approval for mail, file, and calendar access. Use Defender for Cloud Apps to identify over-permissioned or inactive OAuth apps and revoke stale tokens. Monitor for spikes in new consent events and correlate with sign-in anomalies to detect malicious consent phishing.

Conclusion

AI is increasingly embedded into attacks to solve operational challenges that previously required more effort, expertise, or custom engineering. In PromptSpy, AI overcomes device fragmentation in UI automation. LAMEHUG generates adaptive commands on demand, removing the need for hardcoded attack logic. The AiFrame campaign uses AI branding to scale exploitation of user trust. In each case, AI is functional and consequential, directly enhancing the effectiveness, resilience, and reach of the attack. Defenders must adjust their assumptions accordingly. Static analysis and install-time review are no longer sufficient when critical functionality is generated or executed at runtime. Signature-based detection cannot reliably catch command sequences produced dynamically by AI models. User awareness programs are limited when malicious extensions appear “Featured” in official stores. Effective defense requires governance covering AI-adjacent trust channels, telemetry that captures runtime behavior, and detection logic designed to address the variability and adaptability introduced by AI- accelerated attacks.

Further Reading

1. IBM Cost of a Data Breach Report 2025
2. First Android Malware Weaponizes Gemini AI to Evade Detection, Maintain Persistence
3. PromptSpy: First Android malware to use generative AI in its execution flow
4. CERT-UA Discovers LAMEHUG Malware Linked to APT28, Using LLM for Phishing Campaign
5. Novel malware from Russia's APT28 prompts LLMs to create malicious Windows commands
6. New "LameHug" Malware Deploys AI-Generated Commands
7. "AiFrame" Fake AI Assistant Extensions Targeting 260,000 Chrome Users via injected iframes
8. Fake AI Chrome Extensions Exposed 260,000 Users, Targeting Gmail
9. GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools
10. Demystifying Device-specific Compatibility Issues in Android Apps
11. UAC-0001 Cyberattacks on the Security and Defense Sector Using the LAMEHUG Software Tool
12. What is Polymorphic Malware? Examples & Challenges
13. Govern Discovered Apps with Microsoft Defender for Cloud Apps