AI-Powered Ransomware: Automated Variant Proliferation

How LLMs Are Reshaping the Ransomware Threat Landscape

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-15

Key Takeaways

The ransomware threat landscape has entered a new phase. Large language models are being embedded into malware execution itself—not merely in the attack planning stages, but as live code generators that produce unique, functionally novel variants on every run. ESET Research's August 2025 discovery of PromptLock demonstrated a proof-of-concept—assessed by ESET as likely academic in origin—in which a locally hosted LLM generates fresh malicious Lua scripts dynamically rather than deploying a static payload [1]. Months later, IBM X-Force identified Slopoly, what it assessed as the first likely AI-generated malware deployed in an active criminal ransomware campaign—used by the threat group Hive0163 in an Interlock ransomware operation that maintained persistent access to a victim organization for over a week [2]. Underground markets have accelerated this trend through purpose-built criminal LLMs such as WormGPT and FraudGPT, which remove safety guardrails and provide adversaries with on-demand ransomware code generation capabilities [3][4].

The implications for defenders are significant. Every generated variant carries a distinct file hash, eroding the effectiveness of signature-based detection. CrowdStrike's 2026 Global Threat Report documented an 89% year-over-year increase in AI-enabled adversary attacks and an average eCrime breakout time of just 29 minutes, with the fastest observed breakout reaching 27 seconds [5]. IBM X-Force's 2026 Threat Intelligence Index found that active ransomware and extortion groups surged 49% in 2025 compared to the prior year [6]. The convergence of AI-generated code generation, existing ransomware-as-a-service (RaaS) infrastructure, and a rapidly growing ecosystem of criminal AI tooling is restructuring both attacker capability and defender response requirements.

Background

Ransomware has evolved through several distinct generations since the first documented case in 1989. Early ransomware relied on symmetric encryption with keys embedded in the malware itself; by the 2010s, asymmetric key exchange with command-and-control infrastructure had become standard. The RaaS model—in which ransomware operators license their code to affiliates who execute the attacks and

share ransom proceeds—emerged in the mid-2010s and drove a dramatic expansion of the threat actor pool by separating technical development from operational execution [7]. Today, IBM X-Force tracked 109 distinct extortion groups active in 2025, a 49% increase from the prior year [6].

What has changed in 2025 and into 2026 is the introduction of LLMs as a core component of the malware development and execution stack. This differs from earlier uses of AI in cyberattacks—where models primarily assisted with phishing lure drafting or reconnaissance—in a fundamental way: the model is now embedded inside the attack tool itself, generating malicious code at runtime or producing functional exploit components on demand through API calls. The result is a class of malware whose code changes with every execution while its behavior remains consistent, a combination that directly challenges detection architectures built on static signatures and known hashes.

Two distinct models for AI integration have emerged. In the first, criminal threat actors query public or uncensored LLMs to generate ransomware code components outside the malware, then assemble those components into attack tools—the approach facilitated by WormGPT and FraudGPT on underground forums. In the second, the LLM is embedded within or called by the malware at execution time, generating code dynamically so that no two executions produce an identical artifact—the approach demonstrated by PromptLock and theorized in the academic "Ransomware 3.0" research [8]. Both models produce variant proliferation that complicates signature-based detection, as the per-execution hash variation undermines static file matching approaches.

Security Analysis

AI-Embedded Ransomware: How It Works

PromptLock, discovered by ESET Research when samples were uploaded to VirusTotal in August 2025, is among the most thoroughly analyzed public examples of AI-embedded ransomware, given the detail provided in ESET's published research. Written in Golang, PromptLock runs a locally hosted LLM—identified in ESET's analysis as `gpt-oss:20b`, served via the Ollama API—and uses predefined prompts to dynamically generate Lua scripts that determine whether discovered files should be exfiltrated or encrypted [1]. The encryption layer uses SPECK 128-bit encryption. Because the Lua scripts are generated at runtime by the embedded model, the indicators of compromise (IoCs) vary between executions, complicating both threat hunting and incident response workflows that rely on static signatures or file hash matching.

ESET assessed PromptLock as a proof-of-concept associated with academic researchers rather than an active criminal campaign. The significance of the discovery, however, extends beyond the specific sample: it demonstrates that the architectural pattern is viable and deployable. The model integration uses commodity infrastructure—the Ollama API, a widely available open-source LLM serving framework—meaning that the approach is accessible to adversaries without specialized machine learning expertise.

The IBM X-Force identification of Slopoly in early 2026 demonstrated that similar AI-generation techniques had already crossed into live criminal operations. X-Force documented Slopoly as malware deployed by Hive0163, a threat group conducting Interlock ransomware campaigns. Analyzing the malware's code, X-Force assessed that variable naming conventions and structural characteristics suggested intentional malicious design typical of AI-generated code, noting that "any model guardrails, if present, were successfully circumvented" [2]. The malware maintained persistent access to the victim organization for over a week and enabled data exfiltration ahead of the ransomware deployment phase. IBM characterized the finding as "the start of a fundamental shift of dynamics within the threat landscape" [2].

The BlackMamba proof-of-concept published by HYAS in 2023 provides foundational technical context for understanding the polymorphism problem. BlackMamba queries OpenAI at runtime to re-synthesize its malicious keylogging payload on every execution, producing instances that are functionally identical—they perform the same malicious action—but structurally distinct [9]. Every execution presents a different hash. Traditional signature-based defenses identify malware by matching file characteristics to known patterns; when the file characteristics change on every execution while the malicious behavior remains constant, the signature match fails even as the threat succeeds.

The Underground Market for Criminal LLMs

The emergence of purpose-built criminal LLMs has expanded AI-powered ransomware development from a capability requiring significant technical sophistication to one available to a broader range of threat actors. WormGPT, originally built on the open-source GPT-J model and trained on malware-related datasets, is sold on underground forums and functions as a malware template generator. When provided with appropriate prompts, it delivers functional PowerShell scripts for file encryption with configurable parameters [3]. Updated variants have been observed built on xAI's Grok and Mistral's Mixtral models, with development continuing through the latter half of 2025 [3]. FraudGPT operates as a complementary service marketed for phishing automation and malware code generation, reducing the skill requirement for ransomware development further still [4].

These tools represent an extension of the RaaS model's core logic: separating technical development from operational execution to lower barriers to entry for potential affiliates. In the RaaS model, affiliates without malware development expertise can license finished ransomware code from operators. Criminal

LLMs push this further by enabling affiliates to generate customized malware components on demand, producing variants that may evade detection systems tuned to known RaaS families. IBM X-Force noted that its analysts observed adversaries leveraging legitimate GenAI platforms across numerous organizations, injecting malicious prompts to generate credential-stealing commands as a precursor to ransomware deployment—suggesting that commercially available AI tools are now being exploited by criminal actors at meaningful scale [2].

The Polymorphism Detection Gap

The central defensive challenge posed by AI-generated ransomware is the erosion of signature-based detection reliability, as per-execution hash variation neutralizes static file matching. Hash-based identification fails when every sample is unique. Behavioral detection, which identifies malicious activity by what a program does rather than what it looks like, has long been a complement to signature-based approaches; AI-generated polymorphic malware does not eliminate behavioral detection's applicability, but it changes the urgency calculus significantly. Security teams that have relied primarily on signature detection must accelerate the adoption of behavioral approaches to avoid a detection gap.

CrowdStrike's 2026 data indicates that 82% of detections in 2025 were malware-free intrusions—adversaries using living-off-the-land techniques and exploiting legitimate tools rather than deploying easily detected malware binaries [5]. AI-generated ransomware fits into this pattern: when the payload is generated dynamically and does not match any known hash, it operates beneath the threshold of signature-based detection even when it is, technically, a distinct binary artifact. The implication is that defenders relying on behavioral monitoring of file I/O patterns, encryption key generation, network C2 callbacks, and process execution chains—rather than binary signatures—are better positioned to detect AI-generated ransomware than those relying on hash matching alone.

The ENISA Threat Landscape 2025 report documents an additional dimension: the emergence of stand-alone malicious AI systems designed to operate outside easily monitored network paths. ENISA identified Xanthorox AI as an example—a criminal AI platform running locally on servers rather than calling external APIs—an architecture that, whether by design or operational convenience, avoids the network traffic signatures that could alert defenders to LLM-enabled malware activity [10]. The architectural response to detection efforts is itself being automated.

Operational Acceleration and the RaaS Amplification Effect

The attack lifecycle data reported by CrowdStrike—29-minute average breakout times with a 27-second minimum—reflects the broader acceleration of adversary operational tempo that AI tooling may be contributing to, among other factors including credential theft automation and RaaS operational

maturity [5]. AI assistance in reconnaissance, phishing lure personalization, and vulnerability identification compresses the time between initial access and ransomware deployment, reducing the detection window available to defenders. Trend Micro's 2026 threat predictions describe an emerging paradigm in which AI generates hyper-personalized, multilingual phishing lures for initial access and automates the full attack chain via agentic AI frameworks [11].

The RaaS ecosystem amplifies this acceleration because it distributes the benefits of AI tooling across a large pool of affiliates. An operator who integrates AI-powered variant generation into their RaaS platform provides those capabilities to every affiliate using the service, potentially enabling a broader pool of threat actors to deploy detection-evading ransomware regardless of their prior malware development expertise. IBM X-Force's identification of 109 active extortion groups in 2025—a 49% year-over-year increase—reflects the compounding effect of a more capable and more numerous threat actor ecosystem [6].

Recommendations

Immediate Actions

Security operations centers should audit their detection coverage for completeness against behavioral indicators that apply regardless of payload hash. File I/O monitoring for rapid sequential encryption operations, detection of anomalous Windows Shadow Copy deletion commands, network monitoring for C2 callback patterns and unusually large outbound data transfers, and process tree analysis for ransomware-typical execution chains should be validated against current detection rules. Where coverage gaps exist, they should be prioritized given the evidence that AI-generated polymorphism is actively defeating signature-based controls in live campaigns.

Organizations using endpoint protection platforms that rely primarily on signature matching should evaluate the platforms' behavioral detection capabilities against documented AI-generated ransomware tactics. The PromptLock case and IBM's assessment of Slopoly suggest that AI-generated ransomware does not change what malware does—it changes what it looks like. Detection systems that observe behavior rather than match signatures are not defeated by variant proliferation in the same way.

Threat intelligence subscriptions should be reviewed to ensure they include coverage from vendors with documented AI-powered ransomware research, such as ESET, IBM X-Force, and CrowdStrike. Given the pace of developments in this space—from PromptLock's August 2025 discovery to the Slopoly in-the-wild deployment documented in early 2026—point-in-time signature updates are insufficient; continuous threat intelligence on emerging AI-generated families and techniques is necessary.

Short-Term Mitigations

Organizations should accelerate deployment of behavioral detection capabilities that cover the full ransomware attack lifecycle rather than only the encryption phase. AI-generated variants can be detected earlier—during reconnaissance, initial access, lateral movement, or data staging phases—if behavioral monitoring covers those stages. A variant that has never been seen before still traverses the same attack lifecycle as its predecessors; detecting its behavior at earlier stages reduces the blast radius even when signature-based detection fails.

Network segmentation should be reviewed with AI-accelerated breakout times in mind. If a sophisticated adversary can complete lateral movement in minutes rather than hours, network architectures that rely on delayed detection-and-response have reduced effectiveness. Zero Trust network architectures that enforce least-privilege access for every lateral connection—requiring reauthorization rather than assuming that an authenticated internal session is trustworthy throughout its duration—are more resilient against compressed attack timelines.

Security teams should evaluate their backup and recovery architectures against the assumption that ransomware may complete encryption before it is detected. Immutable backups, offline or air-gapped backup copies, and tested recovery procedures that can execute under adversarial conditions—where active ransomware operations may attempt to identify and destroy backup systems—are the primary resilience layer when prevention and detection fall short.

Strategic Considerations

The distinction between signature-based and behavioral detection is not new, but AI-powered variant proliferation has transformed it from a theoretical concern to an operational imperative. Organizations should treat the migration from primarily signature-based to primarily behavioral endpoint and network detection as a strategic priority rather than an optional enhancement. The evidence from PromptLock and IBM's assessment of Slopoly suggests that AI-generated ransomware is already operational in criminal campaigns; the sophistication of those campaigns can be expected to improve as criminal LLM tooling matures.

Threat modeling exercises should explicitly incorporate AI-powered adversary capabilities. Tabletop exercises that assume fixed attack patterns—where a specific ransomware family uses specific tools and leaves specific IoCs—are less valuable than exercises that model adversaries who can generate novel variants and adapt their tooling in response to detection. CSA's Cloud Threat Modeling frameworks provide structured methodologies for this type of exercise and should be applied with the assumption that adversary tooling is now LLM-augmented [13].

Vendor and supply chain risk programs should assess the extent to which third-party software and service providers have hardened their own environments against AI-enhanced ransomware. Given that AI-accelerated breakout times reduce the window for supply chain partners to detect and contain incidents before they propagate, organizations should treat a supply chain partner's detection and response maturity as a direct component of their own resilience posture.

CSA Resource Alignment

The threat patterns described in this research note connect to multiple CSA frameworks and resources that provide guidance for organizations seeking to strengthen their defensive posture.

CSA's **LLM Threats Taxonomy** (2024), produced by the AI Controls Framework Working Group, offers the foundational classification framework for AI-related threats underlying AI-powered ransomware, including Model Manipulation, Insecure Supply Chain, and the abuse of AI systems to generate malicious outputs [14]. The taxonomy's nine primary threat categories and four asset classification dimensions offer a structured basis for assessing AI-related risk components in modern ransomware operations.

CSA's **Agentic AI Red Teaming Guide** (2025) addresses the operational security of AI systems that may be targeted by adversaries or misused in attacks. Organizations that have deployed AI systems internally—including LLM-powered security tooling—should apply red teaming disciplines to ensure those systems cannot be manipulated into generating attacker-useful outputs, a threat class directly relevant given the documented ability of criminal actors to circumvent LLM guardrails in WormGPT and the Slopoly campaign [15].

The **AI Controls Matrix (AICM)** provides a structured control framework applicable to AI system deployments across 18 security domains. AICM controls addressing AI supply chain security, model integrity, and operational monitoring are relevant to organizations seeking to prevent the deployment of AI-enabled malware within their environments, while access control and data protection domains address the data exfiltration precursors documented in the Slopoly campaign.

CSA's **Top Threats to Cloud Computing** research (2025) addresses ransomware as a persistent top-tier cloud threat and maps specific Cloud Controls Matrix (CCM) controls to ransomware mitigation. CCM controls in the Identity and Access Management (IAM) and Threat and Vulnerability Management (TVM) domains are directly applicable: IAM-14 addresses privileged access controls that limit lateral movement during a ransomware campaign, and TVM-02 addresses vulnerability and patch management that reduces the initial access opportunities ransomware operators exploit [16].

CSA's **Protecting Against the Future of Ransomware** guidance emphasizes zero-trust architecture, API access management, and supply chain security documentation as durable controls against evolving ransomware techniques [7]. These controls remain applicable to AI-augmented ransomware: the attack vector evolution does not change the value of eliminating unnecessary trust, limiting lateral movement paths, and maintaining verified and segregated backups.

The broader CSA **AI Safety Initiative** provides the organizational context within which AI-powered ransomware should be assessed. The initiative's work on AI organizational responsibilities, AI risk management frameworks, and agentic AI security is directly relevant to the governance questions that AI-augmented criminal tooling raises: how organizations account for AI capabilities in their threat models, how they evaluate the AI security posture of their vendors, and how they build governance structures capable of adapting to an adversary ecosystem that is itself using AI to evolve at machine speed.

References

[1] ESET Research, "ESET Discovers PromptLock, the First AI-Powered Ransomware," ESET Newsroom, August 27, 2025. https://www.eset.com/us/about/newsroom/research/eset-discovers-promptlock-the-first-ai-powered-ransomware/

[2] Golo Mühr, IBM X-Force, "A Slopoly Start to AI-Enhanced Ransomware Attacks," IBM Security, March 12, 2026. https://www.ibm.com/think/x-force/slopoly-start-ai-enhanced-ransomware-attacks

[3] CSO Online, "WormGPT Returns: New Malicious AI Variants Built on Grok and Mixtral Uncovered," CSO Online, June 18, 2025. https://www.csoonline.com/article/4008912/wormgpt-returns-new-malicious-ai-variants-built-on-grok-and-mixtral-uncovered.html

[4] LevelBlue/SpiderLabs, "WormGPT and FraudGPT: The Rise of Malicious LLMs," LevelBlue, 2024. https://www.levelblue.com/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms

[5] CrowdStrike, "2026 Global Threat Report," CrowdStrike, February 2026. https://www.crowdstrike.com/en-us/global-threat-report/

[6] IBM Security, "IBM 2026 X-Force Threat Intelligence Index," IBM Newsroom, February 25, 2026. https://newsroom.ibm.com/2026-02-25-ibm-2026-x-force-threat-index-ai-driven-attacks-are-escalating-as-basic-security-gaps-leave-enterprises-exposed

[7] Chris Niggel (Okta), "Protecting Against the Future of Ransomware," Cloud Security Alliance SECtember Presentation, 2024.

[8] arXiv, "Ransomware 3.0: Self-Composing and LLM-Orchestrated," arXiv:2508.20444, August 28, 2025. https://arxiv.org/abs/2508.20444

[9] HYAS, "BlackMamba: Using AI to Generate Polymorphic Malware," HYAS Blog, July 31, 2023. https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware

[10] ENISA, "ENISA Threat Landscape 2025," European Union Agency for Cybersecurity, v1.2, January 2026. https://www.enisa.europa.eu/sites/default/files/2026-01/ENISA%20Threat%20Landscape%202025_v1.2.pdf

[11] Trend Micro, "The AI-fication of Cyberthreats: Trend Micro Security Predictions for 2026," Trend Micro, November 2025. https://www.trendmicro.com/vinfo/us/security/research-and-analysis/predictions/the-ai-fication-of-cyberthreats-trend-micro-security-predictions-for-2026

[12] QBE Insurance Group, "Cyber Market Outlook 2026," QBE, December 2025. (Projects publicly named ransomware victims may exceed 7,000 by end of 2026; previously misattributed to Bitsight in earlier drafts. Bitsight historical data available at https://www.bitsight.com/underground/ransomware)

[13] Cloud Security Alliance, "Cloud Threat Modeling 2025," Cloud Security Alliance, 2025.

[14] Siah Burke, Marco Capotondi, Daniele Catteddu, Ken Huang et al. (CSA AI Controls Framework Working Group), "Large Language Model (LLM) Threats Taxonomy," Cloud Security Alliance, 2024.

[15] Ken Huang et al. (CSA AI Organizational Responsibilities Working Group), "Agentic AI Red Teaming Guide," Cloud Security Alliance, 2025.

[16] Jon Michael Brook et al. (CSA Top Threats Working Group), "Top Threats to Cloud Computing Deep Dive 2025," Cloud Security Alliance, 2025.

