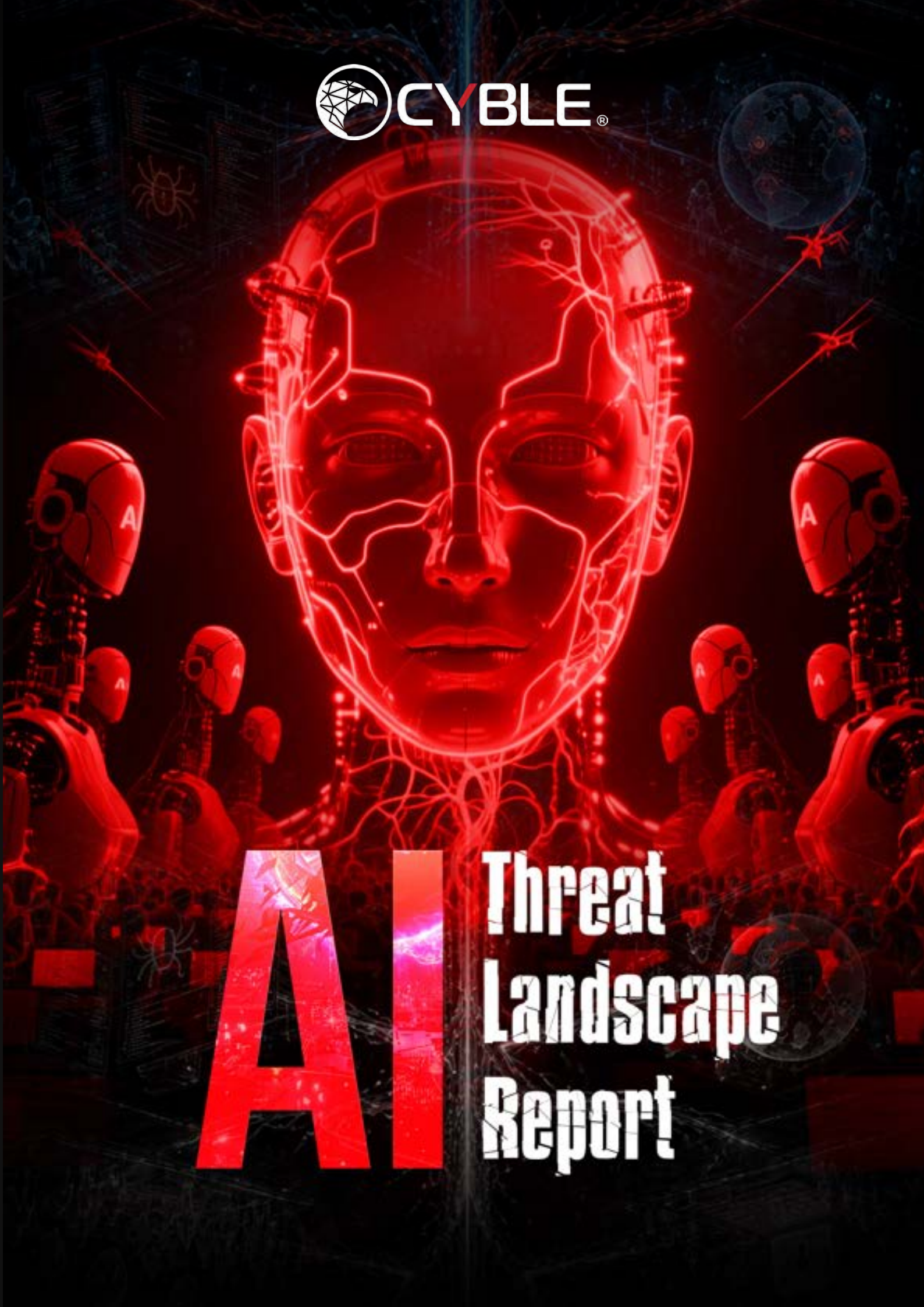

CYBLE
AI
Threat
Landscape
Report

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The cyber threat landscape has entered a new phase, one defined not by the sophistication of individual attackers but by the industrialization of attack operations. AI has moved from an experimental capability into operational infrastructure, and its impact is now visible across every major category of cyber threat.

Threat actors are not reinventing their playbooks; they are compressing the most time-consuming parts. Reconnaissance that once took days, phishing content that once required native-language speakers, malware that once demanded skilled developers, and social engineering that once depended on direct human interaction are all being accelerated by AI tools that are widely accessible, inexpensive, and increasingly indistinguishable from legitimate use.

Across all disrupted operations, no evidence has been found that AI provides threat actors with novel capabilities they could not otherwise obtain, but the acceleration itself has significant consequences for defenders.

This report examines how that shift is playing out across eight interconnected threat areas. AI is being weaponized across the entire attack lifecycle, from AI-generated phishing and deepfake social engineering to LLM-assisted malware development, runtime code generation, and self-rewriting evasion techniques.

Nineteen confirmed threat actor groups, spanning Russian, Chinese, Iranian, North Korean, and financially motivated criminal operators, have been observed integrating AI into their campaigns, using it for reconnaissance, lure generation, vulnerability research, malware development, and trusted-service abuse.

At the same time, the AI supply chain has become an active attack surface, with threat actors poisoning model repositories, compromising CI/CD pipelines, manipulating agent skill marketplaces, and engineering malicious packages specifically designed to deceive AI coding assistants rather than human developers.

AI systems themselves are now primary targets. The vulnerabilities being exploited, prompt injection, jailbreaking, data poisoning, memory manipulation, agentic exploitation, RAG poisoning, and API abuse, do not require exploiting traditional software flaws.

They require understanding how a model processes instructions, retrieves information, stores context, and interacts with connected tools. In parallel, threat actors are exploiting the trust users place in AI brands, impersonating ChatGPT, Claude, DeepSeek, and Gemini through fake installer sites, malicious browser extensions, and shared conversation links hosted on trusted AI platform domains.

Underpinning all of this is a structured underground AI marketplace where purpose-built malicious models, jailbroken commercial wrappers, and specialist attack-as-a-service platforms are sold through dark web forums, Telegram channels, and private networks, using subscription models, lifetime licenses, and one-time flat fees that make AI-assisted attack capability accessible to any operator regardless of technical skill.

What started as a handful of experimental tools has matured into a criminal ecosystem that mirrors the commercial mechanics of legitimate SaaS, with versioned product updates, tiered pricing, and dedicated support communities.

The convergence of these trends defines the current threat environment. AI is not simply adding new tools to attacker playbooks; it is connecting and accelerating the entire operation.

Organizations that have not yet updated their security strategies to account for AI-enabled threats face a landscape where the quality floor of attacks has risen sharply. The technical barrier to entry has dropped significantly, and the defenders' traditional detection signals, such as poor grammar, known malware signatures, and recognizable attack infrastructure, are becoming unreliable.

An effective response requires a fundamental rethink of how security teams detect threats, validate identities, govern AI-integrated workflows, and build resilience against attack chains that combine multiple AI-enabled techniques in a single operation.

AI PHISHING & SOCIAL ENGINEERING

Phishing has long been the most common and costly entry point for cyberattacks. AI has not changed that reality; it has made it significantly worse. What AI introduces is not a new type of attack but a dramatic improvement in the quality, speed, and scale at which existing attacks can be executed.

Defenders that once relied on spotting poor grammar, awkward phrasing, or generic templates now face campaigns that are polished, contextually aware, and personalized at scale.

The scale of AI's impact on phishing is no longer speculative. Data from across the industry paints a consistent picture of acceleration:

AI PHISHING BY THE NUMBERS

Key indicators showing the rapid growth and effectiveness of AI-driven phishing

14X SURGE

In December 2025, AI-generated phishing attacks surged 14x, with their share of all reported attacks rising from 4% to 56% over the holiday season - a trend that continued into 2026.

24-54% MORE EFFECTIVE

AI phishing is now 24-54% more effective than traditional human-crafted methods, with AI-generated emails achieving a 54% click-through rate compared to 12% for manual

82.6%

82.6% of phishing emails now contain AI-generated elements, showing how deeply automated tooling has become embedded in attack workflows.

1,265% SPIKE

Phishing attacks have increased by 1,265% since ChatGPT's launch in late 2022, marking one of the sharpest escalations in the history of the

AI is reshaping phishing at scale - faster, more convincing, and far more effective.

- In December 2025, AI-generated phishing attacks surged 14x, with their share of all reported attacks jumping from 4% to 56% over the holiday season, a trend that held into 2026.
- AI phishing is now 24–54% more effective than traditional human-crafted methods, with AI-generated emails achieving a 54% click-through rate compared to just 12% for manual attempts.
- 82.6% of phishing emails now contain AI-generated elements, reflecting how deeply automated tooling has embedded itself into attack workflows.
- Phishing attacks have spiked 1,265% since ChatGPT's launch in late 2022, marking one of the sharpest escalations in the history of the threat.

HOW THREAT ACTORS ARE USING AI IN PHISHING & SOCIAL ENGINEERING ATTACKS

AI is enhancing phishing, social engineering, and attack preparation

1. Reconnaissance and target development

Threat actors use LLMs to rapidly synthesize open-source intelligence, profile high-value targets, identify key decision-makers, and map organizational hierarchies. This research helps attackers build phishing lures using job titles, project names, reporting relationships, and realistic communication styles, compressing days of manual work into hours.

2. Lure generation and personalization

LLMs help generate hyper-personalized, culturally nuanced lures that mirror the tone of a target organization or local language. This reduces the grammatical and contextual errors defenders have historically relied on to spot phishing attempts.

3. AI-generated phishing kits

In November 2025, a phishing kit tracked as COINBAIT was identified, with development likely accelerated by AI code-generation tools. It impersonated a major cryptocurrency exchange and used a polished React single-page application built with Lovable AI to harvest credentials.

4. Deepfake-enabled social engineering

Threat actors are cloning executive voices and video footage to support fraud and targeted intrusions. In one February 2024 case, deepfake impersonation contributed to a $25.6 million fraud, showing how fake video calls and AI-generated personas are becoming operational tools.

Reconnaissance and target development

AI is transforming the preparation phase that makes phishing attacks credible in the first place. State-sponsored actors are using LLMs to rapidly synthesize open-source intelligence, profile high-value targets, identify key decision-makers within defense sectors, and map organizational hierarchies.

The output of that research feeds directly into phishing lures, personalized with job titles, internal project names, reporting relationships, and communication styles that make messages appear to come from a known and trusted source rather than an unknown attacker.

What once took a team days of manual research can now be completed by a single operator in hours, before a single phishing email is sent.

Lure generation and personalization

LLMs allow threat actors to generate hyper-personalized, culturally nuanced lures that mirror the professional tone of a target organization or local language, largely erasing the grammatical and contextual tells that defenders have historically relied on to identify phishing attempts.

AI-generated phishing kits

In November 2025, a phishing kit, tracked as COINBAIT, was identified, with its construction likely accelerated by AI code-generation tools. The kit masqueraded as a major cryptocurrency exchange to harvest credentials. It was built using the AI-powered platform Lovable AI, wrapping the operation in a full React single-page application with complex state management, a level of sophistication indicative of code generated from high-level prompts.

Deepfake-enabled social engineering

Threat actors have cloned video footage and voice recordings of executives to convince employees to transfer substantial funds, with one incident in February 2024 resulting in a $25.6 million fraud. Deepfake use is now moving beyond isolated incidents to active campaign infrastructure, with threat actors deploying fake video calls and AI-generated personas as standard components of targeted intrusions.

AI Phishing Is Outpacing Defender Capacity

Beyond enabling faster campaign execution, AI-powered phishing is creating a secondary crisis inside Security Operations Centers, one that compounds the threat by degrading the very teams responsible for catching it.

- AI-generated lures now arrive in high volumes with enough variation to appear unique, eliminating the ability to identify and dismiss similar alerts in bulk and forcing individual review of every case.
- Polished impersonation and contextual personalization have removed the quick visual judgments analysts once relied on, making each alert more time-consuming to investigate than before.
- Short-lived infrastructure and freshly registered domains return inconclusive results from reputation-based tools, pushing more of the analytical burden onto human reviewers.
- As the volume of uncertain cases grows, experienced analysts are pulled away from complex, high-risk threats to revisit alerts that were never fully resolved.
- High-risk incidents, credential theft, malware delivery, and targeted intrusions increasingly get buried in the noise, widening the window between initial access and detection.
- The asymmetry is structural: AI reduces the cost and effort of launching attacks while simultaneously raising the cost and effort of defending against them, a gap that widens with every improvement in attacker tooling.

AI-POWERED PHISHING IS CREATING A SECONDARY CRISIS INSIDE SECURITY OPERATIONS CENTERS BY INCREASING ANALYST WORKLOAD AND SLOWING DETECTION.

1 **High-Volume Variation**

AI-generated lures now arrive in high volumes with enough variation to appear unique, preventing analysts from identifying and dismissing similar alerts in bulk and forcing individual review of every case.

2 **Slower Triage**

Polished impersonation and contextual personalization have removed the quick visual judgments analysts once relied on, making each alert more time-consuming to investigate than before.

3 **Weak Reputation Signals**

Short-lived infrastructure and freshly registered domains return inconclusive results from reputation-based tools, pushing more of the analytical burden onto human reviewers.

4 **Analyst Overload**

As the volume of uncertain cases grows, experienced analysts are pulled away from complex, high-risk threats to revisit alerts that were never fully resolved.

5 **High-Risk Incidents Buried**

Credential theft, malware delivery, and targeted intrusions increasingly get buried in the noise, widening the window between initial access and detection.

6 **Structural Asymmetry**

AI reduces the cost and effort of launching attacks while simultaneously raising the cost and effort of defending against them, a gap that widens with every improvement in attacker tooling.

NOTABLE ATTACK CASES

Examples of how threat actors are using AI in phishing and social engineering campaigns

1. **UNC1069 — Cryptocurrency Sector Targeting**
   A North Korean threat actor compromised a Telegram account belonging to a cryptocurrency executive, built rapport with the victim, and sent a Calendly link to a spoofed Zoom meeting. During the call, the victim saw what appeared to be a deepfake CEO and was guided through a ClickFix attack that launched a multi-stage infection chain, ultimately deploying seven malware families to steal credentials, browser data, session tokens, and cryptocurrency wallet information.

2. **APT42 — Iranian State-Sponsored Phishing Augmentation**
   APT42 used generative AI models, including Gemini, to enhance reconnaissance and targeted social engineering. The group used AI to search for official email addresses, research potential business partners, and create personas tailored to each target's biography to improve engagement.

3. **FAMOUS CHOLLIMA — Fake Linkedin Profiles and Job Interviews**
   The DPRK-associated group FAMOUS CHOLLIMA used GenAI to create realistic Linkedin profiles with believable backgrounds and AI-generated profile images to deceive recruiters, and potentially used AI to generate plausible responses during job interviews as part of efforts to infiltrate private corporations.

4. **ClickFix campaign via AI Platforms**
   In December 2025, threat actors abused the public sharing features of generative AI services such as ChatGPT, Copilot, DeepSeek, Gemini, and Grok to host deceptive social engineering content. Attackers manipulated AI tools into generating seemingly legitimate troubleshooting instructions containing malicious commands, then promoted trusted-domain share links through malicious advertising to distribute information-stealing malware targeting Windows and macOS.

5. **UTA0388 — China-Aligned Spear Phishing with LLM Assistance** UTA0388 targeted organizations across North America, Asia, and Europe with phishing emails impersonating senior researchers from fabricated institutions. The campaign delivered a malicious archive containing the GOVERSHELL backdoor. OpenAI later confirmed that the group used ChatGPT to help craft phishing emails and assist in malware development. As the campaign matured, the actor shifted to rapport-building phishing through multi-email conversations before delivering a malicious link.

6. **AI-Assisted Biometric Phishing - Browser Permission Abuse** Active since early 2026, this campaign used lures such as "ID Scanner," "Telegram ID Freezing," and "Health Fund AI" to trick victims into granting browser camera and microphone access. Once approved, malicious scripts captured live images, video, audio, device details, contact information, and approximate location, then exfiltrated the data to attacker-controlled Telegram bots. Emoji-based formatting and structured annotations in the code suggest generative AI may have been used to accelerate script development.

7. **InboxPrime AI - Phishing-as-a-Service Kit**
   First observed in October 2025, InboxPrime AI is a phishing kit built for underground cybercrime networks. Its AI-powered email generator creates phishing emails from simple inputs such as topic, tone, language, and industry. It uses spintax to vary messages, includes a built-in spam checker to reduce detection, and automates sender identity rotation across compromised Gmail accounts to imitate legitimate users, vendors, or executives.

Key takeaway: AI is being used to improve targeting, scale deception, and increase the credibility of social engineering operations.

UNC1069: Targeting the cryptocurrency sector

North Korean threat actor UNC1069 compromised a Telegram account belonging to a cryptocurrency executive. It used it to build rapport with a victim before sending a Calendly link to a spoofed Zoom meeting hosted on attacker-controlled infrastructure. During the call, the victim was presented with what appeared to be a deepfake video of a CEO from another cryptocurrency company.

The threat actor then executed a ClickFix attack, directing the victim to run troubleshooting commands that initiated a multi-stage infection chain, ultimately deploying seven distinct malware families designed to harvest credentials, browser data, session tokens, and cryptocurrency wallet information.

APT42: Iranian state-sponsored phishing augmentation

The Iranian government-backed group APT42 leveraged generative AI models, including Gemini, to significantly augment reconnaissance and targeted social engineering, using AI to search for official email addresses of specific entities, research potential business partners to establish credible pretexts, and craft personas tailored to each target's biography to maximize engagement.

FAMOUS CHOLLIMA: Fake LinkedIn profiles and job interviews

DPRK-associated group FAMOUS CHOLLIMA used GenAI to create realistic LinkedIn profiles complete with believable backgrounds and AI-generated profile images to deceive recruiters, and potentially leveraged AI to generate plausible responses during job interviews as part of a campaign to infiltrate private corporations.

ClickFix campaigns via AI platforms

In December 2025, threat actors began abusing the public-sharing features of generative AI services, including ChatGPT, Copilot, DeepSeek, Gemini, and Grok, to host deceptive social engineering content. By manipulating AI tools into generating seemingly legitimate troubleshooting instructions with embedded malicious commands, attackers created shareable links hosted on trusted AI platform domains, which were then promoted through malicious advertising to distribute information-stealing malware targeting both Windows and macOS systems.

UTA0388: China-aligned spear phishing with LLM assistance

UTA0388 is a China-aligned threat actor targeting organizations across North America, Asia, and Europe. Emails impersonated senior researchers from fabricated institutions, socially engineering targets into clicking links that led to a malicious archive containing the GOVERSHELL backdoor. OpenAI later confirmed that UTA0388 leveraged ChatGPT to craft phishing emails and assist in malware development.

As the campaign matured, the group shifted to rapport-building phishing, establishing benign multi-email conversations with targets before introducing a malicious link. Signs of unsupervised LLM usage were evident throughout: emails simultaneously used three different identities across the friendly name, sender address, and signature block; fabricated phone numbers followed sequential patterns; and some targets received emails entirely in the wrong language.

AI-assisted biometric phishing: Browser permission abuse

Active since early 2026, a widespread social engineering campaign lured victims through themes such as “ID Scanner,” “Telegram ID Freezing,” and “Health Fund AI”, tricking users into granting browser-level camera and microphone access under the guise of identity verification or account recovery.

Once permissions were granted, malicious scripts silently collected live images, video recordings, audio, device specifications, contact details, and approximate location, exfiltrating everything to attacker-controlled Telegram bots. Analysis of the campaign’s code revealed emoji-based formatting and structured annotations embedded within the operational logic, patterns increasingly associated with generative AI being used to accelerate malicious script development.

InboxPrime AI: Phishing-as-a-service kit

First observed in October 2025, InboxPrime AI is a purpose-built phishing kit rapidly gaining adoption across underground cybercrime networks. At its core is an AI-powered email generator that produces fully composed phishing emails from a handful of dropdown inputs, topic, tone, language, and industry, eliminating the need for any copywriting skill.

The AI engine also applies spintax variations, so no two recipients receive identical messages, actively evading signature-based filters. A built-in spam checker then analyzes each generated email and suggests corrections to remove detection triggers before sending. On top of this, the kit automates sender identity rotation across compromised Gmail accounts, mimicking the identities of legitimate users, vendors, or executives with high fidelity.

The convergence of AI-generated lures, deepfake social engineering, and AI-built phishing infrastructure creates an environment where traditional detection signals are no longer reliable. Campaigns that once required days of preparation and left visible traces of error can now be produced in hours by a single operator.

Recommendations

- Enforce hardware-based MFA (FIDO2/passkeys) on all critical systems; stolen credentials alone become useless without the physical key
- Never trust voice or video calls as identity verification; always confirm high-stakes requests like fund transfers through a separate, pre-established contact method
- Retrain staff to understand that polished writing, familiar faces, and trusted platforms (Zoom, Calendly, AI tools) are no longer reliable signals of legitimacy
- Enforce DMARC, DKIM, and SPF on all domains, and replace signature-based email filters with behavioral AI-driven detection that doesn’t rely on spotting errors
- Require dual authorization and out-of-band confirmation for any financial or access-change request, regardless of how legitimate the source appears
- Audit and minimize your organization’s public footprint on LinkedIn and company websites; AI reconnaissance tools harvest this data to craft personalized, convincing lures

AI-assisted Malware Attacks

Malware operations are moving into a more adaptive phase of AI abuse. Earlier, attackers mostly used AI to write phishing lures, translate content, generate simple scripts, or debug commodity malware. Recent reporting shows a clear shift, from AI as a productivity aid to AI as an operational component embedded directly inside malware workflows.

This includes malware that queries LLMs during execution to generate commands, malware whose codebase was largely generated or heavily accelerated by AI, backdoors that abuse AI platforms as command-and-control channels, and mobile malware that uses generative AI to interpret device screens and decide how to manipulate the user interface.

This does not mean every AI-assisted malware family is fully autonomous or entirely AI-generated. In most observed cases, human operators still define the objective, supply prompts, integrate outputs, test functionality, and control deployment.

The important shift is that AI is reducing the time, skill, and manpower needed to build or adapt malware, and it allows malicious logic to be generated at runtime rather than stored statically in the binary, creating new challenges for detection teams that traditionally depend on signatures, static strings, known functions, and predictable behavior.

The ways AI is being embedded into malware fall into several distinct patterns:

- **Malware code generation and rapid development**

  AI is being used to accelerate malware development by generating boilerplate code, implementing modules, building loaders, creating command handlers, and adding evasion routines, producing large codebases faster than a small team could traditionally manage. In the most advanced cases observed, actors appear to use AI not just for coding but for planning the malware architecture itself, including task breakdowns, module specifications, sprint schedules, and implementation guidance, compressing what previously required coordinated teams into a workflow manageable by a single operator.

- **Runtime command and code generation**

  Some malware now queries an LLM during execution to generate commands or scripts only when needed. This reduces the amount of malicious logic stored inside the sample and undermines static detection. Rather than embedding every command or function in the payload, the malware requests generated system commands, scripts, or task-specific logic at runtime, making each execution potentially unique.

- **AI-assisted obfuscation and self-rewriting code**

  AI is being used to rewrite malware source code, generate decoy logic, insert junk code, and alter binaries just in time to evade static signatures. This allows attackers to produce frequently changing variants without manually rewriting the malware, complicating hash-based detection, YARA rule coverage, and static analysis pipelines.

- **AI as command-and-control infrastructure**

  Threat actors are abusing legitimate AI platforms as communication channels. In these cases, the AI service is not generating malware code; it is being abused as trusted infrastructure to store, retrieve, or relay commands between the operator and the compromised host. This blends malicious C2 traffic with legitimate API usage that most network controls are not designed to flag.

- **Context-aware UI manipulation on infected devices**

  Mobile malware is beginning to use generative AI to adaptively interact with devices. Rather than relying on hardcoded UI paths, the malware queries an AI model to interpret on-screen elements and receive step-by-step instructions for navigating Android interfaces, maintaining persistence, or blocking removal, adapting dynamically across different device layouts and OS versions.

- **AI-themed lures and fake AI applications**

  Threat actors are exploiting the popularity of AI tools by disguising malware as AI-enhanced productivity software, automation utilities, or AI-based services. AI is used to generate convincing application interfaces, produce legitimate-looking code, improve documentation, and build professional-looking installers that increase user trust and bypass scrutiny.

NOTABLE MALWARE CASES

AI-ENABLED MALWARE

Examples of how malware is leveraging AI for command generation, evasion, adaptation, and covert operations

LAMEHUG / PROMPTSTEAL

Uses the Qwen 2.5-Coder-32B-Instruct model through the Hugging Face API to turn natural-language attack goals into Windows commands for system discovery, document collection, and data exfiltration.

MalTerminal

Uses GPT-4 at runtime to dynamically generate ransomware code or a reverse shell based on the operator's selected objective.

PROMPTFLUX

Uses a hardcoded Gemini API key and a "Thinking Robot" module to query the LLM hourly and rewrite its VBScript source code to evade static signature detection.

PROMPTLOCK

Uses an LLM at runtime to generate and execute Lua scripts for reconnaissance, data exfiltration, and file encryption across Windows and Linux.

EvilAI

Relies on AI-generated code to build malware disguised as legitimate productivity or AI-enhanced applications, packaged with polished interfaces and valid digital signatures.

VoidLink

Assessed as largely AI-generated using spec-driven development, with AI used to create build plans, sprint tasks, code iterations, and testing for a modular Linux C2 framework.

NGate

A legitimate Android NFC payment app trojanized with AI-generated malicious code, adding NFC card relay and PIN theft functions; emoji-style debug logs suggest AI-assisted development.

PromptSpy

Uses Google Gemini to analyze the live Android device screen and generate step-by-step UI instructions, helping the malware adapt across different device layouts and OS versions.

SesameOp

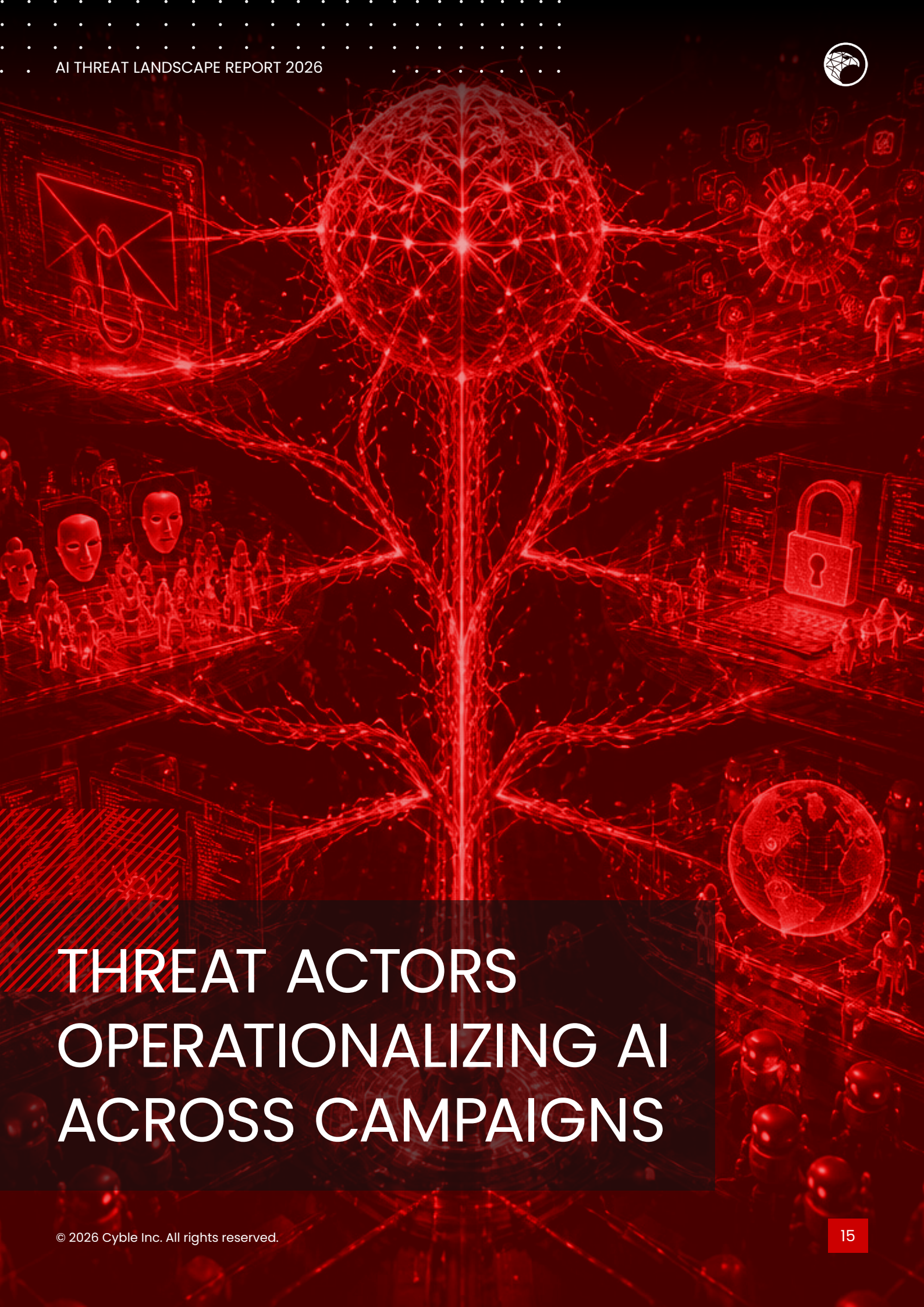
Abuses the OpenAI Assistants API as a covert command-and-control relay, retrieving encrypted commands from API threads, executing them locally, and posting results back.

Malware	How AI Was Used
LAMEHUG / PROMPTSTEAL	Uses the Qwen 2.5-Coder-32B-Instruct LLM via the HuggingFace API to generate Windows commands from natural-language descriptions of attack steps, objectives, collecting system information, harvesting Office and PDF documents, and exfiltrating data.
MalTerminal	Uses OpenAI GPT-4 to dynamically generate ransomware code or a reverse shell at runtime based on operator selection
PROMPTFLUX	Uses a hardcoded Gemini API key to query the LLM hourly via a "Thinking Robot" module, instructing it to rewrite the malware's entire VBScript source code to evade static signature-based detection
PROMPTLOCK	Uses an LLM to dynamically generate and execute Lua scripts for reconnaissance, data exfiltration, and file encryption across Windows and Linux at runtime
EvilAI	Uses AI-generated code to build malware disguised as legitimate productivity or AI-enhanced applications with professional interfaces and valid digital signatures, bundled alongside functional software to avoid suspicion
VoidLink	Assessed as largely AI-generated using Spec Driven Development, AI tasked to produce a structured build plan with sprint schedules and specifications, then used to implement, iterate, and test a modular Linux C2 framework end-to-end, reaching a functional implant in under a week.
NGate	Legitimate Android NFC payment app trojanized with AI-generated malicious code, evidenced by emoji in debug logs typical of AI output, adding NFC card data relay and PIN theft capabilities
PromptSpy	Uses Google Gemini to analyze the live Android device screen and generate step-by-step UI instructions to maintain persistence, adapting dynamically across different device layouts and OS versions
SesameOp	Abuses the OpenAI Assistants API as a covert C2 relay, fetching encrypted commands from OpenAI threads, executing them locally, and posting results back, blending malicious traffic with legitimate API usage

The shift from AI-assisted to AI-integrated malware carries significant implications for detection, attribution, and defense. Signature-based defenses fail when malicious logic is generated at runtime. Network monitoring struggles when C2 traffic routes through trusted AI platforms. Attribution becomes harder as AI-generated code lacks the stylistic fingerprints of a human developer, and single operators can now build malware frameworks that previously required coordinated teams.

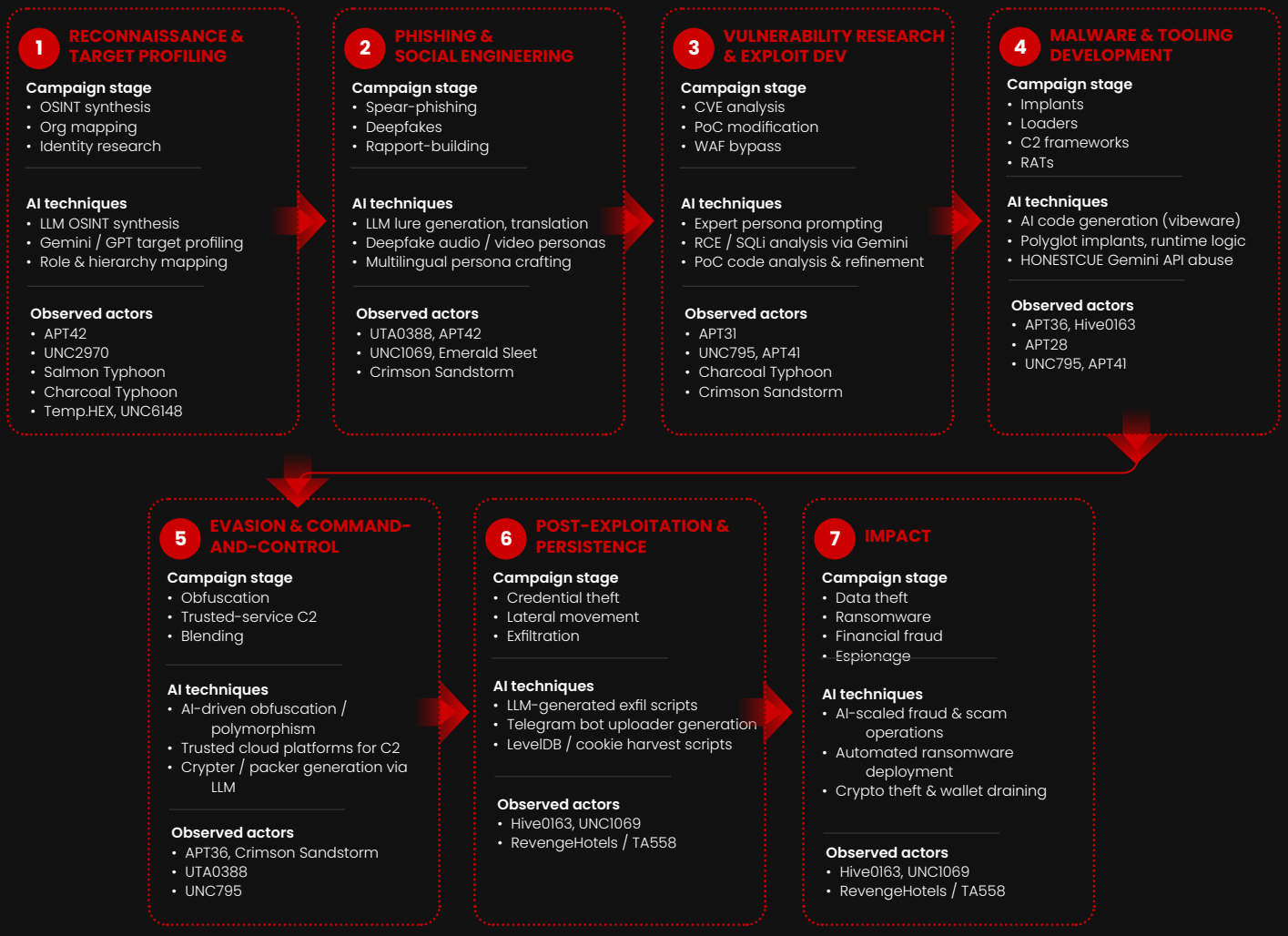
Recommendations

- Shift from signature-based to behavior-based detection, focusing on what malware does rather than what it looks like
- Monitor for anomalous outbound connections to LLM API endpoints, including OpenAI, HuggingFace, Google Gemini, and equivalents
- Enforce script integrity controls such as code signing and file integrity monitoring for scripting environments
- Establish AI API guardrails by allowing only approved teams, users, devices, and applications to access AI services; investigate traffic from non-engineering teams, unmanaged endpoints, unknown apps, or systems with no business need
- Treat AI-generated malware as a growing baseline threat rather than an edge case; the tooling, techniques, and evidence base are now well-established across financially motivated, state-sponsored, and opportunistic threat actors



THREAT ACTORS OPERATIONALIZING AI ACROSS CAMPAIGNS

CAMPAIGN STAGES, AI TECHNIQUES, AND OBSERVED ACTORS BY STAGE



Threat actors are increasingly using AI as an operational accelerator rather than as a completely new attack class. The clearest shift visible across observed reporting is that AI is reducing the time and skill required to conduct high-quality operations.

State-sponsored actors, financially motivated criminals, and loosely organized groups have all been observed integrating AI into their workflows to move faster, target more precisely, and operate at greater scale than their traditional tradecraft allowed.

Threat Actors are using AI to compress the most time-consuming parts of their work. Before the widespread use of AI, Threat Actors had to conduct detailed and time-consuming reconnaissance. Phishing content that once needed native speakers and malware that required skilled developers is now sped up by AI tools that are widely available, cheap, and harder to tell apart from legitimate content.

Threat actors are also growing more sophisticated in how they use AI operationally, combining multiple models across different tasks, adapting their output to remove detectable signs of AI usage, and increasingly using AI not just for a single step but across the entire campaign lifecycle. Importantly, across all disrupted operations, no evidence has been found that AI provided threat actors with novel offensive capabilities they could not otherwise have obtained from publicly available resources. AI is an accelerant, not a capability leap, but the acceleration itself has significant consequences for defenders.

The following patterns represent the primary ways AI is being operationalized across observed campaigns:

Reconnaissance and target profiling: Actors use LLMs to synthesize open-source intelligence, identify organizations and individuals of interest, map roles and hierarchies, research experts, and build target-specific context for phishing or espionage operations. Tasks that previously required hours of manual research can now be completed in minutes, enabling more precise targeting at scale.

Social engineering and phishing lure generation: AI is being used to draft persuasive phishing emails, improve tone and grammar, translate content into multiple languages, create believable personas, and sustain rapport-building conversations across multiple turns before delivering a malicious payload. This is especially impactful for actors operating outside their native-language environment, where AI eliminates the language barriers that previously limited their geographic reach.

Vulnerability research and exploit development: Actors are using LLMs to research publicly disclosed vulnerabilities, understand proof-of-concept exploit code, identify weaknesses in specific technologies, and prototype exploit logic against target systems. APT31 was observed directing Gemini to analyze RCE and WAF bypass techniques against specific US-based targets, explicitly blurring the line between routine security assessment queries and targeted malicious reconnaissance.

Scripting and tooling support: Threat actors are using LLMs to write, debug, and refine scripts for automation, file manipulation, web scraping, command execution, data handling, and operational infrastructure, spanning from basic file operations to complex Windows API interactions and credential-handling routines.

Malware and C2 development: AI is increasingly embedded in malware development workflows beyond simple script assistance. Actors are using AI-assisted development to produce high-volume disposable implants across niche languages, generate functional C2 frameworks, and compress development timelines that previously required coordinated teams down to the work of a single operator. GREYVIBE's custom toolchain, including LegionRelay and associated obfuscators, was assessed with moderate confidence to have been developed with LLM assistance.

Evasion and detection bypass: Actors are experimenting with AI to understand defensive controls, rewrite code to reduce signature-based detection, and develop alternate implementations of known techniques. This includes researching antivirus bypass methods and generating functional code in niche programming languages specifically to reset the detection baseline for security engines unfamiliar with those languages.

Trusted-service abuse and operational blending: AI-assisted tooling increasingly uses legitimate cloud and communication platforms for C2 or exfiltration, blending malicious traffic with normal service activity that most network controls are not designed to flag. Actors leverage LLMs' strong coverage of public SDK documentation to quickly generate stable, functional integration code for platforms such as Slack, Discord, Google Sheets, and cloud storage services with minimal expertise.

Notable Threat Actors Using AI in Their Campaigns

Threat Actor	Description	Use of AI	MITRE ATT&CK Techniques
Forest Blizzard / APT28	Russian GRU-linked actor targeting government, defense, logistics, energy, and NGOs.	Used LLMs for reconnaissance into satellite and radar technologies and for basic scripting tasks, including file manipulation, data selection, regular expressions, and multiprocessing. Activity assessed as early-stage exploration rather than deep operational integration.	T1593, T1059, T1059.001
GREYVIBE	Russia-nexus group targeting Ukrainian military, government, civilian, and business entities since August 2025, assessed to have ties to the broader cybercrime ecosystem.	Demonstrated systematic use of generative AI across all stages of operations, from crafting spear-phishing lures and malicious scripts to full malware development and setting up backend infrastructure. Custom toolchain, including LegionRelay and associated obfuscators, assessed with moderate confidence to have been developed with LLM assistance.	T1566, T1059, T1027, T1105, T1071
Emerald Sleet / Kimsuky	A North Korean actor focused on spear-phishing and intelligence collection targeting experts on North Korea.	Used LLMs to research think tanks and experts, understand publicly known vulnerabilities, troubleshoot web technologies, perform scripting tasks, and generate content likely used in spear-phishing campaigns.	T1593, T1598, T1566, T1059
Crimson Sandstorm / CURIUM	Iranian IRGC-linked actor using watering-hole attacks and social engineering to deliver custom .NET malware.	Used LLMs to generate phishing emails, develop .NET and web code, and research antivirus bypass techniques, including disabling antivirus through registry and Windows policy modifications, and deleting files in directories after application close.	T1566, T1189, T1059, T1027, T1562.001

Charcoal Typhoon / CHROMIUM	China-linked actor targeting government, higher education, communications infrastructure, oil and gas, and IT sectors.	Used LLMs to research technologies, platforms, and vulnerabilities; generate and refine scripts; support translations; and refine operational commands. Activity reflects limited exploration of LLM capabilities rather than full operational integration.	T1595, T1593, T1059, T1105
Salmon Typhoon / SODIUM	China-linked actor targeting US defense contractors, government agencies, and cryptographic technology entities.	Used LLMs for broad intelligence gathering, research on sensitive geopolitical topics and high-profile individuals, code-error troubleshooting, technical translation, and research into file concealment and operational command execution.	T1593, T1059, T1036, T1083
APT31 / Judgement Panda	China-linked actor conducting targeted cyber espionage against government and critical infrastructure with confirmed US targeting.	Employed a highly structured approach by prompting Gemini with an expert cybersecurity persona to automate vulnerability analysis and generate targeted testing plans, directing the model to analyze RCE, WAF bypass techniques, and SQL injection test results against specific US-based targets. Trend Micro	T1595, T1593, T1190, T1059
UNC795	China-based actor integrating Gemini across the entire attack lifecycle	Relied heavily on Gemini throughout their entire attack lifecycle, using it multiple days a week to troubleshoot code, conduct research, develop web shells and PHP scanners, and build an AI-integrated code auditing capability.	T1595, T1059, T1505.003, T1027
APT41	China-linked actor conducting both espionage and financially motivated intrusions globally.	Leveraged Gemini to accelerate development and deployment of malicious tooling, including knowledge synthesis, real-time troubleshooting, and code translation.	T1059, T1105, T1027, T1588
Temp.HEX / Mustang Panda	A China-based actor conducting intelligence-collection operations against targets in Asia and against separatist organizations globally.	Used Gemini to compile information on individuals, including targets in Pakistan, and to collect data on separatist organizations across multiple countries. Similar targets later appeared in an active campaign, linking the AI-assisted research to real-world operations.	T1593, T1589, T1598

APT42	Iranian government-backed actor known for targeted credential theft, persona-based phishing, and social engineering.	Used Gemini to augment reconnaissance, identify official email addresses, research potential business partners, craft target-specific personas, translate content, support debugging, and develop specialized malicious tools. Also attempted to build a data processing agent that converts natural language to SQL to analyze PII, including asset ownership, location, and travel patterns.	T1593, T1589, T1566, T1587
UNC2970	North Korean actor focused on defense-sector targeting and recruiter-themed social engineering, overlapping with Lazarus Group and Operation Dream Job.	Used Gemini to synthesize OSINT, profile high-value targets in defense and cybersecurity sectors, research major companies and technical roles, and gather salary and job context to craft convincing recruitment lures.	T1593, T1589, T1566, T1585
APT36 / Transparent Tribe	Pakistan-linked actor targeting Indian government, diplomatic, defense, and South Asian policy entities.	Pivoted to AI-assisted vibeware development, producing high-volume disposable implants in niche languages including Nim, Zig, Crystal, and Rust. Also abused Slack, Discord, Supabase, Firebase, Google Sheets, and Google Drive for C2 and exfiltration using LLM-generated SDK integration code.	T1059, T1102, T1567.002, T1053.005, T1555.003
Hive0163	Financially motivated ransomware cluster associated with Interlock ransomware and large-scale extortion operations.	Deployed likely LLM-generated Slopoly PowerShell C2 client during a ransomware intrusion, exhibiting AI code patterns including extensive comments, accurate variable naming, error handling, and unused functions. Model guardrails were successfully circumvented.	T1059.001, T1053.005, T1071.001, T1105, T1486
RevengeHotels / TA558	Cybercriminal group active since 2015, focused on stealing payment-card data from hotels and travelers in Latin America.	Used LLM-generated JavaScript loader and PowerShell downloader code in phishing campaigns delivering VenomRAT across Brazilian and Spanish-speaking Latin American hotels, enabling geographic expansion without native-language operators.	T1566.002, T1059.007, T1059.001, T1105, T1219

UTA0388 / UNK_DropPitch	China-aligned actor targeting North America, Asia, and Europe with interest in Taiwan-related and semiconductor sector targets.	Used ChatGPT to craft multilingual spear-phishing emails across English, Mandarin, German, French, and Japanese, impersonating fabricated researchers. Showed signs of unsupervised LLM usage, including inconsistent identities, misdirected emails, and rapport-building sequences before malware delivery.	T1566.002, T1204.002, T1574.001, T1053.005, T1059.001
UNC1069	North Korea-linked actor targeting cryptocurrency and Web3 entities for financial theft.	Used AI-enabled social engineering via compromised Telegram accounts, Calendly lures, spoofed Zoom infrastructure, and a deepfake video of a cryptocurrency CEO, leading victims into ClickFix commands that deployed seven malware families harvesting credentials, session tokens, and wallet data.	T1566, T1204.004, T1059.004, T1059.003, T1059.007, T1555.003
APT45 / Andariel / Onyx Sleet	North Korean state-linked group focused on cyberespionage and financially motivated operations, associated with DPRK's Reconnaissance General Bureau.	Sent thousands of repetitive prompts to Gemini to recursively analyze different CVEs and validate proof-of-concept exploits, building what GTIG described as a more robust arsenal of exploit capabilities that would be impractical to manage without AI assistance. Built custom agentic frameworks to automate vulnerability scanning and CVE analysis at an industrial scale.	T1595, T1588.005, T1190, T1059
APT27 / Emissary Panda	China-linked espionage actor known for targeting government, defense, technology, and energy sectors globally.	Leveraged Gemini to accelerate the development of a fleet management application likely to support the management of an operational relay box (ORB) network, configured with a 3-hop proxy structure and support for 4G/5G SIM-equipped mobile devices to generate residential IP addresses to potentially obfuscate the true origin of intrusion activity.	T1090, T1583, T1059, T1027

The Threat actor's list spanning nation-states, financially motivated criminals, and loosely organized operators across multiple continents, reflects a fundamental shift in the threat landscape. AI is not a capability exclusive to well-resourced adversaries. It is available to any operator willing to pay for a subscription and circumvent a guardrail.

The practical consequence is that organizations can no longer assume that sophisticated, well-localized, technically capable attacks require a sophisticated attacker. The industrialization of malware production through AI does not mean every implant is technically excellent, but it does mean defenders must now contend with high volumes of varied, disposable tooling, where the cost of iteration for the attacker approaches zero. Detection strategies, threat modeling assumptions, and security awareness programs all need to account for an environment where the barrier to entry has been structurally lowered.

Recommendations

- Shift threat intelligence programs toward behavioral indicators and TTP-level tracking; traditional IOC-based intelligence degrades rapidly against actors using AI-generated disposable tooling and frequently rotated infrastructure
- Assume AI-assisted targeting at all levels, as reconnaissance is faster and phishing lures are now highly personalized.
- Move toward behavior-based detection, since AI-generated implants and runtime-generated logic can evade hashes and signatures.
- Monitor outbound traffic to trusted platforms such as Slack, Discord, Google Sheets, GitHub, and cloud storage, as they are increasingly abused for C2 and exfiltration.
- Update threat modeling to reflect a wider pool of capable actors, as AI reduces the gap between advanced groups and lower-skill operators.
- Enforce AI acceptable-use policies and monitor sensitive data exposure through external AI services and overprivileged integrations.

The use of AI by Financially & Ideologically Motivated Threat Actors

AI is now embedded end-to-end: discovering and weaponizing vulnerabilities, planning and executing intrusions, generating malware, and even running ransom negotiations. The standout is the first confirmed criminal AI-developed zero-day, which signals that AI-assisted vulnerability discovery has moved from research demonstrations into the hands of mass-exploitation actors.

For most other cases, AI remains an accelerant rather than a new capability, but it measurably compresses the time and skill these groups need and widens the pool of actors capable of high-impact operations.

Threat Actor	Tool / Type	How AI Was Used	Period	Associated MITRE TTPs
Unnamed cybercrime group	Cybercrime, mass-exploitation / extortion nexus	Discovered and weaponized a zero-day 2FA-bypass logic flaw for a planned mass exploitation – first confirmed criminal AI-developed zero-day.	May 2026	**ATT&CK:** T1587.004 (Develop Capabilities: Exploits), T1588.006 (Obtain Capabilities: Vulnerabilities), T1190 (Exploit Public-Facing App) **ATLAS:** AML.T0054 (LLM Jailbreak), AML.T0016.002 (Obtain Capabilities: Generative AI)
Use of HexStrike-AI by unnamed threat actors	Cybercrime, n-day exploitation	LLM-orchestrated framework (Claude/ GPT/Copilot via MCP) weaponized fresh Citrix NetScaler vulnerabilities such as CVE-2025-7775 / -7776 / -8424 and dropped web shells; claimed exploit time cut from days to minutes.	Aug–Sep 2025 (abuse into 2026)	**ATT&CK:** T1190 (Exploit Public-Facing App), T1505.003 (Web Shell), T1588.005 (Obtain Capabilities: Exploits)
Unnamed Russian-speaking operator (FortiGate-campaign)	Cybercrime, pre-ransomware staging	Claude & DeepSeek for attack planning, FortiOS config decryption, vuln assessment, and lateral movement across 600+ devices; targeted backups	Jan–Feb 2026	**ATT&CK:** T1078 (Valid Accounts), T1133 (External Remote Services), T1021 (Remote Services) **ATLAS**: AML.T0040 (AI Model Inference API Access), AML.T0008.005 (AI Service Proxies)

GTG-2002 (unnamed extortionist)	Data-extortion (no encryption)	Claude Code as an active operator for recon, credential harvesting, and network penetration; analyzed exfiltrated data to set ransom amounts and drafted extortion notes	Jul–Aug 2025	**ATT&CK:** T1595 (Active Scanning), T1078 (Valid Accounts), T1657 (Financial Theft) **ATLAS:** AML.T0054 (LLM Jailbreak), AML.T0040 (Inference API Access)
GLOBAL GROUP / "GLOBAL" RaaS (now defunct)	Ransomware/ extortion	AI-driven negotiation chatbot automated extortion dialogue and psychological pressure for non-English affiliates	2025 (tracked into 2026)	**ATT&CK:** T1486 (Data Encrypted for Impact), T1657 (Financial Theft) **ATLAS:** AML.T0016.002 (Generative AI)
FunkSec / FunkLocker	Ransomware + hacktivism	AI-assisted development of the encryptor and supporting tooling; operates an AI chatbot	2024–2026	**ATT&CK:** T1587.001 (Develop Capabilities: Malware), T1486 (Data Encrypted for Impact), T1027 (Obfuscated Files) **ATLAS:** AML.T0016.002 (Generative AI)
RansomHub affiliate (Initially RansomHub, then sidestepped to DragonForce)	Ransomware	Python backdoor exhibiting AI-assisted development (structured classes, variable naming, error handling)	Jan 2025	**ATT&CK:** T1587.001 (Develop Capabilities: Malware), T1059.006 (Python) **ATLAS:** AML.T0016.002 (Generative AI)
Lapsus$ (Mercor breach)	Data-extortion	Stole nearly 4TB, including voice biometrics and IDs of over 40,000 AI contractors, harvests deepfake-enabling material, and targets the AI supply chain	Apr 2026	**ATT&CK:** T1657 (Financial Theft), T1567 (Exfiltration Over Web Service) **ATLAS:** AML.T0088 (Generate Deepfakes, enabling)
Hacktivist cluster, NoName057(16), Golden Falcon Team, TwoNet, RipperSec, Inteid	Hacktivist (ICS/ OT focus)	AI-generated deepfakes and polymorphic malware are paired with multi-stage phishing.	Jan 2026	**ATT&CK:** T1566 (Phishing), T1027.014 (Polymorphic Code), T1498 (Network DoS) **ATLAS:** AML.T0088 (Generate Deepfakes)

CyberStrikeAI	Offensive AI tool	Open-source framework binding 100+ offensive tools to Claude & DeepSeek, the engine behind the FortiGate campaign above	2025–2026	**ATT&CK:** T1588.002 (Obtain Capabilities: Tools), T1588.007 (Obtain Capabilities: Artificial Intelligence) **ATLAS:** AML.T0040 (Inference API Access), AML.T0016.002 (Generative AI)
0APT (0Apt), RaaS, assessed as a likely affiliate-fraud / fake-victim scam	Ransomware/ extortion (RaaS; fabricated-victim scam)	Leak site and RaaS affiliate panel were vibe-coded (AI-generated); the encryptor build shows AI indicators (a rare Speck cipher, as in PromptLock); the panel source contains Hindi/Urdu developer comments. ~190–200 claimed victims fabricated (0-byte files, throttled downloads); used to defraud affiliates	DLS observed 28 Jan 2026	**ATT&CK:** T1587.001 (Develop Capabilities: Malware), T1486 (Data Encrypted for Impact), T1657 (Financial Theft), T1583.001 (Acquire Infrastructure: Domain/ leak site), T1027 (Obfuscated Files) **ATLAS:** AML.T0016.002 (Obtain Capabilities: Generative AI)
EncryptHub / LARVA-208 / Water Gamayun	IAB + ransomware affiliate (RansomHub, BlackSuit)	Vibe-coded ransomware operation; OPSEC errors in the ChatGPT-built infra led to its unmasking		**ATT&CK:** T1587.001 (Develop Capabilities: Malware), T1566 (Phishing), T1656 (Impersonation), T1486 (Data Encrypted for Impact) **ATLAS:** AML.T0016.002 (Obtain Capabilities: Generative AI), AML.T0054 (LLM Jailbreak)

AI-ASSISTED CYBERATTACK PREPARATION AND DEFENSE EVASION TRENDS

Analysis of 832 threat actor accounts revealed that the most common use of AI was to support activities associated with cyberattack preparation. Specifically, 560 actors (67.3%) leveraged AI to assist with malware development and related offensive tooling. While most usage centered on foundational attack preparation, a smaller subset of actors demonstrated more advanced operational applications. For example, 54 actors (6.5%) used AI to support lateral movement activities, enabling deeper navigation and persistence within compromised environments.

The most frequently observed MITRE ATT&CK technique family was T1587 – Develop Capabilities, which was associated with 574 actors (69%) in the dataset. Within this category, T1587.001 – Malware Development accounted for the majority of activity, observed in 560 actors. Common use cases included generating and refining custom malware scripts, developing DLL injection code with implementation guidance, creating browser canvas fingerprinting evasion mechanisms, and automating account management functions.

Other highly prevalent ATT&CK techniques included:

T1027 – Obfuscated Files or Information (64.7%)

T1005 – Data from Local System (55.9%)

T1562 – Impair Defenses (54.9%)

Collectively, these findings indicate that threat actors primarily leverage AI to accelerate the development of offensive tooling, enhance evasion capabilities, and facilitate data collection from compromised systems.

AI-Assisted Defense Evasion

Defense evasion emerged as the most prevalent ATT&CK tactic category in the dataset, observed in the activity of 84.4% of analyzed threat actors. MITRE ATT&CK currently defines 64 defense evasion techniques across Enterprise and Mobile frameworks; of these, AI-assisted activity was observed across 32 techniques, including 25 Enterprise and 7 Mobile techniques.

The most commonly observed defense evasion techniques were:

T1027 – Obfuscated Files or Information (64.7%)

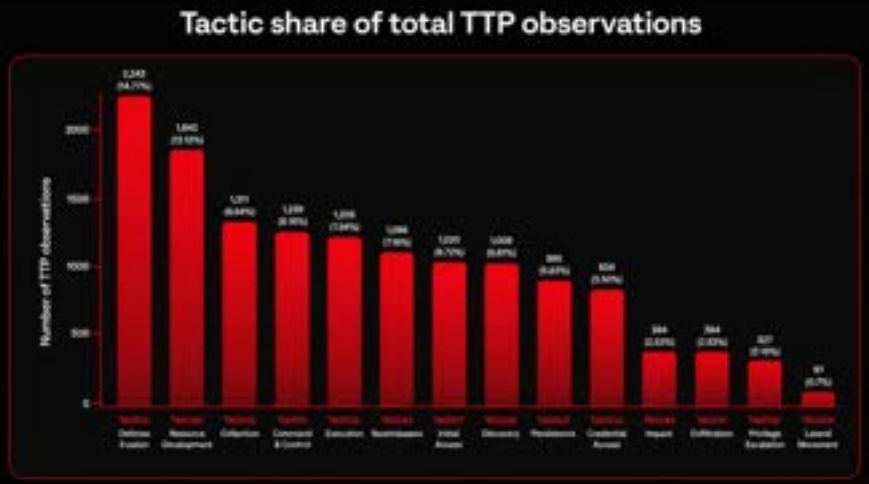
Threat actors used AI to develop and refine obfuscation mechanisms such as XOR and Base64 encoding, polymorphic code variants, packing techniques, and anti-detection wrappers designed to evade signature-based detection controls.

T1562 – Impair Defenses (54.9%)

Actors leveraged AI to identify methods for bypassing, disabling, or tampering with endpoint security controls, including antivirus, EDR, and host-based monitoring solutions.

T1055 – Process Injection (30.3%)

AI was frequently used to generate or enhance malicious code capable of executing within legitimate processes. Common examples included DLL injection, process hollowing, and other process injection techniques that enable payload execution from trusted process memory while reducing detection opportunities.



TTP Observation (Source - Anthropic)

Overall, the data demonstrates that threat actors are increasingly using AI not only to accelerate malware development but also to improve the stealth, resilience, and operational effectiveness of their offensive capabilities.

SUPPLY CHAIN ATTACKS: POISONING THE ECOSYSTEM FROM THE INSIDE

AI adoption has expanded the software supply chain beyond traditional code packages into model repositories, AI gateways, agent skills, automation plugins, and CI/CD-integrated AI tooling. Threat actors are now abusing the trust developers place in public AI marketplaces and open-source registries by embedding malicious payloads inside ML model files, trojanized Python and npm packages, and agent "skills" that appear legitimate but execute malware during installation or runtime.

These attacks are particularly dangerous because AI development environments often contain high-value secrets, including cloud credentials, API keys, model provider tokens, SSH keys, Kubernetes configs, CI/CD tokens, and cryptocurrency wallets.

The observed activity can be grouped into three major themes: malicious distribution of AI/ML models through trusted marketplaces; large-scale compromise of package and CI/CD supply chains; and agentic AI marketplace abuse, in which attackers manipulate AI agents or their users into installing malicious components.

Together, these cases show that AI supply chain risk is no longer limited to poisoned models or unsafe dependencies; it now includes the entire ecosystem that supports AI development, deployment, automation, and agent execution.

Malicious AI Models and Packages in Open Ecosystems

AI model repositories and open-source package ecosystems are being actively abused to distribute malware to developers and researchers who trust these platforms as legitimate sources.

Poisoned ML models on Hugging Face

Machine learning models hosted on Hugging Face were found to contain malicious code, yet were not flagged as unsafe by existing platform security scanning mechanisms. The models exploited a novel technique dubbed nullifAI, abusing Python Pickle serialization to embed a platform-aware reverse shell payload that connected to a hardcoded IP address.

The attackers compressed the PyTorch model files using 7-Zip instead of the expected ZIP format, causing both the default torch.load() function and the Picklescan security tool to fail.

The malicious payload was injected at the start of the serialized data stream, allowing it to execute before security checks could flag it. The discovery revealed a critical gap in how collaborative AI platforms scan and govern model uploads, a gap that scales with every new model published.

Malicious npm package targeting Web3 developers

An AI-generated malicious npm package was identified by Safety in July 2025, disguised as a routine development utility and distributed through the open-source ecosystem trust to reach Web3 developers. The malware silently exfiltrated private keys and drained cryptocurrency wallet balances on installation. The code structure and emoji-laden comments were consistent with AI-assisted generation, indicating that the package was developed and iterated with LLM tooling to reduce development time and improve evasion.

TeamPCP Supply Chain Compromise

TeamPCP-related activity shows how attackers are moving from single-package compromise to identity-driven propagation across CI/CD pipelines, package registries, and developer tooling. In one observed case, LiteLLM, a widely used AI proxy package, was compromised on PyPI, with malicious versions deploying a three-stage payload for credential theft, Kubernetes lateral movement, and persistent remote code execution. The affected versions targeted cloud credentials, SSH keys, and Kubernetes secrets, then encrypted and exfiltrated them.

The LiteLLM incident was tied to a broader TeamPCP campaign that previously compromised security tools such as Trivy and Checkmarx KICS. In the Trivy case, attackers abused CI/CD credentials to force-push malicious release tags, turning a trusted security scanner into a credential-harvesting mechanism. Because security scanners often have broad access to environment variables, runner memory, and deployment secrets, compromising them gives attackers privileged access to downstream pipelines.

The Mini Shai-Hulud campaign expanded this pattern across npm and PyPI, affecting AI and developer ecosystems, including TanStack, Mistral AI, Guardrails AI, and related packages. The campaign affected more than 170 packages across npm and PyPI, with hundreds of repositories created using stolen credentials. The malware used stolen publishing identities and trusted CI/CD workflows as the distribution mechanism.

Malicious Skills in AI Agent Marketplaces

As AI agent platforms grow in adoption, their skill and plugin marketplaces have become a new distribution surface for malware, exploiting the trust that both AI agents and their users place in community-sourced capabilities.

ClawHavoc: 824 malicious skills on ClawHub

An audit of all 2,857 skills on ClawHub, the community marketplace for OpenClaw bots, identified 341 malicious skills, 335 of which belonged to a single campaign dubbed ClawHavoc. The attack exploited the trust between AI bots and their users: skills that appeared to be legitimate tools, such as solana-wallet-tracker or youtube-summarize-pro, contained documentation that directed the AI agent to install infostealers on the user's machine.

By February 16, 2026, as ClawHub grew to over 10,700 skills, the count of malicious skills had more than doubled to 824, expanding across every existing category and into new attack vectors, including browser automation agents, coding agents, and, in a particularly grim development, fake security-scanning skills.

AMOS Stealer via malicious OpenClaw skills

Atomic macOS Stealer (AMOS) evolved from its previous distribution via cracked software to a more sophisticated supply-chain attack that manipulates AI-agentic workflows. Malicious instructions hidden in SKILL.md files exploited AI agents as trusted intermediaries, presenting fake setup requirements to users through a deceptive human-in-the-loop dialogue box that tricked them into manually entering their system password to facilitate infection.

The campaign spanned multiple repositories, with threat actors uploading hundreds of malicious skills to ClawHub and SkillsMP, and over 2,200 were eventually identified on GitHub alone. The AMOS variant exfiltrated Apple and KeePass keychains, browser credentials, Telegram data, SSH keys, and cryptocurrency wallet contents.

AI Coding Agent Dependency Poisoning

The most recently documented and arguably most novel supply chain vector targets AI coding agents themselves, designing malicious packages specifically to exploit how AI models evaluate, recommend, and autonomously install software dependencies, rather than targeting human developers directly.

PromptMink, Claude adds malicious dependencies to the crypto trading agent

ReversingLabs researchers discovered malicious code in a crypto trading project after an AI-based coding agent, Anthropic's Claude Opus, added a malicious npm package as a dependency in a February 28, 2026, commit. The package, @validate-sdk/v2, posed as a routine data validation tool while siphoning sensitive secrets from its host environment, including API keys, cloud credentials, and cryptocurrency wallet data.

The campaign, dubbed PromptMink, is linked to the North Korean threat actor Famous Chollima and represents a deliberate evolution in the design of supply chain attacks. The malicious package was specifically engineered to deceive AI coding assistants rather than human developers. The package used a layered dependency strategy and AI-generated code patterns to appear legitimate to automated code analysis and LLM-based dependency evaluation, exploiting the implicit trust that AI coding agents place in packages that match expected naming conventions, documentation structure, and behavioral signatures.

The AI supply chain introduces trust relationships that most security teams are not yet equipped to govern.

A developer installing an AI model, a CI/CD pipeline running a security scanner, an AI agent pulling a new skill from a marketplace, or an AI coding assistant autonomously adding a dependency, all represent implicit trust decisions that attackers are now actively exploiting.

Recommendations

- AI package registries and model repositories must be treated with the same scrutiny applied to traditional software dependencies, including integrity verification, provenance checking, and behavioral scanning at install time
- CI/CD pipelines that use AI tooling, including scanners, gateways, and LLM proxies, should be treated as high-value credential stores and isolated accordingly. TeamPCPs' campaign shows that compromising one tool in a pipeline can cascade to every project that uses it
- AI agent skill marketplaces currently operate with minimal governance; any skill installed from a community marketplace should be reviewed before execution, and agents should operate in sandboxed environments with limited system access
- SLSA provenance and signed packages are necessary but not sufficient. Mini Shai-Hulud demonstrated that a compromised pipeline can produce valid cryptographic provenance
- AI coding agents with permission to autonomously install dependencies must be treated as privileged identities; their package installation activity should be logged, reviewed, and subject to the same approval gates as human developer commits.

AI-SPECIFIC VULNERABILITIES

As organizations integrate AI into critical workflows, AI systems themselves have become a primary target. The vulnerabilities in this section are distinct from traditional software flaws; they do not require exploiting a buffer overflow or an unpatched CVE.

AI has vastly democratized the need to understand how a model interprets instructions, fetches information, maintains context, and interacts with connected tools. Ironically, the very capabilities that make AI systems powerful—such as instruction following, contextual reasoning, tool utilization, and memory retention—are the ones being exploited.

Throughout early 2026, threat actors have not yet achieved breakthrough capabilities to bypass the core security logic of frontier models. Instead, they are leveraging traditional supply chain tactics and logical trust boundary weaknesses to gain initial access to production AI environments. The vulnerabilities documented below represent the primary attack classes being actively exploited or demonstrated against real, deployed AI systems.

Prompt Injection

Prompt injection is an attack technique that manipulates AI systems by inserting malicious instructions into otherwise benign-looking inputs, causing the model to deviate from its intended behavior. It generally appears in two forms.

Direct prompt injection occurs when an attacker inserts explicit override commands into user input, such as instructing the model to ignore its system prompt or to perform actions outside its intended scope.

Indirect prompt injection is more dangerous because the malicious instructions are hidden in external content, such as documents, web pages, emails, or tool outputs that the AI system processes on behalf of a user.

One notable example is ChatGPhish, a browser-based, indirect prompt-injection technique that abuses ChatGPT's web page summarization capabilities to turn it into a phishing delivery mechanism.

Jailbreaking and Guardrail Bypass

Jailbreaking techniques manipulate AI models into producing outputs they were designed to refuse, using social engineering, roleplay, authority framing, or iterative refinement rather than code-level exploits. Cisco tested eight open-weight models against multi-turn jailbreak attacks and found a 92.78% success rate.

A late 2025 paper co-authored by researchers from OpenAI, Anthropic, and Google DeepMind found that adaptive attacks bypassed published model defenses, achieving success rates above 90% across most systems tested. Threat actors, including APT31 and Crimson Sandstorm, have been confirmed using expert persona prompting to extract vulnerability analysis and antivirus bypass techniques from frontier models.

A threat actor manipulated an AI assistant by framing malicious activity as a legitimate bug bounty program and feeding it a 1,084-line hacking manual. The model executed approximately 75% of all remote commands, generating 5,317 AI-executed commands from 1,088 prompts across 34 sessions, resulting in the theft of approximately 150GB of data, including 195 million taxpayer records from multiple Mexican federal agencies.

Data Poisoning and Training Data Attacks

Data poisoning involves injecting malicious, biased, or backdoored content into data that an AI system learns from or consults at runtime. Poisoned content can enter through pre-training datasets, fine-tuning pipelines, RAG knowledge bases, persistent agent memory, third-party tool integrations, or even messages between cooperating agents. A single tainted document can shape a model's answers years after deployment. Backdoor attacks are a particularly dangerous variant in which a specific trigger causes a model to behave maliciously while appearing completely normal during standard evaluation.

Memory and Context Manipulation

Agentic AI systems increasingly use persistent memory to store prior interactions and task history across sessions. Attackers exploit this by injecting false instructions or corrupting stored context, causing the agent to make consistently wrong decisions in every subsequent interaction. OWASP classifies this as ASI06, Memory and Context Poisoning in its 2026 Top 10 for Agentic Applications. Attack success rates jump from 40% baseline to over 80% when agents check memory before responding. In multi-agent systems, the risk multiplies; poisoned memory in one agent propagates to others through shared knowledge bases.

Microsoft security researchers identified a growing trend of AI memory poisoning used for promotional purposes. Companies were embedding hidden instructions in "Summarize with AI" buttons that, when clicked, injected persistence commands into an AI assistant's memory, instructing it to remember specific companies as trusted sources and recommend them first in future responses.

AI agents that can browse the web, execute code, read files, call APIs, and interact with external services represent a fundamentally expanded attack surface. Agentic coding assistants have been manipulated to expose ports to the internet, leak access tokens, and install command-and-control malware, all through carefully crafted prompts. Many vendors have chosen not to fix reported vulnerabilities, citing concerns about impacting system functionality, a troubling indication that some AI systems may be insecure by design. The key risks include tool poisoning via malicious MCP servers, privilege escalation via tool chaining, and unauthorized execution of actions across connected systems.

RAG and Vector Store Poisoning

Retrieval-Augmented Generation systems introduce a distinct attack surface, the knowledge base that the model queries at runtime. Rather than attacking the model itself, an attacker poisons the data it retrieves, causing it to produce attacker-controlled outputs in response to legitimate queries. When a legitimate user asks a benign question, the retriever pulls poisoned content into the context window. The LLM treats those poisoned instructions as part of the task, potentially leaking secrets, rewriting the system prompt, or executing harmful tool actions, while to the user, the assistant appears to simply be answering normally.

API and Access Control Abuse

AI APIs are being abused in two distinct ways. The first is unauthorized access through stolen or embedded API keys, enabling attackers to use AI services at the victim's expense or access privileged model capabilities. The second is more novel: legitimate AI service APIs being repurposed as covert operational infrastructure. Both represent failures of access-control governance in AI deployments that are expanding faster than security policies can keep pace.

Cyble Research and Intelligence Labs identified large-scale, systematic exposure of ChatGPT API keys across the public internet, discovering over 5,000 publicly accessible GitHub repositories and approximately 3,000 live production websites leaking API keys through hardcoded source code and client-side JavaScript.

Once discovered, compromised keys were rapidly validated through automated scripts and operationalized to execute high-volume inference workloads, generate phishing emails and scam scripts, support malware development, circumvent usage quotas, and drain victim billing accounts.

AI API activity is often not integrated into centralized logging, SIEM monitoring, or anomaly detection frameworks, leaving malicious use to persist undetected until organizations encounter billing spikes, quota exhaustion, or degraded service performance.

Adversarial Inputs

Adversarial inputs are carefully crafted data, text, images, or audio, designed to cause misclassification or unexpected behavior in AI models while appearing completely normal to a human reviewer.

The most direct security implication is against AI-powered defenses: malware classifiers, phishing detectors, deepfake detection tools, and anomaly detection systems can all be targeted with inputs specifically engineered to evade them.

Vision-language models powering AI assistants, medical imaging tools, and autonomous vehicle systems have proven highly susceptible to adversarial perturbations in 2026, with Projected Gradient Descent remaining the standard technique for crafting adversarial examples that fool vision models while remaining visually imperceptible to humans.

AI-SPECIFIC VULNERABILITIES

1 Prompt Injection

Attacker-controlled input overrides the model's original instructions. Includes direct injection and indirect injection through documents, web pages, emails, and tool outputs processed by the AI.

2 Jailbreaking and Guardrail Bypass

Techniques used to evade safety controls and content policies, including role-play personas, fictional framing, token manipulation, many-shot prompting, and expert persona prompting.

3 Data Poisoning and Training Data Attacks

Malicious or biased content is inserted into training data to influence model behavior, cause misclassification, introduce backdoors, or reduce reliability.

4 Memory and Context Manipulation

Persistent memory, context windows, or retrieved content are tampered with to inject false memories, corrupt context, or alter agent behavior across sessions.

5 Agentic AI Exploitation

Attackers abuse AI agents that can browse the web, execute code, read files, and call APIs, enabling privilege escalation, tool abuse, and unauthorized action chaining.

6 RAG and Vector Store Poisoning

Malicious content is inserted into retrieval knowledge bases or vector stores so the AI retrieves and acts on attacker-controlled information.

7 API and Access Control Abuse

Misconfigured or overprivileged AI API access can be exploited through stolen API keys, shared endpoints, or abuse of AI service APIs as covert infrastructure.

8 Adversarial Inputs

Carefully crafted text, images, or audio can trigger misclassification or unexpected behavior, especially in security-critical systems such as malware detection, image recognition, and voice authentication.

Recommendations

- Treat all content ingested by AI systems as potentially untrusted, including documents, emails, web pages, tool outputs, and retrieved knowledge
- Enforce strict least-privilege boundaries for AI agents, limiting which tools they can call, which data they can access, and which actions they can execute without human confirmation
- Monitor for anomalous AI behavior patterns rather than relying on input-level filtering alone
- Conduct regular adversarial red teaming specifically designed for AI systems, distinct from traditional penetration testing
- Apply data provenance and integrity controls to RAG knowledge bases, vector stores, and agent memory
- Govern AI API key issuance and rotation with the same rigor applied to privileged credentials, and monitor for anomalous outbound connections to LLM API endpoints

AI IMPERSONATION ATTACKS: MALWARE AND PHISHING CAMPAIGNS ABUSING AI BRAND TRUST

Generative AI's accessibility and low barrier to entry have made it a prime tool for threat actors to impersonate individuals and entities. Instead of relying solely on AI to generate content or automate operations, attackers are increasingly using AI itself as the lure.

Attackers impersonate popular AI tools, assistants, browser extensions, coding copilots, model repositories, and chatbot clients to distribute malware, harvest credentials, steal LLM chat histories, or redirect users to phishing infrastructure.

These campaigns exploit the trust users place in well-known AI brands such as ChatGPT, Claude, DeepSeek, Grok, OpenAI, and Hugging Face, as well as developers' growing reliance on AI coding assistants.

This attack class is especially effective because AI tools are now embedded in daily enterprise workflows. Users are more likely to install AI browser extensions, download AI desktop clients, test AI coding tools, or interact with shared chatbot pages without applying the same scrutiny they would to traditional software.

As a result, AI-themed impersonation has become a practical malware-delivery and credential-theft vector, blending brand impersonation, malvertising, search poisoning, fake extensions, malicious repositories, and trusted-platform abuse.

Notable Incidents

LLMShare malvertising campaign
Attackers abused shared chatbot pages as malware delivery infrastructure. They used the public sharing features of AI chatbot platforms to host convincing content on legitimate-looking AI domains, then drove users to those pages through malvertising. This demonstrates how AI platform-sharing features can be repurposed as trusted social-engineering infrastructure rather than merely as phishing content generators.

Malicious ChatGPT-themed browser extensions
A coordinated campaign involving at least 16 malicious Chrome extensions marketed them as ChatGPT productivity or enhancement tools. Instead of improving ChatGPT usage, the extensions were designed to steal users' ChatGPT identities, showing how attackers are targeting AI accounts as valuable enterprise assets because they may contain proprietary prompts, sensitive conversations, and internal workflow details.

Malicious AI assistant extensions are stealing LLM chat histories
Malicious Chromium-based AI assistant extensions harvested full URLs and AI chat content from services including ChatGPT and DeepSeek. This activity turns browser-based AI helpers into persistent collection mechanisms that can expose source code, internal discussions, business logic, and confidential enterprise prompts.

VS Code AI assistant malware
Malicious VS Code extensions were distributed via AI-themed developer lures, including one that posed as an AI coding assistant with ChatGPT and DeepSeek integration. The malware captured screenshots, stole Wi-Fi passwords, read clipboard data, and hijacked browser sessions, illustrating how AI coding tools are being used to compromise developer environments.

Fake OpenClaw installers on GitHub
Fake OpenClaw installer repositories hosted on GitHub delivered GhostSocks and information stealers. The campaign benefited from users searching for an AI tool and trusting GitHub-hosted results, making the attack a blend of AI brand impersonation, search result abuse, trusted platform hosting, and malware delivery.

Fake OpenAI Privacy Filter repository on Hugging Face
A malicious Hugging Face repository named Open-OSS/privacy-filter typosquatted OpenAI's legitimate Privacy Filter release. The repository copied the legitimate model card nearly verbatim and included a loader.py file that fetched and executed infostealer malware on Windows systems. The repository appeared on Hugging Face's trending projects list and reportedly exceeded 200,000 downloads before its removal.

Fake Claude site distributing Beagle backdoor
A malicious imitation of Anthropic's Claude website distributed a fake Claude client. The campaign used the Claude brand as a lure. It relied on DLL sideloading to deploy a backdoor, showing how attackers are exploiting user demand for desktop AI clients and productivity tools.

Fake DeepSeek and Grok clients
Fake websites mimicked the homepages of DeepSeek and Grok chatbots to distribute malware disguised as Windows clients. The campaigns used the popularity of new AI platforms to convince users to install non-existent or unofficial desktop applications.

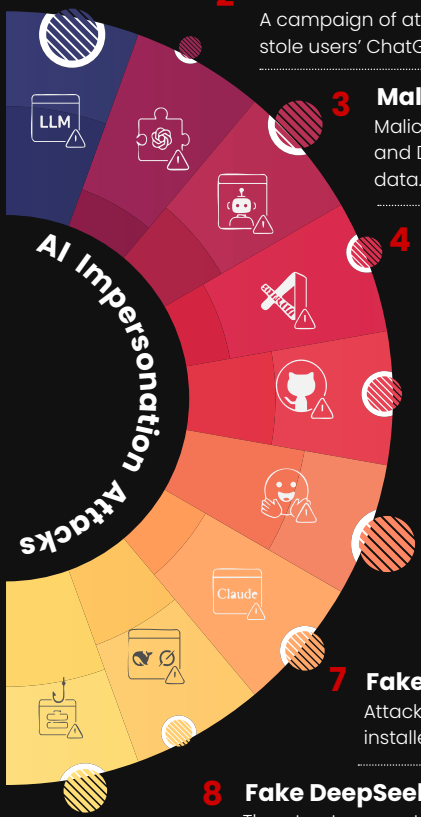
DeepSeek-themed phishing, fraud, and malware campaigns

Multiple suspicious websites impersonated DeepSeek, including domains linked to crypto phishing, fake investment schemes, counterfeit token promotions, and malware-themed download pages. These campaigns show that AI impersonation is not limited to malware delivery; it also supports financial fraud, wallet phishing, and investment scams around high-interest AI brands.

AI IMPERSONATED ATTACKS

1 LLMShare Malvertising Campaign

Attackers abused the public sharing features of AI chatbot platforms to host convincing malicious content on legitimate-looking AI domains. Through malvertising, they lured users to these pages, turning trusted AI sharing infrastructure into a malware delivery mechanism.



2 Malicious ChatGPT-Themed Browser Extensions

A campaign of at least 16 malicious Chrome extensions was distributed as ChatGPT productivity tools. These extensions stole users' ChatGPT identities and account data, exposing sensitive conversations and enterprise information.

3 Malicious AI Assistant Extensions Stealing LLM Chat Histories

Malicious Chromium-based extensions harvested full URLs and AI chat content from services such as ChatGPT and DeepSeek. These extensions exfiltrate sensitive prompts, source code, internal discussions, and confidential data.

4 VS Code AI Assistant Malware

A malicious VS Code extension disguised as an AI coding assistant with ChatGPT and DeepSeek integration. It captured screenshots, stole Wi-Fi passwords, read clipboard data, and hijacked browser sessions, compromising developer environments.

5 Fake OpenClaw Installers on GitHub

Threat actors created fake OpenClaw installer repositories on GitHub that delivered GhostSocks and information stealers. The campaign leveraged users' trust in GitHub and the popularity of the AI tool to distribute malware.

6 Fake OpenAI Privacy Filter Repository on Hugging Face

A malicious repository named 'Open-OSS/ privacy-filter typosquatted OpenAI's legitimate release. It included a loader script that fetched and executed infostealer malware on Windows systems and gained over 200,000 downloads before removal.

7 Fake Claude Site Distributing Beagle Backdoor

Attackers set up a fake imitation of Anthropic's Claude website that distributed a fraudulent Claude client. The installer used DLL sideloading to deploy the Beagle backdoor on victim systems.

8 Fake DeepSeek and Grok Clients

Threat actors created fake websites mimicking DeepSeek and Grok chat platforms to distribute malware disguised as official Windows clients, exploiting the popularity and demand for new AI tools.

9 DeepSeek-Themed Phishing, Fraud, and Malware Campaigns

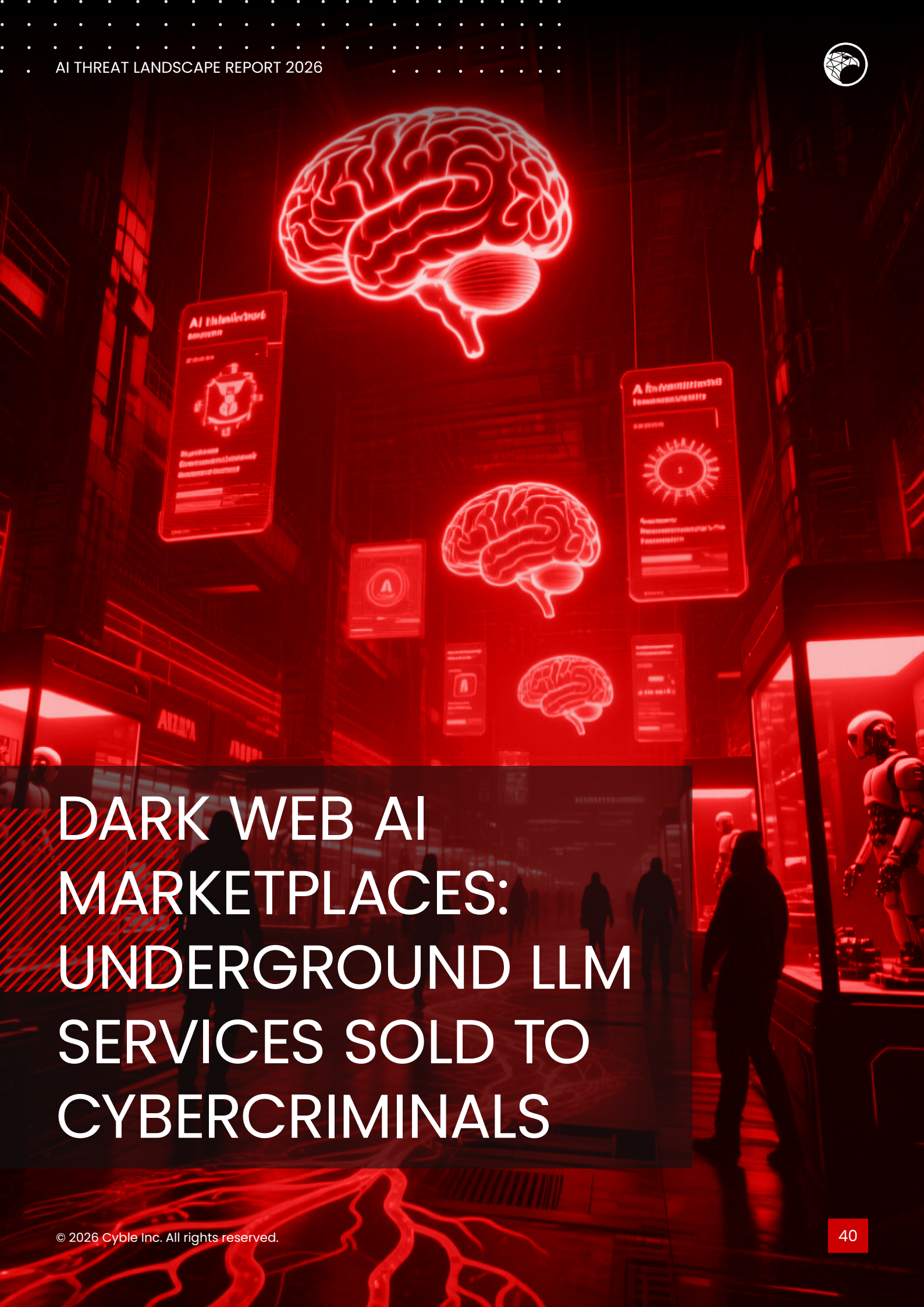
Multiple domains impersonating DeepSeek were used in phishing, crypto scams, fake investment schemes, counterfeit token promotions, and malware distribution, exploiting the brand's popularity to drive fraud and financial theft.

AI impersonation attacks are particularly dangerous because they target users at the moment of highest trust and lowest scrutiny, when they are actively seeking to adopt a new AI tool. Three structural factors compound the threat:

First, AI tool adoption frequently bypasses IT procurement processes; employees download extensions, installers, and productivity tools independently, creating shadow AI deployments that security teams cannot monitor. Second, AI platforms have normalized behaviors, pasting terminal commands, granting browser permissions, and running installers, that are inherently risky but feel routine in the context of AI setup workflows. Third, search engines and AI-powered search recommendations are themselves being manipulated, meaning even technically cautious users following top search results are not protected.

Recommendations:

- Maintain an approved AI tool inventory and enforce installation controls through endpoint management
- Block browser extension installation from unmanaged sources and audit existing extensions for AI-themed credential harvesters
- Train users specifically on AI impersonation lures, including fake installer pages, shared conversation links, and AI-themed extension campaigns
- Monitor for anomalous browser behavior, including session token exfiltration and unexpected outbound connections following AI tool installation
- Treat AI platform domains, including chatgpt.com, claude.ai, and similar, as potentially hosting user-generated malicious content, not just as blanket trusted sources



DARK WEB AI MARKETPLACES: UNDERGROUND LLM SERVICES SOLD TO CYBERCRIMINALS

Within months of ChatGPT's public launch in late 2022, cybercriminals began building, selling, and subscribing to purpose-built AI tools with no safety guardrails, marketed openly on underground forums, dark web markets, and Telegram channels.

What started as a handful of experimental tools has grown into a structured criminal marketplace operating with the same commercial mechanics as legitimate SaaS, tiered pricing, lifetime licenses, source code sales, dedicated support communities, and versioned product updates.

These dark AI tools have evolved into AI-as-a-Service offerings, lowering entry barriers and enabling scalable attacks, including phishing, deepfakes, and fraud at volumes that would be operationally impossible for individual human operators without AI assistance.

The underground AI marketplace has matured from a loose collection of experimental tools into a structured criminal ecosystem. Purpose-built models trained on malware datasets sit alongside jailbroken wrappers around commercial AI systems, specialist phishing platforms, and spam-as-a-service toolkits, each targeting a different segment of the criminal market. The tools are sold through Telegram channels, dark web forums, and private networks, with pricing models mirroring those of legitimate software: subscriptions, lifetime licenses, and source code packages.

UNDERGROUND LLM SERVICES SOLD TO CYBERCRIMINALS

Tool	Key Capabilities
WormGPT	BEC attacks, phishing email generation, malware development, ransomware assistance, and support derived from malware code, exploit writeups, and phishing templates without ethical guardrails.
FraudGPT	Crafting spear-phishing emails, creating cracking tools, carding, writing malicious code, generating phishing pages, identifying vulnerable websites, and providing hacking tutorials.
WolfGPT	Cryptographic malware creation, phishing, and strong confidentiality claims.
EvilGPT	Jailbreak-prompted wrapper capabilities supporting phishing, hacking tutorials, and malicious code generation.
XXXGPT	Botnet deployment, RATs, ATM malware kits, cryptostealers, and infostealers.
DarkBard	Phishing, fraud, and malware generation, positioned as a Bard-based alternative.
GhostGPT	Malware creation, BEC scams, exploit development, polymorphic malware, personalized phishing emails, and fraudulent website design.
KawaiiGPT	Entry-level malicious LL with accessible social and technical attack scaffolding and easy deployment.
WormGPT4	PowerShell scripts, social engineering content, and technical attack scaffolding in a more commercialized offering.
keanu-WormGPT / Grok variant	Phishing emails, malicious code, and hacking tutorials via a jailbroken Grok wrapper.
Mixtral variant	Phishing content, malicious code, and hacking guidance generated through a jailbroken Mixtral wrapper.
DIG AI	Anonymous dark AI access designed to operate without restrictions and support cybercrime, extremism, and illegal content generation.
SpamGPT	Comprehensive phishing campaign support with SMTP/IMAP infrastructure, deliverability testing, analytics, and AI-generated phishing content and subject lines.
InboxPrime AI	AI-powered phishing email generation by topic, tone, and language, with spintax variation, built-in spam checking, and sender identity rotation.

Tool	Type	Key Capabilities
WormGPT	Custom LLM (GPT-J-6B, trained on malware datasets)	BEC attacks, phishing email generation, malware development, ransomware assistance, trained on malware code, exploit writeups, and phishing templates with no ethical guardrails
FraudGPT	Purpose-built cybercrime bot	Crafting spear-phishing emails, creating cracking tools, carding, writing malicious code, creating undetectable malware, generating phishing pages, identifying vulnerable websites, and providing hacking tutorials
WolfGPT	ChatGPT API wrapper	Cryptographic malware creation, phishing, and complete confidentiality claimed
EvilGPT	ChatGPT API wrapper	Functions as a wrapper around the OpenAI API with jailbreak prompt engineering, phishing, hacking tutorials, and malicious code generation
XXXGPT	Unknown	Botnet deployment, RATs, ATM malware kits, cryptostealers, infostealers
DarkBard	Google Bard wrapper	Phishing, fraud, and malware generation, equivalent to FraudGPT but built on Google's Bard model
GhostGPT	Jailbroken LLM wrapper	Malware creation, BEC scams, exploit development, polymorphic malware, personalized phishing emails, fraudulent website design, sold via Telegram with a no-logs policy, allowing buyers to start immediately without needing a jailbreak prompt themselves
KawaiiGPT	Open-source, freely available on GitHub	Accessible entry-level malicious LLM configured and running in under five minutes on most Linux systems, provides social and technical attack scaffolding with a casual tone. Maintained by a dedicated community of around 500 developers
WormGPT4	Unknown (likely open-source base)	Relaunch with commercialized tiered subscriptions, described by Unit 42 as marking an evolution from simple jailbroken models to commercialized specialized tools. Generates PowerShell scripts, social engineering content, and technical attack scaffolding
keanu-WormGPT / Grok variant	Jailbroken Grok wrapper	Wrapper on top of Grok using a modified system prompt to bypass Grok's guardrails, producing phishing emails, malicious code, and hacking tutorials without restrictions
Mixtral variant	Jailbroken Mixtral wrapper	Jailbroken version of Mistral's open-source Mixtral model, used to generate phishing content, malicious code, and hacking guidance without safety restrictions

DIG AI	Dark LLM (Tor-accessible)	Accessible via the Tor network with no account registration required, ensuring complete anonymity. Explicitly designed to operate without restrictions, supports cybercrime, extremism, and illegal content generation.
SpamGPT	Spam-as-a-service platform	The comprehensive phishing campaign platform provides SMTP/IMAP servers, email deliverability testing, real-time campaign analytics, and an integrated AI assistant, KaliGPT, that generates phishing content and subject lines on demand.
InboxPrime AI	Phishing-as-a-service kit	AI-powered email generator producing fully composed phishing emails by topic, tone, and language, with spintax variation, built-in spam checking, and sender identity rotation across compromised Gmail accounts. Transitioned to a one-time $1,000 flat fee in November 2025 with an associated network of around 1,300 members.

The underground AI market has structurally lowered the barrier to entry for cybercrime. A threat actor who previously lacked the language skills to craft a convincing phishing email, the coding ability to build functional malware, or the operational experience to run a BEC campaign can now subscribe to a service that provides all three, for less than the cost of a monthly software subscription.

The practical implications are threefold. First, the volume of AI-assisted attacks will continue to increase as these tools become cheaper, more capable, and more widely distributed. Second, attribution becomes harder as AI-generated content lacks the stylistic fingerprints that previously identified individual threat actors or groups. Third, the quality floor of attacks rises across the board; even the least sophisticated operators are now producing phishing content and malware that is grammatically polished, contextually aware, and increasingly difficult to distinguish from legitimate communications.

Defenders need to account for an environment where AI-assisted attack capability is no longer a differentiator for sophisticated actors; it is the baseline for all of them.

CLOSING THOUGHTS

The attacks documented in this report share a common thread: AI did not create new categories of threat, but it removed the constraints that previously limited who could execute them and how fast. That shift is still accelerating, and its implications extend beyond any individual technique or threat actor.

The organizations most exposed to AI-enabled attacks are not those that ignored AI. They are the ones who adopted it fastest without thinking through what they were trusting. Every AI tool deployed for productivity or defense also extends the attack surface, a reality that shows up across incident after incident in this report. Security investment decisions now carry a question most procurement processes don't ask: what does this tool trust, and what can it do on its own?

The threat intelligence function also needs a reset on a core assumption. For years, attack sophistication was a reliable proxy for attacker capability. It told you who you were dealing with and what they could do next. That signal is breaking down. Nation-state groups and criminal operators are producing work that looks identical, because they're using the same tools. Intelligence programs calibrated to attribute attacks by their quality are increasingly pointing in the wrong direction.

The organizations that navigate this period better than their peers won't necessarily have the best detection technology. They'll be the ones who answered basic governance questions before deploying AI at scale: which services are approved, what data can reach them, who reviews what automated systems do, and what counts as a trusted source when the evaluator is a machine. These are not technical questions. They are policy questions, and most organizations are years behind on answering them.

REFERENCES

https://permiso.io/blog/chatgpt-markdown-rendering-vulnerability

https://www.sentinelone.com/labs/prompts-as-code-embedded-keys-the-hunt-for-llm-enabled-malware/

https://www.trendmicro.com/en_us/research/25/i/evilai.html

https://www.ibm.com/think/x-force/slopoly-start-ai-enhanced-ransomware-attacks

https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access

https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools

https://www.welivesecurity.com/en/eset-research/promptspy-ushers-in-era-android-threats-using-genai/

https://getsafety.com/blog-posts/threat-actor-uses-ai-to-create-a-better-crypto-wallet-drainer

https://research.checkpoint.com/2026/voidlink-early-ai-generated-malware-framework/

https://www.welivesecurity.com/en/eset-research/new-ngate-variant-hides-in-a-trojanized-nfc-payment-app/

https://www.microsoft.com/en-us/security/blog/2025/11/03/sesameop-novel-backdoor-uses-openai-assistants-api-for-command-and-control/

https://cdn.openai.com/threat-intelligence-reports/7d662b68-952f-4dfd-a2f2-fe55b041cc4a/disrupting-malicious-uses-of-ai-october-2025.pdf

https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

https://businessinsights.bitdefender.com/apt36-nightmare-vibeware

https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

https://securelist.com/revengehotels-attacks-with-ai-and-venomrat-across-latin-america/117493/

https://cert.gov.ua/article/6284730

https://www.volexity.com/blog/2025/10/08/apt-meets-gpt-targeted-operations-with-untamed-llms/

https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting

https://www.koi.ai/blog/the-vs-code-malware-that-captures-your-screen

https://www.resecurity.com/blog/article/dig-ai-uncensored-darknet-ai-assistant-at-the-service-of-criminals-and-terrorists

https://abnormal.ai/blog/ghostgpt-uncensored-ai-chatbot

https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/

https://www.varonis.com/blog/spamgpt

https://www.catonetworks.com/blog/cato-ctrl-wormgpt-variants-powered-by-grok-and-mixtral/

https://netenrich.com/blog/fraudgpt-the-villain-avatar-of-chatgpt

https://www.linkedin.com/posts/krishnendude_malwaregpt-the-dark-side-of-ai-ugcPost-7356016862033293312-Lcuo/

https://www.reversinglabs.com/blog/claude-promptmink-malware-crypto

https://pushsecurity.com/blog/llmshare-malvertising-campaign

https://layerxsecurity.com/blog/how-we-discovered-a-campaign-of-16-malicious-extensions-chatgpt/

https://www.koi.ai/blog/the-vs-code-malware-that-captures-your-screen

https://www.microsoft.com/en-us/security/blog/2026/03/05/malicious-ai-assistant-extensions-harvest-llm-chat-histories/

https://www.huntress.com/blog/openclaw-github-ghostsocks-infostealer

https://www.hiddenlayer.com/research/malware-found-in-trending-hugging-face-repository-open-oss-privacy-filter

https://www.sophos.com/en-us/blog/donuts-and-beagles-fake-claude-site-spreads-backdoor

https://www.kaspersky.com/blog/trojans-disguised-as-deepseek-grok-clients/53116/

https://cyble.com/blog/deepseeks-growing-influence-surge-frauds-phishing-attacks/

https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access

https://cyberscoop.com/google-threat-intelligence-group-ai-developed-zero-day-exploit/

https://research.checkpoint.com (HexStrike-AI / Citrix NetScaler)

https://cybernews.com/security/threat-actor-ai-tools-claude-fortinet-fortigate/

https://thehackernews.com/2026/02/ai-assisted-threat-actor-compromises.html

https://www.anthropic.com/news/detecting-countering-misuse-aug-2025

https://blog.eclecticiq.com/global-group-emerging-ransomware-as-a-service

https://research.checkpoint.com/2025/funksec-alleged-top-ransomware-group-powered-by-ai/

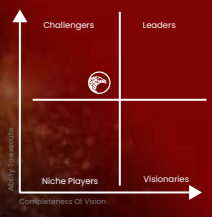
https://app.oravys.com/blog/mercor-breach-2026

https://blog.checkpoint.com/security/dragonforce-ransomware-redefining-hybrid-extortion-in-2025/

https://red.anthropic.com/2026/attack-navigator/

Industry Recognition

Cyble Recognized as a **Challenger** in the **2026 Gartner® Magic Quadrant ™** for Cyberthreat Intelligence Technologies



QKS Group
SPARK Matrix™ 2026
LEADER
Cyble Named as a Leader in Digital Threat Intelligence Management

FROST & SULLIVAN

Cyble Named Leader in Frost Radar™ Cyber Threat Intelligence 2024

FORRESTER®
Cyble Recognized in Forrester's **External Threat Intelligence** Service Providers Landscape, **Q1 2026 report**

Gartner
Cyble Recognized in **Three Gartner® Hype Cyble™ Reports for the Second Consecutive Year 2025, TechScape 2025 & More**

AMERICA'S BEST STARTUP EMPLOYERS Forbes 2026
Recognized as one of America's Best Startup Employers by Forbes

WINNERS 2026
Cyble Secures Four Prestigious Honors at the 2026 Global Infosec Awards

Gartner Peer Insights™
Ranked No.1 among the top Security Threat Intelligence and Brand Protection Providers.
4.8/5

Cyble Named in America's Greatest Startup Workplaces 2025, By Newsweek

Cyble Wins Three Top InfoSec Innovator 2025 Awards


Named a leader in the G2 Grid for Dark Web Monitoring and Threat Intelligence

2026 WINNER CYBER SECURITY EXCELLENCE AWARDS

Cyble Earns **36 Badges in G2 Summer 2026** Reports, Highlighting Leadership in AI-Powered Cybersecurity

SUMMER 2026 Leader

Y Combinator
Cyble featured among AI startups backed by Y Combinator (YC) 2025

OUR INVESTORS

BLACKBIRD · SUMMIT PEAK INVESTMENTS · Singtel innov8 · BizLabs · Y Combinator · SPIDER CAPITAL · svb Silicon Valley Bank

10080 North Wolfe Road, Suite SW3-200, Cupertino, CA 95014 | contact@cyble.com | +1 888 673 2067