REPORT

AI Security 2026

Observations across a customer base of over 625,000 organizations worldwide

SOPHOS

SOPHOS | X-OPS

Letter from John Peterson

In May 2026, Sophos found an attacker running what amounted to a software development operation inside a customer's network. Around a dozen AI agents, coordinated through a commercial coding assistant, were writing and testing malware against Sophos, CrowdStrike, and Windows Defender endpoint protection, each on its own virtual machine. The operation produced nearly 80 modules and more than 70 evasion techniques, every result committed to version control and improved on the next pass. The same operator later used the output to deploy ransomware and steal data.

While the tradecraft itself wasn't new, the agents turned weeks of manual iteration into days of automated iteration, and that change is the subject of this report. AI is changing the speed and shape of security operations more than it is changing the fundamentals of intrusion. Attackers still need initial access, still move laterally, and still exfiltrate through observable channels. What has changed is the clock.

In this report, Sophos will make that case, with data.

We have Managed Detection and Response (MDR) and Incident Response (IR) analysts investigating thousands of incidents a year. We have the Counter Threat Unit (CTU) that monitors underground forums in multiple languages. We have SophosLabs researchers analyzing malware samples at scale. We have a Sophos AI team developing innovative, defensive-oriented techniques. And we have endpoint, firewall, email, cloud, network, and identity telemetry across a global base of over 625,000 customers.

This scale, reach, and expertise give us a vantage point few others hold on the fast-evolving intersection of AI and cybersecurity. We intend to use that perspective to help the industry stay ahead of the clock and ensure our customers remain protected.

J. P. Peterson, CTO, Sophos

Executive summary and key findings

This report draws on Sophos MDR and IR casework, SophosLabs analysis, Sophos CTU intelligence, and endpoint and network observations across a customer base of over 625,000 organizations worldwide. One pattern runs through it all: AI is enabling attackers and defenders to move faster through familiar operations, though evidence of entirely novel attack types remains limited rather than absent – and may grow as capabilities scale.

2026 is an inflection point. Frontier and open-weight models have become useful enough for engineering delegation, while falling inference costs change the economics for both sides. The practical decision is no longer whether AI will be used, but where high-cost frontier systems are justified, where cheaper open-weight models suffice, and how to route work without expanding risk.

The following findings anchor the report:

1. **AI is compressing attack timelines, not inventing new attack types.** The threat actor Sophos tracks as STAC6994 used AI agents to iterate on EDR bypass techniques at a pace that would have taken a human developer weeks. Sophos CTU analysts subsequently confirmed that the operator behind STAC6994 engaged in ransomware deployment and data theft operations.
2. **The credential and identity layer around enterprise AI services is now a harvesting target.** The Salesloft/Drift incident showed how chatbot OAuth tokens translate directly into Salesforce environment compromise. Separately, Sophos CTU analysts observed claims from threat actors on underground forums that AI prompts and generated data may be captured as collateral in cyberattacks.
3. **The underground economy is absorbing AI as infrastructure.** Sophos documented threat actors selling brokered API keys for Claude, ChatGPT, and Grok. Forums have spawned channels dedicated to AI prompt engineering, jailbreaking techniques, and malware development workflows. Some personas are hiring AI prompt engineers. Others remain openly skeptical that AI will change their operations. The adoption curve looks less like a sudden transformation and more like the gradual incorporation of any new tool into existing criminal workflows.
4. **Supply chain attacks are specifically targeting AI development infrastructure.** The Nx Console VS Code extension compromise distributed credential-stealing malware. SophosLabs analysts investigated fake Claude download pages distributing a novel RAT, and a ClickFix campaign that leveraged legitimate ChatGPT shared chat URLs to deliver the MacSync infostealer.
5. **AI-assisted social engineering has moved from experimental to operational.** Sophos CTU documented underground advertisements for voice bots, AI-generated personas for romance fraud, and synthetic profile imagery services. AI's contribution is scale and linguistic quality across languages, applied to the same deception techniques that worked before generative AI.
6. **Exploitation timelines are compressing faster than patching timelines.** CISA's Binding Operational Directive 26-04, issued June 10, 2026, cut mandatory patching to three days for the highest-risk flaws, explicitly citing

AI as a factor narrowing the window between disclosure and exploitation. The directive's first real-world test arrived almost immediately: CVE-2026-10520 (Ivanti Sentry) was exploited within 24 hours of proof-of-concept publication and added to the KEV catalog on June 12. CVE-2026-42208 (LiteLLM) was exploited within 36 hours, with attackers targeting databases containing upstream LLM provider keys.

7. **Proving AI use in an attack remains operationally difficult.** The STAC6994 case was unusual because the development framework was left on a device Sophos could examine. In many incidents, AI assistance is invisible in telemetry. True prevalence of AI-assisted attacks is likely higher than any vendor can directly evidence.

8. **AI reasoning models accelerate investigation.** AI reasoning models compress hours of analyst investigation into minutes for known attack patterns. They do not replace the behavioral detection systems that find a single threat buried in billions of events. The statistical constraints of extreme class imbalance, the need for environment-specific baselines, and the pace of adversary adaptation all require specialized engineering

Part 1:
AI in the attack chain

Sophos X-Ops taxonomy of AI threats

Sophos X-Ops separates AI threats into two top-level categories: malicious use of AI (the threats we protect from), and malicious targeting of AI (the AI we protect). The first covers AI-generated artifacts, AI-augmented runtime capabilities, and AI-orchestrated intrusions. The second covers agent-initiated compromise, AI software impersonation, poisoned LLMs, and attacks on models themselves. The two halves call for different responses: malicious use is a productivity and scale problem; malicious targeting is a trust, supply chain, and governance problem. The taxonomy complements existing frameworks such as MITRE ATLAS, Microsoft's agent failure-mode taxonomy, MIT's AI Risk Repository, and NIST AI 100-2.

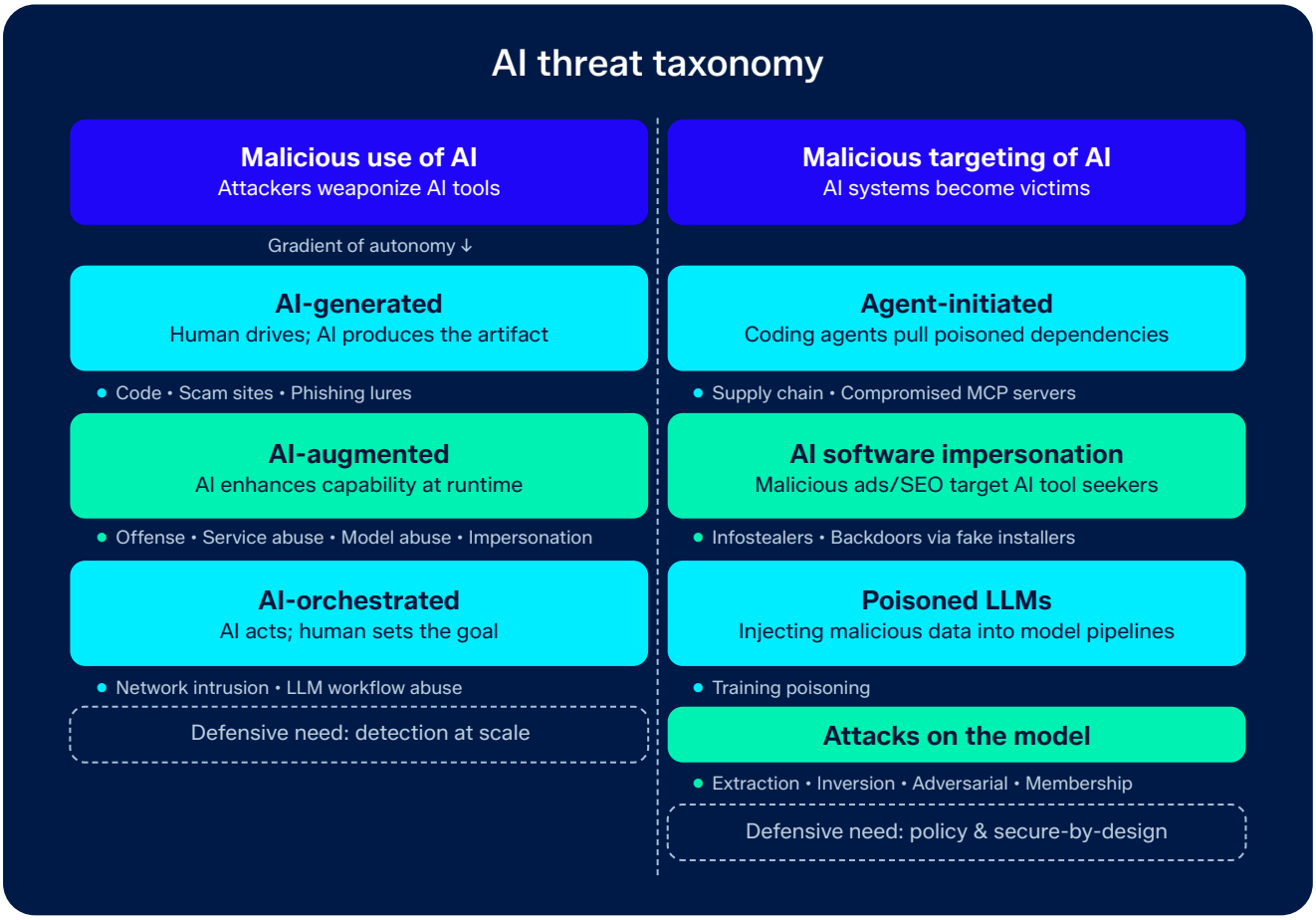


Figure 1: AI Threat Taxonomy Overview

A gradient of autonomy

At the lightest touch, AI-generated attacks involve a human using generative AI to produce an artifact and deploying it manually. A threat actor targeting Mexican government organizations used Claude Code and GPT to generate scripts. Leaked chats from The Gentlemen ransomware group confirmed similar applications. And, in a recent DragonForce ransomware case investigated by the Sophos IR team, the threat actor produced what we strongly suspect to be an AI-generated report during negotiations with the targeted organization, containing a detailed analysis of the exfiltrated data, including PII and legal risks. AI-generated analysis was used as part of an attempt to exert pressure and persuade the organization to pay the ransom.

Further along the gradient, AI-augmented attacks involve a model enhancing a capability at runtime. LameHug, Python-based malware that CERT-UA attributes with moderate confidence to IRON TWILIGHT (APT28), queried a hosted open-weight model to dynamically generate Windows reconnaissance commands per environment, making static analysis far less useful and turning outbound traffic to AI/ML API endpoints into one of the few reliable detection surfaces. AI-augmented impersonation is another active vector: voice cloning and face-swap tooling are now used in real-time fraud, executive impersonation, and KYC bypass.

At the furthest end, Anthropic's November 2025 disclosure documented a Chinese state-sponsored campaign using Claude Code to attempt compromise of roughly thirty targets. These attacks remain rare, but they collapse the time between reconnaissance and action in ways that challenge traditional defensive assumptions.

Case study: *STAC6994*

Sophos analysts observed a threat actor using AI to test EDR evasion tactics in a 'red team' post-exploitation framework, discovering a rogue device running continuous payload generation in a customer environment. The attacker had provisioned virtual machines from Ludus and was using the Cursor IDE with AI agents to develop and test post-exploitation tools.

The framework ran parallel testing across three environments: one VM with Sophos endpoint protection, one with CrowdStrike, and one without any EDR as a control. A fourth VM operated as a Sliver C2 server. Approximately 12 AI agents operated with defined roles.

One agent, running Claude Opus 4.5, handled core operations and rule-setting. Others managed operational security (OPSEC) hardening, documentation, proxy stress testing, VM deployment, and testing tools against the EDR agents. Code commits flowed to Git through Model Context Protocol (MCP).

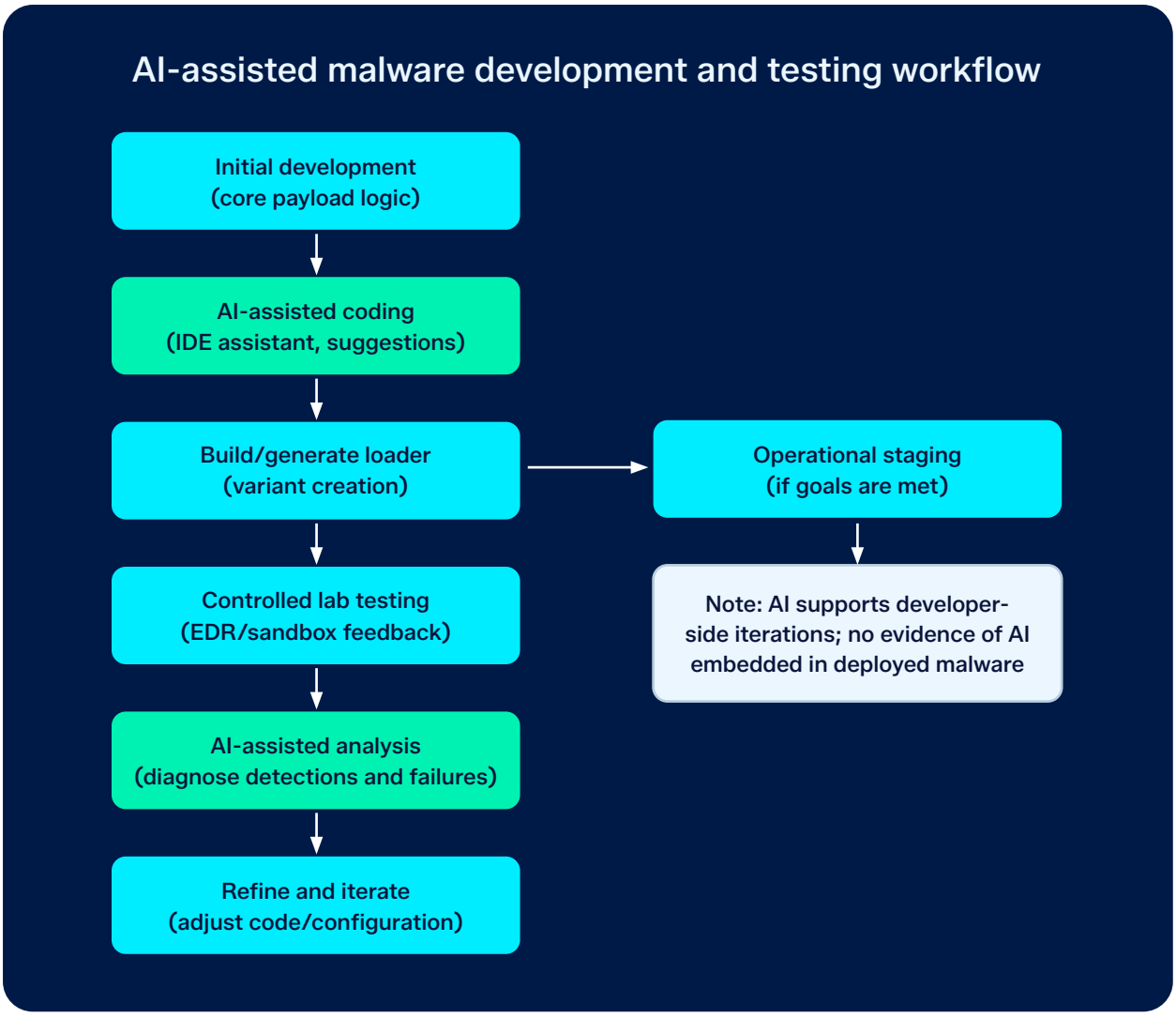


Figure 2: Diagram showing AI's role in the malware development workflow

The playbook was systematic. According to recovered artifacts, the agents read research articles scraped from the SpecterOps blog, extracted offensive techniques, mapped them to the MITRE ATT&CK framework, identified the steps and tools needed to reproduce each technique, prepared the lab environment, executed it, and reported findings.

At the core was a Python-based modular payload generator that wrapped raw payloads in layers of encryption and evasion techniques, producing custom executables in Rust and Go. Nearly 80 different modules testing over 70 different techniques were developed. The actual EDR bypass path was a structured engineering test cycle with human review and iteration. AI coordinated the workflow and supported experimentation; the evidence did not show AI embedded inside deployed malware.

After various iterations, agent documentation claimed near-universal success against the EDR agents, though Sophos analysts noted the evidence did not fully support that conclusion. The crucial operational signal was tempo: the full cycle compressed weeks into days.

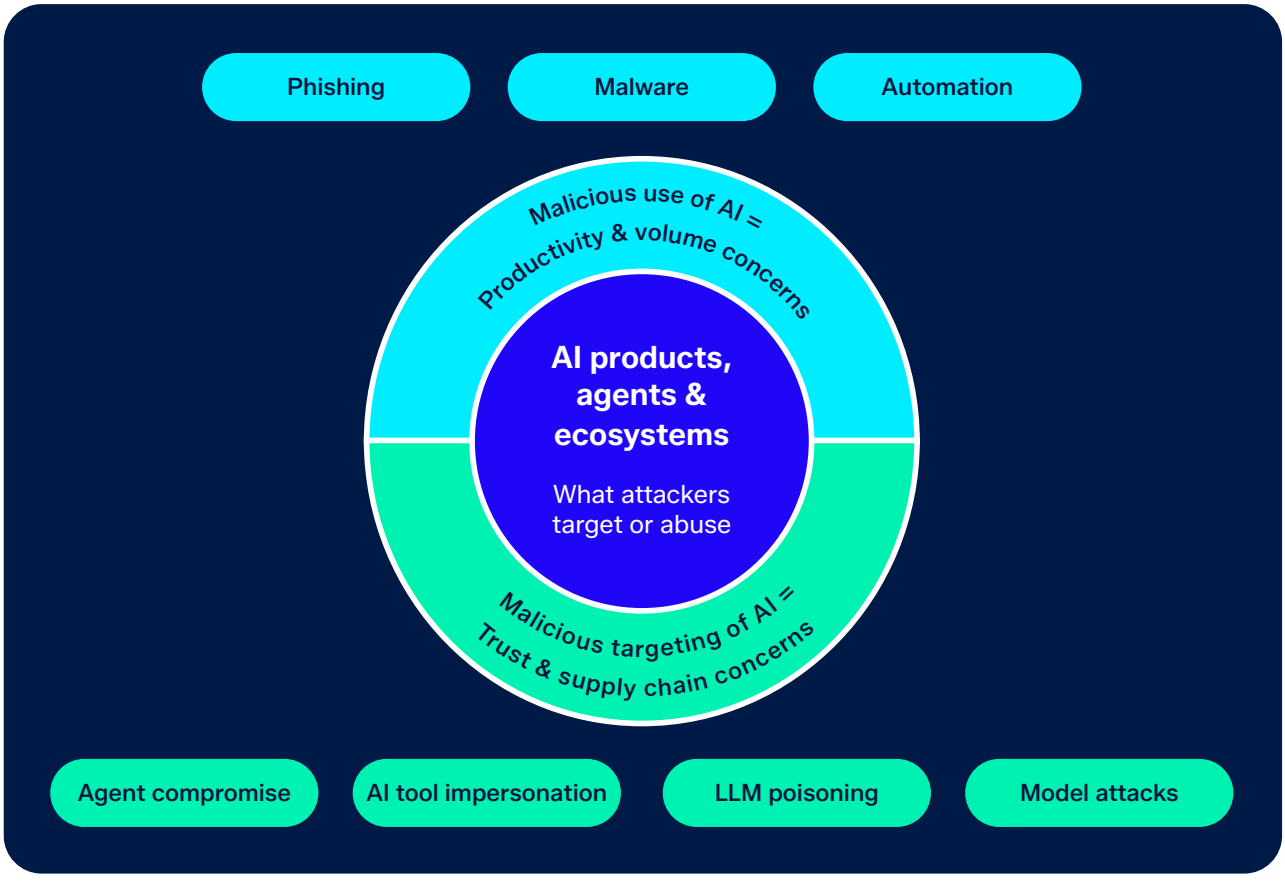
CTU confirmed the STAC6994 operator subsequently engaged in ransomware deployment and data theft; the framework was a production tool for real intrusions.

Malicious targeting of AI

The second part of our taxonomy covers attacks in which AI products, agents, and ecosystems are the target or the attack surface, rather than the tool, and we'll explore this in more detail in Part 2. Agent-initiated compromise is one of the most immediately concerning sub-types: a coding agent pulling down a compromised NPM package or interacting with a poisoned MCP server in the course of doing its job. The concern here lies in the compression of time between publication and execution, with no human-in-the-loop to review the changelog.

Our taxonomy also includes AI software impersonation (where threat actors take advantage of users searching for legitimate AI tools); LLM poisoning (configuring a custom model, or intentionally injecting malicious or misleading content to change how a model behaves); and various attacks on models themselves, such as model extraction, training data inversion, adversarial examples, and membership inference.

Treating the two top-level categories separately provides a clearer picture of the threat. Malicious use of AI is fundamentally a productivity and volume problem; existing detection stacks may catch a lot, but the issue is scale and iteration speed. Malicious targeting of AI is a trust and supply-chain concern, more appropriately addressed with policy, Secure-by-Design frameworks, and similar initiatives.



Underground adoption and skepticism

Sophos has tracked attitudes toward generative AI on criminal forums since 2023, through a 2025 follow-up, and into ongoing monitoring. The picture has evolved significantly. Threat actors are buying brokered API keys for ChatGPT, Claude, and Grok, and forums have spawned dedicated AI channels where personas share prompt templates and jailbreaking techniques. Sophos CTU researchers noted a known underground recruiter previously observed hiring blockchain developers and call center staff for phishing sought to recruit an AI prompt engineer specifically, outsourcing AI expertise the same way threat actors outsource malware development and initial access.

CTU also observed advertisements for AI voice bots designed for vishing and call-based fraud, with positive user reviews suggesting the bots can convincingly emulate tone, cadence, and conversational patterns. The 'HackingRealm' persona advertised an "AI OnlyFans Models" service for romance fraud and social engineering, generating synthetic profile imagery and conversational content to create scalable personas on dating and social media platforms, complete with a supporting website and a feedback page listing positive reviews.

Tools marketed as 'AI-led' are appearing: Leak Bazaar (ML-powered stolen data triage), PolyEngine (polymorphic packer citing Claude Code), and an updated Cobalt Strike with REST API and MCP server integration. Adversaries are attaching LLM integrations to familiar workflows. The Claude reference in the Cobalt Strike advertisement functions partly as a modernity signal, but also reflects how LLM integration is becoming a differentiator in criminal marketplaces, even when it provides greater convenience rather than refined tradecraft.

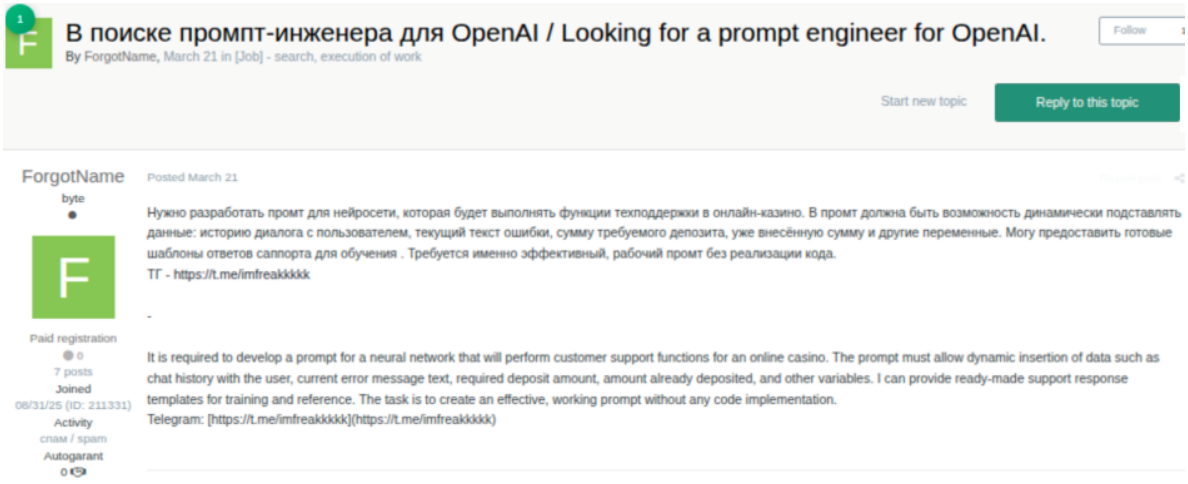
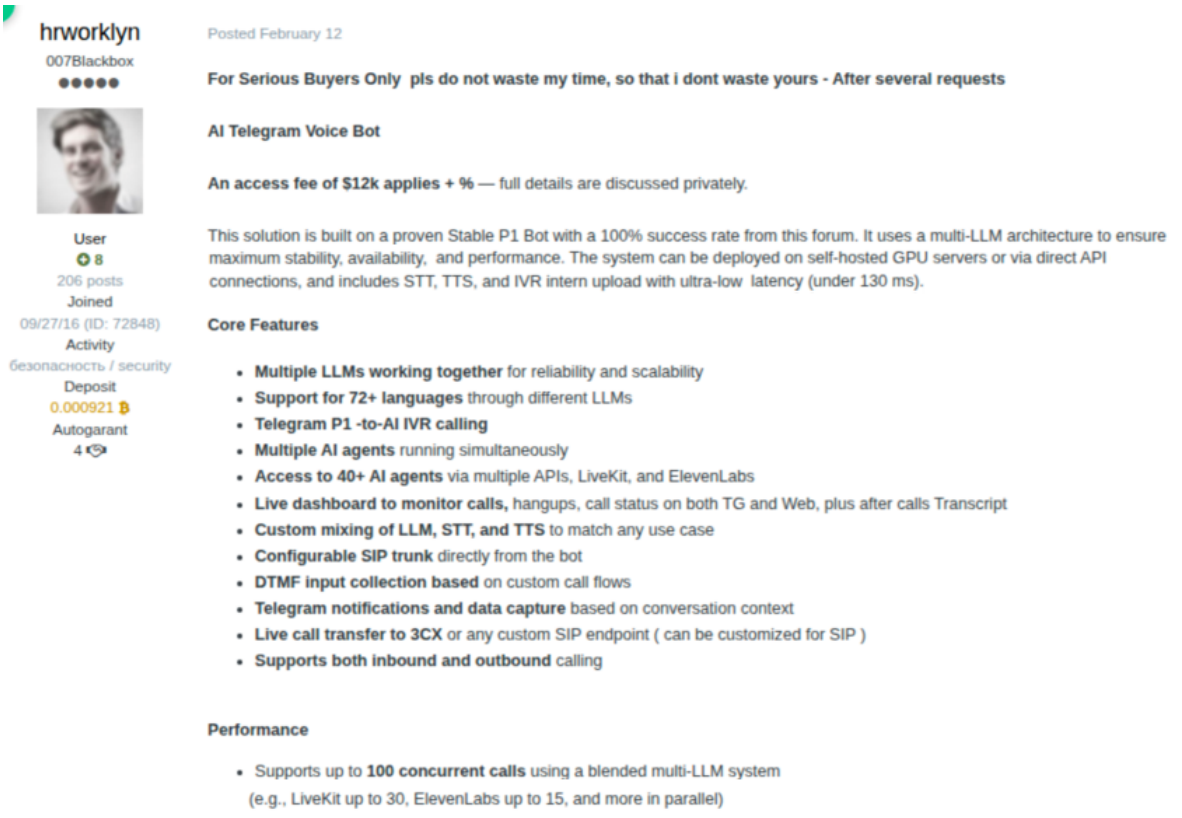


Figure 3: Recruitment post for an OpenAI prompt engineer



hrworklyn
007Blackbox
●●●●●

User
8
206 posts
Joined
09/27/16 (ID: 72848)
Activity
безопасность / security
Deposit
0.000921 ₿
Autogarant
4

Posted February 12

For Serious Buyers Only pls do not waste my time, so that i dont waste yours - After several requests

AI Telegram Voice Bot

An access fee of $12k applies + % — full details are discussed privately.

This solution is built on a proven Stable P1 Bot with a 100% success rate from this forum. It uses a multi-LLM architecture to ensure maximum stability, availability, and performance. The system can be deployed on self-hosted GPU servers or via direct API connections, and includes STT, TTS, and IVR intern upload with ultra-low latency (under 130 ms).

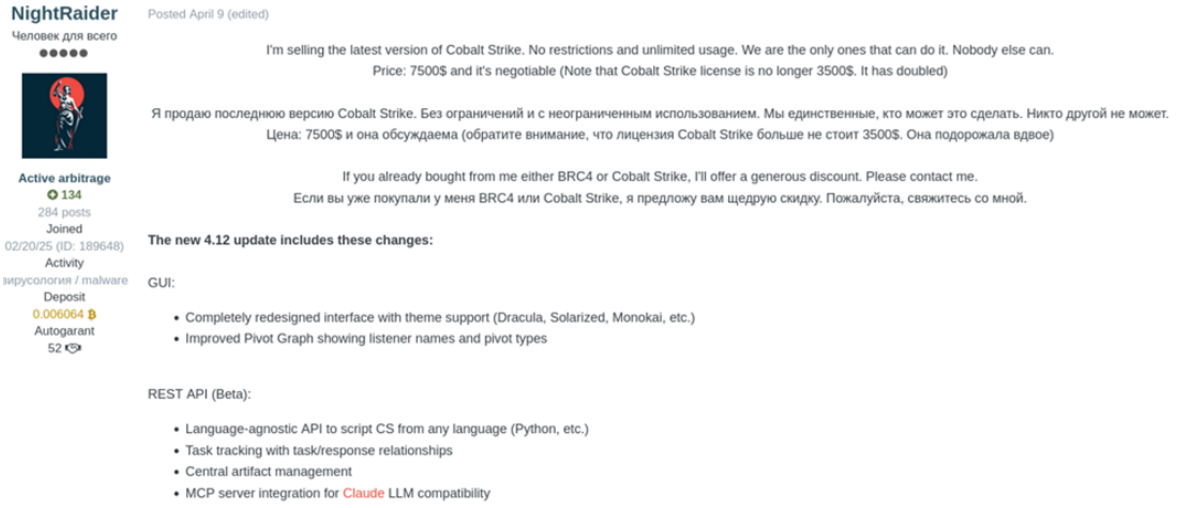
Core Features

- **Multiple LLMs working together** for reliability and scalability
- **Support for 72+ languages** through different LLMs
- **Telegram P1 -to-AI IVR calling**
- **Multiple AI agents** running simultaneously
- **Access to 40+ AI agents** via multiple APIs, LiveKit, and ElevenLabs
- **Live dashboard to monitor calls,** hangups, call status on both TG and Web, plus after calls Transcript
- **Custom mixing of LLM, STT, and TTS** to match any use case
- **Configurable SIP trunk** directly from the bot
- **DTMF input collection based** on custom call flows
- **Telegram notifications and data capture** based on conversation context
- **Live call transfer to 3CX** or any custom SIP endpoint (can be customized for SIP)
- **Supports both inbound and outbound** calling

Performance

- Supports up to **100 concurrent calls** using a blended multi-LLM system (e.g., LiveKit up to 30, ElevenLabs up to 15, and more in parallel)

Figure 4: Advertisement for an AI Telegram voice bot



NightRaider
Человек для всего
●●●●●

Active arbitrage
134
284 posts
Joined
02/20/25 (ID: 189648)
Activity
вирусология / malware
Deposit
0.006064 ₿
Autogarant
52

Posted April 9 (edited)

I'm selling the latest version of Cobalt Strike. No restrictions and unlimited usage. We are the only ones that can do it. Nobody else can.
Price: 7500$ and it's negotiable (Note that Cobalt Strike license is no longer 3500$. It has doubled)

Я продаю последнюю версию Cobalt Strike. Без ограничений и с неограниченным использованием. Мы единственные, кто может это сделать. Никто другой не может.
Цена: 7500$ и она обсуждаема (обратите внимание, что лицензия Cobalt Strike больше не стоит 3500$. Она подорожала вдвое)

If you already bought from me either BRC4 or Cobalt Strike, I'll offer a generous discount. Please contact me.
Если вы уже покупали у меня BRC4 или Cobalt Strike, я предложу вам щедрую скидку. Пожалуйста, свяжитесь со мной.

The new 4.12 update includes these changes:

GUI:

- Completely redesigned interface with theme support (Dracula, Solarized, Monokai, etc.)
- Improved Pivot Graph showing listener names and pivot types

REST API (Beta):

- Language-agnostic API to script CS from any language (Python, etc.)
- Task tracking with task/response relationships
- Central artifact management
- MCP server integration for Claude LLM compatibility

Figure 5: A user on a cybercrime forum advertising a Cobalt Strike update

But while there's clearly evidence of active experimentation and adoption, the community, at least on the forums CTU explored, is not uniformly enthusiastic. Echoing Sophos' previous findings when exploring attitudes to generative AI on criminal forums in 2023 and 2025, CTU observed personas expressing concern that AI will reduce work opportunities, particularly for malware development and scripting. Others dismissed AI outright and encouraged reliance on human skills.

AI-enhanced social engineering and deepfakes

AI is changing how quickly social engineering campaigns scale, how convincingly they read across languages, and how cheaply they can be produced. Mandiant's M-Trends 2026, for example, described a shift from generic phishing to LLM-driven personalized, rapport-building social engineering, with voice phishing climbing to the second-most common initial infection vector at 11%, while email phishing fell to just 6%. The ConnectWise 2026 MSP Threat Report confirmed active use of deepfake-enabled fraud and LLM-generated phishing campaigns, noting that AImade established tactics faster, more scalable, and more convincing rather than creating new attack categories.

Deepfakes have become an operational tool across multiple threat actor categories - most notably with North Korean threat actors, where operatives use AI-generated identities and deepfakes to secure remote employment at Western companies, establishing persistent insider access. Sophos has published a dedicated CISO playbook on North Korean IT worker tactics that covers detection indicators and organizational countermeasures.

Sophos CTU investigated a sha zhu pan scheme involving AI-themed social engineering. A UK-based victim was drawn into a fake AI-powered investment platform called DAIQ Wealth through months of AI-themed 'lessons' on the LINE messaging app. A network of personas built rapport through group chats, each operated by a different individual, suggesting an organized workforce following predefined scripts. The victim lost hundreds of thousands of pounds. When the victim expressed concern about holding almost £20,000 GBP in cash at home, the scammers arranged a physical cash courier to the UK address within 24 hours, complete with a branded receipt from a legitimate financial firm used without its knowledge.

Summing up

As we noted, our STAC6994 case study, underground recruitment of prompt engineers, AI touted as a selling point in cybercrime services-for-sale, and AI-themed fraud campaigns, all point to threat actors adopting AI to complement and enable attacks, while also experimenting with it.

A similar picture emerges across the industry. Claude assisted with attacks against Mexican government networks. Mandiant's AI Risk and Resilience report describes two documented malware families querying LLMs during execution: PROMPTFLUX, an experimental VBScript dropper using the Gemini API for just-in-time self-modification; and PROMPTSTEAL, attributed to IRON TWILIGHT (APT28), marking the first confirmed observation of state-sponsored malware querying an LLM in live operations against Ukraine. The Verizon DBIR documented VoidLink (a malware framework assembled by an AI agent in six days) and PromptLock (AI-powered ransomware first discovered by ESET in August 2025, although this was later revealed to most likely be an academic proof-of-concept, not a malicious in-the-wild sample).

However, fully autonomous AI-driven intrusion campaigns remain unconfirmed at scale. The GuidePoint GRIT 2026 ransomware report stated explicitly that concerns about "super-affiliates deploying fully autonomous ransom-bots" have been overstated, and AI/LLM usage remains rudimentary among less mature threat actors. Recorded Future's AIM3 assessment placed most observed AI malware at maturity levels 1–3 in its AI Malware Maturity Model (Experimenting, Adopting, Optimizing), noting no confirmed examples of truly embedded bring-your-own-AI malware running local models on victim hosts, and concluded that "defenders should prioritize monitoring abuse of legitimate AI services, hardening existing controls, and mapping threats to AIM3 levels rather than overreacting to sci-fi scenarios."

The Anthropic-documented GTG-1002 case (Claude Code automating 80–90% of a multi-target data theft campaign against approximately 30 entities) is the closest to autonomous operation in the public record, but it remains anomalous, rather than illustrative of a trend.

Part 2:
Enterprise AI risk

AI is already inside the enterprise: coding agents write and deploy code with developer credentials, agentic assistants read emails and call APIs, multiple systems are being connected together, and internal teams experiment with open-weight models on managed-cloud infrastructure (open-source, open-weight models are increasingly capable and typically a fraction of the cost of closed-source, closed-weight models, and enable enterprises to own their inferences and retain control of their data, but their provenance, largely, is the People's Republic of China (PRC)). The governance question is no longer whether to permit adoption, but how to secure it before the attack surface hardens around bad defaults.

Even more risk-averse organizations not heavily using AI in sanctioned use cases likely have a shadow AI problem – and, as we'll discuss shortly, fears around shadow AI and data leakage are grounded in reality. We'll discuss the specifics of risk tolerance, controls, and visibility in Part 4.

A distinct trend in 2026 is the targeting of AI-specific development infrastructure, and specifically the trust relationships developers build around AI tools.

AI development infrastructure as an attack target

A distinct trend in 2026 is the targeting of AI-specific development infrastructure, and specifically the trust relationships developers build around AI tools. Unlike most AI-related threats, which remain emerging or largely confined to theoretical proof-of-concept, this is the area where attacks are happening right now.

The **SANDWORM_MODE** npm worm campaign involved at least 19 typosquatted packages designed to mimic legitimate developer utilities and AI coding tools. The packages installed a rogue MCP server and, through embedded prompt injection, coerced legitimate AI assistants into silently retrieving SSH keys and cloud credentials and exfiltrating them without user awareness.

The **Nx Console VS Code extension compromise** harvested credentials from HashiCorp Vault, npm, AWS, GitHub, 1Password, and Anthropic API keys, and appeared to be part of the broader TeamPCP / mini-Shai-Hulud campaign. A compromised extension inside a developer's IDE has direct access to the credentials and repositories that matter most.

The **Anthropic Claude Code source leak** in March 2026, where source was inadvertently published in an npm package, illustrates a different dimension. The code reportedly contained approximately 1,900 files and 500,000 lines of code, including details of unreleased features. While Anthropic stated no customer data was exposed, analysis of the code could enable future identification of bugs and vulnerabilities. AI tooling itself has become an intelligence target.

Our recommendation: any organization with a development team should treat this as their top and immediate risk. Controls here are largely process-driven — careful dependency management, version pinning, and ensuring reviews of new open-source libraries prior to inclusion — coupled with close monitoring of developer endpoints. (In Part 4, we provide a seven-point list of actions defenders can take now to reduce the blast radius of agentic AI deployment).

Where the exposure concentrates

The identity fabric connecting AI services to enterprise systems creates exposure that existing governance was not designed to handle. IBM X-Force claimed that over 300,000 ChatGPT credentials appeared for sale on the dark web during 2025, and that a February 2025 forum post stated over 20 million ChatGPT accounts had been stolen, though that figure remains unconfirmed. The Salesloft/Drift incident demonstrated a concrete mechanism: Drift OAuth tokens were used to breach multiple Salesforce environments, showing how a chatbot's token-based access translates directly into system compromise.

Identity fabric connecting AI services to enterprise systems creates exposure that existing governance was not designed to handle.

BeyondTrust reported a 466.7% increase in active AI agents within enterprise environments over the past year. Those agents introduce specific attack surfaces: prompt injection, malicious tool invocation, and poisoned data inputs. Microsoft's own Copilot vulnerability CVE-2025-32711 (EchoLeak) demonstrated how weak input sanitization can enable zero-click data leakage. When machine identities lack MFA equivalents, carry long-lived secrets, and operate with excessive privilege, they become a viable and attractive target.

Attackers are well aware of these opportunities, and are actively enumerating the attack surface. In January 2026, GreyNoise described campaigns mapping LLM deployments. Researchers documented Operation Bizarre Bazaar: a campaign hijacking exposed LLM and MCP endpoints, validating them, and reselling access through infrastructure linked to silver.inc.

The governance gap

Across the industry, and particularly among senior leadership, there is both concern over the security implications of AI, and a lack of capacity and capability to deal with those implications.

ChatGPT alone generated 410 million DLP policy violations in Zscaler's data. The Saviynt/Cybersecurity Insiders CISO AI Risk Report 2026 found 75% of respondents had discovered shadow AI tools, and 95% doubted they could detect or contain misuse. Only 1% of enterprises have a dedicated AI security budget (Pentera), and according to Splunk's 2026 CISO Report, while there is a consensus that AI can assist with productivity and managing volume, a majority of CISOs are concerned about data leakage, shadow AI, and hallucinations. Interestingly, Splunk found that those fears were higher among CISOs at organizations with greater AI maturity, and suggested that those "who are further along in their generative AI journey are starting to notice some trade-offs."

These trade-offs and fears, particularly around shadow AI, are grounded in reality. In April 2026, Vercel disclosed a breach triggered by an employee's use of Context.ai, a third-party AI productivity tool. The employee registered with their corporate Google Workspace account and granted 'Allow All' permissions to the Context.ai OAuth app. When Context.ai was compromised – reportedly via a Lumma Stealer infection on a Context.ai employee device – attackers inherited those OAuth tokens and pivoted into Vercel's internal environments.

The governance challenge is that AI adoption is outpacing the organizational processes designed to manage it – and the regulatory environment is tightening, while the gap between deployment velocity and governance maturity is widening. The EU AI Act requires high-risk systems to comply with it fully in 2026, carrying penalties up to 7% of global turnover. California's transparency requirements took effect January 1, 2026. Yet Saviynt/Cybersecurity Insiders found that while 71% of large enterprises have deployed AI agents accessing core business systems, only 16% govern that access effectively.

As a starting point for addressing the gap, prompt traffic should be treated as a first-class security log: prompt capture and analysis can expose injection attempts, exfiltration, insider misuse, and unexpected agent behavior. It can begin at the proxy layer for cloud-hosted models or at the harness layer for locally deployed agents, without waiting for full model integration. But prompts are only part of the answer; tools and skills are becoming a fundamental part of the attack surface, and should be treated as such.

AI adoption is outpacing the organizational processes designed to manage it – and the regulatory environment is tightening

Part 3:
How AI is changing defensive operating models

Triage and investigation

Playbook execution, timeline reconstruction, and natural language reporting are where reasoning models can add genuine operational value. For example, Kaspersky's MDR infrastructure processed approximately 400,000 alerts in 2025, with AI-powered detection logic filtering false positives before analyst review; 39,000 alerts were escalated for further investigation. Increasingly, AI models are capable of undertaking those investigations; Prophet Security highlighted two cases in The Hacker News in March 2026, where models performed contextual analysis and investigations of incidents in which malicious activity was not immediately obvious.
The first involved a dormant user with an old API key that suddenly became active; the second was discovering malicious intent in a phishing email that would have passed most, if not all, security and validation checks. In both, Prophet Security notes, no individual data point was suspicious in itself; the threat only became apparent through contextual, multi-point analysis – the kind of analysis that human investigators perform when they have the bandwidth. Of course, the problem is that investigators often don't have that bandwidth, and that's where reasoning models can contribute, by affording every alert the same level of analysis.

However, AI is not a panacea, and it needs careful engineering. Research on overthinking shows that beyond a threshold, additional reasoning makes models reverse correct answers. Negative answer flips outnumbered positive ones after roughly 7,000 tokens, reaching a three-to-one ratio by 12,000. SOC implementations should therefore treat reasoning effort, retry depth, and tool-call loops as tunable engineering parameters.

Broadly, adoption is uneven, and expectations need calibration. Only 40% of CISOs currently employ generative AI in their security functions, according to the Splunk CISO report. Agentic AI shows just 6% active deployment, though 39% are exploring it.

There's also the fact that AI tools aimed at defenders can carry their own risk surface. The Cobalt State of Pentesting report, for example, found that AI and LLM applications harbor high-risk findings at 2.7 times the rate of traditional software, with only 38% of issues resolved during pentesting.

The memory problem

Most production deployments are stuck at what some researchers describe as generation two or three of agentic maturity: tool-augmented assistants that start fresh every session. They retain their training but cannot remember that the same alert pattern appeared yesterday and turned out to be benign, or that blocking a particular IP range last month disrupted the VPN.

The distinction between a senior analyst and a junior one is not necessarily reasoning ability, but what they bring to each event: episodic memory of past incidents, semantic knowledge of the environment, and procedural intuitions about what works. Current AI assistants are more junior than senior, and investigate every alert as if for the first time. Researchers are exploring memory-as-operating-system architectures, external memory-as-middleware services, and knowledge-graph approaches with bi-temporal reasoning. Until these mature, AI systems will stay useful assistants, rather than autonomous operators.

The detection boundary

Behavioral detection systems operate on live telemetry streams at extreme class imbalance. A platform processing trillions of events per day aggregates into millions of deduplicated detections, prioritizes into thousands of high-severity alerts, and produces hundreds of daily investigations that remain critical after human review. The end-to-end ratio approaches one in ten billion. Stefan Axelsson's foundational work on the base rate fallacy in intrusion detection formalized the math: even a detector with 99.9% specificity produces millions of false alarms for every true threat when the prior probability of attack is sufficiently low.

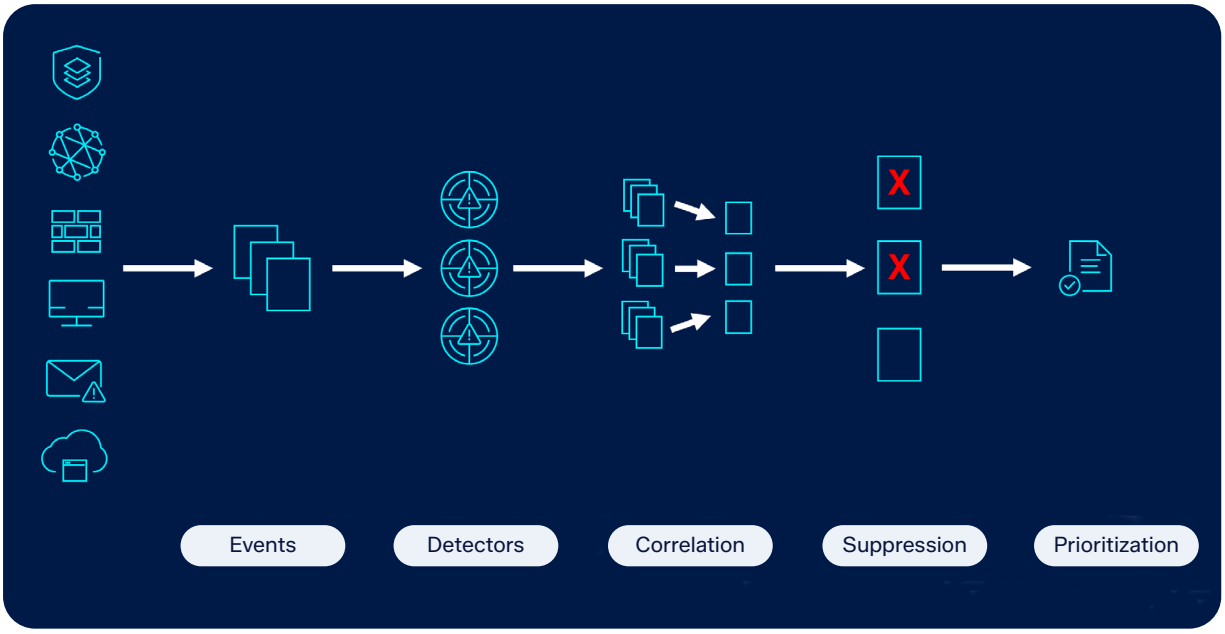


Figure 6: Overview of the multi-step pipeline

Sophos research presented at NorthSec 2026 illustrates why layered detection still matters. A two-week window of telemetry comprising 11.8 trillion events was reduced through detectors of increasing complexity (simple indicator matching, correlation rules, purpose-built ML models, deduplication, correlation, and suppression) to 81,573 high and critical alerts – an average of fewer than 50 per organization. Within those remaining alerts, Sophos identified an infostealer attack linked to the TamperedChef campaign.

Reasoning models belong above this layer. They generate investigation hypotheses, retrieve and synthesize telemetry into evidentiary summaries, and coordinate bounded response actions under explicit guardrails. They do not replace models trained on proprietary incident data, environment-specific baselines, and continuous feedback loops that general reasoning systems do not have access to.

The McKinsey AI Trust Maturity Survey quantified the organizational constraint: security concerns are the top barrier to scaling agentic AI for nearly two-thirds of respondents. Inaccuracy (74%) and cybersecurity (72%) are the leading specific risks. Only about 30% of organizations have achieved maturity level three ("Steps are being taken to develop all necessary responsible AI practices") or higher in AI governance. Organizations with explicit ownership for responsible AI score an average maturity of 2.6, compared to 1.8 for those without clear accountability.

The constraint on AI-assisted defense is organizational readiness. The models are capable enough; the question is whether organizations can operate them at the trust level autonomous security operations require.

Sophos AI research: Defending the AI layer

In addition to the layered detection research described above, the Sophos AI team has developed several other novel defense techniques, each addressing a specific failure mode.

- **The Untrusted Data Scanner (UDS)** inspects external content for prompt injection before it reaches an LLM or agent. Every LLM use case has an instruction channel (what the application tells the model to do) and a data channel (the untrusted content it processes). Prompt injection smuggles instructions into the data channel. UDS is deliberately scoped to this specific problem, as opposed to being a general jailbreak detector. Early results show a low false-positive rate, which matters because security data is full of security language – incident reports, malware write-ups, pentest findings – and a detector that 'cried wolf' on every threat report would be unusable.
- **LLM Scoping** classifies whether queries fall within a feature's intended function, rejecting out-of-scope requests before they reach the LLM. Across evaluated methods, scoping models significantly outperformed system-prompt-based defenses and reduced attack success rates to near zero on multiple public benchmarks. The team also applied Multi-round Automatic Red Teaming (MART) to harden scoping models against adversarial adaptation.

- **CerBERTus**, a 'three-headed' BERT-based model, addresses the brittleness of binary jailbreak detection. A single shared encoder feeds three classification heads: harmfulness (primary), goal category (what the user is trying to do), and framing style (how the request is presented). The auxiliary tasks act as inductive bias, encouraging the representation to separate objective from wrapper. The team built a structured factorial prompt corpus crossing goals (cyberattacks, fraud, explosives, drug synthesis, privacy, extremist propaganda, and more) with adversarial frames (roleplay, fiction, urgency, academic pretext, obfuscation) to train and stress-test this separation.
- **Anomaly detection**: In research presented at Black Hat USA 2025, the team combined anomaly detection with LLMs and found that the method's success came from generating diverse benign command lines, not from locating malicious ones – significantly reducing false-positive rates in command-line classifiers.
- **LLM salting**, presented at CAMLIS 2025, takes inspiration from password salting: introducing targeted behavioral variations so that precomputed jailbreaks cannot be reused across homogeneous LLM deployments.

Deception against autonomous attackers

For two decades, cyber deception never became a primary control. Against autonomous attackers, the calculus changes. AI agents do not get tired and can enumerate every path at machine speed. Deceptive environments, such as those proposed in the Palisade LLM Agent Honeypot and the MANTIS framework, can fill an agent's context with misleading information, waste its inference budget, and impose real economic costs. One such example, HoneyTrap, demonstrated a 149% increase in 'attacker resource consumption' compared to conventional defenses. The field is immature, but the conditions under which deception becomes primary are precisely the conditions offensive AI scaling is creating.

Where the tempo asymmetry bites

An attacker using AI faces no procurement cycle, no compliance review, and no change advisory board. A threat actor can adopt a tool and operationalize results in days. The AI-assisted campaign against FortiGate devices compromised over 600 firewalls across 55 countries in five weeks – the work of a threat actor using commercial AI tools.

On the defender side, a detection fires, AI-assisted investigation produces a containment recommendation in minutes, but the customer's remediation process requires a change ticket, a maintenance window, and a 48-hour SLA. The investigation accelerated but the response did not. This was possibly the case with two vulnerabilities rapidly exploited this past year. As we noted earlier, CVE-2026-10520 (Ivanti Sentry) was exploited within 24 hours of proof-of-concept publication, and CVE-2026-42208 (LiteLLM) within 36 hours.

While AI-assisted triage undoubtedly helps defenders with tempo, work done in advance on security fundamentals can contribute towards addressing the asymmetry here: reducing attack surface, enforcing MFA, and maintaining telemetry complete enough to reconstruct what happened when prevention fails.

Part 4:
What security leaders should do

The evidence in this report supports a small number of conditional decisions, organized by where an organization sits today.

Calibrating risk tolerance for AI adoption

The evidence in this report creates a tension that organizations must find a way to resolve: AI adoption carries real risk, but slow adoption carries risk too. Moving too cautiously while attackers accelerate means falling behind on the defensive tempo that this report argues is the defining challenge. Accepting risk is an inherent part of that; the question is how to take the right risks.

The evidence in this report creates a tension that organizations must find a way to resolve: AI adoption carries real risk, but slow adoption carries risk too.

Sophos's internal risk framework separates good risks from bad ones along two axes: upside potential and downside containment. A good risk has meaningful impact, measures to contain the blast radius, a reversible path (pilot, rollback, or fallback), clear ownership, and defined feedback loops. A bad risk has unbounded harm potential, violates non-negotiables (security, privacy, legal, compliance), lacks a recovery plan, has unclear ownership, or skips human judgement on changes that directly affect customers.

For AI specifically, the framework translates into practical questions before any deployment: what is the worst credible outcome? How big is the potential blast radius? Can we reverse it? Who owns it? Deployments with contained downside, such as a read-only pilot of an agentic triage tool on an isolated dataset, should be pursued aggressively. Those with high downside, giving an agent write access to production infrastructure, need de-risking or scope limits before proceeding. Low-impact, high-downside deployments should be avoided outright.

Non-negotiables apply regardless of circumstance: customer trust, safety, security and privacy, legal and compliance, and operational stability. The CISO's role in an AI-accelerated environment is to ensure that the organization takes smart risks, while avoiding reckless ones.

For organizations deploying agentic AI: blast-radius reduction

Sophos has proposed seven controls, actionable in the next one to six months, to reduce blast radius for organizations seeking to deploy agentic AI:

1. **Agent sandboxing** places a controlled boundary around the agent process. Most harnesses ship with some form of sandboxing, but it is typically opt-in. Where possible, choose remote or cloud-based execution environments over local sandboxes for stronger process separation.

2. **Credential isolation** ensures the agent never sees secrets directly: a separate process resolves credentials from a vault, injects them into the API call, and returns only sanitized responses, keeping long-lived credentials out of LLM context windows.

3. **Sealed tool endpoints** go further. The agent calls a tool by name, but a broker process holds the credential, makes the actual API call per a fixed schema, and enforces per-tool egress allowlists. The agent's only lever is choosing which tool to invoke and what parameters to pass.

4. **Egress restriction and network monitoring** treat the compromised agent as a data exfiltration channel: secret detection in outbound traffic, volume thresholds on outbound requests, and web categorization blocking requests to uncategorized or newly-registered domains.

5. **EDR coverage** must extend to agent host behavior: containers, ephemeral VMs, and unmanaged developer machines are common blind spots.

6. **Human-gated approval** separates the agent's autonomous execution plane from a human-governed control plane: irreversible operations require a cryptographic ceremony, rather than a soft OK button the agent could simulate.

7. **Injection propagation boundaries** address the escalating risk as injections move from session-scoped through agent-level to cross-agent propagation. Treat memory writes and cross-agent handoffs as security events.

When OpenClaw gained traction in early 2026, over 30,000 instances were exposed on the internet and threat actors were already discussing how to weaponize its modular 'skills' for botnet campaigns. Sophos's Red Team tested it directly: 23 actionable findings, AD reconnaissance compressed from three days to three hours, and audit detail that would not be achievable manually in a reasonable timeframe. But the team spent more time building safety boundaries than running the test. The lesson: agentic AI is powerful enough to be worth deploying, but only if the security framework is built first.

Visibility, identity, and patching

We recommend starting with visibility into AI service usage. The AI exposure documented across the Salesloft/Drift and Vercel incidents makes this the highest-leverage first move. Inventory the AI tools in your environment. Identify which ones hold OAuth tokens, API keys, or service accounts connected to enterprise systems. Determine what data flows through them and what permissions they hold. You cannot govern what you cannot see.

The visibility work should produce an AI registry: every AI tool, agent, model, credential it can reach, sandboxing configuration, and expected connectivity requirement. Without that registry, controls such as credential isolation, egress restriction, and sandboxing cannot be applied consistently. Apply identity controls with the same rigor as any other SaaS application: enforce conditional access policies, require MFA for AI service accounts, apply DLP controls to the egress paths AI tools use, and rotate credentials for AI-connected integrations on the same schedule as any other privileged access. The IBM X-Force credential data findings and the Nx Console supply chain attack both confirm that AI service credentials are a harvesting target.

Organizations whose patching cadence is still measured in weeks face worsening risk. The CISA BOD 26-04 three-day timeline is a floor, because AI-assisted exploitation is compressing the disclosure-to-exploitation window to hours. If a critical internet-facing vulnerability cannot be patched within days, that gap carries measurable risk.

AI-assisted exploitation is compressing the disclosure-to-exploitation window to hours.

Evaluate with evidence

Instrument the AI tools your analysts use. Measure validation effort, how much time analysts spend checking AI output. Measure prompting efficiency, how many attempts before useful results. And measure trust calibration — whether analysts over-trust or under-trust the system. Build test sets from real SOC workloads rather than generic benchmarks, and version your procedures so they survive model upgrades.

Secure the AI supply chain

AI introduces supply chain risks that extend beyond traditional software dependencies. Model weights, training data provenance, MCP server integrity, and the inference infrastructure itself all represent trust boundaries. Organizations deploying open-weight models through managed-cloud providers need to confirm the exact model version, serving region, retention period, training-use policy, logging surface, and contract terms before sending private code or data. For models served through PRC-based SaaS endpoints, such as DeepSeek's API, a stricter boundary applies: such models should be used for public or synthetic capability testing, not for proprietary repositories or sensitive engineering context.

Conclusion

The STAC6994 threat actors' framework contained Cobalt Strike profiles, EDR bypass modules, credential harvesting tools, lateral movement utilities, and a C2 channel hidden behind legitimate cloud infrastructure. Every component was familiar to us – except the development process. AI agents iterated on evasion techniques faster than a human developer could have, tested them against specific products in a structured lab, versioned the results through Git, and handed the output to an operator who deployed ransomware and stole data.

Across Sophos MDR and IR casework, CTU intelligence, SophosLabs analysis, Sophos AI research, and the broader security industry, AI is accelerating familiar operations on both sides of the security equation. It helps attackers move faster through known attack chains. It enables defenders to triage and investigate at compressed timescales. It creates new exposure through the identity and governance layer surrounding enterprise AI adoption. It has not produced a fundamentally new intrusion type that existing detection architecture cannot address. The UK AI Security Institute measured a roughly sixfold improvement in autonomous offensive capability over 18 months. Each model generation unlocked steps in multi-stage intrusion scenarios that prior generations could not complete. That trajectory does not permit complacency.

Three themes sharpen the conclusion:

1. **Classification** has to be precise, because lumping all AI threats together obscures the defensive posture each requires.
2. **Evaluation** has to be rigorous, because whether the question is model routing, alert suppression, or agent reasoning depth, the answer should come from measurement.
3. **Transparency** has to be real, because trust accrues to organizations that show their work and share useful failures alongside successes.

What Sophos will watch in the next 12 months: whether autonomous agents move from executing known playbooks to discovering novel attack paths at scale; whether the underground economy develops productized AI offensive tooling that lowers the bar as dramatically as ransomware-as-a-service lowered the bar for extortion; whether enterprise AI governance matures fast enough to close the identity exposure the current adoption wave has created; and whether the exploitation-to-patch window compresses further, forcing a reckoning with organizations that still measure their response cadence in weeks.

The clock has moved. Defenders who do not adjust their tempo will find themselves further behind than they assumed.

Contributors

Nash Borges
SVP of
Engineering and AI

Adarsh Kyadige
Senior Manager
AI Research

Ross McKerchar
CISO

Rafe Pilling
Director of Threat
Intelligence
X-Ops Counter Threat Unit

Matt Stec
Director
Sophos AI

Ryan Westman
Senior Manager Threat
Research
X-Ops Insights

Matt Wixey
Senior Threat Researcher
X-Ops Insights

Our team

Sophos X-Ops is a cutting-edge cybersecurity initiative, bringing together more than 1,000 experts from various specialized security domains within Sophos, including Threat Intelligence, Artificial Intelligence, MDR operations, and internal security operations.

The cross-functional task force strengthens organizational defenses against today's fast-evolving and highly sophisticated cyber threats.

By leveraging the combined expertise of its task force, Sophos X-Ops offers a multidimensional response to cyberattacks, ensuring comprehensive protection, detection, and response capabilities.

This collaborative and innovative approach delivers unparalleled threat response, positioning Sophos as a benchmark for excellence, and a leader in cybersecurity.

Sophos X-Ops leverages the combined expertise of its cross-functional task force. The synergy among Sophos X-Ops' cross-functional teams fuels shared intelligence, enabling them to rapidly adapt to evolving insights – accelerating detection and response times while strengthening overall protection capabilities for Sophos customers.

SOPHOS

To discuss your AI security needs and how Sophos can help, visit our website or speak to an advisor.

United Kingdom and Worldwide Sales
Tel: +44 (0)8447 671131
Email: sales@sophos.com

Australia and New Zealand Sales
Tel: +61 2 9409 9100
Email: sales@sophos.com.au

North America Sales
Toll Free: 1-866-866-2802
Email: nasales@sophos.com

Asia Sales
Tel: +65 62244168
Email: salesasia@sophos.com

2026-07-20 EN (NP) CRE-5820