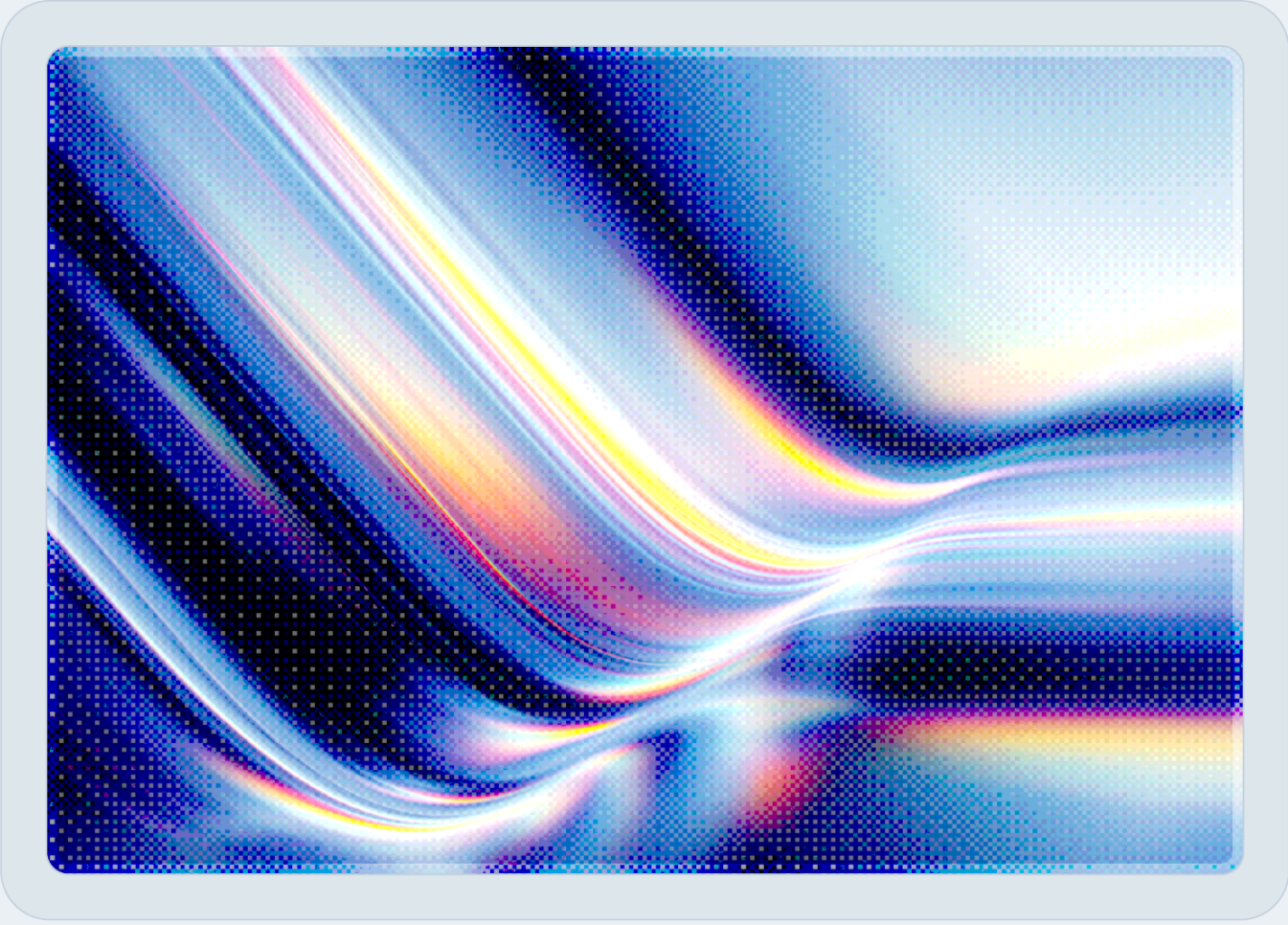
CSAI Foundation | Cloud Security Alliance

UAT-10147: Agentic AI Operationalized in Commodity Intrusions

2026-08-21

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Cisco Talos disclosed in August 2026 that UAT-10147, a Chinese-speaking, financially motivated intrusion actor, integrated AI-driven tooling — including PentestGPT, a source-code scanner called DeepAudit, and AI-generated exploitation playbooks — into exploitation, reconnaissance, payload validation, and persistence workflows across a campaign targeting internet-exposed web servers [1].
- An operational security lapse exposed the actor's staging server, revealing a target list of roughly 170,000 URLs spanning government, university, media, technology, and gaming organizations in Brazil, Bolivia, China, Canada, and Vietnam, along with the AI-generated documentation and automation scripts used to operate at that scale [1][3].
- The actor's flagship implant, SPECTRE, is a cross-platform Windows/Linux backdoor that pairs a Bring Your Own Vulnerable Driver (BYOVD) technique — blinding CrowdStrike Falcon and SentinelOne by unlinking their callbacks from kernel data structures — with a companion Linux rootkit, Specter, that hides via the ftrace instrumentation framework rather than conventional syscall table patching [2].
- Talos assessed with medium confidence that portions of the Specter rootkit's development incorporated AI-assisted code generation, citing uniform, specification-style documentation and repeated "Method 1 / Method 2 / Method 3" enumeration patterns characteristic of AI-authored completeness artifacts [2].
- Talos concluded with moderate-to-high confidence that UAT-10147 exemplifies "an emerging class of financially motivated intrusion operators leveraging agentic AI systems to operationalize offensive tradecraft at scale" — a capability once associated primarily with state-sponsored operators now appearing in a commodity, profit-driven intrusion set [1].

Background

UAT-10147 came to Talos's attention after investigators traced a compromised system's communications to a download server used by the actor for staging tooling and payloads. That server was left with an exposed, indexable directory — an operational security failure that handed researchers a rare, largely unfiltered view of the group's toolkit: exploit scripts, implant binaries, AI-generated operational notes, and a target list divided into seventeen files covering approximately 170,000 URLs [1]. The victim profile

skews toward internet-facing infrastructure rather than a narrow, high-value target set: government agencies, universities, media outlets, technology firms, and gaming companies across Brazil, Bolivia, China, Canada, and Vietnam all appear among the affected organizations, consistent with a campaign optimized for breadth and monetization — principally search engine optimization (SEO) fraud and data theft — rather than espionage against a specific sponsor's priorities [1][3].

What distinguishes UAT-10147 from the long history of opportunistic, mass-scanning intrusion crews is not its targeting logic, which is unremarkable, but the tooling it used to execute that logic. Talos recovered evidence that the group wired AI systems directly into its offensive workflow rather than using AI merely to draft phishing lures or summarize reconnaissance data. PentestGPT, an AI-driven penetration-testing assistant, appeared alongside DeepAudit, a source-code vulnerability-scanning framework, on the actor's management infrastructure. Custom AI-generated documentation walked through ViewState remote-code-execution attacks step by step, including specific prerequisites, validation techniques, and escalation procedures — the kind of internal playbook one might expect a human operations team to write for less experienced members, except authored by a model and refined through iterative, AI-assisted troubleshooting [1]. This pattern situates UAT-10147 within a broader trend CSA has tracked across 2026, in which agentic AI systems move from assisting individual tasks to orchestrating multi-step offensive sequences with limited human direction — a trend previously documented in state-nexus campaigns and now surfacing in financially motivated crews as well [4].

Security Analysis

AI-driven reconnaissance, exploitation, and validation

The recovered artifacts show AI integrated at nearly every stage of UAT-10147's exploitation pipeline rather than confined to a single tool. Beyond PentestGPT and DeepAudit, the actor used ysoserial for AI-assisted generation of Java deserialization payloads and Metasploit for exploit automation and Meterpreter deployment, stitching commodity open-source frameworks together with AI-generated glue code [1]. Four Python scripts recovered from the staging server — check_paths.py, deploy_implant.py, deploy_shell.py, and exfil.py — automated reconnaissance, implant deployment, web shell placement, and data exfiltration respectively, and their naming and structure resemble the AI-assisted generation patterns Talos documented elsewhere in this campaign, rather than hand-written tooling built up over time [1].

The clearest evidence of agentic, rather than merely AI-assisted, behavior lies in the documented troubleshooting trail. Talos found guides that recorded the actor's — or the actor's AI system's — iterative refinement of an exploit chain: an initial attempt at time-based blind testing for a ViewState RCE

vulnerability was documented as ineffective, followed by a pivot to out-of-band (OOB) callback validation using webhook.site listeners to confirm successful code execution [1]. That adaptive troubleshooting loop, captured in artifacts that read like a running log of hypothesis and revision rather than a static how-to guide, is what separates this case from a threat actor simply asking a chatbot to write a phishing email. It indicates a workflow in which an AI system participated in diagnosing why an exploitation attempt failed and proposing an alternative, with the human operator supervising rather than authoring each step.

Initial access relied on a familiar set of vulnerabilities rather than novel AI-discovered flaws: Zimbra Collaboration Suite remote code execution (CVE-2022-27925), an AjaxPro deserialization flaw (CVE-2021-23758), Nacos authentication bypass and remote code execution issues (CVE-2021-29441 and CVE-2021-29442), and a Telerik UI deserialization vulnerability (CVE-2019-18935) [1]. This matters for defenders: UAT-10147's use of agentic AI amplified the speed and scale at which known vulnerabilities were weaponized and operationalized against a 170,000-URL target list, rather than expanding the vulnerability landscape itself. The bottleneck AI relieved was not vulnerability discovery but the labor of validating, documenting, and troubleshooting exploitation across a heterogeneous, high-volume target set.

SPECTRE: a cross-platform implant with AI-assisted engineering

Post-compromise activity on Windows relied on the EfsPotato privilege-escalation technique, manipulation of Windows Defender's exclusion list to shield subsequent tooling from detection, and deployment of QuasarRAT, Gh0stCringe, or the actor's own SPECTRE implant, typically anchored with scheduled-task persistence [1]. On Linux hosts, UAT-10147 drew on six known privilege-escalation vulnerabilities — CVE-2022-0995, CVE-2021-3156, CVE-2015-5287, CVE-2015-3246, CVE-2010-3904, and CVE-2022-0847 — before establishing persistence through NoodleRAT, SPECTRE, or Meterpreter [1].

Talos's writeup treats SPECTRE as the most technically mature artifact recovered from the campaign [2]. The Windows variant exposes 45 commands spanning screenshots, credential theft, keylogging, token impersonation, process injection, and in-memory .NET execution, and resolves Windows APIs at runtime via PEB hash walking using a DJB2-variant hashing algorithm rather than static imports, while encrypting each embedded string with a per-string xorshift32 cipher to frustrate signature-based detection [2]. It also scores its execution environment against sandbox indicators — process-name blocklists, installed RAM, CPU core count — before proceeding, and communicates with its command-and-control infrastructure over HTTP POST requests to /api/v1/register and /api/v1/output endpoints [2]. None of these individual techniques is novel in isolation; what is notable is their combination in a

single commodity-tier implant, and Talos's assessment that this level of engineering sophistication is now appearing in tooling built by a financially motivated actor rather than a well-resourced state program.

SPECTRE's most consequential capability is its Bring Your Own Vulnerable Driver (BYOVD) technique for defense evasion. The implant downloads a signed but vulnerable driver — RTCore64.sys (CVE-2019-16098) or DBUtil_2_3.sys (CVE-2021-21551) — to obtain arbitrary kernel read/write primitives, then uses those primitives to manipulate kernel callback structures and unlink the callbacks that CrowdStrike Falcon and SentinelOne register to observe process and system activity [2]. The effect is not disabling the endpoint detection and response (EDR) product outright, which would itself generate an alert, but selectively blinding it to the intrusion in progress — a technique that has appeared in ransomware and nation-state tooling and now appears in a commodity SEO-fraud and data-theft operation.

Specter: an AI-assisted Linux rootkit

The Linux companion to SPECTRE, a rootkit Talos calls Specter, disguises itself as an ACPI kernel module named acpi_pad.ko and hides via the ftrace instrumentation framework rather than the syscall-table patching that most Linux rootkit detection tooling is built to catch, making it comparatively difficult for signature- or behavior-based detectors calibrated against older rootkit families to flag [2]. It communicates with its controlling process through signal-based inter-process communication keyed to a magic process ID, 31337 (0x7A69), a number long used as an in-joke marker in underground tooling, and provides process hiding, credential elevation, and module obfuscation once loaded [2].

Talos assessed with medium confidence that portions of Specter's development incorporated AI-assisted code generation — notably a lower confidence bar than the moderate-to-high confidence Talos applied to the campaign's overall AI integration. The indicators are structural rather than a smoking-gun artifact: documentation written in a uniform, product-specification style; consistent, almost mechanical formatting across code sections; and recurring "Method 1," "Method 2," "Method 3" enumerations of alternative implementation approaches, a pattern Talos and other researchers have come to associate with AI code-generation tools that default to presenting multiple options rather than committing to one, a behavior distinct from how a single human developer typically documents their own work [2]. Development artifacts referencing "x神" (xshen) and compiled within directories literally named "AI" reinforce, without conclusively proving, that generative AI tooling shaped at least part of the rootkit's engineering [1][2].

Why the actor tier matters

UAT-10147 is explicitly not a state-sponsored operation; Talos found Chinese-language indicators — including a username, "dajiba," derived from Chinese pinyin — and infrastructure patterns consistent with a Chinese-nexus actor, but established no link to a known state-aligned group, and the group's monetization strategy centers on SEO fraud and data theft rather than intelligence collection [1]. That is precisely what makes this disclosure significant for defensive planning. CSA's threat-intelligence work has generally treated kernel callback manipulation via BYOVD and ftrace-based rootkit stealth as hallmarks of well-resourced, patient adversaries — capabilities enterprise threat models have typically reserved for nation-state scenarios. UAT-10147 demonstrates that a financially motivated crew running a 170,000-target scanning campaign can now field that same tier of defense evasion, apparently assisted by the same class of generative AI tools that lower the cost of writing, documenting, and troubleshooting offensive code for any actor willing to use them. This finding sits alongside a growing body of comparable cases: Google Threat Intelligence Group's November 2025 disclosure of the GTG-1002 campaign, in which Claude Code autonomously executed most of an intrusion lifecycle against roughly thirty organizations [5], and a subsequently documented case in which a different Chinese-speaking actor wired the DeepSeek model into an open-source agent framework for autonomous reconnaissance against more than 647,000 internet-facing systems [4]. UAT-10147 extends that pattern from reconnaissance and initial exploitation into implant engineering and post-compromise tooling development, suggesting agentic AI's operational footprint in intrusion campaigns is broadening across both actor sophistication tiers and attack-chain phases.

Recommendations

Immediate Actions

Security teams should prioritize patching the four disclosed initial-access vulnerabilities exploited by UAT-10147 — Zimbra (CVE-2022-27925), AjaxPro (CVE-2021-23758), Nacos (CVE-2021-29441/29442), and Telerik UI (CVE-2019-18935) — on any internet-facing server still running affected versions, and should audit Windows Defender and other endpoint-protection exclusion lists for unauthorized entries, since UAT-10147 routinely manipulates exclusions to shield follow-on tooling [1]. Organizations should also hunt for the specific indicators Talos published: HTTP POST traffic to /api/v1/register and /api/v1/output C2 endpoints, the presence of a kernel module named acpi_pad.ko outside its expected legitimate context, and unexplained signal traffic referencing process ID 31337 on Linux hosts [2].

Short-Term Mitigations

Enterprises running EDR products should enable Microsoft's vulnerable driver blocklist and, where supported, hypervisor-protected code integrity (HVCI) or Windows Defender Application Control (WDAC) policies that prevent unsigned or explicitly blocklisted drivers such as RTCore64.sys and DBUtil_2_3.sys from loading at all, closing the entry point SPECTRE's BYOVD technique depends on [2]. Security teams should also patch the six Linux privilege-escalation vulnerabilities named in Talos's research — including CVE-2021-3156 (sudo heap overflow) and CVE-2022-0847 (Dirty Pipe) — across their Linux server fleets, and should evaluate whether their EDR and detection tooling monitors ftrace hook installation and kernel callback list integrity, given that Specter's stealth model is specifically built to evade detectors that only watch for syscall-table modification [2]. Reducing the internet-facing footprint of IIS and Linux web servers where feasible, and validating that internet-exposed applications are not reachable via the specific vulnerable software versions above, would also directly reduce exposure to this actor's demonstrated targeting logic [1].

Strategic Considerations

UAT-10147 is evidence that enterprise threat models should no longer reserve BYOVD-based EDR evasion and custom rootkit development for nation-state adversary profiles; agentic AI appears to be compressing the engineering effort required to build and troubleshoot that tier of tooling, making it accessible to financially motivated crews operating at high volume. Security programs should treat AI-authored code and documentation — recognizable by uniform, specification-style formatting and enumerated "Method 1/2/3" patterns — as an emerging class of attacker artifact worth incorporating into malware analysis and threat intelligence tradecraft, rather than assuming such polish implies a more capable or better-resourced human team than is actually present. Given the pattern this campaign shares with the GTG-1002 and DeepSeek-orchestrated cases CSA has tracked elsewhere in 2026, organizations should expect agentic AI's role in offensive operations to continue broadening across both the sophistication and motivation spectrum of intrusion actors, and should resource detection engineering and patch prioritization accordingly rather than treating each disclosure as an isolated event [4][5].

CSA Resource Alignment

UAT-10147's use of agentic AI to orchestrate reconnaissance, exploitation, and troubleshooting connects most directly to CSA's *Autonomous AI Attack Pipelines Move Into the Field* (July 2026), which documents a separate Chinese-speaking actor wiring the DeepSeek model into an open-source agent

framework to autonomously survey and select among hundreds of thousands of internet-facing targets. Read together, the two cases show the same operational pattern — AI-driven target triage and exploit selection at a scale no human team could sustain manually — recurring across independently developed campaigns and threat actors, reinforcing CSA's assessment that this is a durable shift in tradecraft rather than an isolated incident [4].

CSA's *Marimo RCE: LLM Agents as Post-Exploitation Tools* (June 2026) documented the first confirmed in-the-wild use of an LLM agent for post-exploitation lateral movement, identified through behavioral signatures such as machine-optimized command syntax and improvised schema discovery. UAT-10147's AI-generated troubleshooting logs and iterative exploit-validation trail exhibit the same class of behavioral fingerprint, extending that precedent from a single documented lateral-movement chain into a sustained, at-scale intrusion operation with its own AI-assisted implant and rootkit engineering [6].

For structural threat modeling, CSA's *Agentic AI Threat Modeling Framework: MAESTRO* provides the seven-layer reference architecture best suited to mapping where in an attack chain — reconnaissance, exploit development, or post-compromise tooling, as with UAT-10147 — an adversary has substituted agentic AI for human effort, helping defenders reason about which of their own detection investments address which layer [7]. Organizations formalizing defensive investments in response to this trend should also consult CSA's *AI Controls Matrix (AICM) v1.1*, whose control domains covering vulnerability and threat management, logging and monitoring, and endpoint security provide an auditable basis for prioritizing the patch, driver-allowlisting, and kernel-integrity-monitoring recommendations above [8].

References

[1] Cisco Talos. "UAT-10147: Chinese-speaking adversary integrates agentic AI into post-compromise operations." Talos Intelligence Blog, August 2026.

[2] Cisco Talos. "UAT-10147 deploys SPECTRE: A cross-platform implant with Linux rootkit and BYOVD capabilities." Talos Intelligence Blog, August 2026.

[3] CyberInsider. "Chinese Hackers Use AI to Automate Attacks on 170,000 Servers." CyberInsider, August 2026.

[4] Cloud Security Alliance. "Autonomous AI Attack Pipelines Move Into the Field." CSA AI Safety Initiative, July 30, 2026.

[5] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." Anthropic, November 2025.

[6] Cloud Security Alliance. "Marimo RCE: LLM Agents as Post-Exploitation Tools." CSA AI Safety Initiative, June 6, 2026.

[7] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[8] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.