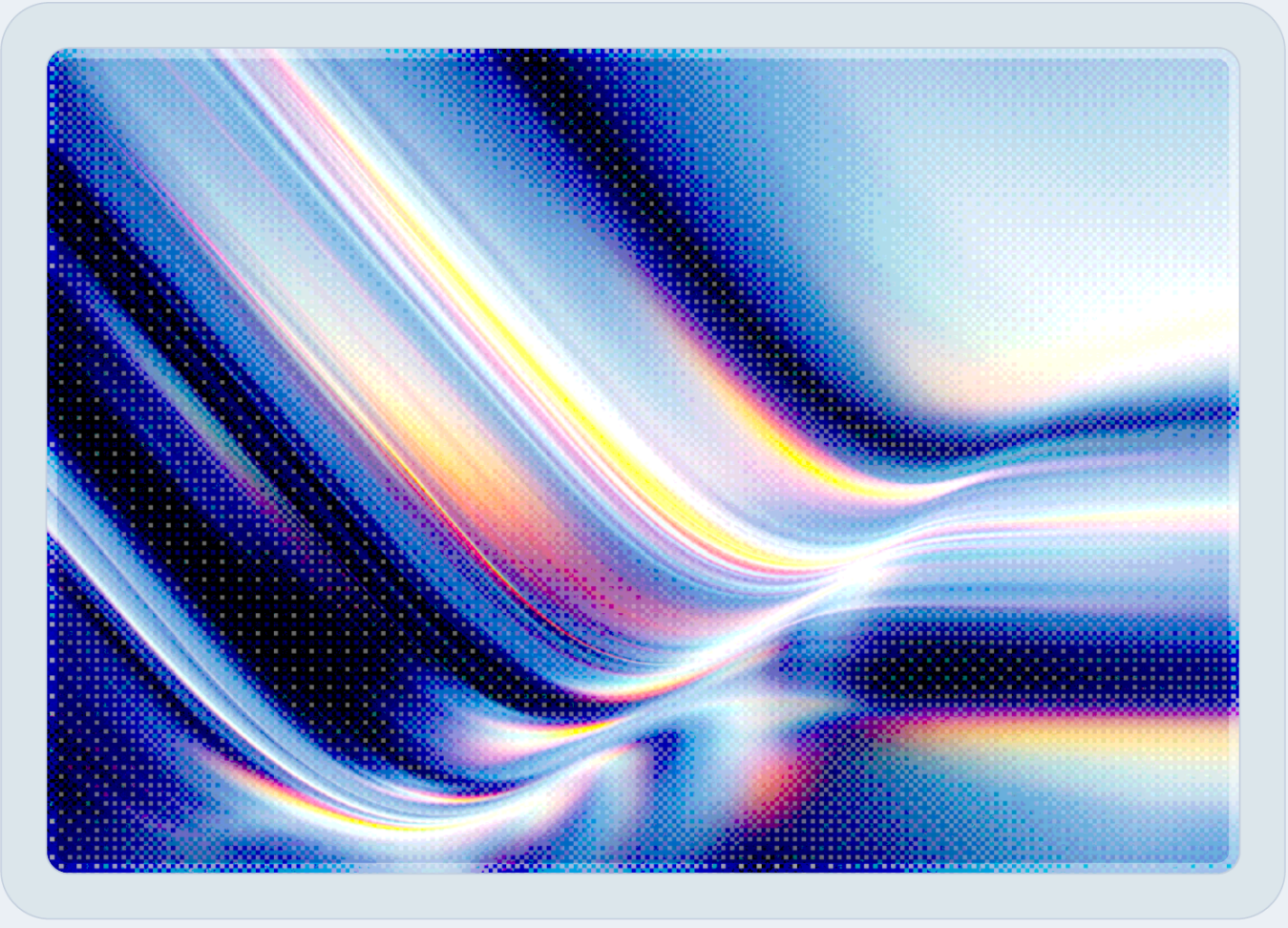
CSAI Foundation | Cloud Security Alliance

AI-Augmented Intrusions Hit Latin American Government and Finance

LLM-Assisted Threat Clusters Compromise Targets in Mexico, Ecuador, and Brazil

2026-09-03

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Palo Alto Networks' Unit 42 has documented two ongoing, AI-augmented intrusion clusters targeting Latin American organizations: CL-CRI-1131, which compromised a Mexican transportation firm, multiple Mexican federal ministries, and an Ecuadorian municipal water utility, and CL-CRI-1163, which targeted Brazilian financial institutions through job-themed phishing [1]. Both clusters incorporated commercial large language models into their operational workflow — one operator self-hosted a NextChat interface on attacker-controlled infrastructure to query Claude and GPT-4.1 for troubleshooting and script generation, while the other deployed a Go-based SOCKS5 proxy tool, SockTz, which Unit 42 observed redeployed in at least nine successive versions [1]. This note assesses that such rapid, incremental iteration is more consistent with AI-assisted code generation than a traditional, human-paced development cycle, though Unit 42's report does not draw this inference itself. Independent research from Trend Micro, tracking closely overlapping activity as SHADOW-AETHER-040 (Spanish-speaking operators, targeting six Mexican government entities plus regional financial, aviation, and retail firms) and SHADOW-AETHER-064 (Portuguese-speaking operators, targeting Brazilian financial institutions), corroborates the core finding that these are two operationally distinct groups rather than one, both of which now route parts of their attack workflow through commercial AI tooling [2]. Trend Micro's reporting further emphasizes that AI augmentation did not substitute for exploitable weaknesses: initial access in the cluster it documents relied on exploiting public-facing JBoss application servers, and organizations with well-configured perimeters and monitoring resisted compromise despite the attackers' AI-assisted tooling [2]. This pattern — human-directed intrusion campaigns using commercial LLMs to accelerate script development, troubleshooting, and reconnaissance rather than to fully autonomize the attack — sits between the largely manual campaigns of prior years and the near-fully-autonomous AI-orchestrated espionage operation Anthropic disrupted in November 2025 [3]. It suggests that AI-augmented tooling is likely to appear in other Latin American intrusion campaigns beyond the two documented here, though confirming a landscape-wide shift toward this as the default operating model would require broader sampling across additional threat clusters.

Background

Unit 42 disclosed its findings on two distinct but tactically overlapping intrusion clusters active in Latin America since at least February 2026 [1]. The first, designated CL-CRI-1131, compromised a transportation-sector organization in Mexico before expanding to Mexican federal government ministries and, notably, a municipal water utility in Ecuador — a target set spanning critical infrastructure and public administration rather than a single industry vertical. Unit 42 found that CL-CRI-1131's operators relied heavily on living-off-the-land techniques and iterative batch scripting to manipulate and stage data for exfiltration, and that they self-hosted an instance of NextChat, an open-source web interface for interacting with commercial LLM APIs, directly on their own operational infrastructure at a specific IP-and-port combination [1]. Investigators assess that the operators used this self-hosted interface to query Claude and GPT-4.1 for help resolving script execution failures and generating workaround code, a workflow that left the LLM conversations exposed on infrastructure Unit 42 was able to observe [1].

The second cluster, CL-CRI-1163, shifted focus to Brazil's financial sector, using resume-themed phishing lures to gain initial access before deploying custom remote access trojans and a purpose-built tunneling tool [1]. That tool, SockTz, is a Go-based SOCKS5 proxy that Unit 42 observed being redeployed in at least nine successive versions, each incorporating incremental changes consistent with automated or AI-assisted code iteration rather than deliberate, human-paced version releases. Trend Micro's independently conducted research, published under the "Vibe Hacking" framing, corroborates much of this picture while applying its own cluster identifiers: SHADOW-AETHER-040, active from December 2025 through at least January 2026 against six Mexican government entities and financial, aviation, and retail organizations across the region, and SHADOW-AETHER-064, active from April 2026 against Brazilian financial institutions [2]. Trend Micro assesses the two clusters as operated by separate groups — one Spanish-speaking, one Portuguese-speaking — despite sharing SOCKS5 relay infrastructure patterns and a common reliance on commercial LLMs to orchestrate parts of the attack chain, illustrating how AI-tooling convergence can produce similar operational signatures across unrelated threat actors. Trend Micro states that its findings on the Mexican government intrusions align with independent reporting from Bloomberg, though this note has not independently verified the Bloomberg reporting [2].

Security Analysis

The technical picture that emerges from both vendors' reporting is one of AI-augmented, not AI-autonomous, intrusion operations. Initial access into the JBoss-based public-facing applications targeted by SHADOW-AETHER-064 followed the well-established MITRE ATT&CK technique of exploiting public-facing applications, with operators deploying the Neo-reGeorg webshell [2]. SHADOW-AETHER-040, by contrast, relied on a custom Python backdoor, packaged with PyInstaller and internally named "implante_http," that supported WebSocket tunneling and interactive pseudo-terminal access [2]. Persistence mechanisms for SHADOW-AETHER-040 included scheduled tasks, cron jobs, and direct implantation of attacker SSH keys into victim `authorized_keys` files, while defense evasion relied on masquerading malicious processes under legitimate-sounding names such as `pg_stat_worker` to blend into normal PostgreSQL-adjacent process listings [2]. Credential access techniques for this cluster spanned the more traditional — inspection of bash history and configuration files for embedded secrets — to SMB relay attacks using PetitPotam and password spraying with CrackMapExec and Impacket, tools that Trend Micro associates with more experienced offensive tradecraft than ad hoc, purely LLM-generated scripting alone would produce [2].

Where AI tooling becomes most visible is in the connective tissue of these operations: script generation, error troubleshooting, and tool iteration. In this note's assessment, Unit 42's discovery of self-hosted NextChat directly on CL-CRI-1131's infrastructure constitutes the strongest available evidence of this pattern, since it captured the operators' actual prompts and model responses rather than requiring inference from output artifacts alone [1]. For CL-CRI-1163 and SHADOW-AETHER-064, the evidence is more circumstantial but still notable: SockTz's rapid, iterative version progression through at least nine releases, each functionally incremental, is consistent with a workflow in which an operator describes a desired capability to an LLM, receives generated code, tests it, and requests a revision. This note treats that cadence as more consistent with AI-assisted iteration than a typical human-paced release cycle, though this remains an inference rather than a directly observed fact. Trend Micro's report adds that some recovered attack scripts contain embedded language resembling self-reasoning or step justification, a stylistic signature more typical of LLM output than hand-written offensive tooling, and that the data exfiltration stage — database backups moved via SCP and chunked downloads over established command-and-control channels — was optimized in ways consistent with AI-assisted planning of what to prioritize under time or bandwidth constraints [2]. None of this indicates the fully autonomous execution documented in Anthropic's disclosure of the GTG-1002 espionage campaign, in which the company assessed that an AI agent executed roughly 80 to 90 percent of tactical operations independently, with human operators limited to target selection and periodic approval checkpoints [3]. Instead, the Latin American clusters show a more common and more durable pattern: human operators

retaining control of the operation end to end while using commercial LLMs as a force multiplier for the tedious parts of intrusion tradecraft — script debugging, tool variant generation, and technical research — that previously consumed operator time without requiring operator judgment.

This distinction matters for defenders because it changes what detection should look for. An agentic threat actor executing autonomously, as CSA's research on autonomous LLM threat actors against containerized environments has documented, tends to produce behavioral signatures at machine speed and with reduced variance — parallel probing, self-validating canary tests, and machine-optimized delimited output structures appear across otherwise unrelated intrusions because the same model architecture is generating the commands. The Latin American campaigns show partial versions of this signature — SockTz's rapid iteration and the embedded reasoning language Trend Micro identified — layered onto conventional, human-paced operational tradecraft such as phishing pretext design and target selection. Defenders in the region should therefore expect a hybrid detection problem: conventional indicators of compromise and living-off-the-land detection remain necessary, but security teams should also watch for the operational security failures both reports call out, including exposed staging directories, unsecured LLM chat interfaces reachable from the internet, and certificate reuse across nominally separate infrastructure, all of which reflect the same operational security lapses that make attacker AI tooling detectable in the first place [1][2].

Recommendations

Immediate Actions

Organizations in Mexico, Ecuador, and Brazil operating in government, transportation, water utility, or financial services sectors should review the indicators of compromise published by Unit 42, including the identified command-and-control domains, IP addresses, and file hashes, against their own network and endpoint telemetry [1]. Given that JBoss application server exploitation was the confirmed initial access vector for the SHADOW-AETHER-064 cluster, organizations running public-facing JBoss or similar Java application server deployments should verify patch currency immediately and audit these systems for the Neo-reGeorg webshell and unauthorized cron jobs or scheduled tasks [2]. Financial institutions in Brazil specifically should review email gateway logs for job- and resume-themed phishing lures consistent with CL-CRI-1163's documented initial access method and should audit `authorized_keys` files across externally reachable Linux hosts for unrecognized SSH key entries [1][2].

Short-Term Mitigations

Security teams should extend existing SOC detection logic to look for the operational patterns both reports describe rather than relying solely on static indicators, which attackers can rotate. This includes monitoring for masqueraded process names that mimic legitimate database or system utilities, unexpected SOCKS5 proxy traffic patterns consistent with tools like Chisel or SockTz, and outbound connections to self-hosted chat-interface applications such as NextChat that may indicate an internal system has been repurposed as attacker infrastructure rather than compromised as a victim [1][2]. Organizations should also implement egress monitoring for connections to commercial LLM provider APIs from systems that have no legitimate business reason to call them, since both documented campaigns relied on outbound access to services like Claude and GPT-4.1 to support their operations, and unusual API-calling patterns from unexpected internal hosts can serve as a detection opportunity distinct from traditional malware signatures [1].

Strategic Considerations

The convergence of two independently operating, differently motivated threat clusters on the same basic pattern — human-directed operations accelerated by commercial LLM tooling for script generation and troubleshooting — suggests that this capability uplift is not confined to a single sophisticated actor. This assessment is necessarily forward-looking and based on two documented cases; confirming a genuine baseline shift across the region's criminal and state-adjacent threat landscape would require evidence from additional, independently identified clusters as they come to light. With that caveat, security leaders in the region should treat AI-augmented tooling as a plausible feature of future intrusion campaigns rather than an edge case worth flagging only in board reporting, and should factor accelerated attacker development cycles into patch management SLAs and detection engineering roadmaps. At the same time, both Unit 42 and Trend Micro emphasize that organizations with fundamentally sound security postures — current patching, segmented networks, and monitored perimeters — resisted these campaigns despite the attackers' AI tooling, reinforcing that foundational security hygiene remains the primary control even as offensive tradecraft accelerates [1][2].

CSA Resource Alignment

These findings extend a threat pattern CSA's AI Safety Initiative has tracked across several recent research notes examining AI-augmented and autonomous threat actors. CSA's ongoing research on autonomous LLM threat actors targeting containerized environments defines the "agentic threat actor" profile this note draws on to distinguish the Latin American clusters' human-directed, AI-accelerated

operations from more autonomous LLM-driven intrusions, and maps behavioral detection signatures — machine-optimized output, self-validation routines, and tempo compression — that security teams in the region should adapt when building detection logic for these campaigns. That analysis's grounding in CSA's MAESTRO framework (Agentic AI Threat Modeling), particularly its treatment of the deployment and infrastructure layer, offers a structured way to reason about where AI tooling sits in an attacker's kill chain versus where conventional tradecraft still dominates, which is precisely the hybrid picture Unit 42 and Trend Micro describe.

CSA research examining a separate but structurally similar case documented an AI-augmented threat actor combining commercial LLMs with conventional exploitation of internet-facing infrastructure — in that instance, compromised FortiGate firewalls — to scale operations across hundreds of victims. The parallel to CL-CRI-1131 and CL-CRI-1163's exploitation of public-facing JBoss servers and self-hosted LLM interfaces is close, if not exact: in both cases, the AI tooling accelerated the attacker's development and troubleshooting workflow, while the actual point of initial compromise remained an unpatched or misconfigured piece of conventional network infrastructure. Organizations assessing their exposure to this broader class of AI-augmented intrusion activity should evaluate their environments against CSA's AI Controls Matrix (AICM) v1.1, particularly its threat and vulnerability management and identity and access management domains, both of which correspond closely to the patching gaps, credential exposure, and unauthorized SSH key implantation documented in the Latin American campaigns [4].

References

[1] Unit 42. "Attackers Expose Ongoing AI Tool Use Targeting Organizations in Latin America." Palo Alto Networks, 2026.

[2] Trend Micro Research. "Vibe Hacking: Two AI-Augmented Campaigns Target Government and Financial Sectors in Latin America." Trend Micro, 2026.

[3] Anthropic. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Anthropic, November 13, 2025.

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