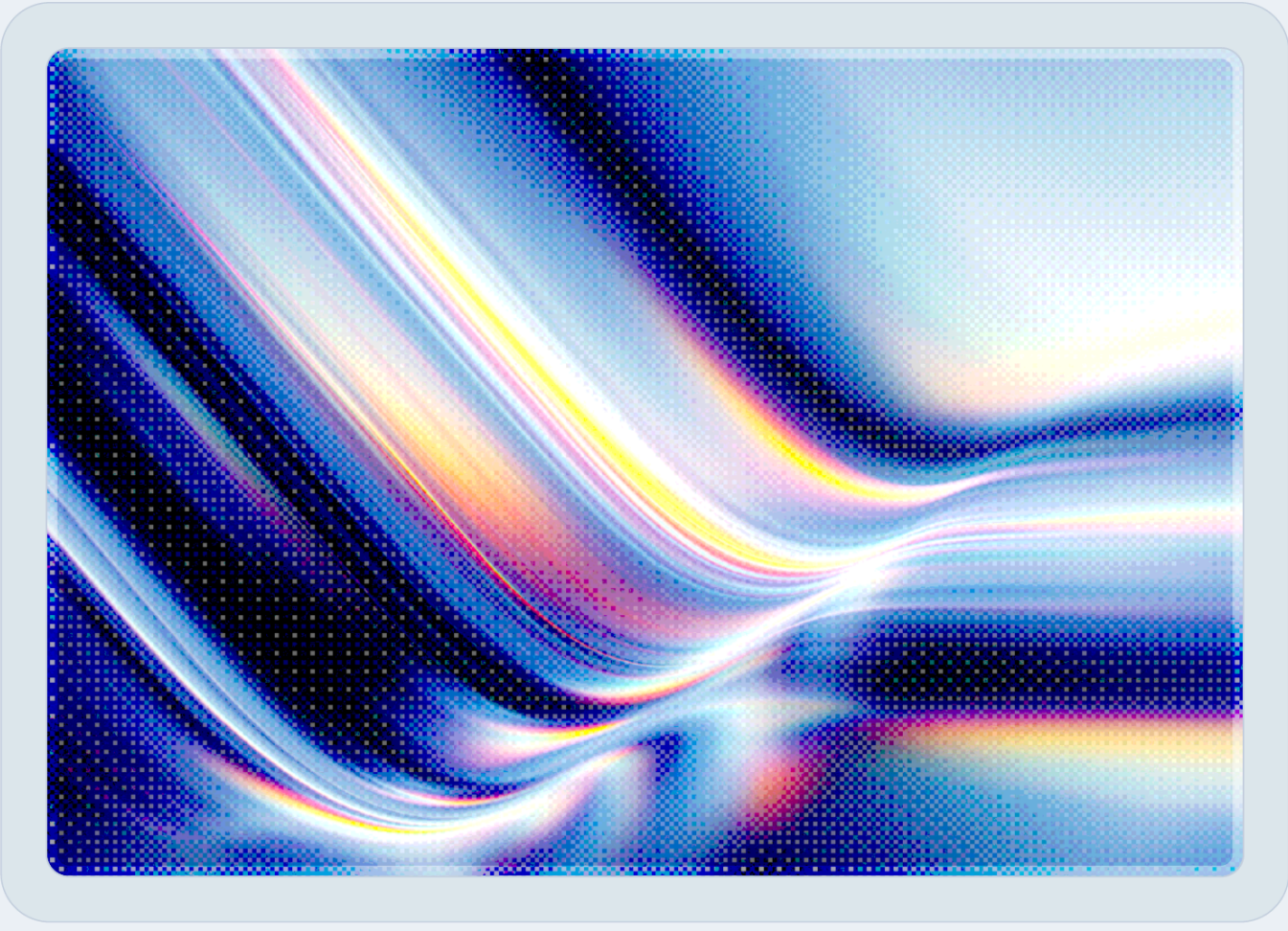
CSAI Foundation | Cloud Security Alliance

VOLTA: Infostealers That Use Commercial LLMs to Appraise Victims

Security Implications of AI-Driven Victim Prioritization in Commodity Malware

2026-11-04

AI-assisted Rapid Research



CSAI | cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

Cisco Talos published an analysis on October 31, 2026 of VOLTA, also labeled VOLTA-AUTODOXER. VOLTA is a self-deleting Windows infostealer that, after harvesting credentials and personal data, sends a structured "evidence bundle" about the victim to a commercial large language model. The model returns a ranked appraisal of what the victim is worth and how weak their security habits are [1]. CSA is not aware of an earlier publicly documented commodity stealer that uses a frontier model as a built-in triage analyst, although none of the cited sources makes a priority claim. The AI layer does not deliver, persist, or evade. Its only job is to replace the human analysis that operators usually perform on stolen logs [1][2].

The AI stage works by calling the OpenRouter chat-completions API. VOLTA requests Anthropic's Claude Opus 4.6 by default and falls back to Moonshot AI's Kimi K2 and then Google's Gemini 2.5 Flash when it detects refusal phrases such as "I can't" or "I'm not able" in a response [1]. Because the LLM request goes out from the victim's own machine before the final exfiltration, unexpected outbound traffic to an LLM gateway from a non-developer endpoint is a practical detection signal [1]. The sample also appears to come from a builder kit with six configuration slots filled at build time, including the Telegram exfiltration channel and API keys. This suggests VOLTA is built for distribution to multiple operators, which would put the AI dossier capability within reach of buyers who lack analytical skills [1][2].

Two qualifications shape how defenders should respond. First, the LLM stage is optional. Without an API key, VOLTA still performs its full collection and exfiltration routine as a conventional stealer, so existing infostealer controls remain the first line of defense [1]. Second, Talos documented a single sample, first seen October 11, 2026, and the delivery vector has not been publicly reported [1]. In CSA's assessment, no wider campaign has yet been independently confirmed in public reporting, and VOLTA is best treated as an instructive design pattern rather than evidence of a large active operation.

Background

Infostealers have long been the base layer of the cybercrime economy. They harvest browser credentials, session cookies, cryptocurrency wallets, and authentication tokens, and they hand those "logs" to operators or resell them in bulk. Prior CSA research has covered this ecosystem through 2026, including ACR Stealer's ClickFix campaigns, which drained Microsoft 365 session tokens [6], and the IronWorm and Miasma worms, which specifically target AI provider API keys and developer publishing tokens [8]. A

persistent constraint in this economy has been analyst labor. A single infection can yield thousands of files, and deciding which victims justify follow-on fraud, account takeover, or extortion takes sustained human attention. Talos frames VOLTA's design as "effort displacement," because it applies AI to exactly that bottleneck [1].

AI-assisted victim appraisal has precedent. In August 2025, Anthropic reported disrupting a data-extortion operation in which an actor used Claude Code against at least 17 organizations. In that case, "Claude analyzed the exfiltrated financial data to determine appropriate ransom amounts," and the ransoms sometimes exceeded $500,000 [5]. In November 2025, Google Threat Intelligence Group (GTIG) described malware families that query LLMs during execution. These included PROMPTSTEAL, which GTIG attributed to APT28 and which used an open model via Hugging Face to generate collection commands, and the experimental PROMPTFLUX, which asked Gemini to rewrite its own code [4]. VOLTA differs from both. Unlike the Anthropic case, the appraisal is not an interactive actor session; it is automated inside a reusable commodity binary. Unlike PROMPTSTEAL and PROMPTFLUX, the model is not used for execution or evasion. It is used for post-theft analytical judgment on behalf of a paying customer [1][4][5].

VOLTA also arrives in a broader setting of LLMjacking, where stolen or misused AI compute becomes part of the attacker's infrastructure. CSA's research on LLMjacking documented how compromised inference capacity has moved from simple cost theft toward use as the reasoning engine of offensive pipelines [7]. In the same week as the VOLTA report, Talos also described NIK C2, a remote access tool that can use a victim's own `OPENAI_API_KEY` environment variable to fund the attacker's AI analysis [2][11]. Taken together with the earlier Anthropic and GTIG findings, these developments suggest that commercial LLM access is emerging as a swappable component of criminal tooling [1][2][4][5]. Purpose-built malicious models are not required.

Security Analysis

How VOLTA Builds and Appraises a Victim Profile

According to Talos, VOLTA's collection stage is extensive but conventional. It takes passwords, cookies, and saved payment cards from Chromium-based browsers, and its decryption path depends on launching Chrome with `--headless` to get around App-Bound Encryption. It also collects Firefox and Thunderbird data, Outlook profiles, Slack workspaces, Steam and Battle.net accounts, Discord, Telegram, and Roblox sessions, Microsoft OAuth tokens, SSH keys, gcloud, OpenVPN, and Hide.me configurations, every saved Wi-Fi password in plaintext, screenshots, and wallet seed phrases. An embedded cracker runs against wallet vaults inline [1]. What sets VOLTA apart is its enrichment stage. Using stolen sessions, it

searches Gmail for receipts from Amazon, DoorDash, Uber Eats, Grubhub, Walmart, and PayPal to corroborate a home address. It also pulls food-delivery order history, reuses the Discord token, and values Roblox inventory through a live lookup against the third-party Rolimons service [1].

The resulting evidence bundle is about a person rather than a machine. It includes geolocation, home address, employer and employer domain, inferred birth year, the last four digits of payment cards, Slack workspaces, document authorship metadata, and indicators of iPhone backups [1][2]. VOLTA sends this bundle, in hand-assembled JSON with a generic Mozilla/5.0 User-Agent, to OpenRouter at a temperature of 0.2. Talos interprets that setting as favoring accuracy and consistency over creativity [1]. The roughly 240-word system prompt casts the model as "a forensic intelligence analyst drafting a target assessment." It instructs the model to "go deep, not wide," to connect sparse details such as names, addresses, or universities to password patterns, and to produce two required sections [1][2]. The first, "WHAT THEY'RE WORTH," ranks valuables into low, medium, and high monetization tiers. The second, "THEIR OPSEC," assesses the victim's security posture and password hygiene. Talos describes the output as a "prioritized and citation-backed dossier" [1][2].

The table below summarizes the staged pipeline Talos reconstructed from the sample.

Stage	Activity	Output / Artifact
1. Harvest	Credential, session, wallet, Wi-Fi, and screenshot collection	`volta_log.zip` exfiltrated; harvest log at `%APPDATA%\volta_debug.txt`
2. Enrichment	Gmail receipt search, delivery-order history, Discord and Roblox valuation	Structured victim evidence bundle
3. AI appraisal	OpenRouter call with refusal-triggered model fallback	`Identity_Report.txt`, `AI_Analysis.txt` in `volta_dox.zip`
4. Network enrichment	Additional results if enrichment succeeds	`volta_doxnet.zip`
5. Cleanup	Self-deletion of the executable	Debug log may survive for responders

Source: Cisco Talos [1]. Exfiltration is reported to occur via a Telegram channel configured at build time [1][2].

Refusal Resilience as a Design Requirement

For the AI safety community, VOLTA's most notable design choice is its refusal handling. The malware checks each model response for refusal phrases. When it finds one, it resubmits the same request to the next model in its chain, up to three models in total [1]. This turns provider-side safety training from a barrier into a routing decision. A refusal from one model simply sends the request to another model with a different policy posture. Because OpenRouter aggregates many providers behind a single API key and a single endpoint, the attacker pays almost nothing in engineering effort to switch models. Talos suggests that refusal resilience may become a standard design requirement for future AI-integrated malware [1].

Public reporting does not establish how often the default model actually refused, or which model produced the dossiers in any live infections. Readers should not take the fallback chain as evidence that any particular model complied. The architecture implies that safeguards evaluated one model at a time are weaker than their individual refusal rates suggest. An adversary needs only one model in the chain to comply, so the effective safeguard is roughly that of the most permissive model available through the gateway. The same design has limits, however. VOLTA's phrase matching on strings such as "I can't" and "I'm not able" will miss refusals worded differently, as well as partial compliance. The same aggregation that makes model-switching cheap also concentrates risk for the attacker: a single gateway key funds the entire chain, so key revocation by the aggregator disables all fallback models at once. For model providers and gateway operators, this points toward detection that focuses on request patterns, such as structured personal-data bundles paired with target-assessment framing, rather than on individual refusals alone.

Commoditization and the Lowered Skill Floor

Talos assesses VOLTA as a builder-based product. Six swappable template slots, covering the Telegram channel, API keys, and other settings, are filled at build time, which suggests the author compiles a custom binary per customer [1]. The binary's credential protection offers little real protection: the encryption key is derived via SHA-256 from two constants stored in the same file [1]. Talos also observed that all of the sample's endpoints are third-party services, with no actor-owned infrastructure [1]. This matters for two reasons. First, the AI analysis is delivered as a feature, so a buyer with little analytical skill receives an automated, prioritized assessment of each victim in their Telegram channel, although the quality of those assessments in live infections has not been publicly evaluated [1][2]. Second, depending entirely on legitimate services (OpenRouter, Telegram, Gmail, and Rolimons) makes infrastructure-based blocking harder and pushes defenders toward behavioral and contextual detection.

The practical risk is not that VOLTA steals more than earlier stealers. Talos characterizes VOLTA as a mature stealer [1], and in CSA's assessment its collection set, centered on browser credentials and session tokens, is broadly comparable to commodity families such as the ACR Stealer operations CSA has previously analyzed [6]. The likely risk is that it shortens the time between theft and targeted exploitation. A ranked dossier that links a person's employer, home address, financial footprint, and password habits makes it easier to pick high-value victims for business email compromise, account takeover, cryptocurrency theft, sextortion, or physical-world harassment. For enterprises, the "employer and employer domain" field, along with Slack workspaces and Microsoft OAuth tokens, means a personal-device infection could lead to corporate compromise, as session-token theft has done in earlier campaigns [6]. These consequences are inferences from the documented capability. Talos has not reported observed downstream abuse.

Detection Opportunities and Evidentiary Limits

Because the LLM call goes out from the infected host, VOLTA creates network telemetry that conventional stealers do not. Talos highlights several detection opportunities: outbound requests to OpenRouter from hosts where that is unexpected, lookups to Rolimons, Chrome launched with `--headless` by unexpected parent processes, and the persistent `%APPDATA%\volta_debug.txt` artifact. That log records what was collected and whether the LLM call succeeded, even after the executable deletes itself [1]. Talos published a SHA-256 hash (`57ef4242666d78c50df145b34023df564f0fad38812d6a580e66433d427e2279`) and two filenames, but no actor-owned network indicators [1].

The evidence base is still limited. Talos analyzed one sample, surfaced through its CAIRN research effort. Its first-seen timestamp is October 11, 2026 and its PE compile timestamp is August 17, 2026 [1]. The initial infection vector is not described in the Talos report or in secondary coverage [1][2][3], and secondary coverage to date has restated the Talos findings without adding independent observations [2] [3]. Until wider deployment is confirmed, VOLTA should be described only by what the primary source shows. Organizations should therefore treat VOLTA as a credible and instructive design pattern rather than evidence of a large active campaign.

Recommendations

Immediate Actions

Security teams should first hunt for the indicators Talos published. That means searching endpoint telemetry for the published hash and filenames, for `volta_debug.txt`, `Identity_Report.txt`, and `AI_Analysis.txt` under user profile directories, and for `volta_*.zip` archives [1]. In parallel, teams should query proxy, DNS, and EDR network logs for connections to `openrouter.ai` and other LLM gateway endpoints from endpoints whose owners have no business reason to call them. Because browser extensions, IDE plugins, and desktop assistants increasingly call such gateways from ordinary devices, teams should baseline sanctioned AI tool usage first; process-level attribution, meaning which binary made the call, is more reliable than destination alone. Unexplained traffic of this kind from finance, HR, or general knowledge-worker devices deserves investigation whatever malware family is involved. Teams should also alert on Chrome or Edge processes launched with `--headless` by parents that are not approved developer or automation tools, because that path supports VOLTA's App-Bound Encryption bypass [1].

If an infection is confirmed, responders should preserve `%APPDATA%\volta_debug.txt` before cleanup and use it to scope what was taken and whether a dossier was generated [1]. Remediation should treat every harvested session as compromised. Responders should revoke Microsoft OAuth, Slack, Discord, and other session tokens rather than relying on password resets alone, rotate SSH keys and cloud CLI credentials, and treat any wallet whose seed phrase was stored in plaintext as fully compromised. Talos recommends moving funds to a new wallet with a seed generated on an air-gapped device [1].

Short-Term Mitigations

Over the coming weeks, organizations should make LLM API egress an explicitly governed category rather than general web traffic. A practical approach is to allow connections to AI providers and aggregators only through a sanctioned enterprise gateway, and to log or block direct calls from unmanaged processes. This also helps against the LLMjacking patterns CSA has documented [7]. Application control policies should restrict which binaries can launch browsers headlessly and should block unsigned executables from user-writable paths.

Credential hygiene deserves renewed attention because VOLTA's value comes from aggregation. Talos recommends hardware-backed credential storage [1]. Organizations should also move high-value accounts to phishing-resistant FIDO2 authentication, shorten session-token lifetimes for SaaS applications that support it, and enable token binding or continuous access evaluation where the platform

supports them. Bring-your-own-device and contractor policies should account for the fact that a personal-device infection can expose corporate Slack workspaces and OAuth tokens. Security awareness programs should also mention that stolen data may now be analyzed automatically to choose targets, which makes follow-on pretexting more convincing and more personal.

Strategic Considerations

For AI model providers and API aggregators, VOLTA shows that refusal-based safeguards can be routed around when an attacker has cheap access to multiple models. Gateway operators are well placed to detect cross-model retry patterns, structured personal-data payloads with target-assessment framing, and API keys used from many residential endpoints, and their ability to revoke a single key that funds an entire fallback chain makes them an effective point of intervention. They should consider shared abuse signals with upstream providers, because no single model provider sees the whole fallback chain. Enterprises that build on aggregators should check that their providers maintain abuse-response processes and key-revocation capabilities, and should include these in AI supplier assessments.

For security programs more broadly, the lesson is that AI is reducing the labor cost of the analytical stages of an attack, not only the technical ones. Threat models that assume a long delay between credential theft and targeted exploitation, because human operators must first sort through logs, should be revisited. Detection engineering should also expect more malware that calls legitimate AI services at runtime, which suggests context-aware egress monitoring will remain relevant beyond any single family [1] [4].

CSA Resource Alignment

CSA's research note **"ACR Stealer's ClickFix Campaigns Drain M365 Tokens"** is the closest prior analysis of the commodity stealer layer VOLTA builds on. That note explains why browser session-token theft, rather than password theft, is the enterprise-critical outcome, and why application control and session hardening matter more than patching against user-executed delivery chains [6]. VOLTA's collection of Chromium sessions and Microsoft OAuth tokens means those recommendations apply directly. The AI appraisal stage adds a prioritization layer on top of the same underlying risk.

CSA's **"LLMjacking Evolved: Stolen AI Compute as Offensive Infrastructure"** documents the trend VOLTA extends: commercial or compromised AI inference becoming a component of offensive tooling [7]. VOLTA differs in that, as reported, it uses operator-supplied API keys rather than stolen compute. It nonetheless reinforces that note's recommendation to treat AI gateways and model-serving infrastructure as high-value assets subject to zero-trust controls and abuse monitoring. CSA's **"Miasma**

and IronWorm: Self-Replicating Worms Targeting AI Credentials" complements this by showing that AI credentials are themselves now primary theft targets [8]. Organizations should therefore assume that LLM API keys harvested by one stealer family may end up funding the analysis stage of another.

From a control perspective, the **AI Controls Matrix (AICM) v1.1** provides the framework for both sides of this problem [9]. Its Identity and Access Management and Threat and Vulnerability Management domains cover session-token governance, credential storage, and detection of anomalous use. Its model-provider and supply-chain controls give AI service providers and aggregators a basis for abuse monitoring, key revocation, and safeguards evaluated across a multi-model gateway rather than one model at a time. For teams modeling AI-integrated malware, CSA's **MAESTRO** agentic threat-modeling framework offers a layered way to analyze where a runtime model call sits in an adversary's pipeline and which ecosystem participants can observe or interrupt it [10].

References

[1] Fetterman, Ryan / Cisco Talos. "Appraisal by AI: Model-driven victim prioritization in the VOLTA infostealer." Cisco Talos Intelligence Blog, October 31, 2026.

[2] Lakshmanan, Ravie. "New VOLTA Infostealer Uses Claude, Gemini, and Kimi AI to Rank Victims by Financial Value." The Hacker News, November 2, 2026.

[3] Paganini, Pierluigi. "VOLTA, the infostealer that asks an AI which victims are worth the effort." Security Affairs, November 2, 2026.

[4] Google Threat Intelligence Group. "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools." Google Cloud Blog, November 5, 2025.

[5] Anthropic. "Detecting and countering misuse of AI: August 2025." Anthropic, August 2025.

[6] Cloud Security Alliance. "ACR Stealer's ClickFix Campaigns Drain M365 Tokens." CSA AI Safety Initiative, July 17, 2026.

[7] Cloud Security Alliance. "LLMjacking Evolved: Stolen AI Compute as Offensive Infrastructure." CSA AI Safety Initiative, June 20, 2026.

[8] Cloud Security Alliance. "Miasma and IronWorm: Self-Replicating Worms Targeting AI Credentials." CSA AI Safety Initiative, June 9, 2026.

[9] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, June 22, 2026.

[10] Cloud Security Alliance. "MAESTRO: Agentic AI Threat Modeling Framework." CSA AI Safety Initiative, n.d. (accessed November 4, 2026).

[11] Cisco Talos. "NIK C2: The new AI-enabled command-and-control framework." Cisco Talos Intelligence Blog, October 29, 2026.