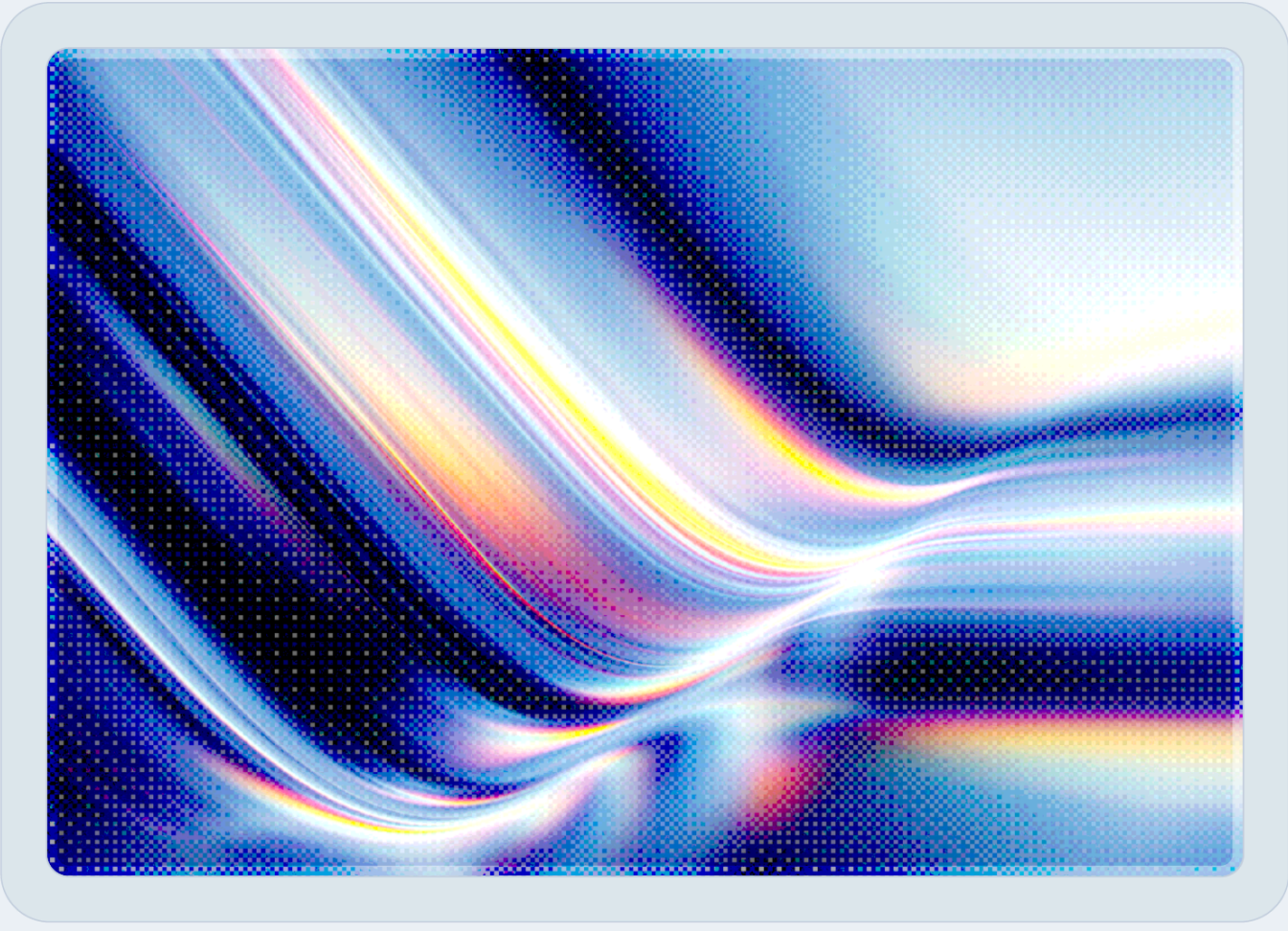
CSAI Foundation | Cloud Security Alliance

Stolen AI Session Tokens Are Bypassing MFA

2026-09-11

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Revision Note

Version 1.1 (September 11, 2026) corrects a scale-inflation problem in the Key Takeaways and Background sections, where Google's and Microsoft's ecosystem-wide SSO token counts had been presented as "AI-specific exposure" without the source's own caveat that these totals span each vendor's full account base. It also corrects a citation mismatch in which a reference about Chrome's device-bound session credentials was misattached to claims about Telegram token-resale marketplaces, replaces a citation that did not substantiate the specific Lumma/Vidar capabilities it was attached to, and softens several instances of unqualified or comparative language flagged in quality review. Two references to CSA research that could not be confirmed as publicly live have been removed; the underlying analytical insight is retained in prose without a numbered citation, consistent with the project's rules on internal and unpublished sources. References have been renumbered accordingly.

Key Takeaways

An infostealer log analysis published by Okta's threat intelligence team on September 9, 2026, found hundreds of live authentication tokens tied specifically to AI services — including Anthropic's Claude, Character.ai, Cursor, and Poe.com — alongside thousands more general-purpose SSO tokens for Google and Microsoft accounts that can also reach AI products like Gemini and Copilot, in a single leaked malware data dump [1][2]. Because these tokens represent an already-completed login, an attacker who replays one gains direct account access without ever supplying a password or satisfying a multi-factor authentication (MFA) challenge, rendering MFA ineffective against this class of intrusion [1]. Of the 44,791 unique JSON Web Tokens (JWTs) Okta identified in the dump, 555 were tied specifically to AI service authentication, and 1,843 JWTs and JSON Web Encryption (JWE) objects across all services remained unexpired on the day the data was released [1][2]. Underground marketplaces have already begun productizing the problem, with vendors on Telegram advertising bundled, discounted access to Claude, ChatGPT, Gemini, and Cursor accounts complete with 24/7 support and money-back guarantees [3]. For enterprises, the exposure extends beyond a single hijacked chat session: stolen tokens can expose conversation histories, uploaded files, connected tools, and — where API keys are involved — direct, billable access to model inference [1][3].

Background

Infostealer malware such as Lumma Stealer and Vidar has, for several years, focused on harvesting saved browser passwords, autofill data, and cryptocurrency wallet contents from infected endpoints [5]. Increasingly, however, these tools are purpose-built to extract live authentication artifacts: browser session cookies, JWTs, JWEs, and API keys sitting in local storage or memory at the moment of infection [1]. Okta's research team, led by threat intelligence director Jeremy Kirk, analyzed a 7 GB infostealer log dump released for free on a Telegram channel on August 2, 2026, containing data harvested from 5,871 infected machines across 162 countries [1][2]. Using pattern matching and the open-source secret-scanning tool TruffleHog, the researchers catalogued the credentials, sessions, and keys the dump exposed for major cloud and AI platforms [1].

Anthropic-specific session tokens numbered 561 in the dump, of which 164 remained unexpired — the clearest AI-specific figure in the report. Google and Microsoft, by contrast, contributed far larger totals (9,829 and 2,491 tokens respectively) through their all-purpose SSO gateways; Okta's own report notes these totals span each vendor's full account ecosystem and are not exclusively tied to AI product access [1][2]. The dataset also included live authentication data for Amazon, Notion, Character.ai, Cursor, Poe.com, Pika AI, and the design tool Gamma, reflecting how thoroughly AI and AI-adjacent services have been woven into everyday knowledge work — and how thoroughly infostealers now target them [1] [3]. Beyond session tokens, Okta recovered 24 still-valid API keys spanning Google Gemini, OpenAI, Groq, and OpenRouter, and found that 17.7% of the JWTs in the dump contained plaintext personally identifiable information such as names, phone numbers, and email addresses [1][2]. This is not a hypothetical risk: Okta's report references prior incidents in which stolen AI credentials and API keys were used to run up roughly $1 million, $25,000, and $600,000 in unauthorized usage charges against victim accounts [1].

This pattern builds on a broader trend CSA's AI Safety Initiative has tracked throughout 2026, in which commodity infostealer campaigns increasingly use AI-branded lures — fake Claude installers, spoofed verification pages, and ClickFix-style social-engineering prompts — specifically to harvest the browser sessions tied to AI accounts. What distinguishes the Okta findings from earlier single-campaign disclosures is scale: rather than remaining incidental byproducts of generic credential theft, the tokens suggest a shift toward being catalogued, filtered by service, and resold as a distinct product line, with buyers using anti-detect browsers such as Camoufox and automation frameworks like SeleniumBase to replay tokens while mimicking the original victim's device fingerprint [1][3].

Security Analysis

The technical reason MFA fails against this attack class is straightforward: MFA protects the login event, not the session that follows it. Once a user authenticates — whether with a password alone or password plus a second factor — the service issues a token (a cookie, JWT, or JWE) that represents "this session is authenticated" for some period of time, often extending to days or weeks for convenience [1]. An infostealer that exfiltrates that token hands the attacker a valid, already-authenticated session; replaying it does not require the attacker to know the password, possess the second factor, or trigger any authentication challenge at all, because from the server's perspective the session is legitimate [1][2]. This shares the same underlying weakness CSA's AI Safety Initiative has documented in earlier 2026 disclosures involving both a ClickFix-style credential-theft campaign against Microsoft 365 accounts and a token-forwarding flaw in an AI agent preview feature: in each case, a stolen or forwarded live session bypassed password resets and fresh MFA challenges alike.

Several factors compound the risk specifically for AI services. First, many AI platforms issue long-lived tokens by default to reduce login friction for a product category built around frequent, informal use, which widens the window during which a stolen token remains replayable [1]. Second, AI accounts frequently hold more than conversational access: connected API keys, workspace integrations, uploaded documents, and agent configurations mean a hijacked session can expose material well beyond a single chat transcript, including source code, legal drafts, or health information a user has pasted into a prompt [3]. Third, the emergence of a dedicated resale market — Okta flagged a Telegram vendor branded "Poison Claude" explicitly marketing access to Anthropic's Opus and Sonnet models — signals that attackers now see standing demand for AI account access specifically, rather than treating it as an incidental find among broader stolen-credential bundles [1][3]. Finally, because a meaningful share of harvested JWTs carry plaintext PII, a single leaked dump can simultaneously enable account takeover and downstream identity fraud, compounding the blast radius of any one infection [1][2].

It is worth being precise about what this data does and does not show. Okta's analysis describes tokens present in a criminal log dump and their theoretical replayability; it does not claim to have observed the volume of tokens actually exploited for account takeover, nor does it attribute a specific breach to this dataset. The financial-loss figures cited come from separate, previously reported incidents rather than from this dump itself. The risk is nonetheless concrete: unexpired, replayable tokens for major AI providers were sitting in a freely distributed criminal dataset, and the existence of an organized resale market suggests demand that will likely outlast this particular leak [1][3].

Recommendations

Immediate Actions

Security teams should treat this disclosure as a trigger for proactive session hygiene rather than a one-time cleanup. Organizations should force a global session and API key rotation for AI platform accounts used by employees, particularly for any account with browser extensions, third-party integrations, or shared/managed devices, where general infostealer research indicates infection risk is elevated. Security teams should also review AI platform sign-in and usage logs for anomalous access patterns — logins or API calls from unfamiliar geographies, device fingerprints, or times inconsistent with the account owner's normal behavior — since these are frequently the only visible trace of a replayed token [1]. Any API keys embedded in code repositories, configuration files, or CI/CD pipelines should be audited and scoped down or revoked, given that 24 live keys for mainstream AI services were recovered from a single leaked dump [1][2].

Short-Term Mitigations

Beyond initial cleanup, organizations should shift toward shorter-lived tokens and continuous verification wherever the AI platform or identity provider supports it. Adopting OAuth 2.0 flows with short-lived access tokens and refresh-token rotation reduces the window in which a stolen token remains useful, and enforcing IP allowlisting or conditional access policies can block replay attempts originating outside expected network ranges even when a token is valid [1]. Endpoint controls are a necessary complement to identity controls here: since the tokens originate from infostealer infections rather than a flaw in the AI platforms themselves, reducing malware dwell time through EDR tooling, restricting execution of unsigned scripts, and training users to recognize AI-branded phishing lures remain first-order defenses [1]. Where available, enabling device-bound session credentials — a mechanism Chrome and Google Workspace have rolled out broadly in 2026 to cryptographically tie a session cookie to the originating device's hardware-backed key store — can make an exfiltrated cookie useless outside the device it was stolen from [4][6].

Strategic Considerations

Longer term, enterprises should treat AI account access as a first-class identity governance problem rather than a shadow-IT afterthought. That means bringing AI platform accounts under the same conditional-access, session-lifecycle, and behavioral-monitoring policies applied to core SaaS and cloud infrastructure, rather than allowing employees to authenticate to AI tools outside centralized identity

management [7]. It also means favoring phishing-resistant authentication such as passkeys over password-plus-MFA wherever an AI vendor supports it, since passkeys remove the credential-theft step almost entirely, even though they do not, on their own, protect an already-issued session token [1]. Finally, organizations should build AI-specific incident response playbooks that assume session and API key compromise as a distinct scenario from password compromise, with clear procedures for token revocation, usage-anomaly investigation, and vendor notification, since password resets alone do nothing to invalidate a live, stolen session.

CSA Resource Alignment

This incident sits squarely within a pattern CSA's AI Safety Initiative has been tracking across several 2026 disclosures: infostealer-driven theft of live session tokens that renders password- and MFA-based defenses ineffective. That broader body of tracking — spanning both credential-stealing campaigns against mainstream enterprise SaaS accounts and session-isolation flaws in AI agent tooling — has consistently pointed to the same corrective: token revocation, not password reset, is the action that actually closes off a stolen live session, since a password change does nothing to invalidate a token the attacker already holds.

On the framework side, CSA's Zero Trust Principles and Guidance for IAM provides the architectural foundation for the continuous-verification approach this incident demands: its emphasis on risk-based, contextual authorization rather than one-time authentication decisions directly addresses why a valid-but-stolen session token should not be treated as equivalent to a trusted user [7]. Finally, the AI Controls Matrix (AICM) v1.1's Identity and Access Management domain offers the control baseline enterprises should use to formalize session-lifecycle management, credential scoping, and token revocation practices for AI accounts specifically, giving security and compliance teams a concrete framework against which to audit their AI platform governance [8].

References

[1] Okta Threat Intelligence. "Signing In Without Actually Signing In." Okta, September 9, 2026.

[2] The Hacker News. "Infostealer Logs Expose Replayable AI Tokens That Can Bypass MFA." The Hacker News, September 9, 2026.

[3] Gizmodo. "There's a New Black Market Just for Stolen ChatGPT and Claude Logins. It's Open 24/7.." Gizmodo, September 2026.

[4] Cybersecurity News. "Google Chrome's Device-Bound Session Credentials Now GA to Block Account Takeovers." Cybersecurity News, 2026.

[5] Microsoft Threat Intelligence. "Lumma Stealer: Breaking Down the Delivery Techniques and Capabilities of a Prolific Infostealer." Microsoft Security Blog, May 21, 2025.

[6] Google Security Blog. "Protecting Cookies with Device Bound Session Credentials." Google, 2026.

[7] Cloud Security Alliance. "Zero Trust Principles and Guidance for IAM." CSA Zero Trust Working Group, 2023.

[8] Cloud Security Alliance. "AI Controls Matrix v1.1." CSA, June 23, 2026.

