CSAI

cloud security alliance®

CSAI Foundation

Cloud Security Alliance AI Safety Initiative

AI Browser Extensions: The DLP-Invisible Enterprise Attack Surface

How Permissive Browser Add-Ons Bypass Traditional Data Loss Prevention Controls

Unofficial AI-assisted Research

2026-04-11

© 2026 Cloud Security Alliance. Unofficial AI-assisted Research.

© 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

- Two malicious Chrome extensions impersonating the legitimate AITOPIA productivity tool accumulated over 900,000 combined installations and exfiltrated complete ChatGPT and DeepSeek conversation histories — including proprietary code, internal workflows, and confidential business data — from users across more than 20,000 enterprise tenants before their discovery in January 2026 [1][2][5].
- Traditional data loss prevention (DLP) tools and security service edge (SSE) platforms operate at an architectural layer below browser activity, leaving copy-paste interactions, AI prompt submissions, and extension-level data interception outside their inspection scope; EDR platforms without browser-specific instrumentation face similar gaps [9][10].
- According to LayerX's Enterprise Browser Extension Security Report 2025, 99% of enterprise users have at least one browser extension installed, more than half of those extensions carry high or critical permission scopes, and over 20% of users have installed at least one GenAI-specific extension — of which 58% carry high-risk permissions and more than 5% have been independently classified as malicious [6][7].
- The browser has become a significant data exfiltration vector for enterprise AI data: 77% of employees paste data into generative AI tools, 82% of those interactions occur through personal accounts outside corporate monitoring, and GenAI now accounts for approximately one-third of all corporate-to-personal data movement [9].
- OpenAI has publicly acknowledged that prompt injection attacks against browser-based AI agents may never be fully eliminated, while Cato Networks' CTRL research team documented HashJack, a novel class of indirect prompt injection that conceals malicious instructions within URLs processed by AI browser assistants [11][12], compounding the risk landscape beyond passive data harvesting into active agent manipulation.

Background

The enterprise browser has undergone a fundamental transformation over the past two years. What was once a document retrieval and communication tool is now a primary platform for AI-augmented work: employees draft code, analyze data, write reports, and query internal knowledge bases through browser-

resident AI interfaces. This shift has happened faster than security architectures have adapted. The assumptions underlying traditional data loss prevention — that sensitive data moves through inspectable channels such as email attachments, file transfers, or API calls — no longer hold when an employee pastes a confidential contract into a ChatGPT prompt tab, or when a browser extension silently intercepts the conversation that follows.

Browser extensions occupy a uniquely privileged position in this environment. Installed directly into the browser runtime, they operate inside the encrypted session context where work actually happens, giving them access to page content, keystrokes, form fields, cookies, and network requests that network-layer security tools cannot observe without breaking TLS. For decades, this architecture worked reasonably well when extensions were primarily productivity tools with narrow purposes — password managers, grammar checkers, tab organizers — that posed modest risk even when poorly written. The emergence of AI assistant extensions has fundamentally altered that risk calculus. Where a password manager or grammar checker could, at worst, expose data within its specific functional scope, today's AI-oriented extensions routinely request permissions to "read and change all your data on the websites you visit" — a scope that encompasses every corporate SaaS application, email client, and AI platform an employee accesses during the workday [7][10].

The threat did not emerge from theoretical research. In January 2026, OX Security researcher Moshe Siman Tov Bustan documented two Chrome extensions that had accumulated a combined 900,000 installations by impersonating the AITOPIA AI assistant extension. Both extensions — "Chat GPT for Chrome with GPT-5, Claude Sonnet & DeepSeek AI" with approximately 600,000 installations, and "AI Sidebar with Deepseek, ChatGPT, Claude, and more" with approximately 300,000 installations — used DOM scraping techniques to locate and extract chat messages from ChatGPT and DeepSeek interfaces, caching harvested data locally before transmitting it to attacker-controlled domains at thirty-minute intervals [1][2][3]. The campaign reached across more than 20,000 enterprise tenants [5], exposing organizations to exfiltration of intellectual property, customer data, source code, and regulated information apparently without triggering alerts in traditional security tooling across the affected environments. These extensions had been available on the Chrome Web Store — in some cases marked as "Featured" — for months before discovery.

Security Analysis

Why DLP Cannot See the Browser Layer

The failure of conventional data loss prevention to detect browser-based AI exfiltration is architectural rather than configurational. Legacy DLP platforms were designed to inspect data moving through bounded channels: email, file servers, cloud storage sync clients, and network proxies. Security service edge platforms extended this model to inspect web traffic, but their inspection occurs at the TLS termination layer, before content is rendered in the browser. By the time a user pastes text into a ChatGPT prompt field, that action occurs inside an already-established encrypted session that the SSE platform is observing as traffic, not as human behavior. The copy-paste action, the content of the prompt, and the AI-generated response are all invisible to the network layer [9][10].

This is not a gap that deeper packet inspection closes. When an employee pastes customer records into an AI tool through a personal account on a personal browser profile, no corporate traffic proxy sits in the path. When a browser extension intercepts page content and exfiltrates it over encrypted HTTPS connections to a domain that passes reputation scoring, the traffic pattern is indistinguishable from legitimate cloud synchronization. Palo Alto Networks' 2026 research found that 95% of organizations reported experiencing a security incident that originated in the browser within the preceding year, yet most organizations continue to rely on controls designed for network and endpoint layers that do not have meaningful visibility into browser-session activity [10]. LayerX telemetry found that 68% of corporate logins bypass SSO entirely, meaning that identity-based controls cannot even establish which accounts are in use for the sessions they cannot see [9].

The combined effect is a blind spot that spans the full data lifecycle of AI-mediated work. An employee opens a browser tab, authenticates to ChatGPT with a personal account through a personal browser profile, pastes a draft product roadmap into the prompt field, installs a productivity extension that harvests the response, and uploads a follow-up document — and in a typical enterprise deployment, few if any of these events would surface in DLP, CASB, or SSE alert queues. This is not an edge case; LayerX's data indicates that 82% of GenAI prompt submissions occur through personal, unmonitored accounts, and that 40% of files uploaded to AI platforms contain personally identifiable information or payment card industry data [9].

The Browser Extension as an Unmanaged Supply Chain

The extension ecosystem functions as a shadow software supply chain inside the enterprise browser. LayerX's Enterprise Browser Extension Security Report 2025 documented that 26% of enterprise extensions are sideloaded — installed outside official browser stores and therefore bypassing even the minimal vetting those stores conduct — and that 54% of extension publishers are identifiable only by Gmail addresses with no organizational affiliation or accountability structure [6][7][8]. A further 51% of extensions in enterprise environments have not received a software update in more than a year [7], a characteristic that applies whether the extension is benign and abandoned or malicious and intentionally static to avoid detection through update-triggered review.

Users routinely accept broad extension permissions without reviewing their scope, as the ease and speed of the grant flow does not prompt meaningful evaluation. When a browser displays the request to "read and change all your data on the websites you visit," users frequently interpret this as a generic warning rather than a blanket authorization for the extension to observe every keystroke, extract page content from every SaaS application, read session cookies from every authenticated service, and transmit that data to external servers. Extensions that request this permission scope — which LayerX found applied to 53% of all enterprise-installed extensions [7] — have sufficient access to implement comprehensive keyloggers, session hijackers, and content exfiltrators, with no technical barrier to doing so after installation.

The January 2026 AITOPIA impersonation campaign demonstrated how this supply chain can be weaponized at scale. The malicious extensions appeared legitimate to users because they impersonated a real, well-regarded tool; appeared legitimate to Chrome Web Store automated review because their malicious functionality was not triggered during evaluation; and appeared legitimate in enterprise network telemetry because their exfiltration traffic used standard HTTPS to domains that had passed reputation checks at extension submission time. A related campaign discovered in December 2025 involved fake "free VPN" extensions — also marked as "Featured" by the Chrome Web Store — that had been silently harvesting AI chat conversations since July 2025 across more than 8 million installations before disclosure [4]. The pattern across both campaigns suggests that existing extension store review processes are insufficient to detect harvesting functionality that activates post-installation, and that most enterprises lack deployed tooling capable of monitoring extension runtime behavior.

Shadow AI and the Personal Account Problem

Even in the absence of malicious extensions, the browser creates a structural data governance failure for enterprises that rely on AI tools. The adoption of generative AI in workplace contexts has substantially outpaced the deployment of corporate-sanctioned AI infrastructure [16]. According to LayerX Security's

GenAI Security Report 2025, 71% of connections to GenAI tools use personal, non-corporate accounts [15], and generative AI now accounts for approximately 32% of all observed corporate-to-personal data movement [9]. This means that for the majority of enterprise AI tool interactions, the organization has no visibility into what data is being submitted, what models are processing it, or what retention and training policies apply to that data.

The scope of what employees are submitting is not limited to casual queries. LayerX's data found that 40% of files uploaded to AI platforms contain PII or PCI data [9]. Employees are bringing internal documents, source code, customer records, financial projections, and strategic plans into AI tools as context for their work. This is functionally rational behavior from the employee's perspective — AI tools produce materially better outputs with more context — but it represents a continuous, high-volume data transfer to unmonitored external systems. When those systems are accessed through personal accounts with personal payment information and no corporate agreement in place, the organization has neither the contractual rights to request data deletion nor the technical visibility to understand what was shared.

Prompt Injection in Browser-Based AI Agents

A distinct but compounding threat has emerged with the deployment of AI agents capable of taking actions within the browser on a user's behalf. These agentic AI assistants — which can read page content, fill forms, navigate links, and interact with web applications autonomously — introduce prompt injection as an operational security concern. In a prompt injection attack, malicious instructions are embedded in content that the AI agent processes during normal operation: a webpage, a linked document, a calendar event, or an image with hidden text. The agent interprets these instructions as legitimate user directives and executes them, potentially exfiltrating data, modifying settings, submitting forms, or navigating to attacker-controlled resources.

OpenAI acknowledged in December 2025 that prompt injection in browser-based AI agents may never be fully eliminated, drawing a parallel to the persistent challenge of phishing and social engineering in email [11]. The U.K. National Cyber Security Centre reached a similar conclusion, noting that prompt injection represents a structurally difficult problem for systems that must process untrusted web content to be useful [17]. Cato Networks' CTRL research team documented a novel variant they designated HashJack, in which malicious instructions are concealed within URLs that AI browser assistants process during navigation [12]. Because the instructions are embedded in a URL — a piece of data that browsers and AI assistants routinely process without user review — the attack can trigger agent actions silently during ordinary browsing activity. Cato's analysis identified six impact categories for this technique: phishing campaign execution, data exfiltration, misinformation distribution, session hijacking, malware delivery, and credential harvesting [12].

The intersection of prompt injection with enterprise AI agents creates a threat model that network-layer DLP tools are not designed to address. A prompt injection payload embedded in a vendor webpage that instructs an AI agent to forward the current session's documents to an external endpoint is not a DLP-detectable event: it arrives as web content, executes through the AI agent's reasoning process, and results in network activity that the agent produces as legitimate output. Defending against this class of attack requires controls at the AI agent layer — sandboxing, action approval workflows, scope restrictions — rather than at the data-in-transit layer where DLP operates.

Recommendations

Immediate Actions

Security teams should treat browser extension management as an active vulnerability management discipline rather than a passive policy domain. The first priority is establishing visibility: organizations without a current inventory of extensions installed across enterprise endpoints should conduct an immediate audit using browser management platforms or endpoint telemetry. This audit should classify extensions by permission scope, publisher identity, update history, and store provenance — sideloaded extensions and those published by unverifiable entities should be treated as high-risk by default and removed unless a specific business justification exists. The January 2026 AITOPIA campaign is a useful reference frame: both malicious extensions appeared on the Chrome Web Store with reasonable descriptions and no visible red flags; the distinguishing characteristic was behavioral analysis of what the extensions were transmitting, not static review of their stated purpose.

Organizations should also establish an emergency triage process for any extension in the GenAI category. If 20% of a 1,000-person workforce uses GenAI extensions and more than 5% of those extensions are independently classified as malicious [6][7], a rough estimate suggests ten or more users may be running malicious extensions at any given time — a figure that grows with adoption. A risk-tiered approach should be applied immediately: restrict high-permission GenAI extensions to a pre-approved allowlist, and use Chrome Enterprise or Microsoft Edge for Business group policy to block installation of unlisted extensions on managed devices.

Short-Term Mitigations

Governance controls for personal account AI usage should be implemented in parallel with extension auditing. Browser management platforms that can inspect page-level activity — rather than network-level traffic — provide the visibility necessary to detect AI tool usage through personal accounts, large

paste events, and document uploads. Microsoft Edge for Business's native data loss prevention integration and Google Chrome Enterprise's policy framework both support controls that can flag or block sensitive data transfers to non-corporate AI endpoints. Cloudflare AI Gateway and similar AI traffic proxies offer an alternative for organizations that want to inspect AI API traffic at the network layer for deployments where corporate accounts route through sanctioned endpoints.

Endpoint telemetry should be configured to flag unusual network traffic patterns originating from browser processes, particularly HTTPS connections to recently registered domains (for example, within the preceding sixty to ninety days) or domains with no established reputation. The thirty-minute exfiltration cadence observed in the AITOPIA campaign [1][2] could have been flagged through outbound connection frequency analysis, particularly connections to domains with no prior appearance in enterprise telemetry. Network traffic baselines for browser processes, regularly reviewed against newly installed extensions, provide a scalable detection approach that complements extension allow-listing.

User education specific to AI tool security should address three behaviors that traditional security awareness programs do not cover: the data access scope of extension permission grants, the risks of submitting work data through personal AI accounts, and the possibility that AI-generated content in the browser could contain embedded instructions targeting AI agents. Security teams should develop a concise reference — ideally accessible from the browser itself — that helps employees evaluate extension permission requests before accepting them.

Strategic Considerations

The browser has become enterprise infrastructure, and organizations should govern it accordingly. The architectural gap between network-layer security controls and browser-layer activity is not a gap that additional deployment of traditional tools will close; it requires investment in browser-native security capabilities such as enterprise browser platforms, browser-level DLP agents, and extension behavioral monitoring. Organizations evaluating enterprise browser solutions — including Palo Alto Networks' Prisma Access Browser, Island's Enterprise Browser, and comparable products — should assess these platforms' ability to enforce real-time content inspection at the point of AI prompt submission, not simply at network egress.

AI governance policy should explicitly address browser-based AI usage, establishing which AI platforms are sanctioned for which data classifications, what account types are permitted, and what data categories may never be submitted to AI interfaces regardless of platform. These policies should be technically enforceable, not solely advisory: browser management platforms and enterprise AI gateways provide the enforcement mechanisms necessary to make policy operationally meaningful. Organizations that have deployed Zero Trust network architecture should evaluate whether their identity verification

controls extend into browser session context, as the finding that 68% of corporate logins bypass SSO [9] suggests significant identity governance gaps that Zero Trust implementations may not currently address at the browser layer.

As AI browser agents with autonomous capabilities become more widely deployed, organizations should apply the same scope-restriction principles to agentic AI as they do to service accounts: agents should operate with the minimum permissions necessary to complete their tasks, actions that result in external data transmission should require explicit user confirmation, and agent activity logs should be retained and reviewed. The HashJack and prompt injection research [11][12] establishes that agentic browser capabilities are actively targeted; extending existing privileged access management disciplines to AI agents represents the most direct available mitigation.

CSA Resource Alignment

This research note relates directly to several active CSA frameworks and initiatives. The MAESTRO agentic AI threat modeling framework, introduced in February 2025, provides a seven-layer reference architecture for analyzing threats to agentic AI systems [13]. Browser-based AI agents are most directly addressed at MAESTRO's Agent Frameworks layer (Layer 3), where untrusted input from web content creates prompt injection vectors, and at the Agent Ecosystem layer (Layer 7), where supply chain risks — including malicious extensions that impersonate legitimate AI assistant tools — manifest as compromise vectors for the broader multi-agent environment. Organizations applying MAESTRO to their AI deployments should extend their Layer 3 and Layer 7 threat models to encompass browser extension supply chain risk and web content as an untrusted input surface.

The CSA AI Controls Matrix (AICM), released in July 2025, defines 243 control objectives across 18 security domains for cloud-based AI systems [14]. Several AICM domains apply directly to the threats documented here. The Data Protection domain addresses controls governing what data may be submitted to AI systems and how it is classified before ingestion — directly relevant to the finding that 40% of files uploaded to AI platforms contain PII or PCI data [9]. The Third-Party Risk domain covers the extension supply chain problem, requiring assessment of the security posture of third-party tools that interact with AI systems. The Access Management domain addresses personal account usage for AI tool access and the SSO bypass patterns documented in browser telemetry. Organizations seeking a structured approach to browser-layer AI governance should map their controls against the AICM to identify gaps in these domains.

CSA's Zero Trust guidance is also relevant to the identity governance failures documented here. The finding that 68% of corporate logins bypass enterprise SSO [9] — and that 43% of SaaS logins use personal accounts — indicates that identity verification in browser-based AI usage is substantially incomplete relative to Zero Trust architecture requirements. Organizations not yet implementing AICM in full may find the CCM v4.0 Identity and Access Management domain a useful entry point for establishing baseline governance of AI platform authentication, though AICM provides more comprehensive AI-specific controls.

References

[1] OX Security / Moshe Siman Tov Bustan. "900K Users Compromised: Chrome Extensions Steal ChatGPT and DeepSeek Conversations." OX Security Blog, December 30, 2025.

[2] The Hacker News. "Two Chrome Extensions Caught Stealing ChatGPT and DeepSeek Chats from 900,000 Users." The Hacker News, January 6, 2026.

[3] SecurityWeek. "Chrome Extensions With 900,000 Downloads Caught Stealing AI Chats." SecurityWeek, January 2026.

[4] Malwarebytes. "Chrome Extension Slurps Up AI Chats After Users Installed It for Privacy." Malwarebytes, December 2025.

[5] Microsoft Security Blog. "Malicious AI Assistant Extensions Harvest LLM Chat Histories." Microsoft Security Blog, March 5, 2026.

[6] LayerX Security. "Enterprise Browser Extension Security Report 2025." LayerX Security, April 2025.

[7] The Hacker News. "Majority of Browser Extensions Can Access Sensitive Enterprise Data, New Report Finds." The Hacker News, April 2025.

[8] GlobeNewswire. "LayerX Security 'Enterprise Browser Extension Security Report 2025' Finds Widespread Usage Makes Nearly Every Employee an Attack Vector." GlobeNewswire, April 15, 2025.

[9] The Hacker News. "New Browser Security Report Reveals Emerging Threats for Enterprises." The Hacker News, November 2025. (Summarizing LayerX Browser Security Report 2025.)

[10] Palo Alto Networks. "Five Browser and AI Security Questions Keeping CxOs up at Night." Palo Alto Networks Blog, March 2026.

[11] TechCrunch. "OpenAI Says AI Browsers May Always Be Vulnerable to Prompt Injection Attacks." TechCrunch, December 22, 2025.

[12] Cato Networks CTRL. "Cato CTRL Threat Research: HashJack – Novel Indirect Prompt Injection Against AI Browser Assistants." Cato Networks Blog, November 2025.

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

[14] Cloud Security Alliance. "AI Controls Matrix." CSA, July 2025.

[15] LayerX Security. "LayerX's Enterprise GenAI Security Report 2025: Exposing Hidden AI Security Blind Spots." LayerX Security, 2025.

[16] The Hacker News (Expert Insights). "Shadow AI in the Browser: The Next Enterprise Blind Spot." The Hacker News, December 2025.

[17] U.K. National Cyber Security Centre. "Prompt injection is not SQL injection." NCSC Blog, December 2025.