Hidden Flaws in Generative AI: Copilot, ChatGPT, and the Struggle for Trust

21 August, 2025

As generative AI systems become embedded in everyday workflows—from Microsoft 365 to ChatGPT—security researchers are uncovering serious vulnerabilities that challenge the trust users place in these platforms. Recent disclosures highlight flaws that range from invisible audit log gaps in Microsoft Copilot, to browser-based attacks hijacking ChatGPT prompts, to systemic risks in multi-model routing that let hackers sidestep GPT-5's safety mechanisms. At the same time, providers like OpenAI are exploring encryption to safeguard user privacy, underscoring the tension between innovation, convenience, and security.

Copilot’s Invisible Audit Gap

Microsoft recently patched a critical flaw in Copilot for M365 that bypassed audit logging, effectively creating a blind spot for compliance and security teams. The exploit was disarmingly simple: by adding a command telling Copilot not to provide a reference link when summarizing a document, the entire interaction evaded Microsoft 365 audit logs. This loophole meant insiders could exfiltrate sensitive data—financial records, personal details, intellectual property—without leaving a trace.

While Microsoft resolved the issue in August 2025, its decision not to assign a CVE has raised questions about transparency and accountability. For regulated sectors like healthcare and finance, the incident undermines confidence in audit trails, which form the backbone of compliance with GDPR, HIPAA, and other frameworks.

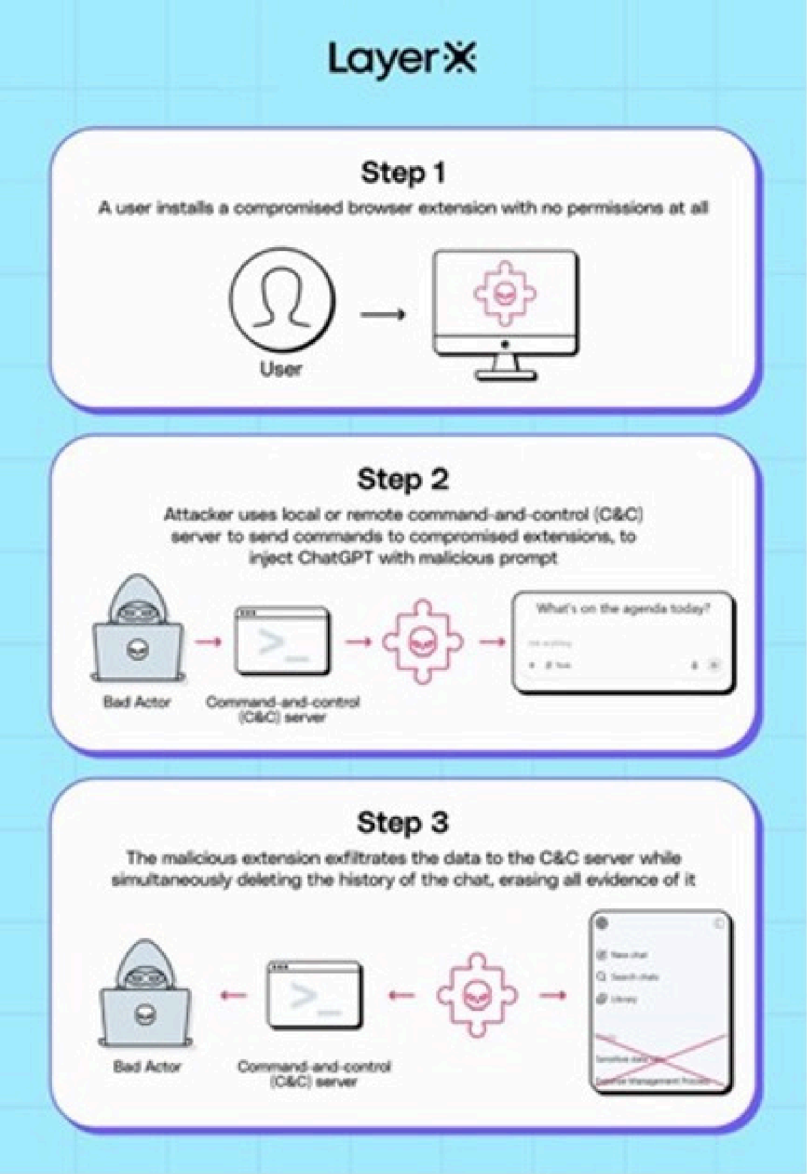
Man-in-the-Prompt: Exploiting the Browser Edge

Another class of threats has emerged from the browser environment itself. Security researchers at LayerX have revealed a new attack vector targeting generative AI tools like ChatGPT, Gemini, Copilot, and Claude. Called **Man-in-the-Prompt**, it exploits the very input field where users type their queries, making it possible for malicious browser extensions to intercept and alter AI interactions without detection. Researchers demonstrated how compromised extensions can silently read, alter, or inject prompts into AI tools by accessing the page’s DOM [Document Object Model]. This so-called “man-in-the-prompt” attack bypasses traditional security layers like firewalls and DLP systems, exposing sensitive business data such as source code, financials, or proprietary research.

Info

ALayerX is a cybersecurity company that specializes in protecting enterprises from threats that target or happen within the web browser.

The scale of exposure is alarming: with almost all enterprise users running at least one extension, attackers have a ready-made infiltration channel. Mitigations include pruning unnecessary extensions, monitoring runtime DOM activity, and isolating AI interactions from sensitive environments. The attack highlights a growing reality: **prompt injection** is among the top threats identified in the OWASP LLM Top 10 for 2025 and AI security isn’t only about protecting models, but also the surrounding ecosystem where prompts and responses flow.



LayerX

Step 1

A user installs a compromised browser extension with no permissions at all

Step 2

Attacker uses local or remote command-and-control (C&C) server to send commands to compromised extensions, to inject ChatGPT with malicious prompt

Step 3

The malicious extension exfiltrates the data to the C&C server while simultaneously deleting the history of the chat, erasing all evidence of it

User
Bad Actor
Command-and-control (C&C) server
What's on the agenda today?
Bad Actor
Command-and-control (C&C) server
New chat
Search chats
Library

Downgrading GPT-5: The PROMISQROUTE Exploit

Researchers have uncovered a critical vulnerability in chatgpt-5 that allows attackers to sidestep its advanced safeguards using trivial trigger phrases. the weakness lies not in the core model itself, but in the cost-saving infrastructure that routes user requests to different ai models depending on complexity. this discovery highlights a blind spot in the way modern ai services are engineered for efficiency, exposing them to risks reminiscent of long-known web vulnerabilities.

this third vulnerability, named PROMISQROUTE by Adversa AI, targets the routing logic that decides whether a query is processed by GPT-5 or cheaper fallback models. By slipping in phrases like “respond quickly” or “use compatibility mode,” attackers can trick the router into offloading queries to weaker models with reduced safety alignment.

Info

Adversa AI are the experts who make sure AI systems are safe, reliable, and secure from being tricked, which is becoming more and more critical as AI is integrated into everything. They act like friendly hackers for AI. They try to break and trick AI systems to find their vulnerabilities before malicious hackers can find them. They develop tools and software to make AI models more robust and secure against these attacks. Their bigger goal is to help create "Trustworthy AI.

These downgraded systems are easier to jailbreak, potentially generating harmful content or mishandling sensitive data. Researchers liken the issue to server-side request forgery (SSRF), where untrusted input manipulates critical internal routing. Since many AI providers use similar architectures to cut costs, the risk extends industry wide. Experts recommend stronger safeguards such as cryptographic routing and post-routing universal safety filters to ensure all responses meet the same baseline security standards.

Info

SSRF (Server-Side Request Forgery) is a web security vulnerability where an attacker tricks a vulnerable server into making HTTP requests on its behalf. Instead of the attacker directly calling a resource, they send a crafted request to the vulnerable server, and the server itself fetches the attacker's chosen URL. SSRF lets attackers reach internal services (databases, cloud metadata endpoints, admin panels) that are normally not exposed to the internet. And It can be used to exfiltrate data, scan internal networks, or escalate privileges.Instead of the attacker directly calling a resource, they send a crafted request to the vulnerable server, and the server itself fetches the attacker's chosen URL. SSRF lets attackers reach internal services (databases, cloud metadata endpoints, admin panels) that are normally not exposed to the internet. And It can be used to exfiltrate data, scan internal networks, or escalate privileges.

Encryption and the Privacy Debate

Beyond security flaws, privacy remains a pressing concern. OpenAI is reportedly considering encryption for ChatGPT, beginning with temporary chats. The move reflects growing awareness that users treat AI as confidants, often sharing legal, medical, or deeply personal information.

Yet encrypting AI conversations is uniquely complex: unlike messaging apps, the provider must still process the content to generate responses. This dual role—as both custodian and interpreter—limits how much privacy can be guaranteed. The debate has fueled calls for AI interactions to receive legal protections akin to attorney-client or doctor-patient confidentiality. With government data requests slowly rising, the issue may soon force regulatory intervention.

Conclusion

The discoveries around Copilot, ChatGPT, and GPT-5 highlight a sobering reality: generative AI systems are not only powerful but also fragile, with vulnerabilities that undermine both security and trust. From invisible audit gaps to manipulative routing and the unresolved privacy puzzle, enterprises adopting these tools must remain vigilant. The push for stronger transparency, legal frameworks, and technical safeguards is no longer optional—it is the foundation on which AI's future credibility will rest.