Man-in-the-Prompt: The invisible attack threatening ChatGPT and other AI

Data: 2025-08-16 17:37:13

Autor: Inteligência Against Invaders

Man-in-the-Prompt: The invisible attack threatening ChatGPT and other AI systems

Man-in-the-Prompt: a new threat targeting AI tools like ChatGPT and Gemini via simple browser extensions, no complex attack needed.

A new type of threat is alarming the world of cyber security: it is called Man-in-the-Prompt and is capable of compromising interactions with leading generative artificial intelligence tools such as ChatGPT, Gemini, Copilot, and Claude. The problem? It does not even require a sophisticated attack: all it takes is a browser extension.

"*LayerX's research shows thatanybrowser extension, even without any special permissions, can access the prompts of both commercial and internal LLMsand inject them with prompts to steal data, exfiltrate it, and cover their tracks.The exploit has been tested on all top commercial LLMs, with proof-of-concept demos provided for ChatGPT and Google Gemini*", explains researcher Aviad Gispan of LayerX. https://layerxsecurity.com/blog/man-in-the-prompt-top-ai-tools-vulnerable-to-injection/

Point Wind credit

What is "Man-in-the-Prompt"?

With this term, LayerX Security experts refer to a new attack vector that exploits an underestimated weakness: the input window of AI chatbots. When we use tools such as ChatGPT from a browser, our messages are written in a simple HTML field, accessible from the page's DOM (Document Object Model). This means that any browser extension with access to the DOM can read, modify, or rewrite our requests to the AI, and do so without us noticing. The extension doesn't even need special permissions.

ChatGPT Injection Proof Of Concept https://youtu.be/-QVsvVwnx_Y

[VÍDEO/IFRAME REMOVIDO]

Point Wind credit

How the attack works

1. The user opens ChatGPT or another AI tool in their browser.

- The malicious extension intercepts the text that is about to be sent.

- The prompt is modified, for example to add hidden instructions (prompt injection) or exfiltrate data from the AI’s response.

- The user receives a seemingly normal response, but behind the scenes, data has already been stolen or the session compromised.

This technique has been proven to work on all major AI tools, including:

- ChatGPT (OpenAI)
- Gemini (Google)
- Copilot (Microsoft)
- Claude (Anthropic)
- DeepSeek (Chinese AI model)

What are the concrete risks?

According to the report, the potential consequences are serious, especially in the business world:

- Theft of sensitive data: if the AI processes confidential information (source code, financial data, internal reports), the attacker can read or extract this information through modified prompts.

- Manipulation of responses: an injected prompt can change the behavior of the AI.

- Bypassing security controls: the attack occurs before the prompt is sent to the AI server, so it bypasses firewalls, proxies, and data loss prevention systems.

According to LayerX, 99% of business users have at least one extension installed in their browser. In this scenario, the risk exposure is very high.

What we can do

For individual users:

- Regularly check installed extensions and uninstall those that are not necessary.
- Do not install extensions from unknown or unreliable sources.
- Limit extension permissions whenever possible.

For businesses:

- Block or actively monitor browser extensions on company devices.
- Isolate AI tools from sensitive data, when possible.
- Adopt runtime security solutions that monitor the DOM and detect manipulation in input fields.
- Perform specific security tests on prompt flows, simulating injection attacks.
- An emerging measure is the use of so-called prompt signing: digitally signing prompts to verify their integrity before sending. “Spotlighting” techniques, i.e., labeling the sources of AI instructions, can also help distinguish reliable content from potential manipulations.

A bigger problem: Prompt Injection

The Man-in-the-Prompt attack falls under the broader category of prompt injection, one of the most serious threats to AI systems according to the OWASP Top 10 LLM 2025. These are not just technical attacks: even seemingly harmless external content, such as emails, links, or comments in documents, can contain hidden instructions directed at the AI.

For example:

- Corporate chatbots that process support tickets can be manipulated with malformed requests.
- AI assistants that read emails can be tricked into sending information to third parties with injected prompts.

What we learn

The LayerX report raises a crucial point: AI security cannot be limited to the model or server, but must also include the user interface and browser environment. In an era where AI is increasingly integrated into personal and business workflows, a simple HTML text field can become the Achilles heel of the entire system.

Credit

About the author: Salvatore Lombardo(**X**@Slvlombardo)

Electronics engineer and Clusit member, for some time now, espousing the principle of conscious education, he has been writing for several online magazine on information security. He is also the author of the book “La Gestione della Cyber Security nella Pubblica Amministrazione”. “Education improves awareness” is his slogan.

Follow me on Twitter:@securityaffairsandFacebookandMastodon

PierluigiPaganini

(SecurityAffairs–hacking,Man-in-the-Prompt)