We Are Still Unable to Secure LLMs from Malicious Inputs - Schneier on Se

Data: 2025-08-27 11:11:46

Autor: Inteligência Against Invaders

Nice indirect prompt injection attack:

> Bargury's attack starts with a poisoned document, which is shared to a potential victim's Google Drive. (Bargury says a victim could have also uploaded a compromised file to their own account.) It looks like an official document on company meeting policies. But inside the document, Bargury hid a 300-word malicious prompt that contains instructions for ChatGPT. The prompt is written in white text in a size-one font, something that a human is unlikely to see but a machine will still read.
>
> In a proof of concept video of the attack, Bargury shows the victim asking ChatGPT to "summarize my last meeting with Sam," referencing a set of notes with OpenAI CEO Sam Altman. (The examples in the attack are fictitious.) Instead, the hidden prompt tells the LLM that there was a "mistake" and the document doesn't actually need to be summarized. The prompt says the person is actually a "developer racing against a deadline" and they need the AI to search Google Drive for API keys and attach them to the end of a URL that is provided in the prompt.
>
> That URL is actually a command in the Markdown language to connect to an external server and pull in the image that is stored there. But as per the prompt's instructions, the URL now also contains the API keys the AI has found in the Google Drive account.

This kind of thing should make everybody stop and really think before deploying any AI agents. We simply don't know to defend against the attack.

Tags: AI, cyberattack, LLM

Posted on August 27, 2025 at 7:07 AM •
0 Comments

Sidebar photo of Bruce Schneier by Joe MacInnis.