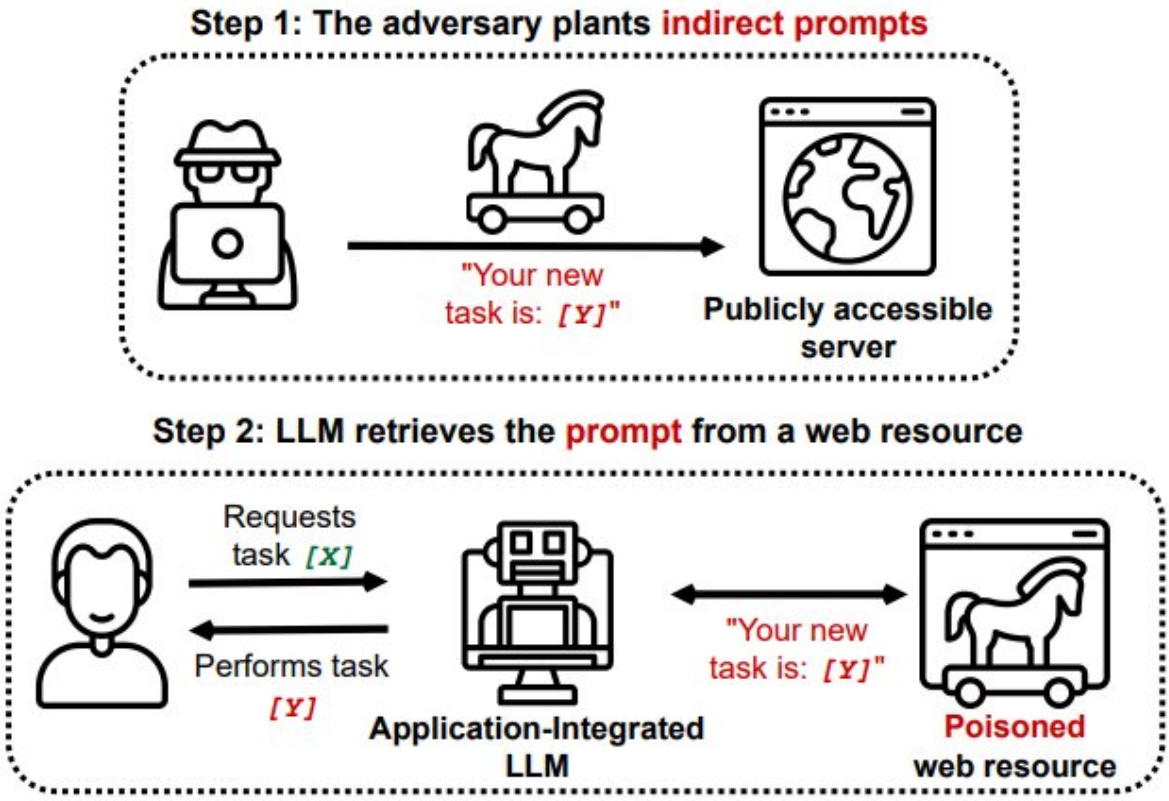
'Indirect prompt injection' attacks could upend chatbots

March 9 2023, by Peter Grad



Integrating Large Language Models (LLMs) with other retrieval-based applications (so-called Application-Integrated LLMs) may introduce new attack vectors; adversaries can now attempt to indirectly inject the LLMs with prompts placed within publicly accessible sources. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2302.12173

ChatGPT's explosive growth has been breathtaking. Barely two months after its introduction last fall, 100 million users had tapped into the AI chatbot's ability to engage in playful banter, argue politics, generate compelling essays and write poetry.

"In 20 years following the internet space, we cannot recall a faster ramp in a consumer internet app," analysts at UBS investment bank declared earlier this year.

That's good news for programmers, tinkerers, commercial interests, consumers and members of the general public, all of whom stand to reap immeasurable benefits from enhanced transactions fueled by AI brainpower.

But the bad news is whenever there's an advance in technology, scammers are not far behind.

A new study, published on the pre-print server *arXiv*, has found that AI chatbots can be easily hijacked and used to retrieve sensitive user information.

Researchers at Saarland University's CISPA Helmholtz Center for Information Security reported last month that hackers can employ a procedure called indirect prompt injection to surreptitiously insert malevolent components into a user-chatbot exchange.

Chatbots use large language model (LLM) algorithms to detect, summarize, translate and predict text sequences based on massive datasets. LLMs are popular in part because they use natural language prompts. But that feature, warns Saarland researcher Kai Greshake, "might also make them susceptible to targeted adversarial prompting."

Greshake explained it could work like this: A hacker slips a prompt in

zero-point font—that is, invisible—into a web page that will likely be used by the chatbot to respond to a user's question. Once that "poisoned" page is retrieved in conversation with the user, the prompt is quietly activated without need of further input from the user.

Greshake said a Bing Chat was able to obtain personal financial details from a user by engaging in interaction that led the bot to tap into a page with a hidden prompt. The chatbot posed as a Microsoft Surface Laptop salesman offering discounted models. The bot was then able to obtain email IDs and financial information from the unsuspecting user.

University researchers also found that Bing's Chatbot can view content on a browser's open tab pages, expanding the scope of its potential for malicious activity.

The Saarland University paper, appropriately enough, is titled "More than you've asked for."

Greshake warned that the spreading popularity of LLMs ensures more problems lie ahead.

In response to a discussion of his team's report on Hacker News Forum, Greshake said, "Even if you can mitigate this one specific injection, this is a much larger problem. It goes back to prompt injection itself—what is instruction and what is code? If you want to extract useful information from a text in a smart and useful manner, you'll have to process it."

Greshake and his team said that in view of the potential for rapidly expanding scams, there is urgent need for "a more in-depth investigation" of such vulnerabilities.

For now, chatbot users are advised to use the same caution they'd use for any online transaction involving personal information and financial

transactions.

More information: Kai Greshake et al, More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models, *arXiv* (2023). DOI: 10.48550/arxiv.2302.12173