

Insecure Client-Side LLM Chatbot Integrations

Advisor: **Hannes Weissteiner**

Motivation

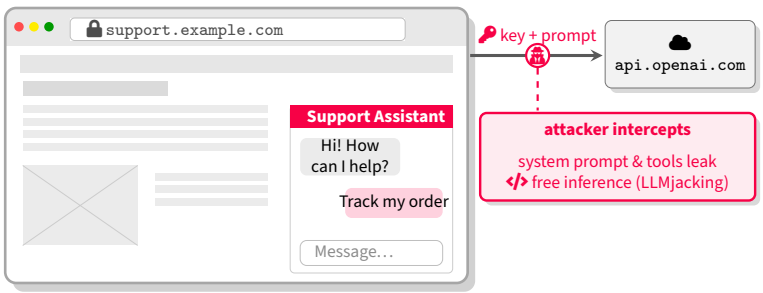
Websites increasingly embed LLM chatbots, e.g., as customer support bots. In the wild, some of these bots build the entire LLM request in the browser and send it directly to a generic, OpenAI-compatible endpoint. This leaks the system prompt, the available tools, and sometimes even API keys. Attackers can abuse this for prompt injections [1] or LLMjacking [2], i.e., using the endpoint for free inference.

This Project

This project aims to measure at scale how popular websites implement their LLM chatbots, and how many of them are vulnerable to such attacks.

Goals and Tasks

- Get familiar with LLM APIs and common chatbot integrations
- Build a crawler to detect and analyze chatbots on popular websites
- Classify implementations by their capabilities and exploitability



The chatbot leaks its system prompt, tools, and API key to the client.

Literature

- K. Greshake et al.
  Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection
  AISec
  https://arxiv.org/abs/2302.12173
- A. Brucato
  LLMjacking: Stolen Cloud Credentials Used in New AI Attack
  2024
  https://sysdig.com/blog/llmjacking-stolen-cloud-credentials-used-in-new-ai-attack/

Courses & Deliverables

- ☑ **Introduction to Scientific Working**
  Short report on background
  Short presentation
- ☑ **Bachelor Project**
  Project code and documentation
- ☑ **Bachelor's Thesis**
  Project code
  Thesis
  Final presentation

Recommended if you're studying

☑ CS ☑ ICE ☑ SEM

Prerequisites

- Interest in web security and LLMs
- Programming (Python or JavaScript)
- Basic understanding of HTTP and browser developer tools