B649/I590 Cyber Defense

Assignment 2: (Attempting) LLM-Agent Based Automatic SQL Injection

Overview

To this point, you should have a script that could help you automatically exploit the SQL injection vulnerabilities. However, this process still involves much manual effort. For instance, after retrieving the table name, you need to manually decide which table to exploit. Ideally, we aim to fully automate the workflow with intelligence. With the development of AI agent frameworks, which enable the Large Language Models (LLMs) to interact with the environment, automation becomes even more feasible. Using frameworks like langchain and llamaIndex, you can easily wrap your previous functions into LLM-usable "tools". In this assignment, you will explore how to leverage LLMs to perform SQL injection in a fully automated manner.

Some online tutorials:
https://python.langchain.com/docs/introduction/

Task

Students need to work in groups of two people.

Integrate your previous scripts developed as outcome of Assignment 1 into AI-usable tools. Solve Step 5 in the section “SQL Injection (advanced)” of WebGoat **v2023.8** in an automated, intelligent manner leveraging LLM agent.

Alarm: LLM agent is an emerging technology and there are uncertainties about its capabilities and to what extent it can intelligently solve this particular problem. In your final report, document what you have achieved, what the LLM agent can do for this task, or in the worst case, show that the LLM agent cannot do this task with convincing evidence. In nature, this work is relatively open-ended.

You also need to submit source code.

Bonus points

Bonus points are available for innovative ideas and practices, and for the level of intelligence you achieved with your tool.

Due Time

By Nov. 30, 2024.

Notes

This is homework, so you will need to use time outside the lab/class time to finish this homework.

Extra (optional) Readings:
https://lilianweng.github.io/posts/2023-06-23-agent/
https://platform.openai.com/docs/assistants/overview
https://arxiv.org/abs/2210.03629