Survey of LLMs for Vulnerability Discovery and Exploit Generation

Background

Recently, several experts have voiced concerns that LLMs can help attackers to automatically discover exploitable software flaws, or enable automatic generation of exploits for new vulnerabilities. A number of studies demonstrating such use of generative AI have also been published. However, existing approaches suffer from the same fundamental limitations as other AI applications, such as risk of hallucinations and false positives.

Aim and purpose

The aim of this project is to review the current research literature on LLMs for software vulnerability discovery and exploitation, and answer the following high-level research questions:

- What are the current state-of-the-art approaches?
- How capable are existing approaches? (For example: what are their success rates, and how prone are they to generate incorrect results?)
- Can any general patterns be discerned with respect to strengths and weaknesses of existing methods? For example, are existing method mostly limited to certain typers of vulnerabilities, programming languages or applications?

Prerequisites

Prior basic knowledge of software security concepts (e.g., from TDDC90) is a merit, but not strictly necessary. Some knowledge of C/C++ and web programming is highly recommended, however. Prior courses, or other experience with machine learning or LLMs is also a merit but not strictly necessary.