Bachelor Semester Project (BSP) Registration

Student Name: Meireles Lopes Steve
Matriculation ID: 022148763b
Email: steve.meireles.001@student.uni.lu

Project Title: Recognizing and Attacking Chatbot Models through Specific Queries

Supervisor: Francois Jérôme
Secondary Language: French
Date: March 13, 2025

1. Abstract

Nowadays, chatbots run on Large Language Models (LLMs) and are increasingly used, causing new openings for attacks. This project answers the question: ”What are the minimal queries to identify the model used?”. This will be done by quering and comparing n-numbers of LLMs, analyzing their results, and identifying their individual differences. We deduct signatures to find the minimal set of queries for a set of given LLMs.

2. Work

The project will be split into three phases:

- White Box phase: In this phase n-numbers of open-source LLMs are compared to each other using random queries. The information gained will be an individual signature for each LLM.
- Black Box phase: In this phase we do not know which of the n-LLMs are used. Using the information of the first phase, specific queries are send in a specific order depending on the answers. After this phase, we should be able to determine the hidden model.
- Follow attack phase: Each model has several vulnerabilities which can be attacked. These vulnerabilities are exploited. Most of them are shown in the ATLAS Matrix [1].

References

[1] MITRE Corporation, “Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS),” 2024, accessed: 2024-03-10. [Online]. Available: https://atlas.mitre.org/matrices/ATLAS