Student Assistant for the Information Leakage Detection on LLMs

Information leakage in Artificial Intelligence systems, particularly in Large Language Models (LLMs), poses significant security and privacy concerns. Sensitive information such as the model's architecture details, system configurations, or proprietary data can unintentionally be exposed through model outputs or interactions. Detecting and mitigating such leaks are crucial for maintaining the integrity and confidentiality of LLM systems.

We seek excellent student assistants motivated to be part of ongoing research in these areas and contribute to cutting-edge research in LLM security. As a student assistant working on information leakage detection, you will contribute to developing methods and tools to identify and prevent unauthorized disclosure of sensitive information in LLM models. This work is essential for enhancing the security of LLM applications and ensuring compliance with privacy regulations.

Your tasks include:

- Research existing methods for detecting information leakage in AI models, focusing on LLMs.
- Collect datasets of LLM model outputs that may contain leaked information; analyze the data to understand patterns or features indicative of information leakage.
- Implement prototype to detect information leakage in LLM outputs; test and refine the algorithms based on performance metrics.

Prerequisites

- Knowledge of machine learning concepts and familiarity with LLMs.
- Proficiency in programming languages such as Python; experience with CUDA is a plus.
- Strong analytical and problem-solving skills.
- Ability to conduct thorough literature reviews and synthesize information.
- Good communication skills for documentation and teamwork.
- Ability to work independently and strong motivation

Contact

System Security Lab is at the forefront of research in system security and AI trustworthiness. If you are intrigued by this subject, please get in touch with Dr. Lichao Wu and Mr. Mohamadreza Rostami at *info@trust.tu-darmstadt.de* to obtain further information. To facilitate the process, kindly include your **CV** and a copy of your **transcripts**.

Technische Universität Darmstadt
FB Informatik – System Security Lab
S2|20 (CYSEC Gebäude)
Pankratiusstraße 2
64289 Darmstadt
https://www.trust.informatik.tu-darmstadt.de

November 2024