A Structured Analysis of Security and Privacy Threats in Large Language Models

Ana Șarapova

Technical University of Moldova, sarapova.ana@isa.utm.md, 0009-0007-5776-4844

Keywords: Large Language Models, Security Threats, Data Privacy

Abstract. Large Language Models (LLMs), such as ChatGPT, have rapidly become integrated into daily life, often without a full understanding of their security and privacy implications. As these models grow more influential, two key groups have emerged: one advocating for the shutdown of LLMs due to their numerous risks, and the other calling for the development of ethical guidelines and security protocols [1]. Most of the research literature categorizes the threats posed by LLMs into four major pillars: security, privacy, trust, and ethical considerations [2]. Despite their seamless integration, LLMs present vulnerabilities that can indirectly lead to malicious attacks, placing users and organizations at risk. The exponential advancements in LLM technology have outpaced security measures, leaving critical issues unresolved. This paper aims to analyze these challenges at a broad level, identifying root causes and exploring potential remedies. The goal is to provide an understanding of LLM risks and promote responsible usage through informed guidelines.

References

[1] FLORIAN, Allwein. ChatGPT-A critical view. In: *IU Discussion Papers-IT & Engineering* [online]. 2024 [cited 14.09.2024]. Available: https://www.econstor.eu/handle/10419/287746

[2] WANG, Y., PAN, Y., YAN M., SU, T., LUAN, T. A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions. In: *IEEE Open Journal of the Computer Society* [online]. Volume 4, No. 1, pp. 280-302, 2023 [cited 14.09.2024]. Available: https://www.computer.org/csdl/journal/oj/2023/01/10221755/1PELXFR2hdS.