Contribution ID: **206** Type: **Paper presentation**

Evaluating Large Language Models Security and Resilience: A Practical Testing Framework

Thursday 18 September 2025 15:15 (15 minutes)

Large Language Models (LLMs) are increasingly used in real-world applications, but as their capabilities grow, so do the risks of misuse. Despite their widespread adoption, the security of these models remains an area with many open questions. This paper explores these issues through a set of applied experiments carried out in a controlled environment designed for testing. A prototype application that allows demonstrating how an LLM security benchmarking tool could function in practice and allowing users to simulate attacks and assess the effectiveness of several defense strategies, for example in-context defense and paraphrase-based approaches was designed. The experimental results show notable differences between the tested methods. Some techniques were able to fully block attacks while maintaining the model's ability to respond accurately to regular prompts. The prototype serves as a practical starting point for further research and can be extended to support more complex evaluation workflows in the field of LLM security.

Author: Mr NIȚESCU, George (National University of Science and Technology POLITEHNICA Bucharest)

Co-authors: Mr OUATU, Andrei (National University of Science and Technology POLITEHNICA Bucuresti); TURCANU, Dinu (Technical University of Moldova)

Presenter: Mr NIȚESCU, George (National University of Science and Technology POLITEHNICA Bucharest)

Session Classification: Doctoral Symposium

Track Classification: Network Security