LLMs and Privacy: from risk assessment to privacy enhancing tools

Nicolas Anciaux*[1]

[1]Concevoir des technologies d'amélioration de la vie privée explicables et efficaces (PETSCRAFT) – Centre INRIA de Saclay, Laboratoire d'Informatique Fondamentale d'Orléans – Campus de l'École Polytechnique 91120 Palaiseau, France

Résumé

The privacy risks associated with Large Language Models (LLMs) are increasingly critical. This talk examines two key areas: the assessment of privacy threats through membership inference attacks (MIAs), including current benchmarking challenges and bias issues; and the use of LLMs as privacy-enhancing tools, such as for anonymization or text rewriting to reduce attribute leakage. These dual perspectives highlight how LLMs can both compromise and protect privacy in AI systems.

*Intervenant