ID de Contribution: **12** Type: **Poster**

Security and Confidentiality in Retrieval-Augmented Generation Systems

Résumé

This Ph.D. research is conducted within the LIMOS laboratory (UMR 6158), Université Clermont Auvergne, and can be part of the DATA programme (I-SITE CAP 20-25), within the domain of Artificial Intelligence for the secure exploitation of data.

The DATA programme aims to address the full lifecycle of data, from acquisition and integration to large-scale storage and value creation through artificial intelligence. In this context, the rapid adoption of generative AI systems has led organizations to increasingly rely on Retrieval-Augmented Generation (RAG) architectures to exploit proprietary and domain-specific data without resorting to costly model fine-tuning.

While RAG enables effective data valorization by combining large language models with external data sources, it also introduces critical challenges related to the confidentiality of sensitive information. This Ph.D. thesis focuses on the analysis and mitigation of unauthorized data disclosure risks in RAG-based systems. The research begins with a structured risk analysis to identify realistic attacker models and data leakage scenarios across the data exploitation pipeline. Based on this analysis, metrics are proposed to quantify sensitive information exposure in RAG architectures. Empirical evaluations assess vulnerabilities, particularly those arising from malicious prompt injection and interaction-based attacks. Finally, the work explores protection mechanisms-such as taint-aware data tracking, secure data handling strategies, and access control-to support trustworthy and secure exploitation of data in AI-driven systems.

Auteur: NOUYEP, Steve (UCA)

Classification de Session: Poster Flash Talks