Secure Multi-Agent Retrieval-Augmented Generation:
Enhancing Cybersecurity in LLM-Driven Agentic AI Systems

1. Context and Background

Recent advances in large language models (LLMs) have revolutionized natural language processing, enabling innovative applications from cybersecurity to automated decision-making. LLMs face significant security challenges, including adversarial attacks that can manipulate model outputs and induce harmful behavior, as well as vulnerabilities such as prompt injection and data poisoning that may compromise the integrity and reliability of generated content. Additionally, ensuring the confidentiality and privacy of sensitive information during both training and inference remains a critical concern, necessitating robust defense mechanisms and continuous security evaluations [1]. Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm by combining LLMs with external knowledge retrieval. However, RAG systems are susceptible to adversarial manipulation, prompt injection, and other vulnerabilities—challenges that become even more complex when these systems are embedded in multi-agent architectures (agentic AI) [2]. In LLM-driven agents, these challenges are compounded by the complexity of autonomous decision-making and inter-agent communication, which expose the system to unique vulnerabilities. Attacking RAG in LLM agents can thus lead to the generation of misleading or harmful outputs by bypassing established context safeguards, which in turn may trigger unintended and potentially dangerous behaviors. Furthermore, such attacks can compromise data integrity, expose sensitive information, and undermine trust in the deployed system, posing severe risks in safety-critical applications and multi-agent environments [3, 4].

Graph mining is widely used in cybersecurity to analyze network structures, detect anomalies, and identify malicious clusters within complex systems [5]. Additionally, probabilistic techniques are essential for risk assessment and threat prediction, enabling security systems to model uncertainty and estimate the likelihood of various attack scenarios, thereby informing more robust defense strategies. Graph mining and probabilistic techniques have shown promise for enhancing the robustness of RAG systems, although their integration remains an emerging research area. For example, graph mining can be used to analyze and validate the relationships among retrieved documents, while probabilistic methods can quantify uncertainty in the generation process to detect and mitigate adversarial manipulation.

This research proposal focuses on developing robust, secure RAG systems in LLM-based multi-agent settings, exploring cutting-edge graph mining and probabilistic techniques to detect and mitigate adversarial attacks.

2. Scientific Challenges

The research will address the following key challenges:

- **Adversarial Vulnerabilities:** RAG systems can be manipulated through deceptively simple adversarial prompts, bypassing contextual safeguards and causing unintended outputs
- **Security in Multi-Agent Architectures:** In a multi-agent system, ensuring secure inter-agent communication and data sharing while preventing coordinated attacks poses a significant challenge.
- **Attacks Detection and Mitigation via Graph Mining:** Leveraging graph mining techniques to model inter-agent relationships and detect anomalous behavior in communication networks.
- **Probabilistic Risk Modeling:** Incorporating probabilistic approaches to quantify uncertainty and assess the robustness of RAG responses against adversarial inputs.

3. Research Statement

The core research question is:

How can we design and develop a secure multi-agent RAG system that leverages graph mining and probabilistic methods to robustly defend against adversarial attacks and ensure resilient, trustworthy outputs in agentic AI applications?

This question will be investigated by analyzing current vulnerabilities, developing novel detection and mitigation mechanisms, and evaluating the system's performance through rigorous experimentation.

4. Research Objectives

To address the research question, the following objectives are defined:

- **Objective 1:** Collect and analyze data from multi-agent case studies to understand attack vectors and inter-agent communication risks.
- **Objective 2:** Develop a robust RAG framework tailored for multi-agent systems with enhanced security features.

- **Objective 3:** Investigate graph mining techniques to detect anomalous agent behaviors and potential adversarial influences.
- **Objective 4:** Employ probabilistic approaches to model uncertainties and evaluate the robustness of the system.
- **Objective 5:** Validate the proposed system through controlled experiments and security analysis.

5. Three-Year Research Plan

Year 1: Foundations and Data Collection

- **State of the Art Review:** Conduct an extensive literature review on cybersecurity vulnerabilities in RAG and multi-agent systems. Map out existing attack methodologies and defense mechanisms.
- **Data Collection and Analysis:** Identify and collect datasets from multi-agent environments and case studies. Analyze real-world scenarios to establish baseline risks.
- **Preliminary Framework Design:** Define initial system architecture and communication protocols for multi-agent RAG.

Year 2: System Development and Experimental Validation

- **Framework Development:** Develop and implement the secure multi-agent RAG framework. Integrate state-of-the-art LLMs with external retrieval mechanisms.
- **Graph Mining Integration:** Explore and integrate graph mining methods to monitor inter-agent communication. Develop algorithms for anomaly detection and mitigation.
- **Probabilistic Modeling:** Incorporate probabilistic techniques to evaluate uncertainties and risk levels.
- **Experimental Evaluation:** Perform experiments and security analysis to assess system performance. Refine the framework based on empirical results.

Year 3: Optimization, Dissemination, and Thesis Completion

- **System Optimization:** Finalize the design and improve the efficiency of detection and mitigation strategies. Enhance interoperability and robustness in dynamic multi-agent scenarios.
- **Dissemination of Results:** Prepare manuscripts for submission to top cybersecurity and AI journals. Present findings at international conferences.
- **Thesis Write-Up and Defense**

6. Intellectual Merit

This research is expected to significantly advance the field of AI cybersecurity by providing:

- A novel, integrated approach combining multi-agent system design with RAG security.
- Groundbreaking contributions in applying graph mining and probabilistic methods for anomaly detection in complex AI systems.
- New insights into the interplay between external knowledge retrieval and adversarial vulnerabilities, which can inform future designs of secure LLM applications.

7. Advising and inter-lab Collaboration

- Dr. **Samia Bouzefrane**, Professeur in Cybersecurity, CNAM-CEDRIC Laboratory
- Dr. **Raphael Fournier**, Associate-Professor/HDR in AI, LIP6 and UFR919 at Sorbonne University.

8. International Collaboration

The research proposed in this thesis will be in collaboration with Dr. **Youakim Badr**, Professor in Data Analytics and Artificial Intelligence (https://youakim.info) from the Pennsylvania State University, USA. His work on trustworthy AI systems, particularly in the application of graphical neural networks and zero-knowledge proofs for intrusion detection, will be explored and integrated into our research framework. This collaboration is expected to significantly enhance the robustness of the proposed multi-agent RAG system and will foster a close research partnership. Moreover, it lays the foundation for jointly submitting research proposals for transatlantic research funds (ANR-NSF), thereby broadening the impact and scope of the project.

9. Bibliographical References

[1] https://arxiv.org/abs/2405.12750
[2] https://arxiv.org/abs/2409.10102
[3] https://arxiv.org/abs/2412.04415
[4] Ammann (2024) Analysis of Risks and Mitigation Strategies in RAG, PhD, OST Ostschweizer Fachhochschule.
[5] https://doi.org/10.1007/s10618-014-0365-y