Secure Communication Protocols for Large Language Models (LLM)-based Multi-Agent Systems

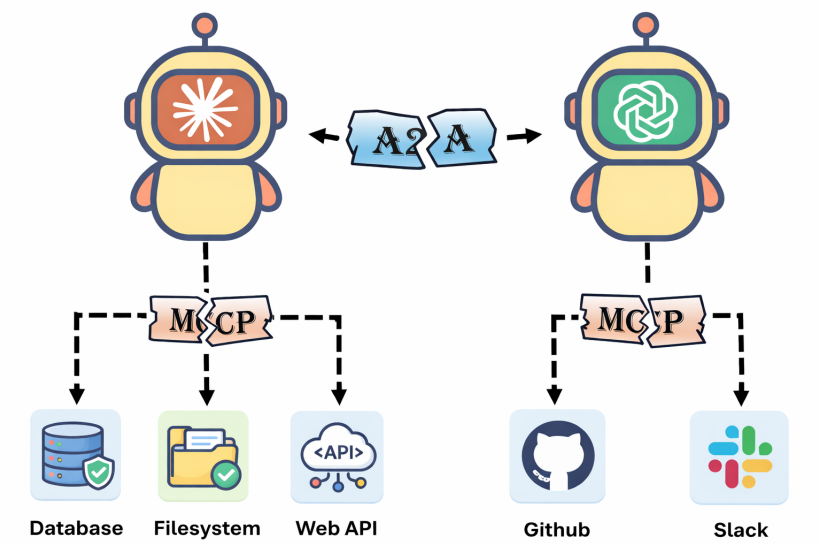


Figure 1: Vulnerability of existing A2A and MCP communication protocols in LLM-based multi-agent systems. Both inter-agent (A2A) and agent-to-resource (MCP) communication channels lack built-in security mechanisms, making them susceptible to prompt injection and other adversarial manipulations.

LLM-based MAS are increasingly integrated into everyday life. They are widely used for general assistance (e.g., ChatGPT, Gemini), software development (e.g., Claude Code), and are currently being explored in more domains.

However, current communication protocols used in LLM-based MAS, such as Agent-to-Agent (A2A) and Model Context Protocol (MCP), lack sufficient security guarantees. For example, A2A communication typically relies on natural language prompts without built-in security properties such as authentication, authorization, or input validation. This design makes LLM-based MAS highly vulnerable to prompt injection attacks and other adversarial manipulations.

This internship project aims to investigate and improve the security of communication protocols in LLM-based MAS. The intern will first perform a systematic analysis of the existing A2A and MCP protocols, identifying potential security weaknesses and attack surfaces. Based on this analysis, the project will design and propose a secure communication protocol that enhances authentication, input control, and robustness against prompt injection and related attacks. The ultimate goal is to strengthen the security of LLM-based multi-agent collaboration systems.

Practicalities

Supervisors:	Xinhai Zou	xinhai.zou@esat.kuleuven.be	office B01.25
	Sayon Duttagupta	sayon.duttagupta@esat.kuleuven.be	office B01.17
	Amit Singh Bhati	amitsingh.bhati@3milabs.tech	office B01.17

 © 2025-2026 COSIC