Université de Rennes

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2026

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : Matisse **POLE DE RECHERCHE : Mathématiques-numérique**

Unité de recherche d'accueil

Unité de Recherche : IRISA, UMR 6074

Equipe : LogicA

Encadrement

Directeur de thèse : Sophie Pinchinat **Unité de recherche** : IRISA (UMR 6074)

Employeur/Établissement d'origine : Université de Rennes

HDR : x oui non **Nb de thèses en cours** : 0

Mail : sophie.pinchinat@irisa.fr **Tél.** : +33 (0)299847254

Co-directeur de thèse : Manuel Maarek **Unité de recherche** : Computer Science Department

Employeur/Établissement d'origine : Heriot-Watt University (Édimbourg, Écosse, Royaume-Uni)

HDR : oui non x has supervised PhD students to completion **Nb de thèses en cours** :1

Mail : m.maarek@hw.ac.uk **Tél.** : +441314513287

Sujet

Titre : Security of interacting AI agent systems

Mots clés : AI agent, interaction, security, prompt-based attacks, instrumentations

Cotutelle internationale: Oui **Si oui, préciser l'établissement et le pays de rattachement** : Heriot-Watt University (Édimbourg, Écosse, Royaume-Uni)

Descriptif (15-30 lignes) :

Multi-agent systems is a paradigm that encompasses all sorts of systems, ranging from classical distributed systems (composed of deterministic components described by pure code) to agentic AI systems, including what we call here "interacting AI agent systems", combining pure code and LLM-based components. While security analysis of classical distributed systems is a well-developed field with a large spectrum of tools such as taint analysis, the analysis of security properties of interacting AI agent systems is still in their infancy because of key main reasons. First, the presence of LLM elements renders the traceability of data challenging because the original data is modified through the attention layers of transformers in a highly complex manner. Second, new security issues have been observed from the composition of LLM elements that would have their security vetted individually, but that may give rise to undesirable emergent overall behaviour.

The thesis is to take place within the LogicA team which offers a relevant research environment with combined theoretical expertise in formal methods for multi-agent systems, experience of development and deployment of LLM-based systems, as well as core AI and GenAI model training and specialisation. Moreover, the thesis is grounded on the ongoing research collaboration between Université de Rennes and Heriot-Watt University providing expertise in synergy with the LogicA team: machine learning and GenAI with Michael Lones, verification of security protocols with Sasa Radomirovic, applied formal methods for the security of software systems with Manuel Maarek.

The topic of this thesis is concerned with the security analysis of interacting AI agent systems. Typically, prompt injection is known to be a high security risk already for a single LLM-based component inference; many solutions have been proposed to mitigate this risk, for instance LLMs can be fine-tuned to maintain the separation between user data and instructions in a prompt. However, there is a strong likelihood of risks being compounded once an LLM interacts with other components of the overall system (for instance, fine-tuning of individual models is unlikely to be a sufficient guard), and therefore protections need to be considered at a higher level. The thesis' main research questions are as follows.

- How effective are existing defences against prompt-based attacks in the context of a system with different forms of LLM composition (effectiveness may be sensitive to the type of architecture)?
- What individual instrumentations of LLMs corroborate the characterisation of secure and insecure LLM inference?
- How can instrumentations of LLMs support compositional reasoning about the security of interacting AI agent systems?
- Which instrumentations of LLMs can serve the design of AI agent interaction protocols to guarantee secure behaviour of an interacting AI agent system?

There will be four main steps in the thesis:

1) Design and implement interacting AI agent systems that evidence the bypassing of solutions against prompt-based attack on individual LLMs.
2) Propose various individual instrumentations of LLMs (for example using metrics derived from the LLM models) to characterise secure and insecure LLM inferences.
3) Run experiments to select individual instrumentations that evidence interaction phenomena between components that reveal security flaws and make the evidence amenable to formal reasoning.
4) Starting from previously selected instrumentations, design remediation (such as interaction protocols) against prompt-based attacks in interacting AI agent systems.

To achieve the four steps, we propose a methodology that takes advantage of the multi-competences of the research environment offered by the LogicA team and the collaboration (cotutelle) with Heriot-Watt University, together keeping up with the fast moving applied and theoretical field of GenAI: (1) targeted implementations of interacting AI agent systems can be hosted by an environment that already exists, where team members have shown their skills on deploying agentic LLMs; (2) instrumentation of LLMs can be supported by team members' knowledge in the training of LLMs and in existing tools on metrics derived from LLM models; (3) expected compositional reasoning is supported by the team expertise in formal verification as well as security protocols.

[Bibliography]

[1] Edoardo Allegrini, Ananth Shreekumar, and Z. Berkay Celik. 2025. Formalizing the Safety, Security, and Functional Properties of Agentic AI Systems. https://doi.org/10.48550/arXiv.2510.14133

[2] ANSSI. 2024. *Recommandations de sécurité pour un système d'IA générative*. Agence Nationale de la Sécurité des Systèmes d'Information. Retrieved July 8, 2025 from https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative

[3] Luca Beurer-Kellner, Beat Buesser, Ana-Maria Creţu, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, Ezinwanne Ozoani, Andrew Paverd, Florian Tramèr, and Václav Volhejn. 2025. Design Patterns for Securing LLM Agents against Prompt Injections. https://doi.org/10.48550/arXiv.2506.08837

[4] Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. 2025. StruQ: Defending Against Prompt Injection with Structured Queries. 2025. 2383–2400. Retrieved January 27, 2026 from https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe

[5] Sarthak Choudhary, Divyam Anshumaan, Nils Palumbo, and Somesh Jha. 2026. How Not to Detect Prompt Injections with an LLM. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security* (*AISec '25*), December 30, 2026. Association for Computing Machinery, New York, NY, USA, 218–229. https://doi.org/10.1145/3733799.3762980

[6] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. 2025. Defeating Prompt Injections by Design. https://doi.org/10.48550/arXiv.2503.18813

[7] Mohamed Amine Ferrag, Norbert Tihanyi, Djallel Hamouda, Leandros Maglaras, and Merouane Debbah. 2025. From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows. https://doi.org/10.48550/arXiv.2506.23260

[8] Shiva Gaire, Srijan Gyawali, Saroj Mishra, Suman Niroula, Dilip Thakur, and Umesh Yadav. 2025. Systematization of Knowledge: Security and Safety in the Model Context Protocol Ecosystem. https://doi.org/10.48550/arXiv.2512.08290

[9] William Hackett, Lewis Birch, Stefan Trawicki, Neeraj Suri, and Peter Garraghan. 2025. Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems. In *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, August 2025. Association for Computational Linguistics, Vienna, Austria, 101–114. Retrieved January 23, 2026 from https://aclanthology.org/2025.llmsec-1.8/

[10] Ronny Ko, Jiseong Jeong, Shuyuan Zheng, Chuan Xiao, Tae-Wan Kim, Makoto Onizuka, and Won-Yong Shin. 2025. Seven Security Challenges That Must be Solved in Cross-domain Multi-agent LLM Systems. https://doi.org/10.48550/arXiv.2505.23847

[11] Jacqueline Mitchell, Brian Hyeongseok Kim, Chenyu Zhou, and Chao Wang. 2025. Understanding Formal Reasoning Failures in LLMs as Abstract Interpreters. In *Proceedings of the 1st ACM SIGPLAN International Workshop on Language Models and Programming Languages* (*LMPL '25*), October 09, 2025. Association for Computing Machinery, New York, NY, USA, 71–83. https://doi.org/10.1145/3759425.3763389

[12] OWASP Gen AI Security Project. 2024. *OWASP Top 10 for LLM Applications 2025*. OWASP. Retrieved from https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[13] OWASP Gen AI Security Project. 2025. *Securing Agentic Applications Guide*. OWASP. Retrieved October 7, 2025 from https://genai.owasp.org/resource/securing-agentic-applications-guide-1-0/

[14] Christian Schroeder de Witt. 2025. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. https://doi.org/10.48550/arXiv.2505.02077

[15] Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R. Zhang. 2025. The Geometry of Reasoning: Flowing Logics in Representation Space. https://doi.org/10.48550/arXiv.2510.09782

[16] Wei Zou, Yupei Liu, Yanting Wang, Ying Chen, Neil Gong, and Jinyuan Jia. 2026. PIShield: Detecting Prompt Injection Attacks via Intrinsic LLM Features. https://doi.org/10.48550/arXiv.2510.14005