CALL FOR APPLICATIONS — PHD POSITION

Beyond the Output: Explaining the Decisions of Multi-Agent AI Systems

Towards Verifiable Explanations for Critical Applications

Supervision: Wassila Ouerdane (Full Professor, MICS, CentraleSupélec) — co-supervision: Zineb Noumir (ALTEN) and Pierre Romet (ALTEN)

"Trustworthy Multi-Agent Systems for Critical Industries" Chair — CentraleSupélec & ALTEN

Context

Multi-agent architectures built on large language models (LLMs) decompose a complex problem into steps assigned to specialized agents, each building on the outputs of the previous ones. In **critical industrial systems**, a decision produced this way can only be accepted if an expert can understand how it was built, what information it relies on, and which requirements it satisfies. Yet a wrongly extracted requirement or erroneous — even hallucinated — information can be passed on and then endorsed by downstream agents, and no single agent holds the full justification of the final decision.

This PhD is part of the "Trustworthy Multi-Agent Systems for Critical Industries" Chair, led by **CentraleSupélec** and **ALTEN**, in close collaboration with ALTEN's Innovation Department. It relies on industrial use cases in requirements-driven engineering, where agents extract, check and exploit requirements from technical documents to support design decisions. Since the constraints involved are explicit and documented, the compliance of a decision and the origin of an error can be checked objectively.

Thesis topic

The goal is to explain the **observable decision process** of a multi-agent architecture — not the internal reasoning of LLMs — by establishing which information **actually determined** a decision, in a way that can be checked against the behavior of the system, and by relating it to the requirements the decision must satisfy. The central question is: *how can we produce an explanation that is faithful to the observable decision process, linked to the requirements it involves, and sufficient to support its audit? In critical settings, not all errors have the same consequences: the thesis will also consider how the criticality of requirements and the severity of potential errors — two related but distinct notions — should shape the level of evidence needed to justify, verify and audit a decision.*

In a chain of agents, the dependencies between agents are known by design. Unlike recent work on failure attribution, which looks for the agent or step responsible for a failure, the aim here is to explain any decision — correct or not — with respect to the requirements it must satisfy, so that it can be audited. Explaining a decision therefore means identifying, among the **information elements** produced or consulted — extracted requirements, verification results, passages of technical documentation, comparison criteria —those that determined it. Defining the granularity of these elements is itself part of the problem: too fine, and the space of candidate explanations becomes intractable; too coarse, and the explanation is of no use to an auditor. The main challenge is **faithfulness**: LLM justifications can be plausible without reflecting the process followed (Turpin et al., 2023). An explanation must therefore be checked against the observable behavior of the system, for instance through **controlled perturbation**: altering or removing an information element and replaying the chain to see whether the decision changes.

- **Core direction — traceability and verification.** Building on recent work on failure attribution in LLM-based multi-agent systems (S. Zhang et al., 2025; G. Zhang et al., 2026), capture executions of the multi-agent architectures studied in a provenance graph inspired by PROV-AGENT (Souza et al., 2025), linked to the requirements of the specification, and use it to identify the information elements a decision depends on. The graph serves a double purpose: it represents the process, and it restricts the perturbations worth testing to the elements upstream of the decision. Their role is then checked through controlled perturbations — removal, alteration, injection of known errors — and repeated replays. Other verification mechanisms may also be explored, for instance, consistency checking against the specification, symbolic verification, or cross-checking between agents. This direction is expected to form the core of the thesis.
- **Second direction — dialogue protocols.** Conditional on the first, this direction turns the chain into genuine collective reasoning: agents exchange their reasons, ask for clarification, validate or object, following a protocol grounded in a formal model of interaction and exchange, in the tradition of dialogue games

(Walton & Krabbe, 1995), with a burden of proof proportionate to the criticality of the requirements involved. The difficulty is that a protocol cannot be expected from an LLM, it must be imposed on it: the orchestrator fixes the admissible moves, agents instantiate argument schemes whose critical questions (Walton et al., 2008) define what each claim commits them to, and conformity is checked outside the model. The dialogue is not a proof of faithfulness: it produces structured claims that can then be checked by these verification mechanisms.

Methodological challenges. Two difficulties cut across both directions. First, LLM agents are not reproducible even at zero temperature, so the natural variability of the system must be measured before any perturbation is interpreted. Second, perturbation-based verification is costly, since isolating the role of each element requires repeated replays; controlling this cost — through the provenance graph, grouped ablations, or approximation methods — is part of the research problem.

Evaluation. It is hierarchical: faithfulness first, then requirement coverage and diagnostic capability — for errors affecting critical requirements — and finally usefulness for audit. Faithfulness is assessed in two complementary ways: against a benchmark of injected known errors, where the ground truth is available, and through controlled perturbation for decisions that are correct, where it is not. The evaluation compares the approaches developed with baselines such as the original chain, the chain with provenance, and the chain with textual justifications. A study with ALTEN engineers, prepared from the second year onwards, will finally check that the explanations improve the audit, not just trust.

Research questions

1. How can the decision process of a multi-agent architecture be represented at the level of the information elements it uses, to relate each decision to the requirements it involves?
2. How can we check, against the observable behavior of the system, that an explanation reflects the information that determined a decision, despite the non-determinism of LLM agents?
3. How can interaction between agents be structured so that the claims it produces are verifiable, with a level of evidence proportionate to the criticality of the requirements at stake, and at what cost?
4. Do these explanations improve the audit of a decision by an expert, and not just their trust?

Environment

- **Host:** MICS laboratory, CentraleSupélec, Université Paris-Saclay (Gif-sur-Yvette); SDS doctoral school.
- **Framework:** CentraleSupélec – ALTEN chair, in close collaboration with ALTEN's Innovation Department.
- **Funding:** doctoral contract funded by the chair (full three-year contract, salary and social security included).
- **Duration and start:** 36 months, starting end January 2027. **Applications close on 11 December 2026**

Candidate profile

A master's degree (M2), an engineering degree, or an equivalent qualification in Computer Science, Artificial Intelligence, or Applied Mathematics. Excellent skills in at least one of the following areas: Generative AI and agentic architectures, multi-agent systems, or Explainable AI (XAI). A background in, or a strong interest in, Knowledge Representation and Reasoning would be considered a strong asset. Good command of Python and scientific English; Autonomy, rigor, a taste for **both formalization and experimentation**, and interest in working with an industrial partner.

How to apply

Applications and questions should be sent to Wassila Ouerdane (wassila.ouerdane@centralesupelec.fr). Applications must include:

- a cover letter explicitly stating the candidate's interest in the proposed topic;
- a curriculum vitae clearly showing how the candidate's profile matches the expected skills;
- Master's (M1 and M2) or engineering school transcripts;
- contact details of at least one referee (two or more would be appreciated).

Optional items: scientific publications, significant internship or project reports, links to relevant work (code, software). Incomplete applications or applications not tailored to the topic will not be considered.

References

Souza, R., Gueroudji, A., DeWitt, S., Rosendo, D., Ghosal, T., Ross, R., Balaprakash, P., & Ferreira da Silva, R. (2025). PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows. *IEEE eScience 2025*, 467–473.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *NeurIPS 2023*.

Walton, D., & Krabbe, E. C. W. (1995). Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. SUNY Press.

Walton, D., Reed, C., & Macagno, F. (2008). Argumentation Schemes. Cambridge University Press.

Zhang, G., et al. (2026). AgenTracer: Who Is Inducing Failure in the LLM Agentic Systems? ICLR 2026.

Zhang, S., Yin, M., Zhang, J., et al. (2025). Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems. ICML 2025.