

Postdoctoral Research Position

Efficient LLM-Augmented Systems for Data and Software Tasks

Axis 1: Core Elements of AI

Supervisor: Prof. Paolo Papotti

Affiliation: EURECOM

Email: paolo.papotti@eurecom.fr

Location: EURECOM, Sophia Antipolis, France

Context and motivation

Large Language Models are increasingly used to execute tasks that involve structured data, text, code, and external tools. However, current solutions often rely on generic agents or fixed prompting pipelines that are expensive, difficult to optimize, and provide limited guarantees on output quality.

This is particularly problematic for data-intensive tasks. A user request may require several operations, such as identifying relevant tables, generating queries, extracting information, validating intermediate results, and combining evidence from structured and textual sources. Similar challenges arise in software engineering, where a natural-language task may require code generation, test creation, execution, debugging, and validation. In both cases, the LLM should not be viewed as a single monolithic component. Instead, it can be integrated into an execution system composed of specialized operators. Each operator may use a different model, prompt, tool, or validation strategy.

Recent systems such as Galois [2,3], LOTUS [4], and Stretto [1] explore this direction for LLM-augmented data processing. They show that the execution of LLM-based workflows raises new optimization problems, since cost, latency, and output quality depend on the choice and organization of operators. The objective of this postdoctoral project is to develop methods and systems for the **efficient, reliable, and task-specific execution of workflows involving LLMs**, with a primary emphasis on structured data processing and an opening toward applications in software engineering.

Research program

The project will investigate how to generate and execute specialized LLM-based workflows from high-level user requests. A central research direction will be the automatic construction of an ad hoc execution harness for a given task. Rather than relying on a generic agent, the system will generate a task-specific pipeline containing the models, prompts, tools, data operators, tests, and validation steps required to solve the request.

For data-processing tasks, such a harness may include operators for:

- schema and table selection;

- SQL generation and execution;
- semantic filtering and information extraction;
- entity matching and data integration;
- aggregation, verification, and result explanation.

For software-engineering tasks, the input may be a natural-language development request, while the generated harness may include operators for:

- code generation or modification;
- repository and dependency analysis;
- test generation;
- program execution;
- debugging and repair;
- functional and non-functional validation.

The research will study whether these specialized harnesses can provide quality comparable to, or better than, generic agentic systems while requiring fewer model calls, tokens, and execution steps. The project will build on principles from database and ML systems, query optimization, program synthesis, and software engineering. The resulting methods will be evaluated primarily on mixed/multimodal data workloads, together with selected software-engineering applications.

Expected outcomes

The project is expected to produce:

- abstractions for LLM-powered operators and task-specific execution harnesses;
- algorithms for quality-, cost-, and latency-aware workflow optimization;
- methods for validating and adapting LLM-generated workflows;
- an open-source research prototype;
- benchmarks comparing specialized harnesses with generic LLM agents;
- publications in leading data-management, artificial-intelligence, and software-engineering venues.

Potential target venues include SIGMOD, VLDB, ICDE, NeurIPS, ACL, EMNLP, ICSE, ASE.

Candidate profile

Applicants should hold, or be close to completing, a PhD in Computer Science, Artificial Intelligence, Data Science, Software Engineering, or a related field. Candidates should have expertise in at least one of the following areas:

- database systems and query optimization;
- large language models and natural language processing;
- data integration and information extraction;
- program synthesis or software engineering automation.

Strong programming skills and experience developing and evaluating research prototypes are required. Experience with SQL engines, LLM agents, code-generation systems, model serving, or automated testing would be an advantage.

Application

Candidates should send the following material to:

Prof. Paolo Papotti - paolo.papotti@eurecom.fr

- curriculum vitae and publication list;
- motivation letter;
- contact information for at least one reference;
- links to representative publications, software, or research artifacts.

References

- [1] G. Sanmartino, M. Urban, P. Papotti, and C. Binnig. *The Stretto Execution Engine for LLM-Augmented Data Systems*. 2026. <https://arxiv.org/abs/2602.04430>
- [2] D. Satriani, E. Veltri, D. Santoro, S. Rosato, S. Varriale, and P. Papotti. *Logical and Physical Optimizations for SQL Query Execution over Large Language Models*. Proceedings of the ACM on Management of Data, 2025.
- [3] M. Saeed, N. De Cao, and P. Papotti. *Querying Large Language Models with SQL*. EDBT, 2024.
- [4] Liana Patel, Siddharth Jha, Melissa Z. Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, Matei Zaharia: *Semantic Operators and Their Optimization: Towards AI-Based Data Analytics with Accuracy Guarantees*. Proc. VLDB Endow. 18(11): 4171-4184 (2025)