Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (SD TIB PAN)

SUBJECT:
Methods and Frameworks for the Multi-Dimensional Evaluation of Large Language Model-Based Systems

SUPERVISOR:
Prof. Gabriella Pasi (main supervisor);
University Milano-Bicocca, Milan, Italy
gabriella.pasi@unimib.it

Wojciech Kusa, PhD (auxiliary supervisor).
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB),
ul. Kolska 12, Warszawa
wojciech.kusa@nask.pl

DESCRIPTION:
Evaluation has become the main bottleneck in the development of LLM-based search systems. Traditional information retrieval measures such as MAP and nDCG assume a ranked list of documents and a fixed set of relevance judgements. Modern systems instead return free-form answers, generated document identifiers, or the result of a multi-step agentic process involving planning, tool calls and self-reflection. Such outputs must be judged along several dimensions at once, including relevance, faithfulness to sources, grounding, completeness, cost and usefulness to the person asking, and each dimension becomes observable at a different level: the retrieved evidence, the individual claim, the final answer, or the full trajectory. A single end-to-end score collapses all of this, showing that a system has failed without indicating which component caused the failure and offering no way to detect gradual degradation as models, corpora and prompts change.
Machine translation has shown that trained neural metrics such as COMET correlate with human judgement far better than surface-overlap measures. This project will adapt that approach to retrieval, designing input representations that combine the user query, the generated response, and the retrieved passages or identifiers, with architectures that incorporate entailment modules and explicit grounding signals. Training such metrics requires data, so the project will create a benchmark of generative retrieval outputs with human judgements of faithfulness, relevance and utility. The resulting metrics will be compared against traditional measures, embedding-based proxies, and current LLM-as-judge and nugget-based protocols, with attention to correlation with human assessment, evaluation cost, and the biases that judge models introduce.
Agentic search pipelines will be decomposed into modular components such as retriever, re-ranker, planner, tool caller and verifier, each with its own inputs, outputs and failure modes. Automated workflow optimisation methods, for example the search-based approach of AFlow or the hierarchical tuning of Cognify, will be used to generate targeted challenge suites that stress individual components. These suites will feed a live evaluation toolkit that tracks component performance over time and flags drift, regressions and newly emerging errors. Assessment will combine established IR metrics with agent-centric measures such as

tool-usage efficiency, number of steps to convergence and cost per completed task, and will draw on recent work on automated failure attribution in multi-agent systems.

The two lines of work are complementary. The learned metrics act as measurement instruments inside the continuous framework, while the traces and failure annotations produced by the framework supply training and validation data for the metrics. Both will be validated prospectively in two high-stakes domains where errors are costly and expert judgement is available: systematic review screening and support for peer review.

REQUIREMENTS:

- MSc degree (or equivalent) in computer science, artificial intelligence, information retrieval, computational linguistics or a related field
- Solid programming skills in Python and willingness to work with machine learning and NLP toolkits
- Experience in NLP, information retrieval or machine learning
- Interest in the evaluation of AI systems and in how agents behave in realistic search tasks
- Motivation to work on trustworthy AI, LLM-based systems and agentic workflows
- Advanced level of English (spoken and written)

BIBLIOGRAPHY:

- Es S. et al. (2024) RAGAS: Automated Evaluation of Retrieval Augmented Generation
- Faggioli G. et al. (2023) Perspectives on Large Language Models for Relevance Judgment
- He Z. et al. (2025) Cognify: Supercharging Gen-AI Workflows with Hierarchical Autotuning
- Pradeep R. et al. (2025) The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models
- Rei R. et al. (2020) COMET: A Neural Framework for MT Evaluation
- Saad-Falcon J. et al. (2024) ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems
- Thakur N. et al. (2021) BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models
- Thakur N. et al. (2025) Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges
- Yao S. et al. (2024) τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains
- Zhang J. et al. (2024) AFlow: Automating Agentic Workflow Generation
- Zhang S. et al. (2025) Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems
- Zheng L. et al. (2023) Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena