**Master’s M2 internship (ARMADA)**

**Provenance- Enabled LLMs LLMs**

**Co-advisors:** Silviu Maniu (Univ. Grenoble Alpes), Noha Ibrahim (Grenoble Polytechnic Institute)

**LIG - Work context**

LIG is a 500-member laboratory with teaching faculty, full-time researchers, PhD students, administrative and technical staff. The mission of LIG is to contribute to the development of fundamental aspects of Computer Science (models, languages, methodologies, algorithms) and address conceptual, technological, and societal challenges. The 24 research teams in LIG aim to increase diversity and dynamism of data, services, interaction devices, and use cases influence the evolution of software and systems to guarantee the essential properties such as reliability, performance, autonomy, and adaptability. Research within LIG is organized into 5 focus areas: Intelligent Systems for Bridging Data, Knowledge and Humans, Software and Information System Engineering, Formal Methods, Models, and Languages, Interactive and Cognitive Systems, Distributed Systems, Parallel Computing, and Networks.

**European ITN ARMADA**

ARMADA us a doctoral network aims at training 15 versatile and interconnected Early Stage Researchers (ESRs) to specialize in the overarching area of Conversational Artificial Intelligence (Conversational AI) and the challenges associated to the recent advances in developing Large Language Models (LLMs), such as ChatGPT and Bard. These specialists will acquire unique knowledge and skills in Artificial Intelligence, Natural Language Processing, Machine Learning, Data Management, and Algorithms Design to improve the reliability of LLMs. A reliable LLM will produce timely, consistent, and verifiable answers, and provide guidance to the user. Due to the highly interdisciplinary aspect, the proposed program will ensure a number of training activities targeted to hone the skills of the trainees. The network provides research training with summer and winter schools on the multidisciplinary aspects of the topic, as well as workshops and courses to foster non-technical social and interpersonal skills, such as scientific writing, innovation, supervision, and management. This program tackles the crucial EU needs for regulating AI by offering to train experts in the area of Conversational AI that will potentially advise EU bodies on technical matters related to the adoption of these technologies in critical disciplines, such as medicine, education, and business intelligence**.** The 8 organizations distributed in 7 countries will form an interoperability platform to share knowledge and skills.

**Scientific context**

Conversational AI systems are Large Language Models (LLMs) that use Transformer Neural Networks. These models are trained on a large amount of text data collected from the web using supercomputers over several days. To give an idea, PaLM, an LLM model by Google, has 540 billion parameters and requires more than a month of training on a specialized computer cluster. The rapid adoption of LLMs has outpaced the development of techniques for evaluating their output quality. This oversight is crucial because LLMs have been shown to be prone to producing what is known as "hallucinations", plausible responses that nonetheless are factually incorrect or inconsistent with the user intent. Therefore relying on LLMs without proper assessment may have severe consequences. Ensuring the quality of LLM output is essential for leveraging the transformative power of these models while mitigating potential risks. By developing robust validation methodologies and incorporating quality-control measures, businesses can harness the benefits of LLMs while safeguarding their decision-making.

**Context -** Large Language Models (LLM) are becoming more and more used for text and image generation, auto-complete assistants for code and text, and planning tasks. While impressive, they are yet to fulfill their potential, mainly for two reasons: they tent to \_hallucinate\_, i.e., invent facts that do not exist, and are not fully capable of \_tracking their reasoning\_, especially in long conversational threads.

To alleviate this, research has focused on adding facts to LLM prompts, thus allowing the models to use actual facts rather than generated ones; this is called Retrieval Augmented Generation (RAG) [1]. In parallel, many approaches to adding tools for explaining outputs of LLMs have been discussed, mostly by borrowing algorithms from research on machine learning explainability [2].

**Objectives**

The overall objective of this internship is to \*\*investigate a complementary approach to explaining and correcting the output of LLMs\*\*, that of \_data provenance\_, inspired from research in explaining the execution of database queries [3]. Provenance operates by adding data annotations to the inputs, along with propagation rules (usually applying mathematical logic). Provenance on top of data retrieved by RAG algorithms has the potential to allow LLMs to explain their output but also \_how\_ they obtained it.

To complete this objective, the following tasks need to be achieved:

1. study the related work on RAG and prompting strategies and data provenance and how they can be used together
2. investigate how to add rules for provenance propagation to RAG-enabled LLMs; for instance, by taking open-source LLMs models such as llama [4] and adding Wikipedia facts to them via RAG libraries [5] along with provenance rules
3. evaluate quantitatively the capability of the LLM to track its outputs to the original data sources via provenance, one-shot and few-shot prompting and compare it against LLMs without such capabilities

**Contact** : "silviu maniu" <silviu.maniu@univ-grenoble-alpes.fr>; "Noha IBRAHIM" <noha.ibrahim@univ-grenoble-alpes.fr>;

**References**

[1]: Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” In \_Proceedings of the 34th International Conference on Neural Information Processing Systems\_, 9459–74. NIPS ’20, 2020.

[2]: Singh, Chandan, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. “Interpretability in the Era of Large Language Models: Opportunities and Challenges.” \_Transactions on Machine Learning Research\_, May 19, 2024.

[3]: Glavic, Boris. “Data Provenance.” \_Foundations and Trends® in Databases\_ 9, no. 3–4 (April 27, 2021): 209–441.

[4]: https://www.llama.com/

[5]: https://huggingface.co/docs/transformers/model\_doc/rag