**Master’s M2 internship (ARMADA)**

**Evaluating Question/Answer Generation with LLMs**

**Co-advisors:** Sihem Amer-Yahia (CNRS, 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.

**Evaluating Question/Answer Generation with LLMs**

**Co-advisors:** Sihem Amer-Yahia (CNRS, Univ. Grenoble Alpes), Noha Ibrahim (Grenoble Polytechnic Institute)

With the rise of large language models (LLM) and the growing availability of large volumes of data, the promise of a future where generative AI tools would enable decision-making on a personal scale is becoming a reality today. It is already possible to ask ChatGPT or LoRA to help us write emails, to initiate a reflection, but also to summarize scientific articles and extract quizzes from them. The purpose of this internship is to develop an LLM-powered approach to generate questions/answers from educational material [2] and a principled evaluation methodology that makes use of LLMs to validate the results.

Multiple-Choice Questions (MCQs) have long been a cornerstone of medical education, providing a standardized means of assessing student knowledge and comprehension. The Uness platform (https://entrainement-ecn.uness.fr/) exemplifies this approach, offering a vast repository of MCQs for medical students to enhance their learning and self-assessment capabilities. However, the creation of high-quality MCQs remains a time-consuming and labor-intensive process for medical educators. This internship will be specifically focused on evaluating the quality of MCQs. It will address the following questions: To what extent do LLM evaluations of existing questions correlate with student ratings? Is it possible to integrate human feedback into the LLM-driven production loop for optimal results?

Given the broad applications of LLMs and the challenges in aligning them with human expectations, researchers have explored using LLMs as validators of other LLM agents’ outputs. Notably, Shankar et al. [5] developed a platform for validating LLM outputs based on LLM-generated criteria and assertions, which are validated based on alignment with human classification of outputs. Several other papers [1,3,4] proposed evaluation metrics based on LLMs, and claimed good correlation with human evaluations, most of the time better than alternative automatic metrics. Evaluation criteria vary across tasks. In [6], both quantitative and qualitative criteria are defined to assess LLM reasoning abilities.

Our real world scenario is based on LISA sheets available to medical students in France on the UNESS platform. UNESS provides access to course booklets and training banks available for free to all medical students in France. It is used by students to practice for various exams, with the option to generate online tests in their chosen medical specialties. For these training sessions, 29,952 isolated questions, 4,757 progressive cases, and 158 critical article readings are available. The platform boasts a community of 60,133 users, with 1,992 being active in the last 7 days. This internship will help teachers enrich this database with new (question,answer) pairs. The focus will be on evaluation methodologies.

**This internship may lead to a PhD in the context of ARMADA.**

**Contact**: "SIHEM AMER YAHIA" <sihem.amer-yahia@univ-grenoble-alpes.fr>;

"Noha IBRAHIM" <noha.ibrahim@univ-grenoble-alpes.fr>;

**References**

[1]  Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. CoRR, abs/2304.00723, 2023.

[2] Guinet G., Omidvar-Tehrani B., Deoras A., Callot L. (2024). Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. [arXiv:2405.13622](https://arxiv.org/abs/2405.13622)

[3]  David Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 15607–15631. Association for Computational Linguistics, 2023.

[4] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Process- ing, EMNLP 2023, Singapore, December 6-10, 2023, pages 2511–2522. Association for Computational Linguistics, 2023.

[5] Shreya Shankar, J. D. Zamfirescu-Pereira, Bjorn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of LLM outputs with human preferences. CoRR, abs/2404.12272, 2024.

[6] Liévin V., Hother C. E., Motzfeldt A. G., Winther O. (2022). Can LLMs Reason about Medical Questions? [arXiv:2207.08143](https://arxiv.org/abs/2207.08143)