Job Offer: PhD Explainable AI for the Correction of Scientific Facts in Large Language Models.

**Starting date: October 1st, 2025 (flexible)**

**Application deadline:** From now until the position is filled

**Interviews (tentative):**  End of June and latter if the position is still open

**Salary:** ~2300€ gross/month (social security included)

**Mission:** research oriented (teaching possible but not mandatory)

**Place of work (no remote):** Laboratoire d'Informatique de Grenoble, CNRS, Grenoble, France
Keywords: natural language processing, explainability, interpretability, large language models, controllability.

 **DESCRIPTION**

Conversational AI systems, built on large-scale language models (LLMs), are trained on vast amounts of web-scraped text data using supercomputers. Despite their rapid adoption, LLMs are prone to "hallucinations"—plausible yet factually incorrect responses—and are known to absorb and even amplify biases present in their training data. This makes them particularly susceptible to propagating erroneous or ideological information, a risk further amplified by the fact that their training data is rarely disclosed. Hence, a significant concern is the impact of training corpus contamination on LLM outputs and internal representations. This issue is especially relevant given the increasing mediation of scientific, particularly medical, knowledge by such tools. For example, a well-know system for generating scientific literature review, which performs Retrieval Augmented Generation (RAG) on scientific literature, when queried with "bosom malignancy" (a "tortured phrase" for breast cancer in malicious scientific literature [Cabanac, 2021]), responded with associated content as if "bosom malignancy" was a valid query. While model association of  "bosom malignancy" to "breast cancer," might be semantically accurate it is scientifically wrong in a formal context. Furthermore, scientific publications can be retracted; however, such publications may continue to be cited long after their retraction [Neupane, 2019; Hsiao 2022]. A recent study by Graña Possamai et al. (2025) showed that retracted articles used in 18 meta-analyses affected the statistical significance of the results.

Addressing these challenges requires methods to explain the models behavior. Some explainable AI (XAI) approaches employs attribution methods, which are broadly categorized into feature and instance attribution. Feature attribution seeks to explain model predictions by identifying the influential features within an input (e.g. SHAP [Scott, 2017]) while instance attribution explains model predictions at the instance level (e.g. Influence functions [Koh, 2017]). While several approaches have been proposed to trace the output generated by large language models (LLMs) back to their originating training examples [Akyurek, 2022; Chang, 2024], classical, model-agnostic retrieval methods like BM25 still outperform them in finding passages that explicitly contain relevant facts.

This Ph.D. thesis aims to investigate methods for tracing the specific training data that LLMs rely on during inference, especially in case of Retrieval Augmented Generation (RAG) in scenarios involving corpus contamination. Additionally, the project will investigate the feasibility of editing models with correct scientific information without requiring a full retraining process. While methods like ROME with Causal Tracing [Meng, 2022] have shown promise in this direction, recent work suggests that simply locating facts within a neural network may not provide sufficient information for effective factual editing [Hase, 2023]. As a Ph.D. student on this project, you'll be expected to develop a strong understanding of explainable AI objectives and techniques [Guidotti, 2018], particularly focusing on data attribution methods.

You will have the opportunity to work on cutting-edge research projects in NLP, contributing to the development of more reliable and interpretable LLMs. It is important to note that the Ph.D. research project should be aligned with your interests and expertise. Therefore, the precise direction of the research can and will be influenced by the personal taste and research goals of the PhD candidate. It is encouraged that you bring your unique perspective and ideas to the table.

**SKILLS**

* Master degree in Natural Language Processing, computer science or data science.
* Mastering Python programming and deep learning frameworks.
* Very good communication skills in English, (French not needed).

**SCIENTIFIC ENVIRONMENT**

The thesis will be supported by the Artificial Intelligence & Language Chair of the MIAI Grenoble Alpes Institute, a national AI excellence center and the ERC (European Research Council) Nanobubbles project. It will be conducted within the Sigma and GETALP teams of the LIG laboratory (<https://lig-getalp.imag.fr/>). The GETALP team has a strong expertise and track record in Natural Language Processing while the sigma team is pioneer in detecting false scientific publications. The recruited person will be welcomed within the team which offers a dynamic, international, and stimulating framework for conducting high-level multi-disciplinary research. The GETALP team is housed in a modern building (IMAG) located on a 175-hectare landscaped campus that was ranked as the eighth most beautiful campus in Europe by Times Higher Education magazine in 2018. The PhD position will be co-supervised by Cyril Labbé (Université Grenoble Alpes), Benoit Favre (Université Aix-Marseille) and François Portet (Université Grenoble Alpes). Moreover, two other PhD positions are open in this project.  The students, along with the partners will closely collaborate. The means to carry out the PhD will be provided both in terms of missions in France and abroad and in terms of equipment. The candidate will have access to the cluster of GPUs of the LIG. Furthermore, access to the national supercomputer Jean-Zay will enable to run large scale experiments.

**INSTRUCTIONS FOR APPLYING**

Applications must contain: CV + letter/message of motivation + master notes (if applicable) + be ready to provide letter(s) of recommendation; and be addressed to Cyril Labbé (cyril.labbe@imag.fr), Benoit Favre (benoit.favre@lis-lab.fr) and François Portet (francois.Portet@imag.fr). We celebrate diversity and are committed to creating an inclusive environment for all employees.

**REFERENCES**

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards Tracing Knowledge in Language Models Back to the Training Data. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chang, T. A., Rajagopal, D., Bolukbasi, T., Dixon, L., & Tenney, I.  (2024). Scalable Influence and Fact Tracing for Large Language Model  Pretraining. arXiv preprint arXiv:2410.17413.

Cabanac, G., Labbé, C., & Magazinov, A. (2021). Tortured phrases: A  dubious writing style emerging in science. Evidence of critical issues  affecting established journals. arXiv preprint arXiv:2107.06751.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv., 51(5), 2018.

Hase, P., Bansal, M., Kim, B., & Ghandeharioun, A. (2023). Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. Advances in Neural Information Processing Systems, 36, 17643-17668.

Hsiao TK, Schneider J. Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. Quant Sci Stud. 2022 Feb 4;2(4):1144-1169.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. Advances in neural information processing systems, 35, 17359-17372.

**J**ayanti Bhandari Neupane, Ram P. Neupane, Yuheng Luo, Wesley Y. Yoshida, Rui Sun, and Philip G. Williams. Characterization of Leptazolines A–D, Polar Oxazolines from the Cyanobacterium Leptolyngbya sp., Reveals a Glitch with the “Willoughby–Hoye” Scripts for Calculating NMR Chemical Shifts. Organic Letters 2019 21 (20), 8449-8453

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1885–1894. PMLR, 06–11 Aug 2017.

Graña Possamai C, Cabanac G, Perrodeau E, Ghosn L, Ravaud P, Boutron I. Inclusion of Retracted Studies in Systematic Reviews and Meta-Analyses of Interventions: A Systematic Review and Meta-Analysis. JAMA Intern Med. 2025 Mar 31:e250256. doi: 10.1001/jamainternmed.2025.0256. Epub ahead of print. PMID: 40163084; PMCID: PMC11959482.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.