3 year PhD position

PhD Thesis: Privacy-Enhancing Tools for Content Sanitization Using Large Language Models – Application to School Bullying and Harassment –

Host. Petrus/Petscraft project-team, Inria Saclay centre at Université Paris-Saclay, Turing building in Palaiseau (near Paris), France.

Supervisors. Nicolas Anciaux (nicolas.anciaux@inria.fr), Adrien Boiret (adrien.boiret@insa-cvl.fr) & Cédric Eichler (cedric.eichler@insa-cvl.fr)

Application. Please contact us if you are interested or need additional details. Application can be sent by email with CV and cover letter.

Objectives of the thesis. The advanced inference capabilities of Large Language Models (LLMs) pose a significant threat to the privacy of individuals by enabling third parties to accurately infer certain personal attributes from their writings [1, 2]. Paradoxically, LLMs can also be used to protect individuals by helping them to modify their textual output from certain unwanted inferences [3, 4], opening the way to new tools. The ultimate objective of this thesis is to work towards an interactive chatbot-like tool for the sanitisation of text, to address applications including two which are especially investigated by our team: production of testimonies in the context of school bullying and work harassment, and participants feedback in participatory platforms. Through a preliminary investigation, we identified guidelines and main difficulties the successful PhD candidate will have to address for the sound development of such a tool:

- A realistic adversary should be used to assess (residual) privacy risks. This poses two main challenges. Firstly, a realistic attacker cannot be generic but must *take into account the vast auxiliary knowledge an attacker may possess* (e.g. through fine-tuning or with the help of a dedicated ontology). Secondly, LLMs tend to always propose a guess which could be as likely as a random guess. Therefore, there is a need for a mechanism to *estimate the likelihood of inferences*.
- Designing and implementing a metric assessing the utility of a text (or the loss of utility due to sanitisation) is no trivial task. Design-wise, a proper metric should evaluate the amount of *information conveyed by a text relevant to its purpose* (e.g. wrt testimonies, whether the victim/perpetrator are identifiable, etc). With regard to implementation, the assessment must be done *automatically without human intervention* (e.g. through a LLM).
- Finally, an LLM-based sanitisation process must be proposed, limiting the capacity of the attacker to make inferences while maintaining the utility of the text. In a chatbot-like application, this process can be iterative and interactive.

Initial roadmap. The PhD project will start by the installation of open source LLMs such as Mistral or Arctic, and the implementation of the guidelines above, before focusing on the specialisation of the anonymisation solution to adapt it to different use cases and datasets.

Potential use-cases. We will focus on two use cases: (1) the anonymous declaration or anonymisation of certain concepts in the context of school, university and work in general. This first use cases will be built with Inria's partners in the context of the services responsible for investigating harassment cases that deal with anonymous witness statements and/or in the context of the labour market and job searches. (2) a second use-case is user feedback in participative platforms aimed at wellbeing, nutrition and health. This use case is still emerging and will be detailed during the PhD project.

Context. This PhD thesis project is part of the French Priority Research Program and Equipment (PEPR) on Cybersecurity, interdisciplinary Project on Privacy (iPoP) project involving several French research teams working on data protection, from Inria, universities, engineering schools and the CNIL (French National Commission on Information Technology and Civil Liberties). The PhD is proposed by Petrus project-team at Inria Saclay and the PETSCRAFT project-team joint between Inria Saclay and INSA CVL, which tightly collaborate in this large initiative on modeling privacy protection concepts and on the design and deployment of *explicable* and *efficient* Privacy-Enhancing Technologies (PETs).

Profile and appreciated skills. Candidates must hold a master (or equivalent) in Computer Science. The following skills are appreciated:

- Basic knowledge in LLMs/ML.
- Basic knowledge in privacy & anonimization.
- Proficiency in programming.

Salary & Benefits.

- Gross monthly salary of 2100 euros
- Possibility of teleworking and flexible organization of working hours
- Social, cultural and sports events and activities
- Access to vocational training

References

[1] Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C., Xu, Z.: User inference attacks on llms. In: Socially Responsible Language Modelling Research (2023)

[2] Staab, R., Vero, M., Balunović, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. arXiv preprint arXiv:2310.07298 (2023)

[3] Staab, R., Vero, M., Balunović, M., Vechev, M.: Large language models are advanced anonymizers. arXiv preprint arXiv:2402.13846 (2024)

[4] Tannier, X., Wajsbürt, P., Calliger, A., Dura, B., Mouchet, A., Hilka, M., Bey, R.: Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. Methods of Information in Medicine (2024)