Offer #2024-07808

PhD Position F/M PhD Thesis: Privacy-Enhancing Tools for Content Sanitization Using Large Language Models – Application to School Bullying and Participatory Testimony Collection

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

The Inria Saclay-Île-de-France Research Centre was established in 2008. It has developed as part of the Saclay site in partnership with **Paris-Saclay University** and with the **Institut Polytechnique de Paris** .

The centre has 39 project teams , 27 of which operate jointly with Paris-Saclay University and the Institut Polytechnique de Paris; Its activities occupy over 600 people, scientists and research and innovation support staff, including 44 different nationalities.

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 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).

Assignment

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 modify their textual output to avoid 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 sanitization of text, to address applications including two which are especially investigated by our team: production of testimonies in the context of school bullying and workplace harassment, and participant 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 sanitization) 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., regarding testimonies, whether the victim/perpetrator are identifiable, etc.). With regard to implementation, the assessment must be done automatically without human intervention (e.g., through an LLM).

- Finally, an LLM-based sanitization 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.

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., Balunovi´c, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. arXiv preprint arXiv:2310.07298 (2023)
[3] Staab, R., Vero, M., Balunovi´c, M., Vechev, M.: Large language models are advanced anonymizers.

arXiv preprint arXiv:2402.13846 (2024)
[4] Tannier, X., Wajsb¨urt, 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)

Main activities

Initial roadmap. The PhD project will start with the analysis of the above-mentioned difficulties, the review of emerging state-of-the-art articles on the subject, and the installation of open-source LLMs such as Mistral or Arctic. The targeted solution should be generic before focusing on the specialization of the anonymization 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 anonymization of certain concepts in the context of school, university, and professional settings in general. This first use case will be developed with Inria's partners within the services responsible for investigating harassment cases that handle anonymous witness statements and/or in the context of the labor market and job searches. (2) A second use case is user feedback on participatory platforms focused on well-being, nutrition, and health. This use case is still emerging and will be detailed during the PhD project.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

1ère et 2ème année :2.082 euros brut

3ème année : 2.190 euros brut

General Information

- **Theme/Domain :** Security and Confidentiality
- **Town/city :** Palaiseau
- **Inria Center :** Centre Inria de Saclay
- **Starting date :** 2024-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-09-30

Contacts

- **Inria Team :** PETSCRAFT
- **PhD Supervisor :**
  Anciaux Nicolas / Nicolas.Anciaux@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.