Offre n°2024-07808

Doctorant F/H Thèse de doctorat : Outils de protection de la vie privée pour la désinfection de contenu à l'aide de grands modèles de langage – Application au harcèlement scolaire et à la collecte participative

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle

Le centre de recherche Inria de Saclay a été créé en 2008. Sa dynamique s'inscrit dans le développement du plateau de Saclay, en partenariat étroit d'une part avec le pôle de l'**Université Paris-Saclay** et d'autre part avec le pôle de l'**Institut Polytechnique de Paris** . Afin de construire une politique de site ambitieuse, le centre Inria de Saclay a signé en 2021 des accords stratégiques avec ces deux partenaires territoriaux privilégiés.

Le centre compte **39 équipes-projets** , dont 27 sont communes avec l'Université Paris-Saclay ou l'Institut Polytechnique de Paris. Son action mobilise **plus de 600 personnes** , scientifiques et personnels d'appui à la recherche et à l'innovation, issues de 54 nationalités.

Le centre Inria Saclay - Île-de-France est un acteur essentiel de la recherche en sciences du numérique sur le plateau de Saclay. Il porte les valeurs et les projets qui font l'originalité d'Inria dans le paysage de la recherche : l'excellence scientifique, le transfert technologique, les partenariats pluridisciplinaires avec des établissements aux compétences complémentaires aux nôtres, afin de maximiser l'impact scientifique, économique et sociétal d'Inria.

Contexte et atouts du poste

Ce projet de thèse de doctorat s'inscrit dans le cadre du Programme et Équipements Prioritaires de Recherche (PEPR) français sur la Cybersécurité, projet interdisciplinaire sur la vie privée (iPoP), impliquant plusieurs équipes de recherche françaises travaillant sur la protection des données, provenant d'Inria, d'universités, d'écoles d'ingénieurs et de la CNIL (Commission Nationale de l'Informatique et des Libertés). La thèse est proposée par l'équipe-projet PETSCRAFT, conjointe entre Inria Saclay et l'INSA CVL, qui collaborent étroitement dans cette grande initiative sur la modélisation des concepts de protection de la vie privée et sur la conception et le déploiement de technologies de protection de la vie privée (PETs) explicables et efficaces.

Avantages:

Mission confiée

Objectifs de la thèse. Les capacités avancées d'inférence des grands modèles de langage (LLMs) posent une menace significative pour la vie privée des individus en permettant à des tiers d'inférer avec précision certaines caractéristiques personnelles à partir de leurs écrits [1, 2]. Paradoxalement, les LLMs peuvent également être utilisés pour protéger les individus en les aidant à modifier leur production textuelle pour éviter certaines inférences indésirables [3, 4], ouvrant ainsi la voie à de nouveaux outils. L'objectif ultime de cette thèse est de travailler à la mise au point d'un outil interactif de type chatbot pour la désinfection de texte, afin de répondre à des applications incluant deux qui sont particulièrement étudiées par notre équipe : la production de témoignages dans le contexte de l'intimidation scolaire et du harcèlement au travail, et les retours des participants sur des plateformes participatives. Certaines difficultés devront être abordées pour la conception et le développement de l'outil envisagé, comme par exemple:

- Un adversaire réaliste doit être utilisé pour évaluer les risques de confidentialité (résiduels). Cela pose deux défis principaux. Premièrement, un attaquant réaliste ne peut pas être générique, mais doit prendre en compte les vastes connaissances auxiliaires qu'un attaquant peut posséder (par exemple, via un ajustement fin ou avec l'aide d'une ontologie dédiée). Deuxièmement, les LLMs ont tendance à toujours proposer une supposition qui pourrait être aussi probable qu'une supposition aléatoire. Par conséquent, il est nécessaire de disposer d'un mécanisme pour estimer la probabilité des inférences.

- Concevoir et mettre en œuvre une métrique évaluant l'utilité d'un texte (ou la perte d'utilité due à la désinfection) n'est pas une tâche triviale. En termes de conception, une métrique appropriée doit

évaluer la quantité d'informations transmises par un texte pertinent par rapport à son objectif (par exemple, par rapport aux témoignages, si la victime/l'agresseur sont identifiables, etc.). En ce qui concerne la mise en œuvre, l'évaluation doit être effectuée automatiquement sans intervention humaine (par exemple, via un LLM).

- Enfin, un processus de désinfection basé sur les LLMs doit être proposé, limitant la capacité de l'attaquant à faire des inférences tout en maintenant l'utilité du texte. Dans une application de type chatbot, ce processus peut être itératif et interactif.

Feuille de route initiale. Le projet de doctorat commencera par l'analyse des difficultés ci-dessus, la lecture des articles de l'état de l'art qui émergent sur le sujet, ainsi que l'installation de LLMs open source tels que Mistral ou Arctic. La seloution visée devra être générique avant de se concentrer sur la spécialisation de la solution d'anonymisation pour l'adapter à différents cas d'utilisation et ensembles de données.

Cas d'utilisation potentiels. Nous nous concentrerons sur deux cas d'utilisation : (1) la déclaration anonyme ou l'anonymisation de certains concepts dans le contexte scolaire, universitaire et professionnel en général. Ce premier cas d'utilisation sera construit avec les partenaires d'Inria dans le cadre des services responsables de l'enquête sur les cas de harcèlement qui traitent des témoignages anonymes et/ou dans le contexte du marché du travail et des recherches d'emploi. (2) Un deuxième cas d'utilisation est le retour d'expérience des utilisateurs sur des plateformes participatives axées sur le bien-être, la nutrition et la santé. Ce cas d'utilisation est encore en émergence et sera détaillé au cours du projet de thèse.

Bibliographie:
[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)

Principales activités

Feuille de route initiale. Le projet de doctorat commencera par l'analyse des difficultés ci-dessus, la lecture des articles de l'état de l'art qui émergent sur le sujet, ainsi que l'installation de LLMs open source tels que Mistral ou Arctic. La seloution visée devra être générique avant de se concentrer sur la spécialisation de la solution d'anonymisation pour l'adapter à différents cas d'utilisation et ensembles de données.

Cas d'utilisation potentiels. Nous nous concentrerons sur deux cas d'utilisation : (1) la déclaration anonyme ou l'anonymisation de certains concepts dans le contexte scolaire, universitaire et professionnel en général. Ce premier cas d'utilisation sera construit avec les partenaires d'Inria dans le cadre des services responsables de l'enquête sur les cas de harcèlement qui traitent des témoignages anonymes et/ou dans le contexte du marché du travail et des recherches d'emploi. (2) Un deuxième cas d'utilisation est le retour d'expérience des utilisateurs sur des plateformes participatives axées sur le bien-être, la nutrition et la santé. Ce cas d'utilisation est encore en émergence et sera détaillé au cours du projet de thèse.

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail (après 6 mois d'ancienneté) et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

Rémunération

1ère et 2ème année :2.082 euros brut

3ème année : 2.190 euros brut

Informations générales

- **Thème/Domaine :** Sécurité et confidentialité
- **Ville :** Palaiseau
- **Centre Inria :** Centre Inria de Saclay
- **Date de prise de fonction souhaitée :**2024-10-01
- **Durée de contrat :**3 ans
- **Date limite pour postuler :**2024-09-30

Contacts

- **Équipe Inria :** PETSCRAFT
- **Directeur de thèse :**
  Anciaux Nicolas / Nicolas.Anciaux@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.