Thème de recherche	Systèmes d'IA fiables basés sur les graphes de connaissances et les LLM : Attaques et défenses
Poste (H/F)	Doctorant
Référence de l'offre	SN/MÖ/trustworthyAI/032025
Département de Recherche	Sécurité Numérique (SN)
Date de publication	11/03/2025
Date d'embauche	Poste à pourvoir de suite
Durée du contrat	Durée de la thèse

Description

La performance des technologies d'IA repose sur l'accès à de grands ensembles de données de bonne qualité et sur l'entraînement d'un modèle précis. Cette dépendance à l'égard de grandes quantités de données rend les systèmes d'IA vulnérables aux attaques de confidentialité qui peuvent divulguer des informations sensibles à la confidentialité, aux attaques adverses qui peuvent manipuler les entrées ou les paramètres du modèle afin de falsifier le processus d'entraînement, et aux attaques d'équité qui visent à modifier le comportement existant du modèle pour induire un biais. La littérature présente diverses propositions pour atténuer chaque catégorie d'attaques. Néanmoins, ces solutions fonctionnent généralement uniquement avec des réseaux neuronaux et nécessitent parfois des modifications considérables et lourdes en termes de calcul des algorithmes existants. L'objectif de la thèse est d'étudier et de concevoir de nouvelles attaques dans chaque catégorie dédiées aux systèmes d'IA utilisant des graphes de connaissances (KG), de les évaluer dans des contextes centralisés ou distribués et, enfin, de proposer des stratégies de défense dédiées. Plus spécifiquement :

- Les attaques de vie privée et de confidentialité visent à apprendre des informations sur l'ensemble de données, le graphe et/ou le modèle. Parmi les attaques potentielles contre la confidentialité des données, on trouve l'inférence d'appartenance, l'inférence de propriété, la fuite de données et les attaques par fuite rapide, etc. Certaines autres attaques, comme les attaques par fuite de gradient, ciblent le modèle réel et apprennent des informations sur ses gradients, donc ses caractéristiques. L'objectif est d'enquêter sur les attaques existantes contre les systèmes d'IA autonomes ou collaboratifs qui utilisent des KG et/ou des modèles de langage volumineux (LLM). Le candidat étudiera et mettra également en œuvre de nouvelles technologies d'amélioration de la confidentialité (PET) en conséquence. Celles-ci peuvent être basées sur la confidentialité différentielle (DP) ou d'autres méthodes empiriques.
- Les attaques adverses ciblent l'intégrité et l'efficacité des systèmes d'IA. Le candidat enquêtera sur les attaques d'empoisonnement de modèles ou de porte dérobée existantes pendant la formation ou les exemples adverses au moment de l'inférence (en considérant les modèles adverses de type boîte noire, boîte grise, boîte blanche) et en développera de nouvelles adaptées aux KG ou aux LLM. Les attaques potentielles comprennent la perturbation des graphes et les attaques par injection de mots/phrases pour les LLM. L'objectif ultime de la recherche est, une fois de plus, de développer des stratégies de défense qui peuvent utiliser des mécanismes DP et/ou la génération de données synthétiques.
- Les attaques d'équité visent à manipuler les informations pour porter atteinte à l'équité des systèmes d'IA. Le candidat étudiera la transition des attaques par porte dérobée aux attaques contre l'équité du modèle, en supposant différents types d'acteurs (non intentionnels ou malveillants) et les évaluera dans divers contextes, comme par exemple lorsque les données des clients ne sont pas indépendantes ou échantillonnées à partir de la même distribution (c'est-à-dire, un paramètre non IID) ou dans des contextes distribués. Des stratégies d'atténuation seront également explorées sur la base de mécanismes DP ou de solutions empiriques personnalisées pour les graphes de connaissances et/ou les LLM exploitant les interrelations entre l'équité et la confidentialité différentielle.

References

[1] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks against Machine Learning Models," in *IEEE Security and Privacy*, 2017.
[2] N. Carlini, F. Tramèr, E. Wallace, M. Hajielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea and C. Raffel, "Extracting Training Data from Large Language Models," in *Usenix Security Symposium*, 2021.
[3] O. Zari, C. Xu, J. Parra-Arnau, A. Ünsal, M. Önen, "Link inference attacks in vertical federated graph learning", in *ACSAC*, 2024.
[4] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of Machine Learning Research*, 2020.

[5] W. Zou, R. Geng, B. Wang, J. Jia, “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models”, in *Usenix Security Symposium,* 2024.
[6] N. Furth, A. Khreishah, G. Liu, N. Phan and Y. Jararweh, "Unfair Trojan: Targeted Backdoor Attacks Against Model Fairness. In H," *Handbook of Trustworthy Federated Learning. Cham: Springer International Publishing,* pp. 149-168, 2024.
[7] I. Gallegos, R. A. Rossi, Barrow, J., M. Tanjim, S. Kim, F. Demoncourt, T. Z. R. Yu and N. Ahmed, "Bias and Fairness in Large Language Models: A Survey," arXiv:2309.00770, 2024.
[8] F. Fioretto, C. Tran, P.V. Hentenryck, K. Zhu, “Differential Privacy and Fairness in Decisions in Learning Tasks: A Survey”, September 2022, https://arxiv.org/abs/2202.08187

Prérequis

- Niveau d'études / diplôme : MS.c
- Domaine / spécialité : informatique avec une expérience dans les technologies d'amélioration de la confidentialité et l'apprentissage automatique.

Dossier de candidature

Les candidatures doivent être accompagnées de :

- Un curriculum détaillé,
- Relevé(s) de notes,
- Lettre de motivation,
- Le nom et l'adresse d'une personne de référence pour la recommandation.

Les candidatures doivent être envoyées par courrier électronique à l'adresse secretariat@eurecom.fr en indiquant la référence : SN/MÖ/trustworthyAI/032025

A propos d’EURECOM

EURECOM est une grande école d’ingénieurs et un centre de recherche en sciences du numérique fondé en 1991 sous la forme d’un GIE, dans la technopole internationale de Sophia Antipolis. L’Institut Mines-Télécom est membre fondateur du GIE. Les activités d'enseignement et de recherche sont organisées autour de 3 thématiques porteuses : sécurité numérique, systèmes de communication et Data Science.

L’institution accueille 150 salariés, chercheurs et administratifs et 400 étudiants internationaux dans ses locaux situés sur le Campus Sophia Tech, le plus grand campus en sciences et technologies de l’information des Alpes Maritimes. EURECOM bénéficie d’un environnement géographique privilégié sur la Côte d’Azur, entre mer et montagne, au cœur d’un écosystème dynamique et pluridisciplinaire qui encourage l’innovation scientifique et technologique de haut niveau.

Avantages sociaux

- Environnement international et multiculturel
- Salaire attractif - Épargne salariale
- Retraite par capitalisation (100% employeur)
- Accord d’Intéressement
- Mutuelle d’entreprise (contrat familial - hauts niveaux de garanties) - 60% employeur
- Prime annuelle de performance
- Titres-restaurant (60% employeur)

EURECOM fait partie des meilleures écoles d'ingénieurs européennes en sciences des technologies numériques. Elle est située au cœur de la Côte d'Azur, au sein de la Silicon Valley européenne (Tech Park Sophia-Antipolis). Les équipes de recherche d'EURECOM évoluent dans un environnement international et multiculturel.

EURECOM mène une politique dynamique en termes d’inclusion et de qualité de vie au travail. Nous nous engageons pour la diversité et accordons la même considération à toutes les candidatures, sans discrimination. Nous recherchons avant tout la compétence et l’esprit d’équipe.

Tous nos postes sont ouverts aux personnes en situation de handicap. EURECOM est doté d’un référent handicap afin d’accompagner, de conseiller, d’organiser les éventuels aménagements et de prendre des engagements positifs en faveur d’une intégration personnalisée.

EURECOM, dans le cadre de son plan d'égalité femmes/hommes, encourage la mixité dans ses équipes. Notre plan d'action en faveur de cette mixité prévoit que nous encouragions les candidatures masculines pour les postes administratifs, postes traditionnellement occupés par des femmes, et les candidatures féminines dans les postes en informatique et recherche, postes traditionnellement occupés par des hommes.

EURECOM mène des actions positives dans le cadre de sa politique RSE. Un référent RSE pilote la politique d'EURECOM en matière de RSE et de transition énergétique (bornes de recharge électrique, panneaux solaires, tri sélectif…).

Site web EURECOM : https://www.eurecom.fr/fr/eurecom/presentation
EURECOM en VIDEO : https://www.youtube.com/watch?v=uIlFcgNijnM
Expériences collaborateurs :
https://www.youtube.com/watch?v=BHv9zIduzuQ
https://www.youtube.com/watch?v=hvbzzCBups8