Graduate School of Information Science, University of Hyogo

14th International Research Seminar

PRIVACY IN FINE-TUNING AND PROMPTING FOR LARGE LANGUAGE MODELS: ATTACKS, DEFENSES, AND FUTURE DIRECTIONS

Wed. 23 July 2025 (13:00 ~ 14:00) JST

IN-PERSON/ONLINE SEMINAR

Fine-tuning and Prompting have emerged as a critical process in leveraging Large Language Models (LLMs) for specific downstream tasks, enabling these models to achieve state-of-the-art performance across various domains. However, the fine-tuning and prompting process often involves sensitive datasets, introducing privacy risks that exploit the unique characteristics of this stage. In this tutorial, I will provide a comprehensive view of privacy challenges associated with fine-tuning and prompting LLMs, highlighting vulnerabilities to various privacy attacks, including membership inference, data extraction, and backdoor attacks. We further review defense mechanisms designed to mitigate privacy risks, such as differential privacy, federated learning, and knowledge unlearning, discussing their effectiveness and limitations in addressing privacy risks and maintaining model utility. By identifying key gaps in existing research, we highlight challenges and propose directions to advance the development of privacy-preserving methods for leveraging LLMs, promoting their responsible use in diverse applications.

Guest Speaker

Yang Cao

Associate Professor

Institute of Science Tokyo, Japan

Institute of SCIENCE TOKYO

Yang Cao is an Associate Professor at the Department of Computer Science, Institute of Science Tokyo (Science Tokyo, formerly Tokyo Tech), and directs the Trustworthy Data Science and AI (TDSAI) Lab. He is passionate about studying and teaching on algorithmic trustworthiness in data science and AI. Two of his papers on data privacy were selected as best paper finalists in top-tier conferences IEEE ICDE 2017 and ICME 2020. He received the IEEE Computer Society Japan Chapter Young Author Award 2019, Database Society of Japan Kambayashi Young Researcher Award 2021. His research projects were/are supported by JSPS, JST, MSRA, KDDI, LINE, WeBank, etc.

Register here (free)

https://shorturl.at/ZC5u2

Contact: rashed@gsis.u-hyogo.ac.jp

Kobe Campus for Information Science, Computational Science Center Building, Large Lecture Hall (720), 7th Floor

https://www.u-hyogo.ac.jp/about/access/

For more details:

https://yangcao888.github.io/