Frontier of Artificial Network

Department of Data Science
香港城市大學
City University of Hong Kong

Differential Privacy in LLM Fine-Tuning: What It Protects, What It Costs, and What It Doesn't

Invited Speaker
Yang Cao
Institute of Science Tokyo

Date: Mar 24, 2026 (Tue)
Time: 9:30am (HKT)
Zoom Meeting: 801 137 0362

Biography

Yang Cao is an Associate Professor at the Department of Computer Science, Institute of Science Tokyo (Science Tokyo, formerly Tokyo Tech), and directing 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 was a recipient of 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.

Abstract

Large language models are often fine-tuned on small and sensitive datasets, where individual training examples can strongly influence model behavior and lead to memorization and privacy leakage. This talk focuses on differential privacy in LLM fine-tuning and explains how DP acts as a learning constraint that limits the influence of individual samples, thereby mitigating such risks. We introduce the core intuition behind DP without assuming prior background, review practical DP fine-tuning mechanisms, and discuss how privacy and utility should be evaluated together. We also clarify what DP fine-tuning can and cannot protect, and outline open challenges in privacy-aware LLM adaptation.