Research Statement

This is a brief statement about my research experiences, research interest, and future research plans.

Research Experience

Since the summer of 2023, I have been interning in Prof. Hongxin Wei's machine learning group at our department (Department of Statistics and Data Science). My primary research experience focuses on ML privacy and large language models (LLMs), specifically on membership inference attacks (MIA) and in-context learning for LLMs.

Under the guidance of Professor Hongxin Wei, I have been involved in three research projects. The first focused on the vulnerability differences of data under membership inference attacks, with an attempt to mitigate these disparities, which could be identified as outliers in terms of features. Although this work did not lead to a published paper due to various reasons, including the immature MIA settings, it greatly enhanced my understanding of MIA at the data level. The second project was related to content risk control for large language models (LLMs) using top-k in-context learning, with the goal of establishing a benchmark for LLM risk control in domestic contexts. Most recently, I proposed PAST, a method for defending against membership inference attacks through adaptive sparsification, which has been submitted to ICLR 2025.

I also deeply appreciate Prof. Guanhua Chen, whose courses on NLP and Spark (with NLP being a graduate-level course) provided me with a profound understanding of various LLM tasks and sparked my interest in areas such as distributed training, efficient fine-tuning, model quantization, retrieval-augmented generation (RAG).

Research Interest

I'm generally interested in machine learning and NLP. Currently, my focus is on Trustworthy and Efficient AI, as well as various NLP tasks (e.g., in-context learning). Other intriguing applications of AI may also capture my interest.

Future Research Plans

For the near future, I have some tentative plans focusing mainly on privacy and LLMs and very likely combined with Efficient AI:

1. Is it possible to build upon existing sparse methods in transformers and consider the parameter significance for privacy defense, in order to achieve trustworthy and efficient NLP?
2. Given the efficiency of the pre-training and fine-tuning paradigm for LLMs, where a few epochs can yield excellent results on new datasets, more privacy might be leaked. I plan to investigate whether the data used for fine-tuning poses significant privacy risks and how to mitigate them.
3. As previously mentioned, the current settings for MIA are quite disorganized. A recent study summarized the errors as assessing privacy at an aggregate level using weak inference attacks. In the corrected setting, heuristic defense methods fail to achieve the level of differential privacy. Thus, building on this, we can explore robust defenses based on the nature of robustness, as the vulnerability assessment of individual samples is strongly correlated with robustness.