Zhiqi Wang (Eli)

Email: eliwang2332@outlook.com
Linkedin: https://www.linkedin.com/in/zhiqi-wang-6841b4251/
Mobile: +1-518-391-5692
Github: ZhiqiEliWang
homepage: https://zhiqiwang.notion.site/info

Research Interests

- **ML Privacy**: Auditing membership privacy and its connection with memorization.
- **AI Safety**: Exploring vulnerabilities of AI models and systems, defining them, and finding ways to fix them.

Education

- **Pennsylvania State University** PA, United States
  Ph.D. in Informatics *Aug 2025 - Present*
  Advisor: Dr. Yuchen Yang
- **Rensselaer Polytechnic Institute** NY, United States
  Bachelor of Computer Science *Sep 2021 - Dec 2024*
  Courses: Machine Learning from Data, Machine Learning Seminar, ML and Optimization, Rensselaer Center for Open Source, Operating Systems, Intro to Algorithm, Principle of Software, Data Structures, Parallel Computing, Linear Algebra etc.

  Dean's Honor List every year.

Publications

- **Membership Inference Attacks as Privacy Tools: Reliability, Disparity and Ensemble**: Exploring the performance of MIAs with the perspective of sample-level analysis; First Author; ACM CCS 2025
- **Evaluating the Dynamics of Membership Privacy in Deep Learning**: Tracking per-sample membership vulnerability during model training; In-submission

Academic Experience: Research and Instruction

- **DSPLab (github.com/RPI-DSPlab)** May. 2023 - Dec. 2024
  Advisor: Lei Yu (Assistant Professor) *Undergraduate Research Assistant, Project Lead*
  Key Contributions:
  - **Framework Development**: Designed and implemented a Python framework for systematically testing and comparing various membership inference attacks under consistent experimental conditions.
  - **Attack Implementations**: Developed APIs for seven membership inference attacks, ranging from basic methods like the Loss Threshold Attack to advanced techniques such as LiRA.
  - **Evaluation Tools**: Built evaluation metrics, including TPR@low FPR and Venn diagrams, to assess attack reliability and visualize overlaps in predictions.
  - **Leadership**: Led a team to integrate Membership Inference Attacks against large language models (LLMs) into the framework, extending its capabilities to analyze LLM privacy risks.
- **LLM Safety Research** Sept. 2024 - Present
  Advisor: Jinghuai Zhang (PhD Student), Yuan Tian (Associate Professor) *Collaborator*
  - **Poisoning Attack Against Long Context LLM:**: Explored methods to inject adversarial tokens into long context prompts to increase LLMs' attention on the adversarial instructions.
  - **Identifying Reward Model:**: Given LLMs' output token, we can detect which reward model is being used to guide LLMs' generation.
  - **Enhance Adversarial Attacks on LLMs' Transferability by Leveraging Alignment**: Investigating the connection between alignment and transferability of adversarial prompts.
- **RPI Intro to AI**
  Undergraduate Teaching Assistant *Jan. 2024 - May. 2024*
  - **Responsibilities**: Assisted students with coursework questions, proctored and graded exams, and held office hours.
- **RPI Intro to CS**
  Undergraduate Student Mentor *Sept. 2022 - Dec. 2024*
  - **Responsibilities**: Assisted students with coursework questions, proctored and graded exams, and held office hours.

Projects

- **MIAE (Membership Inference Attacks Evaluation)**: [Prof. Lei Yu] (github.com/RPI-DSPlab) (*May. 2023 - Dec. 2024*)
  A Python library designed to facilitate the evaluation of membership inference attacks on machine learning models, providing a comprehensive framework for implementing, testing, and comparing various attack types.

  The package supports consistent experimental conditions across multiple attacks by standardizing API inputs. It implements 7 membership inference attacks, ranging from basic methods like the *Loss Threshold Attack* to advanced techniques like *LiRA*. The framework also integrates evaluation tools such as TPR@low FPR and Venn diagrams to visualize overlap in predictions. MIAE is designed to be easily extensible, encouraging researchers to implement and test new attacks. The package was used in our paper, *"Membership Inference Attacks as Privacy Tools: Reliability, Disparity, and Ensemble"*.

  Currently, I am leading a team to integrate Membership Inference Attacks against large language models (LLMs) into the package, which will expand the perspectives mentioned above on LLMs.

- **NYC Subway Challenge**: [Rensselaer Center for Open Source] github.com/RPI-Subway-Challenge/subwayChallenge) (*Jan. 2023 - May. 2023*)
  NYC Subway Challenge uses a greedy algorithm and a BFS search to find the optimal path through the representation of the New York City subway system. Data was provided by MTA and web-scraped from Google Maps. Written in C++ and Python. I implemented a route optimization algorithm for the NYC subway system using BFS.

Presentations and Talks

- **Invited Poster Talk: 2023 RHC Academic Showcase**:
  New York, United States — Oct 15, 2023
  Presented on “Understanding the Dynamics of Membership Privacy in Deep Learning”, highlighting risks and evaluation methods for privacy in AI. Discussed ongoing research on selective data point analysis and demonstrated the potential of the MIAE framework to extend privacy assessments to Large Language Models (LLMs).

- **IBM-RPI Future of Computing Research Collaborations (FCRC) Seminar Series**:
  New York, United States — Nov 12, 2024
  Introduced Membership Inference Attacks (MIAs) and their significance in evaluating privacy. Presented our paper “Membership Inference Attacks as Privacy Tools: Reliability, Disparity, and Ensemble”, with a focus on evaluating unlearning reliability.

Professional Services

- **Reviewer**: AAAI 2026

Leadership & Extracurricular Activities

- **Resident Assistant**: Student Living and Learning, RPI *Sept. 2023 - Dec. 2023*
  - Fostered a supportive and inclusive community among 30+ residents, promoting student engagement and well-being.
  - Served as a key resource for academic and personal guidance, organizing events to build connections and inclusivity.
  - Managed conflicts, provided crisis support, and upheld residence policies, developing strong leadership, communication, and problem-solving skills.
  - Gained experience in creating positive, inclusive environments, preparing for future leadership roles.

- **Bridge Scholar Program Mentor**: Bridge Scholar Program, RPI *May 2024 - Aug. 2024*
  - Provided academic and social support to incoming underrepresented STEM students during an intensive two-week on-campus program.
  - Led workshops on academic success strategies, career development, and effective use of campus resources, equipping students with essential skills for success at Rensselaer.
  - Fostered a welcoming and inclusive environment, contributing to the program’s goal of increasing retention and success of underrepresented students in STEM.

Awards

- ACM CCS 2026 Young Scholars Development Program (YSDP) Travel Grant