Rongzhe Wei

☎ Tel: (765)-746-9720 ✉ Email: rongzhe.wei@gatech.edu
📍 Address: 12th Floor, Coda Building, 756 W Peachtree St NW, Atlanta, GA 30308

EDUCATION BACKGROUND

Georgia Institute of Technology, Atlanta, Georgia Currently, Anticipated Graduation Date: 05/2026
Ph.D. in Machine Learning
School of Electrical and Computer Engineering
Advisor: Pan Li

Xi'an Jiaotong University (XJTU), Xi'an, China 09/2017 - 07/2021
Bachelor of Science in Mathematics (Overall GPA: **3.89**/4.0)
Honors Math Program by Ministry of Education, P.R. China
Advisor: Qinghua Zheng (President of Tongji University; Former Vice Chancellor of XJTU)

RESEARCH INTERESTS

- Trustworthy Machine Learning (Machine Unlearning, Privacy, Model Safety)
- Large Language Models, Agentic AI (Planning, Decision-making, Search Mechanism)
- Graphs Analysis (Graph Diffusion, Graph Neural Networks)

SELECTED PUBLICATIONS (Sorted by Year)

Published:

- **[NeurIPS'25]** **Rongzhe Wei***, Peizhi Niu*, Hans Hao-Hsun Hsu, Ruihan Wu, Haoteng Yin, Mohsen Ghassemi, Yifan Li, Vamsi K. Potluru, Eli Chien, Kamalika Chaudhuri, Olgica Milenkovic, Pan Li. Do LLMs Really Forget? Evaluating Unlearning with Knowledge Correlation and Confidence Awareness. In *Advances in Neural Information Processing Systems*.
  Summary: *In this project, we propose a novel knowledge unlearning definition with corresponding evaluation framework. Our framework accurately captures the implicit structure of real-world knowledge by representing relevant factual contexts as knowledge graphs with associated confidence scores.*

- **[NeurIPS'25]** Yinan Huang*, Haoteng Yin*, Eli Chien, **Rongzhe Wei**, Pan Li. Node-level Differential Private Relational Learning on Graphs. In *Advances in Neural Information Processing Systems*.
  Summary: *In this project, we consider the problem of node-level privacy-preserving relational learning on graphs with novel privacy amplification analysis.*

- **[ICML'25]** **Rongzhe Wei**, Mufei Li, Mohsen Ghassemi, Eleonora Kreacic, Yifan Li, Xiang Yue, Bo Li, Vamsi K. Potluru, Pan Li, Eli Chien. Underestimated Privacy Risks for Minority Populations in Large Language Model Unlearning. In *International Conference on Machine Learning*.
  Summary: *In this project, we identify a critical flaw that the privacy risks faced by minority groups within the training data are often significantly underestimated in large language model unlearning.*

- **[ICML'25]** Haoyu Wang, Shikun Liu, **Rongzhe Wei**, Pan Li. Generalization Principles for Inference over Text-Attributed Graphs with Large Language Models. In *International Conference on Machine Learning*.
  Summary: *In this project, we address the challenges of applying large language models to text-attributed graph learning by proposing the LLM-BP framework, which integrates task-adaptive embeddings and a generalizable graph information aggregation mechanism.*

- **[COLM'25]** Haoteng Yin, **Rongzhe Wei**, Eli Chien, Pan Li. Privately Learning from Graphs with Applications in Large Language Model Finetuning. In *Conference on Language Modeling*.
  Summary: *In this project, we propose a privacy-preserving relational learning pipeline that ensures differential privacy in fine-tuning large language models on sensitive graph data using a tailored DP-SGD.*

- **[EMNLP Findings 2025]** Tianchun Li, Tianci Liu, Xingchen Wang, **Rongzhe Wei,** Pan Li, Lu Su, Jing Gao. Towards Universal Debiasing for Language Models-based Tabular Data Generation. In *Conference on Empirical Methods in Natural Language Processing* findings.
  Summary: *In this project, we developed a universal debiasing framework for large language model-based tabular data generation that mitigates fairness issues by minimizing group-level dependencies, combining mutual information estimation with DPO and targeted debiasing techniques.*

- **[NeurIPS'24] Rongzhe Wei**, Eli Chien, Pan Li. Differentially Private Graph Diffusion with Applications in Personalized PageRanks. In *Advances in Neural Information Processing Systems*.
  Summary*: In this project, we propose a novel graph diffusion framework with a Wasserstein Distance tracking method that extends beyond traditional Privacy Amplification by Iteration analysis, eliminating the diameter assumption and achieving state-of-the-art privacy-utility trade-offs in personalized PageRank applications.*

- **[TMLR'24 → ICLR'25] Rongzhe Wei**, Eleonora Kreačić, Haoyu Wang, Haoteng Yin, Eli Chien, Vamsi K Potluru, Pan Li. On the Inherent Privacy Properties of Discrete Denoising Diffusion Models. In *Transactions on Machine Learning Research*, selected to *International Conference on Learning Representations* Poster.
  Summary: *In this project, we analyze and demonstrate the weak inherent privacy guarantees of discrete denoising diffusion models over discrete data and outliers suffers from higher privacy leakages.*

- **[TKDE'24]** Yizhou Wang, Can Qin, **Rongzhe Wei**, Yi Xu, Yue Bai, & Yun Fu. SLA^{2}P: Self-supervised Anomaly Detection with Adversarial Perturbation. In *IEEE Transactions on Knowledge and Data Engineering.*

- **[WWW'24]** Tianyi Zhang*, Haoteng Yin*, **Rongzhe Wei**, Pan Li, Anshumali Shrivastava. Learning Scalable Structural Link Representations with Bloom Signatures. Proceedings of the *ACM Web Conference*.

- **[NeurIPS'22] Rongzhe Wei**, Haoteng Yin, Junteng Jia, Austin R. Benson, Pan Li. Understanding Non-linearity in Graph Neural Networks from the Bayesian-Inference Perspective. In *Advances in Neural Information Processing Systems*.
  Summary: *In this project, we provide a theoretical analysis of the benefits of non-linearity over linear functions in graph neural networks, focusing on strength of attribute and structural information under the contextual stochastic block model.*

Preprint：

- **Rongzhe Wei***, Peizhi Niu*, Xinjie Shen*, Tony Tu, Yifan Li, Ruihan Wu, Eli Chien, Pin-yu Chen, Olgica Milenkovic, Pan Li. The Trojan Knowledge: Bypassing Commercial LLM Guardrails via Harmless Prompt Weaving and Adaptive Tree Search. *Submitted to ICML 2026.*
  Summary: *This paper proposed a novel LLM jailbreaking paradigm based on adaptive, dynamic knowledge decomposition. Our method can significantly outperform all SOTA jailbreak methods and can jailbreak all current SOTA closed-source commercial LLMs with more than 95% success rate!*
  Note: *This paper received coverage from multiple bloggers and X media accounts within five days of its arXiv release, surpassing 1,000 views, and I am invited to present the work at* ***IBM Research*** *in January 2026.*

- Yupu Gu, **Rongzhe Wei**, Andy Zhu, Pan Li. MoEEdit: Efficient and Routing-Stable Knowledge Editing for Mixture-of-Experts LLMs. *ICLR 2026 Under-review.*

- Andy Zhu, **Rongzhe Wei**, Yupu Gu, Pan Li. PRISM: Algorithm-Agnostic Machine Unlearning for Mixture-of-Experts via Geometric Router Constraints. ACL ARR October *Under-review.*

- Shuaiqi Wang, **Rongzhe Wei**, Mohsen Ghassemi, Eleonora Kreacic, Vamsi K. Potluru. Guarding Multiple Secrets: Enhanced Summary Statistic Privacy for Data Sharing. In *ICLR'24 PML Workshop.*
  Summary: *In this project, we propose a framework to define, analyze, and protect multi-secret summary statistics privacy by designing tailored privacy metrics and release mechanisms, balancing privacy and data distortion, and evaluating their effectiveness on real-world data.*

PROFESSIONAL SERVICES

- **Conference Reviewer:** NeurIPS’22-25, ICML’24-25, ICLR’25-26, AISTATS’23-24, ISIT’25, AAAI’24, LoG’22-24
- **Journal Reviewer:** Computers and Mathematics with Applications, TMLR
- **Graduate Teaching Assistant –** ECE 6720 Convex Optimization (Graduate Level) / ECE 3077 Introduction to Probability and Statistics for ECEs / ECE 8003 Conversational AI (Graduate Level)

INDUSTRIAL EXPERIENCES

- **Amazon** **Seattle, WA**
  AI Research Intern May 2025 – August 2025
  Project: LLM Agent for Decision-making
  Collaborators: Leman Akoglu, Christos Faloutsos, Sarthak Ghosh, Vaibhav Gorde, Na Zhang
  Paper under internal review: Entropy-Guided Branching for Long-Horizon Plan Execution in Large Tool Spaces.

- **JP Morgan Chase & Co.** **Manhattan, NYC, NY**
  AI Research Intern June 2023 – August 2023
  Project (Patent): A General Framework for Graph Data Generation Control via Margin Relaxed Schrodinger Bridges.
  Mentors: Eleonora Kreačić and Vamsi K Potluru

SELECTED HONORS & SCHOLARSHIP

◆ Georgia Tech CSIP Outstanding Research Award	2025
◆ Lambda's Research Grant	2025
◆ OpenAI API Researcher Access Program Grant	2025
◆ Travel Award for NeurIPS 2022	2022
◆ Student Award in 2019 IEEE International Conference on BigData	2019
◆ The First Prize of “*Zhufeng*” Scholarship, established for “*Pilot Scheme of Top-notch Talent Cultivation in Basic Disciplines*”, Ministry of Education for three times	2017 – 2018, 2018 – 2019, 2019 - 2020
◆ Outstanding Student Award for three times, XJTU	2017 – 2018, 2018 – 2019, 2019 - 2020
◆ First-class Scholarship for three times, XJTU	2017 – 2018, 2018 – 2019, 2019 - 2020

TECHNICAL SKILLS

- **Languages:** Chinese (native), English
- **Programming Languages:** Python, C#, Matlab, C, SQL, HTML