# RONGZHE WEI

✉ rongzhe.wei@gatech.edu | ☎ +1 (765) 746-9720 | Atlanta, GA, USA

LinkedIn | GitHub | Google Scholar | Personal Website

## EDUCATION

**Georgia Institute of Technology** Atlanta, USA

*Ph.D. in Machine Learning, School of Electrical and Computer Engineering* Aug. 2021 – Present (Expected Aug. 2026)

- Advisor: Prof. Pan Li
- GPA: 4.0/4.0

**Xi'an Jiaotong University (XJTU)** Xi'an, China

*B.S. in Mathematics, Honors Math Program, Qian Xuesen College* Sep. 2017 – Jul. 2021

- GPA: 3.89/4.0 · Honors Math Program by Ministry of Education, P.R. China
- Advisor: Prof. Qinghua Zheng (President of Tongji University; Former Vice Chancellor of XJTU)

## RESEARCH INTERESTS

- LLM Safety & Alignment: Red-Teaming, Multi-turn Jailbreak & Defense, Machine Unlearning, Data Privacy
- Agentic AI: Planning, Tool-Calling, Search & Exploration
- Graph-Structured Learning: Graph Neural Networks, Graph Diffusion, Privacy-Preserving Graph Methods

## RESEARCH HIGHLIGHTS

**Structured Internal Knowledge for LLM Trustworthiness** GCoM Lab, Georgia Tech, 2024 – Present

> ***Core finding:*** *A fundamental vulnerability in LLM trust and safety stems from ignoring the interconnected nature of internal knowledge. We advocate structured knowledge modeling from both* ***trust*** *(unlearning) and* ***safety*** *(red-teaming) perspectives.*

- ***Evaluating Trust through Knowledge Correlation.*** Proposed an automated deep unlearning evaluation framework that captures complex inference patterns via knowledge graphs. Revealed that existing unlearning methods significantly overestimate forgetting efficacy; correlated knowledge probing exposes poor performance-unlearning trade-offs. → **NeurIPS'25**
- ***Analyzing Safety via Knowledge Weaving.*** Developed CKA-Agent, an automated jailbreak agent that decomposes harmful queries into benign correlated sub-questions, leverages target LLM responses as a knowledge oracle, and performs adaptive tree search. Achieved **∼95% attack success** on GPT-5.2, Gemini-3-Pro, and Claude-4.5-Haiku; demonstrated that current alignment fails to detect distributed multi-turn harmful intent. **GitHub 150+ stars, 40+ forks.** → **Sub. ICML'26**
- ***Ongoing: Robust Guardrails via Multi-Turn RL.*** Designing reinforcement-learning-based defense that detects malicious intent distributed across conversation turns.

**Agentic Long-Horizon Plan Execution in Large Tool Spaces** Amazon AI Research Intern, 2025

- Built SLATE, a **1,000-tool benchmark** with hierarchical multi-step plans for automated plan-level evaluation of tool-augmented LLM agents. Proposed Entropy-Guided Branching (EGB), an uncertainty-aware search algorithm that branches at high-entropy decision points; achieved **54% execution success** on Claude-Sonnet-4 (vs. 29.3% ReAct, 36.5% LATS) with **73% fewer tokens**. → **Sub. ACL'26**

**Tight Non-Divergent Privacy Tracking of Iterative Algorithms** GCoM Lab, Georgia Tech, 2023 – 2024

- Improved Privacy Amplification by Iteration (PABI) analysis by introducing a Wasserstein distance tracking method within shifted Rényi divergence framework, eliminating the restrictive diameter assumption. Applied to differentially private graph diffusion for Personalized PageRank, achieving significantly tighter privacy-utility trade-offs. → **NeurIPS'24**

## INDUSTRY EXPERIENCE

**Amazon** Seattle, WA

*AI Research Intern* May 2025 – Aug. 2025

- Project: LLM Agent for Decision-Making
- Collaborators: Christos Faloutsos (CMU), Leman Akoglu (CMU), Sarthak Ghosh, Vaibhav Gorde, Na Zhang
- Paper under review: *Entropy-Guided Branching for Long-Horizon Plan Execution in Large Tool Spaces.*

**JP Morgan Chase & Co.** New York, NY

*AI Research Intern* Jun. 2023 – Aug. 2023

- Mentors: Eleonora Kreačić and Vamsi K Potluru.
- ICLR'24 PML / U.S. Patent: *System and Method for Enhanced Summary Statistic Privacy for Data Sharing* (2026/0023874).
- U.S. Patent (Pending): *Method and System for Data Generation Control via Margin-Relaxed Schrödinger Bridges*.

## PROFESSIONAL SERVICES

- **Conference Reviewer:** NeurIPS'22–25, ICML'24–26, ICLR'25–26, AISTATS'23–24, ISIT'25, AAAI'24, LoG'22–24

- **Journal Reviewer:** Transactions on Machine Learning Research, Computers and Mathematics with Applications
- **Graduate TA:** ECE 8003 Conversational AI / ECE 6720 Convex Optimization / ECE 3077 Intro to Probability & Statistics

## SELECTED PUBLICATIONS

### Published

1. **ICLR'26** Yupu Gu, **Rongzhe Wei**, Andy Zhu, Pan Li. "MoEEdit: Efficient and Routing-Stable Knowledge Editing for Mixture-of-Experts LLMs." *ICLR 2026.*
2. **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." *NeurIPS 2025.*
3. **NeurIPS'25** Yinan Huang*, Haoteng Yin*, Eli Chien, **Rongzhe Wei**, Pan Li. "Node-level Differential Private Relational Learning on Graphs." *NeurIPS 2025.*
4. **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." *ICML 2025.*
5. **ICML'25** Haoyu Wang, Shikun Liu, **Rongzhe Wei**, Pan Li. "Generalization Principles for Inference over Text-Attributed Graphs with Large Language Models." *ICML 2025.*
6. **COLM'25** Haoteng Yin, **Rongzhe Wei**, Eli Chien, Pan Li. "Privately Learning from Graphs with Applications in Large Language Model Finetuning." *COLM 2025.*
7. **EMNLP'25** Tianchun Li, Tianci Liu, Xingchen Wang, **Rongzhe Wei**, Pan Li, Lu Su, Jing Gao. "Towards Universal Debiasing for Language Models-based Tabular Data Generation." *EMNLP Findings 2025.*
8. **NeurIPS'24** **Rongzhe Wei**, Eli Chien, Pan Li. "Differentially Private Graph Diffusion with Applications in Personalized PageRanks." *NeurIPS 2024.*
9. **TMLR → 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." *TMLR 2024, selected to ICLR 2025 Poster.*
10. **ICLR'24 PML** Shuaiqi Wang, **Rongzhe Wei**, Mohsen Ghassemi, Eleonora Kreačić, Vamsi K. Potluru. "Guarding Multiple Secrets: Enhanced Summary Statistic Privacy for Data Sharing." *ICLR 2024 Workshop on Privacy in Machine Learning (PML).*
11. **TKDE'24** Yizhou Wang, Can Qin, **Rongzhe Wei**, Yi Xu, Yue Bai, Yun Fu. "SLA$^2$P: Self-supervised Anomaly Detection with Adversarial Perturbation." *IEEE TKDE 2024.*
12. **WWW'24** Tianyi Zhang*, Haoteng Yin*, **Rongzhe Wei**, Pan Li, Anshumali Shrivastava. "Learning Scalable Structural Link Representations with Bloom Signatures." *WWW 2024.*
13. **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." *NeurIPS 2022.*

### Preprint / Under Review

14. **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.*
15. Mufei Li*, Shikun Liu*, Xinnan Dai*, **Rongzhe Wei***, Haoyu Wang, Xinjie Shen, Siqi Miao, Jiliang Tang, Pan Li. "Beyond the Sequence – Graph Learning as the Blueprint for Trustworthy Large Language Models." *Submitted to ICML 2026.*
16. **Rongzhe Wei**, Ge Shi, Min Cheng, Leman Akoglu, Vaibhav Gorde, Sarthak Ghosh. "Long-Horizon Plan Execution in Large Tool Spaces through Entropy-Guided Branching." *Submitted to ACL 2026.*
17. Andy Zhu, **Rongzhe Wei**, Yupu Gu, Pan Li. "PRISM: Algorithm-Agnostic Machine Unlearning for Mixture-of-Experts via Geometric Router Constraints." *Submitted to ICML 2026.*

## INVITED TALKS & RESEARCH PROPOSALS

- **Invited Talk:** "From Atomic Facts to Structured Internal Knowledge: Rethinking Unlearning and Jailbreaking in LLMs" · *@IBM Research*, Jan. 2026; *@Google Research Seminar*, Feb. 2026.
- **Proposal Lead Author:** "Structured-Knowledge-Guided Agentic LLM Jailbreaking and Defense" · NAIRR Pilot Award, Feb. 2026. Recognized for "high degree of alignment with national AI strategic focus."

## HONORS & AWARDS

| | |
|---|---|
| Georgia Tech CSIP Outstanding Research Award | 2025 |
| Lambda's Research Grant; OpenAI API Researcher Access Program Grant | 2025 |
| Travel Award for NeurIPS 2022 | 2022 |
| Student Award, IEEE International Conference on BigData | 2019 |
| First Prize "Zhufeng" Scholarship (*Top-notch Talent Cultivation*, Ministry of Education) | 2017–2020 |
| Outstanding Student Award & First-class Scholarship, XJTU | 2017–2020 |

## SKILLS

**Programming:** Python, C#, MATLAB, SQL, LaTeX

**LLM Frameworks:** PyTorch, Hugging Face Transformers, vLLM, DeepSpeed

**Languages:** Mandarin (Native), English (Fluent)