Zheyuan Liu

Personal Information

Zheyuan Liu
University of Notre Dame
Holy Cross Dr, 46556, Notre Dame, IN

Tel.: +1 (857)-971-0949
Email: zliu29@nd.edu
GitHub: https://github.com/franciscoliu
Google Scholar: https://scholar.google.com/citations?user=NLA-nSUAAAAJ&hl=en
Linkedin: https://www.linkedin.com/in/zheyuan-frank-liu-371738185/
Website: https://franciscoliu.github.io/

Education

09/2019–05/2023 **B.S Computer Science**
B.S Applied Mathematics (double-major)
Brandeis University, Waltham, MA, USA
Cumulative GPA: 3.87

09/2023– Current **PhD Computer Science**
University of Notre Dame, Notre Dame, IN, USA
Cumulative GPA: 3.92
Advisor: Prof. Meng Jiang

Research Interest

Trustworthy Generative AI: (Multimodal) Large Language Models Safety, Agentic Safety, Machine Unlearning, Data Privacy

Knowledge-based Model Editing (KME): Knowledge Update, Knowledge Conflict, Model Editing

Data-Centric Problem and Learning: Data Augmentation, Data Generation

Publications

1. **Liu, Zheyuan**, Xu, Z., Dou, G., Yuan, X., Tan, Z., Poovendran, R., and Jiang, M. Steering multimodal large language models decoding for context-aware safety. *arXiv preprint arXiv:2509.19212v1 (In Submission)* (2025)
2. **Liu, Zheyuan**, Dou, G., Yuan, X., Zhang, C., Tan, Z., and Jiang, M. Modality-aware neuron pruning for unlearning in multimodal large language models. In *ACL Main* (2025)

3. **Liu, Zheyuan**, Maharjan, S., Wu, F., Parikh, R., Bayar, B., Sengamedu, S. H., and Jiang, M. Disentangling biased knowledge from reasoning in large language models via machine unlearning. In *ACL Main* ***(Oral, Top 8 %)*** (2025)
4. **Liu, Zheyuan**, Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., and Jiang, M. Protecting privacy in multimodal large language models with mllmu-bench. In *NAACL Main* ***(Oral)*** (2025)
5. **Liu, Zheyuan**, Dou, G., Tan, Z., Tian, Y., and Jiang, M. Towards safer large language models through machine unlearning. In *ACL Findings* (2024)
6. **Liu, Zheyuan**, He, X., Tian, Y., and Chawla, N. Can we soft prompt llms for graph learning tasks? In *The Web Conference (WWW) Short Paper* (2024)
7. **Liu, Zheyuan**, Dou, G., Tian, Y., Zhang, C., Chien, E., and Zhu, Z. Breaking the trilemma of privacy, utility, efficiency via controllable machine unlearning. In *The Web Conference (WWW)* (2024)
8. **Liu, Zheyuan**, Zhang, C., Tian, Y., Zhang, E., Huang, C., Ye, Y., and Zhang, C. G-FAME: Fair graph representation learning via diverse mixture of experts. In *The Web Conference (WWW)* (2023)
9. **Liu, Zheyuan**, Dou, G., Tan, Z., Tian, Y., and Jiang, M. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516 (In Submission of CSUR)* (2024)
10. Wu, W., **Liu, Zheyuan**, Gao, C., Ren, W., and Ding, K. Beyond sharp minima: Robust llm unlearning via feedback-guided multi-point optimization. *arXiv preprint arXiv:2509.20230 (In Submission)* (2025)
11. Xu, G., Duan, Y., **Liu, Zheyuan**, Li, X., Jiang, M., Lemmon, M., Jin, W., and Shi, Y. Incorporating rather than eliminating: Achieving fairness for skin disease diagnosis through group-specific expert. In *MICCAI* (2025)
12. Wang, Z., **Liu, Zheyuan**, Ma, T., Li, J., Zhang, Z., Fu, X., Li, Y., Yuan, Z., Song, W., Ma, Y., et al. Graph foundation models: A comprehensive survey. *arXiv preprint arXiv:2505.15116* (2025)
13. Tan, Z., Zeng, Z., Zeng, Q., Wu, Z., **Liu, Zheyuan**, Mo, F., and Jiang, M. Can large language models understand preferences in personalized recommendation? *arXiv preprint arXiv:2501.13391* (2025)
14. Dou, G., **Liu, Zheyuan**, Lyu, Q., Ding, K., and Wong, E. Avoiding copyright infringement via machine unlearning. In *NAACL Findings* (2025)
15. Tan, Z., **Liu, Zheyuan**, and Jiang, M. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *EMNLP Main* (2024)
16. Liang, Z., Liu, G., **Liu, Zheyuan**, Cheng, J., Hao, T., Liu, K., Ren, H., Song, Z., Liu, J., Ye, F., and Shi, Y. Graph learning for parameter prediction of quantum approximate optimization algorithm. In *Design Automation Conference (DAC)* (2024)
17. Tan, Z., Zeng, Q., Tian, Y., **Liu, Zheyuan**, Yin, B., and Jiang, M. Democratizing large language models via personalized parameter-efficient fine-tuning. In *EMNLP Main* (2024)

18. Zhang, C., Tian, Y., Ju, M., **Liu, Zheyuan**, Ye, Y., Chawla, N., and Zhang, C. Chasing all-round graph representation robustness: Model, training, and optimization. In *ICLR* (2023)

19. Wu, J., Zhang, C., **Liu, Zheyuan**, Zhang, E., Wilson, S., and Zhang, C. GraphBERT: Bridging graph and text for malicious behavior detection on social media. In *ICDM* (2022)

20. Yuan, X., Zhang, C., Liu, Z., Shi, D., Vosoughi, S., and Lee, W. Superficial self-improved reasoners benefit from model merging. *arXiv preprint arXiv:2503.02103* (2025)

21. Ni, B., **Liu, Zheyuan**, Wang, L., Lei, Y., Zhao, Y., Cheng, X., Zeng, Q., Dong, L., Xia, Y., Kenthapadi, K., et al. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872* (2025)

22. Yang, T., Dai, L., **Liu, Zheyuan**, Wang, X., Jiang, M., Tian, Y., and Zhang, X. Cliperase: Efficient unlearning of visual-textual associations in clip. *arXiv preprint arXiv:2410.23330* (2024)

23. Tian, Y., Zhang, C., Kou, Z., **Liu, Zheyuan**, Zhang, X., and Chawla, N. Ugmae: A unified framework for graph masked autoencoders. *arXiv preprint arXiv:2402.08023* (2024)

24. Wang, Y., Peng, H. M., Sha, L., **Liu, Zheyuan**, and Hong, P. State-level covid-19 trend forecasting using mobility and policy data. *medRxiv* (2021)

Invited Talks

Talk at Amazon @ Amazon Benchmarking, Chime, July 2025

Talk at Brandeis @ PhooD Seminar, Zoom, January 2025

Industry Experience

05/2025 – 08/2025 **Amazon**, Palo Alto, CA
Applied Scientist Intern (Rufus Team)

- Developed self-reflective tool-use benchmarks for LLMs, evaluating their performance across domains such as e-commerce, finance, and healthcare.

05/2024 – 08/2024 **Amazon**, Seattle, WA
Applied Scientist Intern (PXTCS Team)

- Worked on addressing fairness/bias issues in Large Language Models via Machine Unlearning techniques.
- Proposed a new prototype that alleviates internal model bias while preserving its reasoning ability. The prototype was later approved by the **United States Patent**.

Teaching Experience

09/2021 – 05/2023 **Brandeis University**, Waltham, MA
Teaching Assistant

- Acted as teaching assistant for **COSI 10a** (Python), **COSI 12b** (JAVA), **COSI 103a** (Fundamentals of Software Engineering) and **COSI 131a** (Operating System class).

09/2023 – 05/2024 **University of Notre Dame**, Notre Dame, IN
Teaching Assistant

- Acted as teaching assistant for **CSE-40923** (Case Studies in Computing-Based Entrepreneurship class) and **CSE-30353** (Signals Processing Fundamentals).

Honors, Awards & Scholarships

03/2025	Graduate School Professional Development Award (750 dollars)
12/2024	US Patent: **Disentangling biased knowledge from reasoning in large language models via machine unlearning.**
03/2024	Conference Presentation Grant (300 dollars)
03/2024	Zahm Professional Development Fund (1250 dollars)
05/2023	Molly W. and Charles K. Schiff Memorial Award (Top 3 %)
06/2022	Provost's Research Fellowship (5000 dollars)
12/2019	Dean's List (Every semester)
09/2017	Patent of a new type of packing tool

Service

Journals

IEEE Transactions on Big Data Reviewer
IEEE Transactions on Neural Networks and Learning Systems (TNNLS) Reviewer
IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI) Reviewer
IEEE Transactions on Audio, Speech and Language Processing Reviewer
TKDE Reviewer (2023, 2024)

Conferences

ICLR 2026 Reviewer
Program Committee of AAAI'2026

Program Committee of CIKM'2024 (Applied Research Track)
Program Committee of CIKM'2025 (Full Paper Track)
NeurIPS Dataset and Benchmark Track Reviewer (2024, 2025)
ARR Reviewer
ICDM 2024 MLoG Workshop Reviewer
ACL 2024 Workshop KnowledgeNLP Reviewers

Current Mentored Students

1. **John Kim,** *Undergraduate Student at Notre Dame* (Since Feb 2025)
2. **Katherine O'Roark**, *Undergraduate Student at Saint Mary's College* (Since Feb 2025)
3. **Han Yan**, *Undergraduate Student at The Chinese University at HongKong, Shenzhen* (Since March 2025)
4. **Wenhan Wu**, *Master Student at Wuhan University, Wuhan* (Since March 2025)

Grants and Gifts

"Inference-Time Safety Calibration for Specialized AI Agents"
June 30, 2025 – June 30, 2026
Funding Vehicle: OpenAI Researcher Access Program
Amount: **$1,000**

"Machine Unlearning for GenAI Safety"
June 24, 2024 – December 24, 2024 *(Completed)*
Funding Vehicle: OpenAI Researcher Access Program
Amount: **$10,000**

Award Proposal Writing Experience

"Calibrating LLM Refusals for Trustworthy Mental Health Support"
July 16, 2025 – July 16, 2026
Funding Vehicle: NAIRR Pilot
Amount: 5,780 GPU-hours
Role: Lead Contributor to Writing and Conceptual Development
PI: Prof. Meng Jiang