

# Zheyuan Liu

## PERSONAL INFORMATION

---

Zheyuan Liu  
University of Notre Dame  
Holy Cross Dr, 46556, Notre Dame, IN  
Tel.: [+1 \(857\)-971-0949](tel:+18579710949)  
Email: [zliu29@nd.edu](mailto:zliu29@nd.edu)  
GitHub: <https://github.com/franciscoliu>  
Google Scholar: <https://scholar.google.com/citations?user=NLA-nSUAJ&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:2509.19212* (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. JIA, M., ZHANG, Z., CASES, I., **LIU, ZHEYUAN**, JIANG, M., AND QI, P. Autorubric-rlv: Rubric-based generative rewards for faithful multimodal reasoning. *arXiv preprint arXiv:2510.14738* (2025)
11. TAN, Z., ZHANG, Z., WEN, H., LI, Z., ZHANG, R., CHEN, P., MO, F., LIU, Z., ZENG, Q., YIN, Q., ET AL. Instant personalized large language model adaptation via hypernetwork. *arXiv preprint arXiv:2510.16282* (2025)
12. HAN, Y., **LIU, ZHEYUAN**, AND JIANG, M. Dual-space smoothness for robust and balanced llm unlearning. *arXiv preprint arXiv:2509.23362* (2025)
13. 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* (2025)
14. ZHANG, C., OUYANG, Z., DIAO, X., **LIU, ZHEYUAN**, AND VOSOUGHI, S. Knowing more, acting better: Hierarchical representation for embodied decision-making. In *EMNLP Findings* (2025)
15. 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)
16. 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)
17. 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)

18. DOU, G., **LIU, ZHEYUAN**, LYU, Q., DING, K., AND WONG, E. Avoiding copyright infringement via machine unlearning. In *NAACL Findings* (2025)
19. TAN, Z., **LIU, ZHEYUAN**, AND JIANG, M. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *EMNLP Main* (2024)
20. 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)
21. 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)
22. 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)
23. WU, J., ZHANG, C., **LIU, ZHEYUAN**, ZHANG, E., WILSON, S., AND ZHANG, C. Graph-BERT: Bridging graph and text for malicious behavior detection on social media. In *ICDM* (2022)
24. YUAN, X., ZHANG, C., **LIU, ZHEYUAN**, SHI, D., VOSOUGHI, S., AND LEE, W. Superficial self-improved reasoners benefit from model merging. In *EMNLP Main* (2025)
25. 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)
26. 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)
27. 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)
28. 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

## Tutorial/Workshop Organizer

1. Socially Responsible and Trustworthy Foundation Models (ResponsibleFM), Workshop on NeurIPS 2025 at Mexico City, Mexico ([Link](#)).
2. Multilinguality in the Era of Large Language Models (MeLLMs), Workshop on ACL 2026 at San Diego, USA.
3. Data Security and Privacy in Machine Unlearning: Recent Advances, Challenges, and Future Perspectives, Tutorial on ICDM 2026 at Washington DC, USA ([Link](#)).

## 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 CUHK, Shenzhen* (Since March 2025)
4. **Wenhan Wu**, *Master Student at Wuhan University, Wuhan* (Since March 2025)
5. **Yiru Wang**, *PhD Student at CUHK, Hongkong* (Since July 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

**“Probing-Guided Robust Unlearning”**

January 1, 2026 – June 30, 2027

Funding Vehicle: Open Philanthropy and Good Ventures Foundation

Amount: **\$281,628**

Role: Lead Contributor to Writing and Conceptual Development

PI: Prof. Meng Jiang