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 and Engineering**
University of Notre Dame, Notre Dame, IN, USA
Cumulative GPA: 3.92
Advisor: Prof. Meng Jiang
Thesis Committee: Prof. Nitesh Chawla (Notre Dame), Prof. Xiangliang Zhang (Notre Dame), Prof. Kai-Wei Chang (UCLA)

Research Interest

Trustworthy Generative AI: (Multimodal) Large Language Models Safety, Vision-Language Action (VLA) Safety, Agentic Safety

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

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

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. In *EMNLP Main* (2026)

2. **Liu*, Zheyuan**, Kim*, D., Wan, Y., Yuan, X., Tan, Z., Mo, F., and Jiang, M. Mtmcs-bench: Evaluating contextual safety of multimodal large language models in multi-turn dialogues. In *ACL Findings* (2026)
3. **Liu, Zheyuan**, Xiao, L., Li, Y., Yun, H., Li, L., Zhang, C., and Jiang, M. Do llms catch their own mistakes? a comprehensive benchmark for reflective tool use llms. In *ACL Findings* (2026)
4. **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)
5. **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)
6. **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)
7. **Liu, Zheyuan**, Dou, G., Tan, Z., Tian, Y., and Jiang, M. Towards safer large language models through machine unlearning. In *ACL Findings* (2024)
8. **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)
9. **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)
10. **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)
11. **Liu, Zheyuan**, Wang, C., Guo, J., Huang, Y., Yuan, X., Ma, N., and Jiang, M. Teaching vision-language-action models to stop safely. *arXiv preprint arXiv:2611.26239* (2026)
12. **Liu, Zheyuan**, Zhao, W., Yuan, X., Ma, N., Huang, Y., and Jiang, M. Knowledge-verified emergent deception in llm agents under conflicting incentives. *arXiv preprint arXiv:2608.26372* (2026)
13. **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)
14. Kang*, D., **Liu*, Zheyuan**, Ma, N., Huang, Y., Tan, Z., and Jiang, M. Prompt-activation duality: Improving activation steering via attention-level interventions. In *NeurIPS* (2026)
15. Yan*, H., **Liu*, Zheyuan**, and Jiang, M. Dual-space smoothness for robust and balanced llm unlearning. In *ICLR* (2026)
16. Zhang*, J., **Liu*, Zheyuan**, Zhu, Y., and Jiang, M. Strength controls change how interventions compare across sparse autoencoder codes. *arXiv preprint arXiv:2610.26489* (2026)

17. Cui*, M., **Liu*, Zheyuan**, Zhu, Y., Zhang, Z., and Jiang, M. Learning from runtime feedback through failure-bank self-evolution for vision-language-action models. *arXiv preprint arXiv:2609.39820* (2026)

18. Chen*, B., **Liu*, Zheyuan**, Zhang, J., Ding, K., Ma, N., Huang, Y., and Jiang, M. Forgotten in weights, recovered by tools: Agentic tool unlearning for llm agents. *arXiv preprint arXiv:2608.21544* (2026)

19. Bai*, Y., **Liu*, Zheyuan**, Yan, H., Xu, Z., Wan, Y., Chen, C., Wang, Z., Yuan, X., Huang, Y., Dou, G., Zhang, Y., Zhu, H., Li, Z., Li, M., Zhang, X., Bansal, M., Koyejo, S., Chang, K.-W., Zhang, Y., and Jiang, M. Inference-time control for trustworthy large language models. *Preprints* (May 2026)

20. Ni, B., **Liu, Zheyuan**, Lei, Y., Wang, L., Zhao, Y., Cheng, X., Zeng, Q., Dong, X., Xia, Y., Kenthapadi, K., Rossi, R., Dernoncourt, F., Tanjim, M., Ahmed, N., Liu, X., Fan, W., Blasch, E., Wang, Y., Jiang, M., and Derr, T. Towards trustworthy retrieval augmented generation for large language models: A survey. *ACM Computing Surveys 58*, 15 (2026), 1–36

21. Yuan, X., Shi, D., Zhang, C., **Liu, Zheyuan**, Yao, S., Vosoughi, S., and Lee, W. Behavior knowledge merge in reinforced agentic models. In *ACL Main* (2026)

22. Mo, F., Su, Z., Hui, Y., Zhang, J., Sun, J. A., **Liu, Zheyuan**, Zhang, C., Sakai, T., and Nie, J.-Y. Opendecoder: Open large language model decoding to incorporate document quality in rag. In *WWW* (2026)

23. Diao, X., **Liu, Zheyuan**, Zhang, C., Wu, W., Kong, K., Shi, L., Ding, K., Vosoughi, S., and Gui, J. Addressing overthinking in large vision-language models via gated perception-reasoning optimization. In *ACL Findings* (2026)

24. Jia, M., Zhang, Z., Cases, I., **Liu, Zheyuan**, Jiang, M., and Qi, P. Autorubric: Rubric-based generative rewards for faithful multimodal reasoning. In *ACL Findings* (2026)

25. Tan, Z., Zhang, Z., Wen, H., Li, Z., Zhang, R., Chen, P., Mo, F., **Liu, Zheyuan**, Zeng, Q., Yin, Q., et al. Instant personalized large language model adaptation via hypernetwork. In *ACL Main* (2026)

26. Sarwar, N., Dipta, R. S., **Liu, Zheyuan**, and Patil, V. Multimodal unlearning across vision, language, video, and audio: Survey of methods, datasets, and benchmarks. In *ACL Findings* (2026)

27. Tian, Y., Zhang, C., Kou, Z., **Liu, Zheyuan**, Zhang, X., and Chawla, N. Adaptive and context-rich generative self-supervised learning on graphs. In *The 40th Annual AAAI Conference on Artificial Intelligence* (2026)

28. Yuan, X., Zhang, C., **Liu, Zheyuan**, Shi, D., Pan, L., Vosoughi, S., and Lee, W. Superficial self-improved reasoners benefit from model merging. In *EMNLP Main* (2025)

29. 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)

30. 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 experts. In *MICCAI* (2025)

31. Dou, G., **Liu, Zheyuan**, Lyu, Q., Ding, K., and Wong, E. Avoiding copyright infringement via machine unlearning. In *NAACL Findings* (2025)

32. Yang, T., Dai, L., **Liu, Zheyuan**, Wang, X., Jiang, M., Tian, Y., and Zhang, X. Cliperase: Efficient unlearning of visual-textual associations in clip. In *ACL Main* (2025)

33. Tan, Z., **Liu, Zheyuan**, and Jiang, M. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *EMNLP Main* (2024)

34. 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)

35. 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)

36. 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)

37. 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)

38. Huang, Y., Wang, Y., Li, L., Chen, Y., Haupt, A., Luo, X., **Liu, Zheyuan**, Choi, J., Juefei-Xu, F., Zhang, X., Poovendran, R., Sra, M., Liu, Y., Burtenshaw, B., Wolf, T., Pentland, A., Chen, Z., and Xu, Z. RSI Arena: Towards recursive self-improvement with human feedback. *arXiv preprint arXiv:2610.33624* (2026)

39. Zhang, J., **Liu, Zheyuan**, Zhu, Y., and Jiang, M. Training negatives still matter after calibrating answerability readouts. *arXiv preprint arXiv:2610.26461* (2026)

40. Huang, Y., Xu, Z., Luo, X., **Liu, Zheyuan**, Li, L., Wang, Y., Song, P., Ramos, E. G., Xie, E., Isayev, O., Liang, P., Dziri, N., Sun, H., Gao, J., Liu, Y., Sra, M., Zhang, X., Pentland, A., and Chen, Z. Epi: Towards a persistent, self-improving agent for autonomous research. *arXiv preprint arXiv:2610.24944* (2026)

41. Phan, L., Yang, S. K., Lim, J. J., Mazeika, M., Zhang, W., **Liu, Zheyuan**, Ren, R., Meng, J., Tan, Y., Zhao, W., Wu, A., Anghel, M., and Hendrycks, D. Cheatbench: Measuring reward gaming in ai agents. *arXiv preprint arXiv:2609.36308* (2026)

42. Huang, Y., Xu, Z., Ma, Y., Wang, W., **Liu, Zheyuan**, Xu, Z., Chen, P.-Y., Galley, M., Lin, Z., Feuerriegel, S., Poovendran, R., Sra, M., Pentland, A., Zhang, X., and Chen, Z. Reward hacking challenges oversight of autonomous research agents. *arXiv preprint arXiv:2609.28614* (2026)

43. Huang, Y., Wang, W., Bao, H., Ma, Y., Luo, X., Nian, Y., Zhuang, H., **Liu, Zheyuan**, Zhao, Y., and Zhang, X. Memoharness: Agent harnesses that learn from experience. *arXiv preprint arXiv:2607.14159* (2026)

44. Xu, Z., Chen, J., Huang, Y., Jiang, D., Chen, J., Hua, H., Wu, Z., **Liu, Zheyuan**, He, Z., Li, L., et al. Autolab: Can frontier models solve long-horizon auto research and engineering tasks? *arXiv preprint arXiv:2606.05080* (2026)

45. Meng, N., **Liu, Zheyuan**, Jiang, M., and Lao, Y. 3deditsafe: Defending 3d editing pipelines from unsafe generation. *arXiv preprint arXiv:2605.15398* (2026)

46. Ding, C., Wu, J., Luo, Y., **Liu, Zheyuan**, Yuan, Y., and Wang, X. Teaching large language models when not to know: Learning temporal critique for ex-ante reasoning. *arXiv preprint arXiv:2605.14636* (2026)

47. Wu, W., **Liu, Zheyuan**, Gao, C., Wang, R., and Ding, K. Beyond sharp minima: Robust llm unlearning via feedback-guided multi-point optimization. *arXiv preprint arXiv:2509.20230* (2025)

48. 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)

49. 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)

50. Phan, L., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249* (2025). Contributor: **Zheyuan Liu**

51. 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 Nankai University, Zoom, April 2026

Talk at UCLA @ UCLA NLP Group, Los Angeles, March 2026

Talk at Stanford University @ STAIR Lab, Zoom, March 2026

Guest Lecture for Large Language Models: Basics and Practice @ University of Notre Dame, December 2025

Talk at Amazon @ Amazon Benchmarking, Chime, July 2025

Talk at Brandeis @ PhooD Seminar, Zoom, January 2025

Industry Experience

05/2026 - 08/2026 **Center for AI Safety**, San Francisco, CA
Research Engineering Intern

- Worked on model welfare and reward hacking.

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.

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)

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
ACM Computing Survey 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
IEEE Transactions on Information Forensics & Security (T-IFS) Reviewer
TKDE Reviewer

Conferences

The Web Conference (WWW) 2027 Area Chair (Research Track)
ICLR 2026 Reviewer
Program Committee of AAAI
Program Committee of CIKM (Applied Research Track)
Program Committee of CIKM (Full Paper Track)
NeurIPS Main Conference + Dataset and Benchmark Track Reviewer
ARR Reviewer
ICDM MLoG Workshop Reviewer
ACL Workshop KnowledgeNLP Reviewers

Tutorial Organizer

[T1] Knowledge Control for Responsible Generative AI: Bridging Academia, Industry, and Society (Link, Paper) 2026

- **Zheyuan Liu**, Yixin Wan, Kai-Wei Chang, Meng Jiang, Jieyu Zhao, Nouha Dziri, Yuning Mao, Jia-Chen Gu, Jindong Gu
- Tutorial at ACL 2026, San Diego, USA.

[T2] Data Security and Privacy in Machine Unlearning: Recent Advances, Challenges, and Future Perspectives (Link) 2025

- Aobo Chen, Wei Qian, **Zheyuan Liu**, Shagufta Mehnaz, Tianhao Wang, Mengdi Huai
- Tutorial at ICDM 2025, Washington DC, USA.

Workshop Organizer

[W1] Socially Responsible and Trustworthy Foundation Models (Link) 2025

- Canyu Chen, Yue Huang, **Zheyuan Liu**, Yilun Zhao, Zhaorun Chen, Haoyue Bai, Xuandong Zhao, Yiyou Sun, Junyuan Hong, Xuefeng (Sean) Du, Jindong Gu, Arman Cohan, Xiangliang Zhang, Manling Li
- Workshop at NeurIPS 2025, Mexico City, Mexico.

[W2] Multilinguality in the Era of Large Language Models (Link) 2026

- Fengran Mo, Xue Zhang, Pinzhen Chen, Kaiyu Huang, **Zheyuan Liu**, Zhiqi Huang, Koel Dutta Chowdhury, Barry Haddow, Meng Jiang, Alexandra Birch, Jian-Yun Nie
- Workshop at ACL 2026, San Diego, USA.

[W3] Physical AI Working Group (Link) 2026

- Student Organizers: **Zheyuan Liu**, Adedayo Jigida
- Faculties: Meng Jiang, James Schmiedeler, Jane Cleland-Huang, Mengxue Hou, Fanxin Kong, Hai Lin, Margaret McGuinness, Patrick Wensing
- Workshop at University of Notre Dame, Notre Dame, USA.

Current Mentored Students

1. **Mingyue Cui,** *Undergraduate @ PolyU, Hongkong* (2026)
2. **Jinchen Zhang,** *Undergraduate @ Tsinghua University, China* (2026)
3. **Zhiguang (Justin) Han,** *Undergraduate @ NTU, Singapore* (2026)
4. **Haiwen (Heaven) Yi,** *Undergraduate @ U Toronto, Canada* (2026)

Past Mentored Students

1. **Baicheng Chen,** *Undergraduate @ CUHK, Shenzhen* (2026)
2. **John Kim,** *Undergraduate @ Notre Dame > CS PhD @ Northwestern University (NSF GRFP Fellowship)* (2025-2026)
3. **Katherine O'Roark**, *Undergraduate @ Saint Mary's College* (2025-2026)
4. **Han Yan**, *Undergraduate @ CUHK -> CS PhD @ NYU Shanghai* (2025-2026)
5. **Wenhan Wu**, *MS @ Wuhan University -> CS PhD @ Georgia Tech* (2025-2026)
6. **Diancheng Kang,** *Undergraduate @ Southern University of Science and Technology* (2025-2026)
7. **Yiru Wang**, *PhD Student @ CUHK, Hongkong* (2025-2026)

Grants and Gifts

"Robust Multimodal Hazard Identification for Safety-Critical Embodied AI"
May 4, 2026 – May 4, 2027
Funding Vehicle: Lambda Research Grant
Amount: **$2,000**

"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

Other Writing Experiences

NSF SaTC, Schmidt Sciences, Google Research Awards