Zirui (Ray) Liu

✉zrliu@umn.edu Google scholar LinkedIn profile Home page

Research Interest

Machine Learning Algorithm & System
- LLM efficiency & scalability
- Large-scale graph learning

Core problems in LLMs
- Retrieval augmented generation and long-context ability
- LLM safety

Education

- **Rice University** — Houston, TX
  Ph.D. Candidate in Computer Science, Co-advised by Dr. Xia (Ben) Hu and Prof. Vladimir Braverman — *Aug. 2021 – Aug. 2024*
- **Texas A&M University (Transfer Out)** — College Station, TX
  Ph.D. Student in Computer Science, Advised by Dr. Xia (Ben) Hu — *Aug. 2019 – Aug. 2021*
- **Harbin Institute of Technology** — Harbin, China
  Electrical Engineering, Master of Engineering — *Sep. 2016 – Jun. 2018*
- **Harbin Institute of Technology** — Harbin, China
  Electrical Engineering, Bachelor of Science — *Sep. 2012 – Jul. 2016*

Publication

1. [**ICML'24**] **Zirui Liu**, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, Xia Hu "*KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache* ". The 41st International Conference on Machine Learning, 2024 (Acceptance rate: 27%). PDF

2. [**ICML'24**] **Zirui Liu**, Zhaozhuo Xu, Beidi Chen, Yuxin Tang, Jue Wang, Kaixiong Zhou, Xia Hu, Anshumali Shrivastava "*Compress, Then Prompt: Improving Accuracy-Efficiency Trade-off of LLM Inference with Transferable Prompt* ". The 41st International Conference on Machine Learning, 2024 (Acceptance rate: 27%). PDF

3. [**ICML'24**] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, **Zirui Liu**, Chia-Yuan Chang, Huiyuan Chen, Xia Hu "*LLM Maybe LongLM: Self-Extend LLM Context Window Without Tuning* ". The 41st International Conference on Machine Learning, 2024 (Spotlight. Acceptance rate: 3.5%.) PDF

4. [**ICML'24**] Shaochen Zhong, Duy Le, **Zirui Liu**, Zhimeng Jiang, Andrew Ye, Jiamu Zhang, Jiayi Yuan, Kaixiong Zhou, Zhaozhuo Xu, Jing Ma, Shuai Xu, Vipin Chaudhary, Xia Hu "*GNNs Also Deserve Editing, and They Need It More Than Once*" . The 41st International Conference on Machine Learning, 2024 (Acceptance rate: 27%). PDF

5. [**ICML'24**] Duy Le, Shaochen Zhong, **Zirui Liu**, Shuai Xu, Vipin Chaudhary, Kaixiong Zhou, Zhaozhuo Xu "*Knowledge Graphs Can be Learned with Just Intersection Features*". The 41st International Conference on Machine Learning, 2024 (Acceptance rate: 27%). PDF

6. [**ICML'24**] Guanchu Wang, Yu-Neng Chuang, Fan Yang, Mengnan Du, Chia-Yuan Chang, Shaochen Zhong, **Zirui Liu**, Zhaozhuo Xu, Kaixiong Zhou, Xuanting Cai, Xia Hu "*TVE: Learning Meta-attribution for*

Transferable Vision Explainer ". The 41st International Conference on Machine Learning, 2024 (Acceptance rate: 27%). PDF

7. [**NeurIPS'23**] **Zirui Liu**, Guanchu Wang, Shaochen Zhong, Zhaozhuo Xu, Daochen Zha, Ruixiang Tang, Zhimeng Jiang, Kaixiong Zhou, Vipin Chaudhary, Shuai Xu, and Xia Hu"*Winner-Take-All Column Row Sampling for Memory Efficient Adaptation of Language Model* ", The 37th Conference on Neural Information Processing Systems, 2023 (Acceptance rate: 26%). PDF

8. [**NeurIPS'23**] Ruixiang Tang, Jiayi Yuan, Yiming Li, **Zirui Liu**, Rui Chen, and Xia Hu"*Setting the Trap: Capturing and Defeating Backdoor Threats in PLMs through Honeypots* ", The 37th Conference on Neural Information Processing Systems, 2023 (Acceptance rate: 26%). PDF

9. [**NeurIPS'23**] Shaochen Zhong, Zaichuan You, Jiamu Zhang, Sebastian Zhao, Zachary LeClaire, **Zirui Liu**, Vipin Chaudhary, Shuai Xu, and Xia Hu"*One Less Reason for Filter Pruning: Gaining Free Adversarial Robustness with Structured Grouped Kernel Pruning*", The 37th Conference on Neural Information Processing Systems, 2023 (Acceptance rate: 26%). PDF

10. [**ICML'23**] **Zirui Liu**, Shengyuan Chen, Kaixiong Zhou, Daochen Zha, Xiao Huang, and Xia Hu. "*RSC: Accelerating Graph Neural Networks Training via Randomized Sparse Computations*", The 40th International Conference on Machine Learning, 2023. (Acceptance rate: 28%). PDF

11. [**ICML'23**] Guanchu Wang, **Zirui Liu**, Zhimeng Jiang, Ninghao Liu, Na Zou, and Xia Hu. "*DIVISION: Memory Efficient Training via Dual Activation Precision*", The 40th International Conference on Machine Learning, 2023 (Acceptance rate: 28%). PDF

12. [**MLSys'23**] Daochen Zha, Louis Feng, Liang Luo, Bhargav Bhushanam, **Zirui Liu**, Yusuo Hu, Jade Nie, Yuzhen Huang, Yuandong Tian, Arun Kejariwal, and Xia Hu. "*Pre-train and Search: Efficient Embedding Table Sharding with Pre-trained Neural Cost Models*", The 6th Conference on Machine Learning and Systems, 2023 (Acceptance Rate: 22%). PDF

13. [**ICLR'22**] **Zirui Liu**, Kaixiong Zhou, Fan Yang, Li Li, Rui Chen, and Xia Hu. "*EXACT: Scalable Graph Neural Networks Training via Extreme Activation Compression*", The 10th International Conference on Learning Representations, 2022 (Acceptance Rate: 33%). PDF

14. [**ICLR'22**] Zhimeng Jiang, Kaixiong Zhou, **Zirui Liu**, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. "*An Information Fusion Approach to Learning with Instance-Dependent Label Noise*", The 10th International Conference on Learning Representations, 2022 (Acceptance Rate: 33%). PDF

15. [**NeurIPS'22**] Daochen Zha, Louis Feng, Qiaoyu Tan, **Zirui Liu**, Kwei-Herng Lai, Bhargav Bhushanam, Yuandong Tian, Arun Kejariwal, and Xia Hu. " *DreamShard: Generalizable Embedding Table Placement for Recommender Systems*", The 36th Conference on Neural Information Processing Systems, 2022 (Acceptance Rate: 26%). PDF

16. [**NeurIPS'22**] Keyu Duan, **Zirui Liu**, Wenqing Zheng, Peihao Wang, Kaixiong Zhou, Tianlong Chen, Zhangyang Wang and Xia Hu. "*A Comprehensive Study on Large-Scale Graph Training: Benchmarking and Rethinking*", The 36th Conference on Neural Information Processing Systems, 2022 (Acceptance Rate: 26%). PDF

17. [**KDD MLG'23**] **Zirui Liu**, Zhimeng Jiang, Shaochen Zhong, Kaixiong Zhou, Li Li, Rui Chen, Soo-Hyun Choi and Xia Hu. "*Editable Graph Neural Network for Node Classifications*", The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining on Mining and Learning with Graphs Workshop, 2023. PDF

18. [**TMLR'23**] **Zirui Liu**, Kaixiong Zhou, Zhimeng Jiang, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. "*DSpar: Embarrassingly Simple Strategy for Efficient GNN training and inference via Degree-based Sparsification*", Transactions on Machine Learning Research, 2023 (Acceptance Rate: 62%). PDF

19. [**TMLR'23**] Xiaotian Han, Zhimeng Jiang, Hongye Jin, **Zirui Liu**, Na Zou, Qifan Wang, Xia Hu. "*Retiring ΔDP: New Distribution-Level Metrics for Demographic Parity*", Transactions on Machine Learning Research, 2023 (Acceptance Rate: 62%). PDF

20. [**IJCAI'22**] Kaixiong Zhou, **Zirui Liu**, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. "*Table2Graph: Transforming Tabular Data to Unified Weighted Graph*", The 31st International Joint Conference on Artificial Intelligence, 2022 (Acceptance Rate: 15%). PDF

21. [**WWW'21**] Ruixiang Tang, Mengnan Du, Yuening Li, **Zirui Liu** and Xia Hu. "*Mitigating Gender Bias in Captioning Systems*", The 32nd Web Conference, 2021 (Acceptance Rate: 20%). PDF

22. [**ICCV'21**] **Zirui Liu**, Haifeng Jin, Ting-Hsiang Wang, Kaixiong Zhou and Xia Hu. "*DivAug: Plug-in Automated Data Augmentation with Explicit Diversity Maximization*", The 18th International Conference on Computer Vision, 2021 (Acceptance Rate: 26%). PDF

23. [**NeurIPS'21**] **Zirui Liu**, Qingquan Song, Kaixiong Zhou, Ting-Hsiang Wang, Ying Shan, and Xia Hu. "*Detecting Interactions from Neural Networks via Topological Analysis*", The 35th Conference on Neural Information Processing Systems, 2021 (Acceptance Rate: 26%). PDF

24. [**RecSys'21**] Ting-Hsiang Wang , Qingquan Song, Xiaotian Han, **Zirui Liu**, Haifeng Jin, and Xia Hu. "*AutoRec: An Automated Recommender System*", The 15th ACM Conference on Recommender Systems (Demo track), 2021. PDF

Research & Industry Experiences

- **Research Intern at Linkedin** — Sunnyvale, CA
  Advertisement AI team, advised by Mr. Aman Gupta. — *May. 2023 – Aug. 2023*
  - Improving the inference efficiency of language models via dynamic sparsity.

- **Research Intern at Meta** — Sunnyvale, CA
  Capacity Engineering and Analysis team. — *May. 2022 – Aug. 2022*
  - Incorporating semi-supervised learning for improving advertisement models.

- **Research Intern at Samsung Research America** — Remote
  Advertisement AI team, advised by Dr. Li Li and Dr. Rui Chen. — *May. 2021 – Aug. 2021*
  - Incorporating the graph structure into the App embeddings for improving the click-through-rate.

- **Research Intern at 4Paradigm** — Beijing, China
  Advised by Mr. Xiawei Guo. — *Mar. 2019 – Aug. 2019*
  - Developed the automated machine learning system for tabular data.

Recognition & Awards

- Rice Graduate Fellowship — 2021
- The GMCC's Scholarship (Corporation Scholarship) — 2015
- The MOONS' Scholarship (Corporation Scholarship) — 2013
- First Prize of Excellent Student Scholarship of HIT — 2012, 2013

Teaching

- Guest Lecturer, COMP 640: Graduate Seminar In Machine Learning, Rice University — 2023
- Guest Lecturer, COMP 631: Introduction to Information Retrieval, Rice University — 2023
- Teaching Assistant, COMP 631: Introduction to Information Retrieval, Rice University — 2022, 2023

Selected Mentorship

- **Keyu Duan** — **Fall 2021 – Present**
  Ph.D. Student, National University of Singapore — *Large-Scale Machine Learning, Neurips 2022*
- **Guanchu Wang** — **Fall 2021 – Present**
  Ph.D. Student, Rice University — *Memory-Efficient Machine Learning, ICML 2023*
- **Shengyuan Chen** — **Summer 2022 – Present**
  Ph.D. Student, The Hong Kong Polytechnic University — *Efficient Machine Learning, ICML 2023*
- **Henry Zhong** — **Fall 2022 – Present**
  Ph.D. Student, Rice University — *Efficient Machine Learning, NeurIPS 2023*
- **Jiayi Yuan** — **Fall 2022 – Present**
  Ph.D. Student, Rice University — *Machine Learning for Healthcare, AMIA*
- **Duy Le** — **Fall 2023 – Present**
  Undergraduate, Case Western Reserve University — *Large-Scale Machine Learning, SDM'24 in submission*

Service

- **Program Committee Member**: NeurIPS, ICML, ICLR, KDD, ICCV
- **Journal Reviewers**: TPAMI, TKDD, TMLR