Zirui (Ray) Liu

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

Research Interest

Long-Term Memory System
– Long context ability, reasoning, and retrieve
Efficient Machine Learning & MLSys
– Algorithm and system support for tuning and deploying foundation models

Appointments

- **Assistant Professor at University of Minnesota** — Minneapolis, MN
  Computer Science Department — *Aug. 2024 (Now)*
- **Research Intern at Linkedin** — Sunnyvale, CA
  Core AI team. — *May. 2023 – Aug. 2023*
- **Research Intern at Meta** — Sunnyvale, CA
  Capacity Engineering and Analysis team. — *May. 2022 – Aug. 2022*
- **Research Intern at Samsung Research America** — Remote
  Advertisement AI team. — *May. 2021 – Aug. 2021*

Education

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

Project Impact & Recognition

- KV Cache quantization framework used in Hugginface [Public Doc] 2024
- Long context extension method deployed in Llama.cpp [Official Pull Request] 2024
- Google I/O session research highlight [YouTube Link] 2024
- RL-based RecSys model sharding system deployed in Meta Production [Public Report] 2022

Publications

- **[Neurips'25]** Jiayi Yuan*, Hao Li*, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, **Zirui Liu**. "*Give Me FP32 or Give Me Death? Challenges and Solutions for Reproducible Reasoning* ", The Conference on Neural Information Processing Systems, 2025. (Oral. Talk and Slides)PDF

- **[EMNLP'25]** Wenya Xie*, Shaochen Zhong*, Hoang Anh Duy Le, Zhaozhuo Xu, Jianwen Xie, **Zirui Liu**. "*Word Salad Chopper: Reasoning Models Waste A Ton Of Decoding Budget On Useless Repetitions, Self-Knowingly for Reproducible Reasoning* ", The 2025 Conference on Empirical Methods in Natural Language Processing (Oral.)

- **[ICML'25]** Mingyu Jin*, Kai Mei*, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, **Zirui liu***, Yongfeng Zhang* "*Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding* ", The 14th International Conference on Mahcine Learning, 2025. (Acceptance rate: 27%) PDF

- **[ICLR'25]** Zeru Shi, Kai Mei, Mingyu Jin, Yongye Su, Chaoji Zuo, Wenyue Hua, Wujiang Xu, Yujie Ren, **Zirui Liu**, Mengnan Du, Dong Deng, Yongfeng Zhang "*From Commands to Prompts: LLM-based Semantic File System for AIOS* ", The 13th International Conference on Learning Representations, 2025. (Acceptance rate: 31%) PDF

- **[ICLR'25]** Wentao Guo, Jikai Long, Yimeng Zeng, **Zirui Liu**, Xinyu Yang, Yide Ran, Jacob R. Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, Beidi Chen, Zhaozhuo Xu. "*Zeroth-Order Fine-Tuning of LLMs with Transferable Static Sparsity*", The 13th International Conference on Learning Representations, 2025. (Acceptance rate: 31%) PDF

- **[NeurIPS'24]** Jimeng Jiang, **Zirui Liu**, Xiaotian Han, Qizhang Feng, Hongye Jin, Qiaoyu Tan, Kaixiong Zhou, Na Zou, Xia Hu. "*Gradient Rewiring for Editable Graph Neural Network Training*", The 38th Conference on Neural Information Processing Systems, 2024. (Acceptance rate: 25.8%) PDF

- **[EMNLP'24]** Guanchu Wang, Yu-Neng Chuang, Ruixiang Tang, Shaochen Zhong, Jiayi Yuan, Hongye Jin, **Zirui Liu**, Vipin Chaudhary, Shuai Xu, James Caverlee, Xia Hu. "*Secured Weight Release for Large Language Models via Taylor Expansion*". The 2024 Conference on Empirical Methods in Natural Language Processing. (Acceptance rate: 20.8%) PDF

- **[EMNLP'24] (Finding)** Jiayi Yuan* , Hongyi Liu*, Shaochen Zhong*, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, **Zirui Liu**, Xia Hu "*KV Cache Compression, But What Must We Give in Return? A Comprehensive Benchmark of Long Context Capable Approaches*". The 2024 Conference on Empirical Methods in Natural Language Processing. (Acceptance rate: 37.7%) PDF

- **[EMNLP'24] (Finding)** Chuang Zhou, Junnan Dong, Xiao Huang, **Zirui Liu**, Kaixiong Zhou, Zhaozhuo Xu "*QUEST: Efficient Extreme Multi-Label Text Classification with Large Language Models on Commodity Hardware*". The 2024 Conference on Empirical Methods in Natural Language Processing. (Acceptance rate: 37.7%) PDF

- **[EMNLP'24] (Finding)**Junda Su, **Zirui Liu**, Zeju Qiu, Weiyang Liu, Zhaozhuo Xu "*In Defense of Structural Sparse Adapters for Concurrent LLM Serving* ". The 2024 Conference on Empirical Methods in Natural Language Processing. (Acceptance rate: 37.7%) PDF

- **[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%, Used in Huggingface Transformers). PDF

- **[ICML'24]** Zhaozhuo Xu*, **Zirui Liu***, 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

- **[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%. Used in Llama.cpp, highlighted during Google I/O Session) PDF

- **[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

- **[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

- **[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

- **[NAACL'24]** Chuang, Yu-Neng, Tianwei Xing, Chia-Yuan Chang, **Zirui Liu**, Xun Chen, and Xia Hu "*Learning to Compress Prompt in Natural Language Formats* ". The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024 (Acceptance rate: 23.2%). PDF

- **[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

- **[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

- **[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

- **[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

- **[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

- **[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

- **[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

- **[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

- **[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

- [**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

- [**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

- [**SDM'23**] Kaixiong Zhou, Soo-Hyun Choi, **Zirui Liu**, Ninghao Liu, Fan Yang, Rui Chen, Li Li, Xia Hu. "*Adaptive label smoothing to regularize large-scale graph training*", SIAM International Conference on Data Mining, 2023 (Acceptance Rate: 27%). PDF

- [**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

- [**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

- [**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

- [**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

- [**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

- [**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

- [**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

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
- Instructor, CSCI 8980: Large Language Model System, University of Minnesota 2025 Spring

Professional Services

- **NSF IIS/III panel reviewer**
- **Program Committee Member**: NeurIPS, ICML, ICLR, KDD, ICCV
- **Journal Reviewers**: TPAMI, TKDD, TMLR