JIAMU **ZHANG**

216-280-7699 | **mz81@rice.edu**, jxz1217@case.edu| Google Scolar | LinkedIn Profile | Personal Webpage

Summary

Ph.D. student in Computer Science focusing on efficient and scalable deep learning and agentic systems, with interests in model compression, system-aware efficiency optimization, and evaluation.

Education

Ph.D. Student in Computer Science — **08/2024 - Present**
Rice University — Houston, TX

Bachelor in Computer Science — **05/2024**
Case Western Reserve University — Cleveland, OH

Awards

Academic Awards: — **05/2024**
- Dean's High Honor List & Latin Honors — Cleveland, OH

Research Awards: — **05/2024**
- The Computer and Data Science Research Award — Cleveland, OH

Scholarship / Fellowship: — **01/2024**
- Swanger Graduate Fellowship — Cleveland, OH

Research Interest

Efficient Models and Inference Systems

- System-aware model compression and algorithm–implementation co-design for efficient inference
- Structured pruning, sparsification, and quantization for compute- and memory-efficient models
- Adaptive computation and scalable architectures (e.g., Mixture-of-Experts, conditional routing)

Agentic and Reasoning-Centric Systems

- Efficient model reasoning and test-time computation control
- Agentic and multi-agent system–aware compression and optimization
- Reducing redundant computation and improving inter-agent communication efficiency
- Evaluation of LLM/Agentic systems beyond accuracy

Publication

- **[CVPR'25] Jiamu Zhang***, Shaochen Zhong*, Andrew Ye, Zirui Liu, Sebastian Zhao, Kaixiong Zhou, Li Li, Soo-Hyun Choi, Rui Chen, Xia Hu, Shuai Xu, Vipin Chaudhary. "*Flexible Group Count Enables Hassle-Free Structured Pruning*", Conference on Computer Vision and Pattern Recognition, 2025. (Acceptance rate: 22.1%) PDF
- **[ACL'25** (Findings)**] Jiamu Zhang**, Jiayi Yuan, Andrew Wen, Hoang Anh Duy Le, Yu-Neng Chuang, Soo-Hyun Choi, Rui Chen, Xia Hu. "*ReasonerRank: Redefining Language Model Evaluation with Ground-Truth-Free Ranking Frameworks*", The 63rd Annual Meeting of the Association for Computational Linguistics, 2025. PDF
- **[Under Review] Jiamu Zhang**, Alessandro Mason, Ning Xie, Aarav Swami, Ashley Chen, Shuai Xu, Vipin Chaudhary, Hanjie Chen. "*Recognizing and Restructuring Latent Experts for Model Compression*"
- **[Under Review] Jiamu Zhang*,** Shaochen Zhong***,** Hoang Anh Duy Le, Wenya Xie, Yifan Lu, Xintong Sun, Mohsen Hariri, Hongyi Liu, Guanchu Wang, Zhaozhuo Xu, Jiarong Xing, Zirui Liu, Shuai Xu, Ning Xie, Li Li, Rui Chen, Ruixiang Tang, Vipin Chaudhary, Xia Hu. "*Sweeping Promptable Spoofs under the DirtyRAG: A Practical, Query-Blind RAG Attack*"
- [**IEEE SPW'25**] **Jiamu Zhang**, Shaochen Zhong, Hoang Anh Duy Le, Xia Hu. "*In-Progress: Structured Pruning in the Wild: Benchmarking Practical Robustness Under Real-World Corruptions*", 2025 IEEE Security and Privacy Workshops (SPW). IEEE, 2025. PDF

- **[TMLR'25]** Yang Sui, Yu-Neng Chuang, Guanchu Wang, **Jiamu Zhang**, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, Xia Hu. "*Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models*", arXiv:2503.16419 [cs.CL] 23 Apr 2025. PDF
- **[ICML'24]** Shaochen Zhong*, Hoang Anh 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 41th International Conference on Machine Learning, 2024. PDF
- **[NeurIPS'23]** Shaochen Zhong, **Jiamu Zhang*,** Zaichuan You*, Sebastian Zhao*, Zachary LeClaire, Zirui Liu, Vipin Chaudhary, Shuai Xu, 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. PDF
- [**npj Digital Medicine**] Andrew Wen, Qiuhao Lu, Yu-Neng Chuang, Guanchu Wang, Jiayi Yuan, **Jiamu Zhang**, Liwei Wang, Sunyang Fu, Kurt D Miller, Heling Jia, Steven D Bedrick, William R Hersh, Kirk E Roberts, Xia Hu, Hongfang Liu. "*Context Matching is not Reasoning: Assessing Generalized Evaluation of Generative Language Models in Clinical Settings*", preprint: PMC12408041 [cs.AI] PDF
- **[arXiv]**Jiayi Yuan, **Jiamu Zhang**, Andrew Wen, Xia Hu. "*The Science of Evaluating Foundation Models*", arXiv:2502.09670v1 [cs.CL] 12 Feb 2025. PDF
- **[Under Review]** Hoang Anh Duy Le, Shaochen Zhong, Jerry Xiao, **Jiamu Zhang**, Yu-Neng Chuang, Li Li, Rui Chen, Shuai Xu, Zirui Liu, Kaixiong Zhou, Vipin Chaudhary, Zhaozhuo Xu, Xia Hu. "*Graph Transformers Get the GIST: Graph Invariant Structural Trait for Refined Graph Encoding*", OpenReview: Ck6WljG6ZM.

Professional Services

- **Conference Reviewers**
  - NeurIPS
- **Journal Reviewers**
  - IEEE TPAMI, IEEE TCDS, ACM TIST, ACM TCH, and Neurocomputing

Teaching

Teaching Assistant

- COMP 584: Natural Language Processing — 2026 Spring
- COMP 631: Introduction to Information Retrieval. Rice University — 2025 Spring
- CSDS 302: Discrete Mathematics. Case Western Reserve University — 2022, 2023, 2024
- CSDS 386: Quantum Computing, Information, and Devices. Case Western Reserve University — 2023 Spring