

JIAMU (MORRIS) ZHANG

Houston, TX | 216-280-7699 | mz81@rice.edu | MorrisZJ.github.io | [Google Scholar](#) | [LinkedIn](#) | [GitHub](#)

RESEARCH

Efficient and scalable LLM and agentic systems — *system–method co-design* — and rigorous ways to measure whether the efficiency is real.

EDUCATION

Ph.D. in Computer Science

08/2024 – Present

Rice University, Houston, TX

- Advisor: Prof. Hanjie Chen, Chili Lab (08/2025 – Present)
- Advisor: Prof. Xia “Ben” Hu, Data Lab (08/2024 – 08/2025)

B.S. in Computer Science

05/2024

Case Western Reserve University, Cleveland, OH

- Department of Computer and Data Science. Advisors: Prof. Shuai Xu and Prof. Vipin Chaudhary.

RESEARCH INTERESTS

- **Memory-bound LLM inference.** KV cache compression, quantization and transform coding for long-context and agentic workloads.
- **Conditional computation.** Mixture-of-Experts and training-free routing that turn static redundancy in pretrained models into input-dependent sparsity — and the serving systems that realize it under real memory budgets.
- **Efficient and verifiable agentic systems.** Cutting redundant computation and communication in LLM agents and multi-agent systems — and rigorously evaluating whether reported efficiency gains are real.

EXPERIENCE

Research Intern, AI Research & Development

05/2026 – Present

Nokia, Sunnyvale, CA

- Efficiency and system–method co-design for large-model inference; led the design and open-source release of AnyJev (below).

OPEN SOURCE

AnyJev — lead author and principal maintainer

1,492★ · 171 forks

2026

*Nokia Applied Research · github.com/nokia-applied-research/AnyJev · `pip install anyjev` · [technical report](#)
[arXiv:2610.00831](#)*

- Turns any LLM into a typed decision model with calibrated probabilities, training-free by default — served through vLLM as a pooling endpoint plus a few kilobytes of closed-form head.
- Cuts option-order sensitivity 0.230→0.073 and calibration error 0.240→0.095 with zero labels; raises the share auto-decidable at 5% risk from 7.7% to 52.0% with 100–500 labels.
- Passed 1,000 GitHub stars in its first ten days; released under Apache-2.0 with a published PyPI package.
- Highest-placed training-free system on the community Jev Decision Index (v0.3.2): joint 4th of 148, ahead of the Jev reference itself.

PUBLICATIONS

EMNLP'26

Rethinking the Evaluation of Efficiency Methods for Multi-Agent Systems Outstanding Paper Award

Jiamu Zhang*, Lingxi Zhang*, Pengjun Lu*, Qiyue Zhang, Yu-Neng Chuang, Zhengchen Li, Shuai Xu, Vipin Chaudhary, Hanjie Chen.

Conference on Empirical Methods in Natural Language Processing (EMNLP), Main Conference, Oral, 2026. [arXiv]

CVPR'25

Flexible Group Count Enables Hassle-Free Structured Pruning

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.

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. Acceptance rate 22.1%. [PDF]

Preprint

BellowsKV: One Latent for Many Layers, Expanded on Demand

Jiamu Zhang, Hanjie Chen.

arXiv:2610.35023, 2026. [arXiv]

Preprint

SPECTRA: Pushing the KV Cache Beyond the 2-Bit Cliff via Spectral Transform Coding

Jiamu Zhang, Liang Wu, Kelly Wan, Hanjie Chen, Liangjie Hong.

arXiv:2608.07915, 2026. [arXiv]

In Review

Recognizing and Restructuring Latent Experts for Large Language Model Compression

Jiamu Zhang, Alessandro Mason, Ning Xie, Aarav Swami, Ashley Chen, Shuai Xu, Vipin Chaudhary, Hanjie Chen.

Under review, 2026.

Preprint

WiSP: A Working-Set View of Mixture-of-Experts Serving on Extremely Low-Resource Hardware

Jiamu Zhang, Liang Wu, Mayank Darbari, Liangjie Hong.

arXiv:2606.21868, 2026. [arXiv]

ACL'25

ReasonerRank: Redefining Language Model Evaluation with Ground-Truth-Free Ranking Frameworks

Jiamu Zhang, Jiayi Yuan, Andrew Wen, Hoang Anh Duy Le, Yu-Neng Chuang, Soo-Hyun Choi, Rui Chen, Xia Hu.

Findings of the Association for Computational Linguistics: ACL 2025, pp. 13623–13639. [PDF]

NeurIPS'23

One Less Reason for Filter Pruning: Gaining Free Adversarial Robustness with Structured Grouped Kernel Pruning

Shaochen Zhong, **Jiamu Zhang***, Zaichuan You*, Sebastian Zhao*, Zachary LeClaire, Zirui Liu, Vipin Chaudhary, Shuai Xu, Xia Hu.

Advances in Neural Information Processing Systems (NeurIPS), 2023. [PDF]

arXiv

The Science of Evaluating Foundation Models

Jiayi Yuan, **Jiamu Zhang**, Andrew Wen, Xia Hu.

arXiv:2502.09670, 2025. [arXiv]

TMLR'25

Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models 603 citations · 791★

Yang Sui, Yu-Neng Chuang, Guanchu Wang, **Jiamu Zhang**, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, Xia Hu.

Transactions on Machine Learning Research (TMLR), 2025. [arXiv][Code]

ICML'24

GNNs Also Deserve Editing, and They Need It More Than Once

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.

International Conference on Machine Learning (ICML), 2024. [PDF]

Context Matching is not Reasoning: Assessing Generalized Evaluation of Generative Language Models in Clinical Settings

Andrew Wen, Qiuhaio 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.

npj Digital Medicine, 2025.

* denotes equal contribution. Citation counts from Google Scholar; star and download counts from GitHub and PyPI. Figures as of October 2026.

HONORS AND AWARDS

Outstanding Paper Award — EMNLP 2026	2026
Swanger Graduate Fellowship — Case Western Reserve University	2024
The Computer and Data Science Research Award — Case Western Reserve University	2024
Dean's High Honor List & Latin Honors — Case Western Reserve University	2024

PROFESSIONAL SERVICE

Conference Reviewer — NeurIPS, ICML, ACL

Journal Reviewer — IEEE TPAMI, IEEE TCDS, ACM TIST, ACM TCH, Neurocomputing

Languages — English, Mandarin Chinese

TEACHING

Guest Lecture, COMP 640 — Graduate Seminar in ML: Large Language Models · Rice University	2026 Fall
“A Simple Intro to Test-Time Scaling & Reasoning: Thinking longer, thinking better”	
TA, COMP 584 — Natural Language Processing · Rice University	2026 Spring
TA, COMP 631 — Introduction to Information Retrieval · Rice University	2025 Spring
TA, CSDS 302 — Discrete Mathematics · Case Western Reserve University	2022, 2023, 2024
TA, CSDS 386 — Quantum Computing, Information, and Devices · Case Western Reserve University	2023 Spring