Jiayi Yuan

Website: https://jy-yuan.github.io/ ◇ Email: jy101@rice.edu

RESEARCH INTERESTS AND HIGHLIGHTS

I aim to build **efficient machine learning algorithms and systems (MLSys)** through methods like *quantization*, *sparsity*, *re-parameterization*, while enhancing system **robustness** and **security**.
Recently, I have been focusing on **LLM post-training**: efficiency, long-context, agent, multimodal.

Highlights:

- **20+ papers** published in prestigious venues (NeurIPS, ICML, ACL, EMNLP, ISCA, etc.) [Google Scholar]
- Our **Nondeterminism** work (*NeurIPS 25' Oral*) became a heated topic: e.g. featured in Thinking Machines Lab's blog.
- **KIVI** (*ICML 24'*) used in KV cache quantization in Huggingface Transformers. [Code]
- **LLM-PTM**: patient-trial matching using LLMs wins AMIA 2023 Best Student Paper Award. [Paper]
- **Stop Overthinking**: first comprehensive survey on efficient reasoning for LLMs. [GitHub]
- IEEE Micro Top Pick 2023 and multiple other awards.

EDUCATION

Rice University — Houston, TX
Ph.D. in Computer Science (Advisor: Dr. Xia "Ben" Hu) — *Aug. 2022 - Present*

Tsinghua University — Beijing, China
B.Eng. in Computer Science — *Aug. 2017 - Jul. 2021*

PROFESSIONAL EXPERIENCES

NVIDIA — Santa Clara, CA
Deep Learning Intern (Mentors: Huizi Mao, Kai Xu) — *May 2025 - Present*

- In Modelopt team: prototyping, developing, and optimizing model-optimization methods and platforms.
- Proposed a universal drop-in sparse attention mechanism for both training and inference.

Amazon — Seattle, WA
Applied Scientist Intern (Mentors: Na Xu, Yang Liu) — *May 2024 - Aug. 2024*

- Developed and implemented LLM agents for Amazon stores, focusing on product recommendations and customer interactions, with the goal of fully automating operations that's over $3M Opex.
- Built a RAG-enhanced multimodal ICL pipeline, increasing the model's accuracy from 60% to 90%+.

PUBLICATIONS

Conference Publications

[C1] *__J. Yuan__, *H. Li, X. Ding, W. Xie, Y. Li, W. Zhao, K. Wan, J. Shi, X. Hu, Z. Liu. "Give Me FP32 or Give Me Death? Challenges and Solutions for Reproducible Reasoning", *NeurIPS Oral, in the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025*

[C2] H. Liu, S. Zhong, X. Sun, M. Tian, M. Hariri, Z. Liu, R. Tang, Z. Jiang, **J. Yuan**, Y. Chuang, L. Li, S. Choi, R. Chen, V. Chaudhary, X. Hu. "LoRATK: LoRA Once, Backdoor Everywhere in the Share-and-Play Ecosystem", *in findings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP Findings), 2025*

[C3] J. Zhang, **J. Yuan**, A. Wen, D. Le, Y. Chuang, S. Choi, R. Chen, X. Hu. "ReasonerRank: Redefining Language Model Evaluation with Ground-Truth-Free Ranking Frameworks", *in findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL Findings), 2025*

[C4] *Y. Wang, ***J Yuan**, Y. Chuang, Z. Wang, Y. Liu, M. Cusick, P. Kulkarni, Z. Ji, Y. Ibrahim, X. Hu. "DHP Benchmark: Are LLMs Good NLG Evaluators?", *in findings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL Findings), 2025*

[C5] *__J. Yuan__, *H. Liu, *S. Zhong, Y. Chuang, S. Li, G. Wang, D. Le, H. Jin, V. Chaudhary, Z. Xu, Z. Liu, X. Hu. "KV Cache Compression, But What Must We Give in Return? A Comprehensive Benchmark of Long Context Capable Approaches", *in findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP Findings), 2024*

[C6] G. Wang, Y. Chuang, R. Tang, S. Zhong, __J. Yuan__, H. Jin, Z. Liu, V. Chaudhary, S. Xu, J. Caverlee, X. Hu. "Taylor Unswift: Secured Weight Release for Large Language Models via Taylor Expansion", *in the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024*

[C7] *__J. Yuan__, *Z. Liu, H. Jin, S. Zhong, Z. Xu, V. Braverman, B. Chen, X. Hu. "KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache", *in the Forty-first International Conference on Machine Learning (ICML), 2024*

[C8] S. Zhong, D. Le, Z. Liu, Z. Jiang, A. Ye, J. Zhang, __J. Yuan__, K. Zhou, Z. Xu, J. Ma, S. Xu, V. Chaudhary, X. Hu. "GNNs Also Deserve Editing, and They Need It More Than Once", *in the Forty-first International Conference on Machine Learning (ICML), 2024*

[C9] *__J. Yuan__, *R. Tang, Y. Li, Z. Liu, R. Chen, X. Hu. "Setting the Trap: Capturing and Defeating Backdoors in Pretrained Language Models through Honeypots", *in the Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023*

[C10] __J. Yuan__, R. Tang, X. Jiang, X. Hu. "Large language models for healthcare data augmentation: An example on patient-trial matching", *Best Student Paper, in AMIA Annual Symposium Proceedings (AMIA), 2023*

[C11] C. Chang, __J. Yuan__, S. Ding, Q. Tan, K. Zhang, X. Jiang, X. Hu, N. Zou. "Towards Fair Patient-Trial Matching via Patient-Criterion Level Fairness Constraint", *in AMIA Annual Symposium Proceedings (AMIA), 2023*

[C12] Q. Feng, __J. Yuan__, F.B. Emdad, K. Hanna, X. Hu, Z. He. "Can Attention Be Used to Explain EHR-Based Mortality Prediction Tasks: A Case Study on Hemorrhagic Stroke", *in the 14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB), 2023*

[C13] Z. Yu, Y. Fu, __J. Yuan__, H. You, Y. Lin. "NetBooster: Empowering Tiny Deep Learning By Standing on the Shoulders of Deep Giants", *in Proceedings of the 60th ACM/IEEE Design Automation Conference (DAC), 2023*

[C14] Y. Fu, Y. Yuan, S. Wu, __J. Yuan__, Y. Lin. "Robust Tickets Can Transfer Better: Drawing More Transferable Subnetworks in Transfer Learning", *in Proceedings of the 60th ACM/IEEE Design Automation Conference (DAC), 2023*

[C15] *Y. Fu, *Z. Ye, __J. Yuan__, S. Zhang, S. Li, H. You, Y. Lin. "Gen-NeRF: Efficient and Generalizable Neural Radiance Fields via Algorithm-Hardware Co-Design", *in the 50th IEEE/ACM International Symposium on Computer Architecture (ISCA), 2023*

[C16] *__J. Yuan__, *C. Li, *W. Chen, Y. Lin, A. Sabharwal. "ERSAM: Neural Architecture Search for Energy-Efficient and Real-Time Social Ambiance Measurement", *in 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023*

[C17] Y. Fu, H. Yang, __J. Yuan__, M. Li, C. Wan, R. Krishnamoorthi, V. Chandra, Y. Lin. "DepthShrinker: A New Compression Paradigm Towards Boosting Real-Hardware Efficiency of Compact Neural Networks", *in the Thirty-ninth International Conference on Machine Learning (ICML), 2022*

[C18] *H. You, *Y. Zhao, *Z. Yu, *C. Wan, Y. Fu, __J. Yuan__, S. Wu, S. Zhang, Y. Zhang, C. Li, V. Boominathan, A. Veeraraghavan, Z. Li, Y. Lin. "EyeCoD: Eye Tracking System Acceleration via FlatCam-Based Algorithm and Accelerator Co-Design", *IEEE Micro Top Pick, in the 49th IEEE/ACM International Symposium on Computer Architecture (ISCA), 2022*

Journal Publications

[J1] Y. Sui, Y. Chuang, G. Wang, J. Zhang, T. Zhang, __J. Yuan__, H. Liu, A. Wen, S. Zhong, H. Chen, X. Hu. "Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models", *in Transactions on Machine Learning Research (TMLR)*

[J2] A. Wen, Q. Lu, Y. Chuang, G. Wang, __J. Yuan__, J. Zhang, L. Wang, S. Fu, K.D. Miller, H. Jia, S.D. Bedrick, W.R. Hersh, K.E. Roberts, X. Hu, H. Liu. "Context Matching is not Reasoning: Assessing Generalized

Evaluation of Generative Language Models in Clinical Settings", *in npj Digital Medicine*

Preprints

[P1] **J. Yuan**, J. Zhang, A. Wen, X. Hu. "The Science of Evaluating Foundation Models"

[P2] *F. Luo, *Y. Chuang, G. Wang, H. Le, S. Zhong, H. Liu, **J. Yuan**, Y. Sui, V. Braverman, V. Chaudhary, X. Hu, "AutoL2S: Auto Long-Short Reasoning for Efficient Large Language Models"

[P3] X. Wu, **J. Yuan**, W. Yao, X. Zhai, N. Liu. "Interpreting and Steering LLMs with Mutual Information-based Explanations on Sparse Autoencoders"

[P4] *Y. Chuang, *S. Li, ***J. Yuan**, *G. Wang, *K. Lai, L. Yu, S. Ding, C. Chang, Q. Tan, D. Zha, X. Hu. "Understanding Different Design Choices in Training Large Time Series Models"

OTHER EXPERIENCES

Rice University Houston, TX
Research Assistant *Aug. 2021 - Present*

- Efficiency problems of long-context LLMs. [P2] [J1] [C5] [C7]
- LLM post-training: finetune, RL, and evaluation. [C1] [P1] [C3] [C4]
- RAG, LLM safety, LLM Agent, LLM Routing. [C2] [C6] [C9]
- Before LLM era: efficient machine learning. [C13] [C14] [C16] [C17]
- Applications in healthcare informatics. [C10] [C11] [C12]

Baidu Inc. Beijing, China
Research Engineer Intern *Dec. 2020 - Jul. 2021*

- Worked in Content Technology Architecture Group and took charge of processing large-scale data streams.
- Developed and optimized fingerprinting algorithms on massive real-world data. Focused on the image and video deduplication problems in both industry and academia.

Tsinghua University Beijing, China
Research Assistant *Jan. 2019 - May. 2021*

- Designed a diagnosis system for solar panel: computer vision for automated defect detection in the industry.
- Used generative models to improve Deepfake detection (forgery detection of face images).
- Built an efficient and highly scalable distributed approximate graph mining system. [Code]

HONORS AND AWARDS

NeurIPS 2025 Oral (77 out of 21575 submissions), *by NeurIPS* *Sep. 2025*
Rice Engineering Alumni Graduate Student Travel Grant, *by Rice University* *Oct. 2024*
D2K Research Mentoring Fellowship, *by Rice University* *Sep. 2024*
SDM'24 Doctoral Forum Travel Award, *by SIAM* *Mar. 2024*
NeurIPS 2023 Scholar Award, *by NeurIPS* *Nov. 2023*
AMIA Best Student Paper, *by AMIA* *Nov. 2023*
AMIA 2023 KDDM Student Innovation Award, *by AMIA* *Oct. 2023*
IEEE Micro Top Picks, *by ACM* *Jul. 2023*

MISC

Teaching

COMP 631 - Information Retrieval, *Guest Lecturer and Teaching Assistant*
COMP 640 - Graduate Research Seminar in Machine Learning, *Guest Lecturer*
COMP 556 - Introduction to Computer Networks, *Teaching Assistant*
COMP 549 - Applied Machine Learning & Data Science Projects, *Research Mentor*

Service

Reviewer: ICML, NeurIPS, ICLR, ACL, EMNLP, NAACL, AISTATS, AMIA, ICHI, TCDS