

# Jiayi Yuan

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

## EDUCATION

---

### Rice University

*Ph.D. in Computer Science* (Advisor: Dr. Xia “Ben” Hu)

Houston, TX

*Aug. 2022 - Present*

### Tsinghua University

*B.Eng. in Computer Science*

Beijing, China

*Aug. 2017 - Jul. 2021*

## RESEARCH INTERESTS

---

### Efficient and Trustworthy Machine Learning

Natural Language Processing (LLMs), Computer Vision, Health Informatics

## PUBLICATIONS

---

### Conference Publications

- [C1] \***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
- [C2] 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
- [C3] \***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
- [C4] 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
- [C5] \***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
- [C6] **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
- [C7] 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
- [C8] 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
- [C9] 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
- [C10] 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
- [C11] \*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

- [C12] \***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
- [C13] 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
- [C14] \*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

## Preprints

- [P1] \*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?”
- [P2] \*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”
- [P3] H. Liu, Z. Liu, R. Tang, **J. Yuan**, S. Zhong, Y. Chuang, L. Li, R. Chen, X. Hu. “LoRA-as-an-Attack! Piercing LLM Safety Under The Share-and-Play Scenario”
- [P4] C. Chang, S. Ding, **J. Yuan**, K. Zhang, X. Jiang, X. Hu, N. Zou. “Fair Patient-Trial Matching for Underrepresented Groups”

## EXPERIENCE

### Rice University

Houston, TX

#### Graduate Research Assistant

Sep. 2022 - Present

- Working on Large Language Models (LLMs): efficient and trustworthy finetuning and inference.
- LLM inference efficiency with KV cache compression. [C1] [C3]
- LLM safety: attack and defense. [C2] [C5] [P3]
- Worked on several projects regarding health informatics. [C6] [C7] [C8] [P4]

### Amazon.com, Inc.

Seattle, WA

#### Applied Scientist Intern

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%+.

### Rice University

Houston, TX

#### Research Assistant

Aug. 2021 - Aug. 2022

- Proposed re-parameterization-based efficient training and inference algorithms. [C9] [C13]
- Proposed a NAS pipeline for real-time social ambiance measurement. [C12]
- Worked on several projects regarding machine learning algorithms and systems co-design. [C11] [C14]
- Took part in efficient computer vision challenges: LPCVC-UAV, DAC-SDC.

### 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

---

Rice Engineering Alumni Graduate Student Travel Grant, <i>by Rice University</i>	<i>Oct. 2024</i>
D2K Research Mentoring Fellowship, <i>by Rice University</i>	<i>Sep. 2024</i>
SDM'24 Doctoral Forum Travel Award, <i>by SIAM</i>	<i>Mar. 2024</i>
NeurIPS 2023 Scholar Award, <i>by NeurIPS</i>	<i>Nov. 2023</i>
AMIA Best Student Paper, <i>by AMIA</i>	<i>Nov. 2023</i>
AMIA 2023 KDDM Student Innovation Award, <i>by AMIA</i>	<i>Oct. 2023</i>
IEEE Micro Top Picks, <i>by ACM</i>	<i>Jul. 2023</i>

## MISC

---

### Teaching

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

### Service

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