Yichuan Wang

yichuan_wang@berkeley.edu | yichuanmistygrass@gmail.com | yichuan-w.github.io

Education

University of California, Berkeley — Berkeley, California
Ph.D. in EECS, Sky Computing Lab — *Aug. 2024 - Present*
- Advised by Matei Zaharia & Joseph E. Gonzalez
- Closely collaborating with Sewon Min

Shanghai Jiao Tong University — Shanghai, China
B.Eng. in Computer Science, ACM Honors Class — *Sept. 2020 - Jun. 2024*

Research Interests

Machine Learning Systems; Large Language Models (LLMs) & Foundation Models; Agentic AI Systems; Scalable Vector Databases & Retrieval-Augmented Generation (RAG); Post-Training & Inference Optimization; Multi-Modal Embeddings & Semantic Search; Efficient Model Serving & Scheduling; GNN at Scale.

Experience

Shanghai Jiao Tong University — Shanghai, China
Undergraduate Researcher in Machine Learning Systems, advised by Prof. Quan Chen — *June 2022 - June 2024*

New York University — New York, NY, USA
Research Assistant in Machine Learning Systems, advised by Prof. Jinyang Li — *Jan. 2023 - Apr. 2024*

LMSYS (SGLang Project) — Remote
Major Developer — *Jul. 2024 - Dec. 2024*

Selected Open-Source Projects

LEANN – Low-Storage Vector Index for Personal RAG (GitHub) — Aug. 2025 – Present
- Creator and lead maintainer of LEANN, a storage-efficient vector index for on-device RAG that reduces index size to under 5% of raw data while maintaining high recall and efficiency.
- Adopted by thousands of developers; the repository has **5k+** GitHub stars with **20+** active external contributors and **40k+** downloads.

Publications

LEANN: A Low-Storage Vector Index
***Yichuan Wang**, Shu Liu, Zhifei Li, Yongji Wu, Ziming Mao, Yilong Zhao, Xiao Yan, Zhiying Xu, Yang Zhou, Ion Stoica, Sewon Min, Matei Zaharia, Joseph E. Gonzalez*
- Preprint (short version in VecDB@ICML 2025) [pdf]
- Adopted by thousands of developers; **5k+** GitHub stars, **20+** active external contributors, **40k+** downloads [GitHub]

DS SERVE: A Framework for Efficient and Scalable Neural Retrieval
Jinjian Liu, **Yichuan Wang***, Xinxi Lyu, Rulin Shao, Joseph E. Gonzalez, Matei Zaharia, Sewon Min*
- * indicates equal contribution
- Accepted by **AAAI 2026** (Demo) [pdf]
- Deployed the largest public vectorstore over pretraining data; serving **1k+ queries/day** [Demo] [Blog]

Locality-aware Fair Scheduling in LLM Serving
Shiyi Cao, **Yichuan Wang***, Ziming Mao, Pin-Lun Hsu, Liangsheng Yin, Tian Xia, Dacheng Li, Shu Liu, Yineng Zhang, Yang Zhou, Ying Sheng, Joseph E. Gonzalez, Ion Stoica*
- * indicates equal contribution
- Preprint [pdf]
- Production deployment: integrated into **Yandex**'s LLM serving infrastructure and **SGLang**'s open-source serving framework

DiskGNN: Bridging I/O Efficiency and Model Accuracy for Out-of-Core GNN Training
Renjie Liu, **Yichuan Wang***, Xiao Yan, Zhenkun Cai, Minjie Wang, Haitian Jiang, Bo Tang, Jinyang Li*

- * indicates equal contribution
- Accepted by **SIGMOD 2025** (Oral) [pdf]
- Collaborated with AWS AI Lab

Optimizing Dynamic Neural Networks with Brainstorm
*Weihao Cui, Zhenhua Han, Lingji Ouyang, **Yichuan Wang**, Ningxin Zheng, Lingxiao Ma, Yuqing Yang, Fan Yang, Jilong Xue, Lili Qiu, Lidong Zhou, Quan Chen, Haisheng Tan, Minyi Guo*

- Accepted by **OSDI 2023** [pdf]
- Collaborated with MSRA

Forming Scalable, Convergent GNN Layers that Minimize a Sampling-Based Energy
*Haitian Jiang, Renjie Liu, Zengfeng Huang, **Yichuan Wang**, Xiao Yan, Zhenkun Cai, Minjie Wang, David Wipf*

- Accepted by **ICLR 2025** [pdf]

The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks
*Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, **Yichuan Wang**, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, Joseph E. Gonzalez*

- Preprint [pdf]

Autellix: An Efficient Serving Engine for LLM Agents as General Programs
*Michael Luo, Xiaoxiang Shi, Colin Cai, Tianjun Zhang, Justin Wong, **Yichuan Wang**, Chi Wang, Yanping Huang, Zhifeng Chen, Joseph E. Gonzalez, Ion Stoica*

- Accepted by **NSDI 2026** [pdf]

Invited Talks

- [10/2025] **Lightspeed interview** – LEANN: Towards Lightweight Vector Search and RAG Everything on PC
- [09/2025] **Bytedance Data Team** – LEANN: Towards Lightweight Vector Search and RAG Everything on PC
- [07/2025] **SIGMOD 2025** – DiskGNN: Bridging I/O Efficiency and Model Accuracy for Out-of-Core GNN Training
- [01/2025] **LMsys** – SGLang-FLPM [Video]

Services

- SIGMOD'25 Artifact Evaluation Committee (04/2025)
- MLSys'25 Artifact Evaluation Committee (03/2025)
- EuroSys'25 Artifact Evaluation Committee (02/2025)
- SIGCOMM'24 Artifact Evaluation Committee (07/2024)
- OSDI'24 Artifact Evaluation Committee (04/2024)
- USENIX ATC'24 Artifact Evaluation Committee (04/2024)

Technical Skills

Programming Languages: Python, CUDA, Triton, Java, Rust, Verilog
Reference: GitHub Profile