Yifan Yu

Email: yifanyu4@illinois.edu| Phone:470-845-7356| Github: github.com/yifan1130

EDUCATION

University of Illinois Urbana-Champaign, Urbana IL **08/2024 - 05/2028**
Doctor of Philosophy | Computer Science
Advisor: Fan Lai
Research Keywords: GenAI Systems, RL training, Muti-agent Systems, Model Compression
Research Impact: IC-Cache is being adopted inside Google; LoftQ is integrated in PEFT and LLaMA-Factory library

Georgia Institute of Technology, Atlanta GA **08/2020 - 05/2024**
Bachelor of Science | Computer Science
Honors: Faculty Honor, Presidential Undergraduate Research Award

PUBLICATION

IC-Cache: Efficient Large Language Model Serving via In-context Caching
Yifan Yu*, Yu Gan*, Nikhil Sarda, Lillian Tsai, Jiaming Shen, Yanqi Zhou, Arvind Krishnamurthy, Fan Lai, Henry M. Levy, David Culler
Accepted by The 31st Symposium on Operating Systems Principles (S0SP), 2025

CORRECT: COndensed eRror RECognition via knowledge Transfer in multi-agent systems
Yifan Yu, Moyan Li, Shaoyuan Xu, Jinmiao Fu, Xinhai Hou, Fan Lai, Bryan Wang
In submission to International Conference on Learning Representations (ICLR), 2026

XRPO: Pushing the limits of GRPO with Targeted Exploration and Exploitation
Udbhav Bamba*, Minghao Fang*, **Yifan Yu***, Haizhong Zheng, Fan Lai
In submission to International Conference on Learning Representations (ICLR), 2026

LoftQ: LoRA-Fine-Tuning-Aware Quantization for Large Language Models
Yixiao Li*, **Yifan Yu***, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, Tuo Zhao
Accepted by International Conference on Learning Representations (ICLR), 2024, ***Oral Presentation***

LoSpase: Structured Compression of Large Language Models based on Low-Rank and Sparse Approximations
Yixiao Li*, **Yifan Yu***, Qingru Zhang, Chen Liang, Pengcheng He, Weizhu Chen, Tuo Zhao
Accepted by International Conference on Machine Learning (ICML), 2023

OPPO: accelerating ppo-based rlhf via pipeline overlap
Kaizhuo Yan*, Yingjie Yu*, **Yifan Yu**, Haizhong Zheng, Fan Lai
In submission to International Conference on Learning Representations (ICLR), 2026

Single-agent or Multi-agent Systems? Why Not Both?
Mingyan Gao*, Yanzi Li*, Banruo Liu*, **Yifan Yu**, Phillip Wang, Ching-Yu Lin, Fan Lai
In submission to ACL Rolling Review

WORK EXPERIENCE

Applied Scientist Intern in Amazon **05/2025 – 08/2025**
- Worked in automatic error attribution for multi-agent systems.
- Proposed a thought-template based method that can significantly improve the error detection accuracies; synthesizing datasets based on Magnetic One to support further research on automatic error attributions.

Student Researcher in Systems Research@Google **05/2024 – 07/2024**
- Worked in efficient systems for serving LLM facing millions of requests leveraging the abundance of users' queries.
- Implemented experiments on millions of real users' requests using Huggingface and vLLM backends; observed significant quality and throughput improvements on Alpaca, Open Orca, and LMSys-chat.

Intern Data Scientist in The Home Depot **05/2022 – 07/2022**
- Worked in the online search team with DSI-based retrieval model as the first undergraduate ever in the team.
- Implemented DSI with PyTorch and HuggingFace; proposed an encoding method based on Pecos clustering for the DSI

indexing process; improved over 15 percent recall rate over baseline, matching the current production search performance.

RESEARCH

Research Assistant at AISys Lab. **08/2024 – Present**

- Worked on efficient serving of LLM facing substantial numbers of requests advised by Prof. Fan Lai
- Implemented a system prototype that can significantly improve the latency and throughput of LLM servings as presented in IC-Cache paper leveraging the in-context learning theory; supported vllm, Gemini backend, and Langchain backends.

Research Assistant at FLASH (Foundations of LeArning Systems for Alchemy) **10/2021 – 05/2024**

- Worked on improving the efficiencies of large language models with Prof. Tuo Zhao.
- Combined the pruning methods and low rank decomposition as presented in LoSparse paper; implemented the codes and conducted the experiments with Huggingface and PyTorch
- Proposed and implemented an alternative optimization method to minimize the quantization error as presented in the LoftQ paper; surpassed QLoRA on different precisions; adopted by PEFT and LLaMA-Factory

SKILLS

Programming: Python (Tensorflow2, Pytorch), Java, C, HTML, CSS, JavaScript, SQL, MATLAB, Mathematica
Related course: GenAI system, Text mining for LLM, Machine Learning, Computer Organization and Programming,