

# Yifan Yu

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

## EDUCATION

University of Illinois Urbana-Champaign, Urbana IL

08/2024 - 05/2029

Doctor of Philosophy | Computer Science

Georgia Institute of Technology, Atlanta GA

08/2020 - 05/2024

Bachelor of Science | Computer Science

Honors: Faculty Honor, Presidential Undergraduate Research Award

## PUBLICATION

EchoLM: Accelerating LLM Serving with Real-time Knowledge Distillation

**Yifan Yu\***, Yu Gan\*, Lily Tasi, Nikhil Sarda, Jiaming Shen, Yanqi Zhou, Arvind Krishnamurthy, Fan Lai, Henry M. Levy, David Culler

*Submitted to 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25)*

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***

LoSparse: 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*

## RESEARCH

Research Assistant at AISys Lab.

08/2024 – Present

- Worked on efficient serving of LLM facing large volume of requests advised by Prof. Fan Lai
- Implemented experiments with vLLM and LMCache backend to reduce the TTFT of the requests on Phi-3 series and Mistral-series models
- Implemented a system prototype that can significantly improve the quality and throughput of LLMs on various datasets leveraging the in-context learning ability of LLMs; supported Huggingface, vLLM, and Gemini API requests 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; further incorporated our method with knowledge distillation and embedded our method with CoFi pruning method; achieved results that match or surpass the state-of-the-art performance on GLUE benchmarks.
- Proposed and implemented an alternative optimization method to minimize the quantization error as presented in the LoftQ paper; surpassed QLoRA on different precisions and showed a large edge over QLoRA on low-precision settings on various benchmarks (GLUE, XSum, WikiText, etc.) using DeBERTa, BART, LLAMA-2

## WORK EXPERIENCE

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.

## SKILLS

**Programming:** Python (Tensorflow2, Pytorch), Java, C, HTML, CSS, JavaScript, SQL, MATLAB, Mathematica

**Related course:** GenAI system, Text mining for LLM, Machine Learning, Deep learning, Natural language Processing, Computer Organization and Programming, Systems and Networks, Design & Analysis of Algorithms