Guanyu Yao

gyao@ucsb.edu • (805) 331-8690 • Github • Homepage

EDUCATION

University of California, Santa Barbara | Sanata Barbara, CA **Sep 2024 – June 2029**
Ph.D in Computer Science| **GPA:** 4.0/4.0
Advisor: Prof. Shiyu Chang

Beijing Normal University | Beijing, China **Sep 2020 – June 2024**
Bachelor of Engineering in Artificial Intelligence | **GPA:** 3.5/4.0

RESEARCH EXPERIENCE

Undergraduate Research Intern | The Hong Kong University of Science and Technology , Hong Kong
Mar 2023 – Mar 2024

ADVISED BY PROF. TONG ZHANG

- Contributor of an open source training/inference platform LMFlow(Github 8k 🌟)
- Implemented multi-GPU batch inference for LMFlow with DeepSpeed; optimized multiple models with flash attention 2.0, and extended LLaMA's inference length 4–8× using Linear & NTK scaling
- Built a pipeline to expand LLaMA's vocabulary for domain/multilingual data, enabling fine-tuning on extended token sets

PUBLICATIONS

An Improved Autoregressive Evaluation Paradigm for Large Language Models **In Submission**
Jipeng Zhang, Rui Pan, Yuzheng Hu, KaShun SHUM, **Guanyu Yao**, Xiang Liu, Renjie Pi, Hanze Dong, Shizhe Diao, Yong Lin, Han Zhao, Tong Zhang

AWARDS & HONORS

Second prize of the 38th National College Students' Physics Competition **Dec 2021**
Third Prize of 14th National College Mathematics Competition **Apr 2023**

TEACHING EXPERIENCE

- **Grader:** CS190I: Introduction to Deep Learning **Fall 2024**
- **Teaching Assistant:** CS 16: Problem Solving with Computers 1 **Winter 2025**

SKILLS

- **Programming Languages:** Python, C, C++, Bash
- **Tools and Technologies:** Git, Pytorch, Deepspeed, Hugging Face Transformers, Linux
- **Machine Learning:** Model Evaluation, NLP, Large Language Models, Distributed Training
- **Performance Optimization:** Multi-GPU acceleration, Flash Attention, Memory Optimization