

Bei Ouyang

+86 180-7965-9312 | ouyb9@mail2.sysu.edu.cn | <https://paprika0741.github.io/>

EDUCATION

Sun Yat-Sen University

M.S. in Computer Science and Technology, GPA 4.0/4.0

China

Sep. 2023 – Present

Sun Yat-Sen University

B.E. in Computer Science and Technology, GPA 4.0/4.0

China

Sep. 2019 – Jun. 2023

PUBLICATIONS & MANUSCRIPTS

Published Papers

Shengyuan Ye, **Bei Ouyang**, Liekang Zeng, Tianyi Qian, Xiaowen Chu, Jian Tang, Xu Chen. *Jupiter: Fast and Resource-Efficient Collaborative Inference of Generative LLMs on Edge Devices*. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications (INFOCOM'25)*.

Bei Ouyang^{*}, Shengyuan Ye^{*}, Liekang Zeng, Tianyi Qian, Jingyi Li, Xu Chen. *Pluto and Charon: A Time and Memory Efficient Collaborative Edge AI Framework for Personal LLMs Fine-Tuning*. In *Proceedings of the 53rd International Conference on Parallel Processing (ICPP'24)*.

Submitted Manuscripts

Shengyuan Ye, **Bei Ouyang**, Jiangsu Du, Liekang Zeng, Tianyi Qian, Wenzhong Ou, Xiaowen Chu, Deke Guo, Yutong Lu, Xu Chen. *Resource-Efficient Collaborative Edge Transformer Inference with Hybrid Model Parallelism*. (Under review, IEEE Transactions on Mobile Computing).

Jingyi Li, Wenzhong Ou, **Bei Ouyang**, Shengyuan Ye, Liekang Zeng, Lin Chen, Xu Chen. *Revisiting Location Privacy in MEC-Enabled Computation Offloading*. (Under review, IEEE Transactions on Information Forensics and Security).

RESEARCH EXPERIENCE

Collaborative Inference of Generative LLMs on Edge Devices

May. 2024 – Nov. 2024

Sun Yat-sen University

Advisor: Prof. Xu Chen

- Addressed the challenge of pipelined parallel acceleration during the prefilling phase by proposing an intra-sequence pipeline parallelism method.
- Achieved parallel acceleration of the autoregressive decoding phase by integrating speculative decoding.
- Conducted experiments on the Llama2 series using NVIDIA edge devices (Nano, TX2, NX), comparing with state-of-the-art parallel LLM inference methods.

Collaborative LLM Fine-tuning Technique for Edge Environments

Sep. 2023 – May. 2024

Sun Yat-sen University

Advisor: Prof. Xu Chen

- Demonstrated the inefficiency of existing PEFT techniques on resource-constrained edge devices.
- Proposed a time and memory-efficient collaborative edge AI framework for in-situ fine-tuning.
- Achieved up to 8.64× acceleration in fine-tuning and reduced memory usage by 88.16% without compromising performance compared to state-of-the-art methods.

Tackling Data Heterogeneity in Federated Learning

Oct. 2022 – Jun. 2023

Sun Yat-sen University

Advisor: Prof. Xu Chen

- Introduced Gini coefficient as a reliable metric to quantify data disparity in federated learning.
- Modeled the process of determining sampling weights as a Markov Decision Process, enabling the use of reinforcement learning to regulate the sampling weights dynamically.
- Developed a data balancing plug-in for FL leveraging deep reinforcement learning.

PROJECTS

CoG 2022 football competetion | Python, Pytorch

March. 2022 – July. 2022

- Implemented the multi-agent reinforcement learning algorithm QMIX for 5v5 soccer scenarios.

Database Project | Go, Gorm, SQL

Nov. 2021 – Jan. 2022

- Developed a web app with a Vue 3 + Element Plus front-end and a Gin + Gorm back-end.

HONORS AND AWARDS

Tencent Special Scholarship (Top 1%), Tencent, 2024
First-Prize Postgraduate Scholarship, Sun Yat-sen University, 2023, 2024
Second-Price Undergraduate Scholarship, Sun Yat-sen University, 2021,2022
Second Prize in the Guangdong Province of the National College Student Mathematical Contest in Modeling, 2021
First-Prize Undergraduate Scholarship, Sun Yat-sen University, 2020

TECHNICAL SKILLS

Languages: Python, C/C++, SQL, GO
Frameworks: PyTorch, OpenMP, CUDA
Developer Tools: Git, Docker, VS Code, Visual Studio, PyCharm
Other: LaTeX

TEACHING ASSISTANT EXPERIENCE

DCS215: Teaching Assistant for Digital Circuits and Logical Design

Spring 2024