Pingzhi Li

Email: pingzhi@cs.unc.edu **Website**: pingzhili.github.io **Google Scholar**:

Education

The University of North Carolina at Chapel Hill (UNC) Chapel Hill, NC
Ph.D. in Computer Science, GPA: 4.0 Aug. 2024 – Jun. 2028
Advisor: Prof. Tianlong Chen

University of Science and Technology of China (USTC) Hefei, China
B.E. in Computer Science, GPA: 3.6 Sep. 2019 – Jul. 2023

Experience

Apple AI/ML (Foundation Models) Cupertino, CA
Research Intern May – Aug. 2025
Advisor: Dr. Xianzhi Du

MIT CSAIL Remote
Research Intern June 2023 – July 2024
Advisor: Dr. Tianlong Chen

USTC Hefei, China
Teaching Assistant Sep. 2022 – Jan. 2023
Course: Computer Programming A (C/C++)

USTC Hefei, China
Undergrad Intern July 2021 - Aug. 2022
Advisors: Prof. Qi Liu, Prof. Enhong Chen

Publications

The *selected publications* are listed below. A full publication list can be found [here].
(* = Equal Contribution) (^ Equal Supervision)

S. Luo, **P. Li**, J. Peng, Y. Zhao, Y. Cao, Y. Cheng, and T. Chen, "Occult: Optimizing Collaborative Communications across Experts for Accelerated Parallel MoE Training and Inference", *International Conference on Machine Learning* (**ICML**), 2025. [Code] [PDF]

P. Li*, M. Zhang*, J. Peng, M. Qiu, and T. Chen, "Advancing MoE Efficiency: A Collaboration-Constrained Routing (C2R) Strategy for Better Expert Parallelism Design", *Annual Conference of the North American Chapter of the Association for Computational Linguistics* (**NAACL**), 2025. **(SAC Award)** [Code] [PDF]

R. Shahroz, **P. Li***, S. Yun*, Z. Wang, S. Nirjon, C. Wong, and T. Chen, "PortLLM: Personalizing Evolving Large Language Models with Training-Free and Portable Model Patches", *International Conference on Learning Representations* (**ICLR**), 2025. [Code] [PDF]

P. Li*, X. Zhao*, G. Sun*, R. Cai*, Y. Zhou*, P. Wang*, B. Tan, Y. He, L. Chen, Y. Liang, B. Chen, B. Yuan, H. Wang^, A. Li^, Z. Wang^, and T. Chen^, "Model-GLUE: Democratized LLM Scaling for A Large Model Zoo in the Wild", *Conference on Neural Information Processing Systems* (**NeurIPS**), 2024. [Code] [PDF]

P. Li*, Y. Zhang*, J. Hong*, J. Li*, Y. Zhang, W. Zheng, P.-Y. Chen, J. D. Lee, W. Yin, M. Hong, Z. Wang, S. Liu, and T. Chen, "Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark", *International Conference on Machine Learning* (**ICML**), 2024. [Code] [PDF]

P. Li, Z. Zhang, P. Yadav, Y.-L. Sung, Y. Cheng, M. Bansal, and T. Chen, "Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy", *International Conference on Learning Representations* (**ICLR**), 2024. **(Spotlight)** [Code] [PDF]

Awards

SAC Award (Low-resource Methods for NLP track), NAACL 2025 — May 2025
1st Place of ACM/IEEE Quantum Computing for Drug Discovery Challenge — Nov. 2023
Outstanding Graduates Scholarship, USTC — June 2023
Silver Medal in Kaggle Feedback Prize - Evaluating Student Writing — March 2022
Outstanding Student Scholarship, USTC — Nov. 2020/21/22

Services

Conference Reviewer: NeurIPS (2024-), ICLR (2025-), ICML (2025-), CVPR (2025-), CPAL (2025-), COLM (2025-), AISTATS (2025-)
Journal Reviewer: IEEE TSP, npj Quantum Information
Tutorial Organizer: Zeroth-Order ML (**AAAI** 2024), MoE (**ICML** 2024)
Lecture: Attention & Transformers (UNC COMP-560)

Mentees

Haoran Wang (Finance@SJTU, Math@USTC) — June 2024 - present
Shuqing Luo (ECE@PKU) — Aug. 2024 - present

Skills

Languages: Mandarin (native), English (professional), German (junior)
Programming Languages: Python, C/C++, Bash
Deep Learning: PyTorch, Jax/Flax