

Pingzhi Li

Email: pingzhi@cs.unc.edu

Website: pingzhili.github.io

Google Scholar: [🔗](#)

Research interests

I pursue *computational parsimony through algorithm-system co-design* to build and understand intelligence. My recent work focuses on:

- *Algorithmic foundations* that uncover structural parsimony in large models, e.g., *sparse mixture-of-experts, low-rank structure*.
- *Systems realization* where the algorithmic insights translate into co-designed computing stacks, e.g. *distributed training, kernel-level acceleration*.
- *Broader impact* where the co-design philosophy generalizes from *language* and *multimodal* models to *scientific discovery*.

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 (Expected)
Advisor: Prof. [Tianlong Chen](#)

University of Science and Technology of China (USTC) Hefei, China
B.E. in Computer Science, GPA: 3.6 Aug. 2019 - Jun. 2023

Experience

Foundation Models team, Apple Cupertino, CA
Research Intern May - Sep. 2026
w. [Xianzhi Du](#)
Research on LLM mid-training.

Foundation Models Team, Apple Cupertino, CA
Research Intern May - Dec. 2025
w. [Xianzhi Du](#), [Yun Zhu](#), [Yihao Feng](#), [Ke Ye](#), [Tao Lei](#)
Research on (1) efficient reasoning [ICLR'26]; (2) multimodal mid-training [production]

MIT CSAIL Remote
Research Intern Jun. 2023 - Jul. 2024
Advisor: [Tianlong Chen](#)

Selected publications

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

#Thrust 1: Efficient Mixture-of-Experts

P. Li*, S. Luo*, Y. Han*, J. Qin*, J. Peng, Y. K. Zhao, Y. Cao, and T. Chen, “Mozart: Modularized and Efficient MoE Training on 3.5D Wafer-Scale Chiplet Architectures”, *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. **(Spotlight)** [\[Code\]](#) [\[PDF\]](#)

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)** [\[Code\]](#) [\[PDF\]](#)

P. Li, X. Jin, Z. Tan, Y. Cheng, and T. Chen. QuantMoE-Bench: Examining Post-Training Quantization for Mixture-of-Experts. *arXiv preprint 2024 arXiv:2406.08155*. [\[Code\]](#) [\[PDF\]](#)

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)** [\[Code\]](#) [\[PDF\]](#)

#Thrust 2: Scalable LLM Training & Adaptation

P. Li, B. Hou, Y. Zhu, Y. Feng, K. Ye, T. Lei, Z. Chen, T. Chen, and X. Du, “Adaptive Thinking: Large Language Models Know When to Think in Latent Space”, *International Conference on Learning Representations (ICLR)*, 2026. [\[PDF\]](#)

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. [\[Code\]](#) [\[PDF\]](#)

#Thrust 3: Efficient AI for Science

P. Li, H. Li, Z. Liu, X. Lin, and T. Chen, “FlashSchNet: Fast and Accurate Coarse-Grained Neural Network Molecular Dynamics”, *International Conference on Machine Learning (ICML)*, 2026. [\[Code\]](#) [\[PDF\]](#)

Awards **SAC Award** (Low-Resource Methods for NLP track), NAACL 2025 May 2025
ICML 2024 **Travel Award** Jun. 2024
1st Place of ACM/IEEE Quantum Computing for Drug Discovery Challenge Nov. 2023
Outstanding Graduates Scholarship, USTC Jun. 2023
Silver Medal in Kaggle Feedback Prize - Evaluating Student Writing Mar. 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-), AAAI (2026-), WACV (2026-)
Journal Reviewer: IEEE TSP (2025-), npj Quantum Information (2025-)
Workshop Co-Organizer: [Lock-LLM](#) (NeurIPS 2025)
Tutorial Co-Organizer: [Zeroth-Order ML](#) (AAAI 2024), [MoE LLM](#) (ICML 2024)
Teaching: *Attention & Transformers* Lecture (UNC COMP-560), *C/C++ Programming* TA (USTC CS-1001)

Mentees *Haoran Wang* (Math@USTC→Finance@SJTU) Jun. 2024 - Apr. 2025
Shuqing Luo (ECE@PKU→CS@UNC) Aug. 2024 - Jan. 2025

Skills **Languages:** Mandarin (native), English (professional), German (junior)
Programming: Python, CUDA, C/C++
Deep Learning: PyTorch (GPU), Jax (TPU)