Pingzhi Li

Email: pingzhili86@gmail.com **Website**: pingzhili.github.io **Google Scholar**:

Research interests

My research centers on **algorithm–system co-design for scalable foundation models**, pursuing **computational parsimony** across neural network stacks. I work on:

- *Scalable algorithms* that uncover structural parsimony in large models, *e.g.*, *sparse mixture-of-experts (MoE), low-rank structure.*
- *Scalable systems* that translate algorithmic insights into high-performance computing stacks, *e.g.*, *kernel-level acceleration, distributed training/inference.*

Education

The University of North Carolina at Chapel Hill (UNC) Chapel Hill, NC
Ph.D. in Computer Science Aug. 2024 - Present
Advisor: Prof. Tianlong Chen

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

Experience

Apple Foundation Models team Cupertino, CA
Research Intern May - Aug. 2026
w. Xianzhi Du, Chao Jiang, Simon Wang, Jun Xu, Zhifeng Chen
Midtraining research

Apple Foundation Models team Cupertino, CA
Research Intern May – Dec. 2025
w. Xianzhi Du, Yun Zhu, Bairu Hou, Yihao Feng, Ke Ye, Tao Lei, Zhifeng Chen
Adaptive thinking [ICLR'26]; Multimodal pretraining [AFM 3]

Selected pubs.

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

#Thrust 1: 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 Communication 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

Rising Star in AI, 2026 AI Symposium at UMich — Sep. 2026
Y Combinator AI Startup School — July 2026
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 Graduate & Student Scholarships, USTC — 2020, 2021, 2022, 2023
Silver Medal in Kaggle Feedback Prize - Evaluating Student Writing — Mar. 2022

Services

Conference Reviewer: NeurIPS (2024-), ICLR (2025-), ICML (2025-), CVPR (2025-), CPAL (2025-), COLM (2025-), AISTATS (2025-), AAAI (2026-), WACV (2026-)
Conference Area Chair: IEEE ICASSP (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)
Mentoring: Haoran Wang (USTC→SJTU), Shuqing Luo (PKU→UNC)