

HAOZE HE

Pittsburgh, PA | (646) 673-2552 | haozeh@cs.cmu.edu

Homepage: <https://hectorrhz.github.io> | GitHub: <https://github.com/HectorHHZ>

Haoze He is currently pursuing a Ph.D. degree at Carnegie Mellon University, School of Computer Science, specializing in machine learning and computer software engineering. His research focuses on **Large Language Models (LLMs), LLM post-training, LLM inference, and distributed machine learning systems**. Haoze's work aims to develop infrastructure and algorithms to optimize LLM performance and system efficiency during post-training and inference.

EDUCATION

Carnegie Mellon University (CMU)

Doctor of Philosophy, Specialized in Machine Learning and Computer Software Engineering
Cumulative GPA: 4.19/4.0

School of Computer Science
Aug 2023 – Present

New York University (NYU)

Master of Science, Specialized in Computer Engineering
Cumulative GPA: 3.93/4.0

Tandon School of Engineering
May 2023

The Chinese University of Hong Kong (CUHK)

Bachelor of Science, Specialized in Computer Science and Engineering

School of Science and Engineering
May 2020

WORK EXPERIENCE

Google DeepMind, Gemini Long Context, Student Researcher, Summer 2026

Amazon, AWS AI Labs, Applied Research Scientist Internship, Summer 2025

PUBLICATIONS

- **Haoze He**, Xinkai Zou, Xuan Jiang, Xingyuan Ding, Ao Qu, Juncheng Billy Li, Heather Miller, “[Less is MoE: Trimming Experts in Domain-Specialist Language Models](#)”, Conference on Empirical Methods in Natural Language Processing (EMNLP Main Conference **Oral**-2026), Published.
- **Haoze He**, Xingyuan Ding, Xuan Jiang, Alex Cheng, Yibo Zhao, Juncheng Billy Li, Heather Miller, “[Preserving Long-Tailed Expert Information in Mixture-of-Experts Tuning](#)”, Conference on Language Modeling (COLM-2026), Published.
- Xinkai Zou, Xuan Jiang, Ruikai Huang, **Haoze He**, Parv Kapoor, Hongrui Wu, Yibo Wang, Jian Sha, Xiongbo Shi, Zixun Huang, Jinhua Zhao, “[Towards Generalizable Context-aware Anomaly Detection: A Large-scale Benchmark in Cloud Environments](#)”, Conference on Language Modeling Workshops (COLM-2026), Published.
- Yuhan Tang, Kangxin Cui, Jung Ho Park, Yibo Zhao, Xuan Jiang, **Haoze He**, Dingyi Zhuang, Shenhao Wang, Jiangbo Yu, Haris Koutsopoulos, Jinhua Zhao, “[RAST-MoE-RL: A Regime-Aware Spatio-Temporal MoE Framework for Deep Reinforcement Learning in Ride-Hailing](#)”, International Conference on Machine Learning (ICML-2026), Published.
- **Haoze He**, Juncheng Billy Li, Xuan Jiang, Heather Miller, “[Sparse Matrix in Large Language Model Fine-tuning](#)”, International Conference on Learning Representations (ICLR-2025), Published.
- Xuan Jiang, Xuanyu Zhou, Yibo Zhao, Shangqing Cao, **Haoze He**, Jinhua Zhao, Mark Hansen, Raja Sengupta, “[From Patchwork to Network: A Comprehensive Framework for Demand Analysis and Fleet Optimization of Urban Air Mobility](#)”, IEEE Transactions on Intelligent Transportation Systems (IEEE T-ITS-2025), Published.
- **Haoze He**, Jing Wang, Anna Choromanska, “[Adjacent Leader Decentralized Stochastic Gradient Descent](#)”, European Conference on Artificial Intelligence (ECAI **Oral**-2024), Published.
- Haoran Zhu, **Haoze He**, Anna Choromanska, Satish Ravindran, Binbin Shi, Phil Chen, Colin Decourt, “[Multi-View Radar Autoencoder for Self-Supervised Automotive Radar Representation Learning](#)”, IEEE Intelligent Vehicles Symposium (IEEE IV-2024), Published.

- **Haoze He**, Parijat Dube, “[RCD-SGD: Resource-Constrained Distributed SGD in Heterogeneous Environment via Submodular Partitioning](#)”, International Conference on Computer Vision Workshops (ICCV-2023), Published.
- Chaoxun Guo, Zhixing Jiang, **Haoze He**, David Zhang, “[Pulse Signal Acquisition and Analysis for Disease Diagnosis: A review](#)”, Computers in Biology and Medicine (CBM-2021), Accepted.

Preprints

- **Haoze He**, Yu Gan, Greg Ganger, Arvind Krishnamurthy, Majid Hadian, Samira Khan, “*Don’t Drop Tokens, Dim Them: Bit Allocation for Multi-Turn KV-Cache Compression*”, manuscript under Google internal review prior to submission.

ACADEMIC BLOG POST

- Peter Zhong, **Haoze He**, Omar Khattab, Christopher Potts, Matei Zaharia, Heather Miller, “[A Guide to Large Language Model Abstractions](#)”, Jan. 2024

AWARDS

- **Presidential Fellowship** in Computer Science at CMU, awarded \$85,000 in total, Nov. 2024

TEACHING

- [Large Language Model System](#) (Head TA of CMU-11868, Spring 2025)
- [Large Language Model Applications](#) (TA of CMU-11766, Spring 2026)

SERVICE

- **Reviewer:** International Conference on Learning Representations (ICLR 2025, 2026, 2027)
- **Reviewer:** Association for the Advancement of Artificial Intelligence (AAAI 2025, 2026)
- **Reviewer:** Conference on Neural Information Processing Systems (NeurIPS 2026)
- **Reviewer:** International Joint Conference on Neural Networks (IJCNN 2025)
- **Reviewer:** International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022, 2023, 2024, 2025, 2026)
- **Reviewer:** International Conference on Computer Vision (ICCV 2023) Workshops

SKILLS

- **Programming Languages:** Proficient in Python, Java, C, Cuda, C++, MATLAB, and R
- **Open-Source Software Libraries:** PyTorch, DeepSpeed, Transformers, Accelerate, Hugging-Face TRL, Unsloth, and MPI4PY

PATENTS

- Zhixia Zheng, **Haoze He**, Saiqin Huang, *Anode Bonding Device*, China, 201720512308.3, 2018-06-12.