

Yiqi Zhang (张懿麒)

Updated July 9, 2026

Email: limitviscent@gmail.com

Phone: +65 98979657

GitHub: [@ExtremeViscent](#)

WeChat: LimitViscent

Homepage: viscent.dev

X: [@YiqiZhangV](#)

Profile

PhD researcher in machine learning systems and high-performance AI, working on algorithm-system co-design for large-scale model training, inference, and post-training. Work spans distributed post-training systems, GPU efficiency, heterogeneous memory, and scalable infrastructure for modern AI workloads.

Skills

Systems: PyTorch, CUDA, NCCL, FSDP, VeRL, Megatron, vLLM/SGLang, Ray, Docker, Linux
Areas: Distributed training, LLM post-training, Parallel algorithms, inference serving, GPU scheduling, Heterogeneous compute

Experience

Tencent

Jun 2026 – Present

Research Intern, Tencent UP (青云) Program

- Contributing to infrastructure for large-scale generative multimodal LLM post-training.
- Building and maintaining internal framework components for reinforcement-learning-based model training workflows.
- Supporting reliability and scalability improvements in production-oriented LLM training pipelines.

Shanghai Qiji Zhifeng Co., Ltd.

Jun 2025 – Jan 2026

- Designed *PlexRL*, a cluster-level orchestration system for multiplexing unified LLM services in RLVR workloads.
- Initiated and served as the main contributor to *Weaver*, a Train-as-a-Service platform for LLM post-training and reinforcement learning workloads.
- Contributed to *NexRL* and *Nex-NI*, an LLM post-training framework and agentic model series for tool-use, coding, and real-world agent workflows.
- Improved RLVR training efficiency by reclaiming cross-job idle GPU capacity through cluster-level multiplexing, affinity-aware scheduling, and centralized model-state management.

Microsoft Research Asia

Jan 2025 – Jun 2025

Mentor: Dr. Yuqing Yang.

- Proposed a high-efficiency rollout scheduling scheme for LLM RL training.
- Developed a batch-centric scheduling strategy that eliminated approximately 70% of rollout scheduling bubbles.
- Reduced steps-to-target reasoning performance by approximately 50% in LLM RL training experiments.

Huawei Singapore Research Center / NUS HPC-AI Lab

Jun 2023 – Dec 2024

Research Collaborator, LLM Systems

- Built high-performance systems for LLM training and inference with CPU/NVMe offloading and heterogeneous memory management.
- Redesigned the LLM offload execution graph, achieving up to 10× fine-tuning speedup and up to 3× inference speedup over state-of-the-art offloading baselines.
- Improved runtime efficiency for large-scale model workloads by reducing memory fragmentation, optimizing data movement, and overlapping computation with I/O.

Earlier Research Experience

Jan 2019 – Jan 2023

- Built early research experience across high-performance computing, GPU acceleration, medical imaging, and computational neuroscience at King’s College London and SUSTech.
- Designed an explainable super-resolution framework for ultra-low-field MRI with Bayesian model selection, achieving over 11% performance improvement.
- Optimized scientific computing and AI workloads including Relion, Presto, DLRM, LD-DMM registration, and numerical simulation kernels; achieved up to 16× GPU speedup in simulation and 2× acceleration for image registration.
- Contributed to medical image augmentation and domain generalization research published in *IEEE Transactions on Medical Imaging*.

Selected
publications

PlexRL: Cluster-Level Orchestration of Serviceized LLM Execution for RLVR

Yiqi Zhang, Fangzheng Jiao, Tian Tang, Boyu Tian, Hangyu Wang, Qiaoling Chen, Guoteng Wang, Zhen Jiang, Peng Sun, Ping Zhang, Xiaohe Hu, Ziming Liu, Menghao Zhang, Yanmin Jia, Yang You, Siyuan Feng
arXiv preprint arXiv:2605.20863, 2026.

SortedRL: Accelerating RL Training for LLMs through Online Length-aware Scheduling

Yiqi Zhang, Huiqiang Jiang, Xufang Luo, Zhihe Yang, Chengruidong Zhang, Yifei Shen, Dongsheng Li, Yuqing Yang, Lili Qiu, Yang You
arXiv preprint arXiv:2603.23414, 2025.

SpeedLoader: An I/O-efficient Scheme for Heterogeneous and Distributed LLM Operation

Yiqi Zhang, Yang You

Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 34637–34655, 2024.

NeX: Experts as services for Efficient, Scalable, and Robust Large-scale MoE Model Inference

Ziming Liu, Boyu Tian, Guoteng Wang, Zhen Jiang, Peng Sun, Zhenhua Han, Tian Tang, Xiaohe Hu, Yanmin Jia, Yan Zhang, **Yiqi Zhang**, et al.
Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2026.

MoNE: Replacing Redundant Experts with Lightweight Novices for Structured Pruning of MoE

Geng Zhang, Yuxuan Han, Yuxuan Lou, Wangbo Zhao, **Yiqi Zhang**, Yang You
Advances in Neural Information Processing Systems (NeurIPS), 2025.

Region-adaptive Sampling for Diffusion Transformers

Ziming Liu, Yifan Yang, Chengruidong Zhang, **Yiqi Zhang**, Lili Qiu, Yang You, Yuqing Yang
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026.

Nex-N1: Agentic Models Trained via a Unified Ecosystem for Large-scale Environment Construction

Yuxuan Cai, Lu Chen, Qiaoling Chen, Yuyang Ding, Liwen Fan, Wenjie Fu, Yufei Gao, Honglin Guo, Pinxue Guo, Zhenhua Han, et al.
arXiv preprint arXiv:2512.04987, 2025.

Education

National University of Singapore

PhD in Data Science

Singapore

Jan 2023 – Present

Advisor: Dr. Yang You.

King's College London

MSc in Neuroscience

London, United Kingdom

Sept 2021 – Sept 2022

Advisor: Prof. Rosalyn Moran.

Southern University of Science and Technology

BSc in Bioscience

Shenzhen, China

Sept 2018 – Jun 2022¹

Advisor: Prof. Shengtao Hou.

Honors and
scholarships

President's Graduate Scholarship, National University of Singapore	2023
Summa Cum Laude, Southern University of Science and Technology	2022
3 rd Place, ISC22 Student Cluster Competition (HPC-AI Advisory Council)	2022
1 st Prize, 4 th APAC HPC-AI Competition (HPC-AI Advisory Council)	2021
AI Special Prize, 4 th APAC HPC-AI Competition (HPC-AI Advisory Council)	2021
Highest Linpack Benchmark Winner, SC21 Virtual Student Cluster Competition (ACM SIGHPC/IEEE)	2021
1 st Prize, ASC20–21 Student Supercomputer Challenge (ASC Committee)	2021
3 rd Prize, 3 rd APAC HPC-AI Competition (HPC-AI Advisory Council)	2020
19 th Place, SC20 Virtual Student Cluster Competition (IEEE)	2020

Earlier
publications

AADG: Automatic Augmentation for Domain Generalization on Retinal Image Segmentation

Junyan Lyu, **Yiqi Zhang**, Yijin Huang, Li Lin, Pujin Cheng, Xiaoying Tang
IEEE Transactions on Medical Imaging, vol. 41, no. 12, pp. 3699–3711, 2022.

40 Hz Light Flicker Alters Human Brain Electroencephalography Microstates and Complexity Implicated in Brain Diseases

Yiqi Zhang, Zhenyu Zhang, Lei Luo, Huaiyu Tong, Fei Chen, Sheng-Tao Hou
Frontiers in Neuroscience, vol. 15, p. 777183, 2021.

¹Joint program of Southern University of Science and Technology and King's College London.