Zhongzhu (Charlie) Zhou

Senior Research Scientist, Together AI, Turbo Team · Founder & PI, FutureMLS Lab

Research: Efficient LLM Training · Efficient Inference · Agentic AI Systems

2 Townsend St., San Francisco, CA 94107 · (+1) 415-310-9490 · zhongzhu@together.ai · zhongzhu.zhou@sydney.edu.au

Homepage · Google Scholar · GitHub · LinkedIn · DBLP · August 2026

Work

- **Future Machine Learning & Systems Lab (FutureMLS)** — 2026 – Present
  Founder & Principal Investigator (PI) · [*futuremls.org*] — *San Francisco, CA*
  Open-source algorithm–system co-design to make ML efficient and practical at scale.
- **Together AI, Turbo Team** — May 2024 – Present
  Senior Research Scientist — *Remote & San Francisco, CA*
  Efficient ML Architecture; Efficient ML Engines, Systems & Infrastructure; Agents, Coding & Science.
- **Dolby** — Mar 2024 – Sep 2024
  Research Intern — *Sydney, Australia*
  Efficient model training and inference research for production media AI workloads.
- **Microsoft, DeepSpeed Team** — Mar 2023 – Feb 2024
  Research Intern — *Sydney, Australia*
  Large-scale RLHF systems and end-to-end training efficiency for ChatGPT-like models.
- **The University of Sydney, FSA Lab** — Mar 2022 – Feb 2026
  Ph.D. Student & Visiting Scholar — *Sydney, Australia*
  LLM systems, RL theory, and AIGC; advisors Shuaiwen Leon Song and Chang Xu.
- **Sun Yat-sen University, School of Computer Science and Engineering** — Sep 2018 – Mar 2022
  Research Associate — *Guangzhou, China*
  Parallel / distributed systems (metadata I/O, hybrid CPU–GPU scheduling, and VFS small-file storage); advisors Dan Huang and Yutong Lu.
- **SYSU, Institute of Advanced Networks and Computing Systems** — Oct 2018 – Mar 2019
  Research Intern — *Guangzhou, China*
  Robotics digital-twin and RL sim-to-real systems; advisor Hejun Wu.
- **Tencent, Weixin Group & UIUC, CS Department** — Jul 2018 – Jul 2020
  Research Intern, Technical Architecture — *Guangzhou & Champaign, IL*
  Code analysis and plagiarism detection for online mini-game platforms; advisors Tao Xie and Yuetang Deng.
- **Microsoft (China) Co., Ltd., Guangzhou Branch** — Sep 2018 – Feb 2019
  Project Assistant to Senior Cloud Architect — *Guangzhou, China*
  Azure cloud architecture and domain NLP Q&A deployment for textile industry applications.
- **SYSU, Computational Medical Imaging Laboratory** — Jul 2016 – Aug 2017
  Research Intern — *Guangzhou, China*
  Medical imaging platforms and mobile image analysis; advisor Yao Lu.

Education

- **The University of Sydney** — Sydney, Australia
  Ph.D., Computer Science / Engineering — *Oct 2022 – Feb 2026*
  GPA 4.0/4.0 (HD). Thesis: *Efficient Compression Algorithm–System Co-Design for Large-Scale Model Training and Inference.*
- **Sun Yat-sen University** — Guangzhou, China
  B.Eng., Computer Science and Technology — *Sep 2015 – Jun 2019*
  GPA 3.9/4.0.

Professional Service

- **Professional Memberships:** IEEE, ACM, and China Computer Federation (CCF)
- **Web Chair:** 32nd IEEE International Symposium on High-Performance Computer Architecture (HPCA 2026)
- **Program Committee:** SELVA 2026 at ACL 2026; Computer Science Research Methods (CSRM 2023), University of Sydney; ASPLOS'24 Artifact Evaluation Committee
- **Conference Reviewer:** NeurIPS 2023, 2025, 2026; ACL 2026; EMNLP 2026
- **Journal Reviewer:** IEEE Transactions on Computers
- **Teaching:** COMP3520 Operating Systems Internals, Tutor, University of Sydney, Fall 2023

Publications

Full author lists: zhongzhuzhou.org. **2026**

- ***Sketched Linear Contrastive Learning: Approximation, Optimization, and Statistical Scaling*** — **arXiv 2026** · [Paper]
- ***OSCAR: Offline Spectral Covariance-Aware Rotation for 2-bit KV Cache Quantization*** — UNDER REVIEW · **ACL 2026** (co-first) · [Paper] · [Code] · 500+ stars
- ***From One-Pass SGD to Data Reuse: Mini-Batch Scaling Laws in Sketched Linear Regression*** — **arXiv 2026** · [Paper]
- ***I-DLM: Introspective Diffusion Language Models*** — **COLM 2026** · [Paper] · [Code]
- ***Squeeze Evolve: Unified Multi-Model Orchestration for Verifier-Free Evolution*** — **COLM 2026** · [Paper] · [Code]
- ***SAW-INT4: System-Aware 4-Bit KV-Cache Quantization for Real-World LLM Serving*** — UNDER REVIEW · 2026 · [Paper] · [Code]
- ***CARE: Covariance-Aware and Rank-Enhanced Decomposition for Enabling Multi-Head Latent Attention*** — **ICLR 2026** (first) · [Paper] · [Code]
- ***Aurora: When RL Meets Adaptive Speculative Training — A Unified Training-Serving System*** — **ICML 2026** · [Paper] · [Code]
- ***KITTY: Accurate and Efficient 2-bit KV Cache Quantization with Dynamic Channel-wise Precision Boost*** — **MLSys 2026** · [Paper] · [Code]
- ***CoderForge-Preview: SOTA open dataset for training efficient coding agents*** — **Together AI Blog 2026** · [Blog] · core lead
- ***Understanding and Steering the Cognitive Behaviors of Reasoning Models at Test-Time (CREST)*** — UNDER REVIEW · 2026 · [Paper] · [Code]
- ***Taylor-Calibrate: Principled Initialization for Hybrid Linear Attention Distillation*** — UNDER REVIEW · 2026 · [Code]
- ***Efficient Compression Algorithm–System Co-Design for Large-Scale Model Training and Inference*** — UNDER REVIEW · **Ph.D. Thesis**, University of Sydney, 2026
- ***LCFS: Hierarchical Performance Isolation for Distributed LLM Serving with LLM Completely Fair Scheduler*** — UNDER REVIEW · 2026
- ***Bio-Inspired LLM-Based Multiagent Systems*** — UNDER REVIEW · 2026
- ***AgentGo: Agent Self-Guided Optimized Program Scheduling for Tool-Using Large Language Models*** — UNDER REVIEW · 2026

2025

- ***Ladder-Residual: Parallelism-Aware Architecture for Accelerating Large Model Inference*** — **ICML 2025** · [Paper]
- ***RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Models*** — **IEEE TPAMI 2025** · [Paper]
- ***Imitate Optimal Policy: Prevail and Induce Action Collapse in Policy Gradient*** — UNDER REVIEW · arXiv 2025 · [Paper]

2024

- ***CorDA: Context-Oriented Decomposition Adaptation of LLMs for Task-Aware PEFT*** — **NeurIPS 2024** · [Paper]
- ***Quant-LLM: FP6-Centric Algorithm-System Co-Design for LLM Serving on Modern GPUs*** — **USENIX ATC 2024** · [Paper]
- ***Flash-LLM: Cost-Effective and Highly-Efficient Large Generative Model Inference with Unstructured Sparsity*** — **VLDB 2024** · [Paper]

2023

- ***DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales*** — **arXiv 2023** · [Paper]
- ***DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery through Sophisticated AI System Technologies*** — **arXiv 2023** · [Paper]

2021

- ***Binary Neural Network for Automated Visual Surface Defect Detection*** — **Sensors MDPI 2021** · [Paper]

2020

- ***JSidentify: A Hybrid Framework for Detecting Plagiarism Among JavaScript Code in Online Mini Games*** — **ICSE 2020** (SEIP) · [Paper]

Book

- ***C Language Programming (Chinese)*** — **Tianjin University Press** · ISBN: 9787561847251

Talks

- **Aurora: Adaptive Speculative Training as Online Reinforcement Learning** — Cornell RL Seminar Weekly, Aug 2026
- **Compression-Centric Systems for Large-Model Post Training & Inference** — Amazon, Jul 2026
- **I-DLM: Introspective Diffusion Language Models** — Ant Group, Apr 2026 · **Scaling Agentic RL & Verified Reasoning** — AI-tonomy Summit, Jun 2026
- **CARE** — ICLR 2026 · **Panel** — Cross-Border Digital Technologies, Nov 2025 · **JSidentify** — ICSE 2020

Honors and Awards

- **USYD Progress Evaluation:** satisfactory or excellent, 2023–2025
- **USYD Scholarships:** APR Intern Program Scholarship (SC3600), 2024; JD Research Scholarship in Artificial Intelligence, 2022
- **SYSU Academic Honors:** National Scholarship (Top 1); Research Honor Degree; two First-Class Scholarships; Second-Class Scholarships
- **Selected Competitions:** COMAP MCM Meritorious Winner; prizes in Chinese Mathematics Competitions, ACM-ICPC, software innovation, and Microsoft Hackathon
- **UIUC:** Illinois Computer Science Summer Research Program, 2018