Zhongzhu / Charlie Zhou

Mar, 2026

Senior Research Scientist, Turbo Team, Together AI

(+86) 13508482354 & (+61) 0414390906 & (+1) 4153109490

zhongzhu@together.ai & zhongzhu.zhou@sydney.edu.au & zhouzhzh8@mail2.sysu.edu.cn

Homepage Linkedin Github Scholar DBLP

82 Waterloo Rd Macquarie Park, Sydney, 2113, NSW, Australia

"Let everything happen to you. Beauty and terror. Just keep going. No feeling is final."
- Rainer Maria Rilke

Work

- **Together.AI** — May. 2024 - Present
  Senior Research Scientist, Turbo Team — *Remote & San Francisco, United States*
- **Dolby** — Mar. 2024 - Sep. 2024
  Research Intern — *Sydney, Australia*
- **DeepSpeed Team, Microsoft** — Mar. 2023 - Feb. 2024
  Research Intern — *Sydney, Australia*
- **WeChat, Tencent & UIUC** — Jul. 2018 - Jul. 2020
  Research Intern, Technical Architecture — *Champaign, IL, United States & Guangzhou, China*
- **Microsoft (China) Co., Ltd., Guangzhou Branch** — Sep. 2018 - Feb. 2019
  Project Assistant to Senior Cloud Architect — *Guangzhou, China*

Education

- **The University of Sydney (USYD)** — Sydney, Australia
  Doctor of Philosophy (Ph.D.) — *Oct. 2022 - Feb. 2026*
  - Accumulated GPA: **4.0/4.0, High Distinct (HD)**
  - **Honor and Awards**:
    Progress Evaluation: satisfactory or excellent, USYD, 2023, 2024, 2025
    APR Intern Program Scholarship (SC3600), USYD, 2024
    The Jingdong Technology (JD) Co Ltd Research Scholarship in Artificial intelligence, USYD, 2022
- **The University of Sydney (USYD)** — Sydney, Australia
  Visiting Scholar — *Mar. 2022 - Oct. 2022*
- **Sun Yat-sen University (SYSU)** — Guangzhou, China
  Research Associate — *Sep. 2019 - Mar. 2022*
  - Accumulated GPA: **3.41/4.0**
  - **Honor and Awards**:
    SYSU Overseas Visiting and Collaborative Research Program Funding Plan, SYSU, 2021
    The Third Class Scholarship X 3, SYSU, 2020, 2021, 2022
    The Second Class Scholarship (**Top 15%** of the major), SYSU, 2019
- **University of Illinois Urbana-Champaign (UIUC)** — Remotely & Champaign, IL, United States
  Summer Seission Student — *Jun. 2018 - Sep. 2018*
  - **Honor and Awards**:
    Illinois Computer Science Summer Research Program, UIUC, 2018
- **Sun Yat-sen University (SYSU)** — Guangzhou, China
  Bachelor of Engineering in Computer Science and Technology — *Sep. 2015 - Jun. 2019*
  - Overall GPA: **3.9/4.0**
  - **Honor and Awards**:
    National Scholarship (**Top 1** of the major), China, 2016
    Research Honor Degree, SYSU, 2019
    The First Class Scholarship X 2 (**Top 5%** of the major), SYSU, 2015-2016, 2017-2018
    The Second Class Scholarship (**Top 15%** of the major), SYSU, 2016-2017
    Meritorious Winner, COMAP's Mathematical Contest in Modeling, United States, 2017
    The Second Prize, The Chinese Mathematics Competitions, 2016
    The Third Prize, The Chinese Mathematics Competitions, 2017
    The Third Prize, ACM-ICPC, SYSU, 2017
    The Second Prize, Student Innovation Software Development Competition, SYSU, 2017
    The Third Prize, Microsoft Hackthon, South China, 2018
- **Changjun High School** — Changsha, China
  Senior Student — *Sep. 2012 - Jun. 2015*
  - **Honor and Awards**:
    The Second Prize, National Olympiad in Informatics in Provinces

Professional Experience

- **Turbo Team, Together.AI** — May. 2024 - Present
  Senior Research Scientist — *Remote & San Francisco United States*

  Advisor: *Ben Athiwaratkun* (Senior Research Scientist, Together.AI), *Shuaiwen Song* (Vice President of Research, Together.AI)
  Industry Projects:

  - **Efficient ML Algorithms**
    - **Ladder-Residual**
      PAPER LINK, CODE LINK
      Motivation: Large-model inference under tensor parallelism often suffers from communication stalls and weak overlap between communication and computation; we sought an architecture-runtime co-design that improves throughput without sacrificing model quality.
      Contributions:
      1. Co-conceived the parallelism-aware residual design and helped shape the paper's system and evaluation story.
      2. Implemented and optimized the `gpt-fast` inference path with CUDA Graphs and PyTorch compile ("reduce-overhead") for large-model serving.
      3. Benchmarked performance across model scales (1B–405B) and TP world sizes (1, 2, 4, 8, 16), validating up to 30% end-to-end throughput improvement on 70B models with P2P enabled and up to 60% with P2P disabled.
    - **CREST (Turbo-reasoning)**
      PAPER LINK, CODE LINK
      Motivation: Reasoning models often under-think or over-think at test time, wasting tokens or missing correct solutions; we sought a training-free intervention that could be deployed in mainstream serving stacks.
      Contributions:
      1. Co-developed the core idea of a training-free test-time steering method that identifies and modulates cognitive attention heads, improving accuracy by up to 17.5% and reducing token usage by 37.6% across reasoning benchmarks.
      2. Designed deployment paths for integrating CREST into `vLLM` and `SGLang`.
    - **CARE**
      PAPER LINK, CODE LINK
      Motivation: MLA-style attention can improve serving efficiency, but most pretrained checkpoints use GQA/MHA and cannot directly benefit; we sought a practical conversion path that preserves quality while lowering inference cost.
      Contributions:
      1. Developed the core idea and empirical framing for upgrading pretrained attention into MLA-compatible forms.
      2. Proposed a conversion pipeline that upgrades pretrained attention (e.g. GQA) into multi-head latent attention (MLA) for faster inference without increasing KV-cache size.
      3. Ran the full experimental suite and carried out `vLLM` integration and theoretical analysis.
    - **SQUEEZE THINK**
      Motivation: Recursive self-aggregation improves reasoning quality, but uniform compute allocation across generation and aggregation wastes cost on easy subsets and under-allocates recovery on hard subsets.
      Contributions:
      1. Helped develop a multi-model orchestration view of recursive self-aggregation, routing generation and aggregation between large and small models based on cross-model confidence.
      2. Owned coding-benchmark execution and evaluation pipelines, especially for LiveCodeBench V6, and supported ablations on routing thresholds and aggregation behavior across AIME 2025 and HMMT 2025.
      3. Demonstrated 30–40% compute reduction at matched accuracy or 5–7 point accuracy gains at equivalent compute.
    - **Agent Evolve**
      Motivation: Current LLM-based multiagent systems are largely static after deployment and lack mechanisms for continual adaptation across agents, skills, and populations.
      Contributions:
      1. Built a bio-inspired LLM multiagent framework with pheromone-style memory, evolutionary division of labor, and skill inheritance for open-ended population adaptation.
      2. Studied population-level adaptation through competition, selection, and cross-generation strategy transfer.
    - **Explored integration of LEXICO compression techniques.**
    - **Prototyped vocabulary-pruned speculators, Mix Architecture Speculator designs**
    - **Explored diffusion LLMs that interleave self-verification with token generation.**
    - **Investigated diffusion-style MoE routers for smoother expert selection.**
    - **Investigated diffusion-style speculator design**
  - **Efficient ML Systems**
    - **Training System: XoRL (RL Training System), Axolotl (SFT Training System)**
      Motivation: Building an RL and SFT training stack for coding and reasoning agents required more than model fine-tuning: it needed an end-to-end system that coupled sandboxed environments, distributed rollout workers, and multi-node training plus serving infrastructure while staying stable under long-context, MoE, and rapidly changing model variants.

Contributions:
1. Built much of the training-side RL framework, including agent PPO trainers, asynchronous rollout and pipeline-training paths, and the execution flow that converts multi-turn agent-environment interaction into PPO and GRPO training batches.
2. Owned the training pipeline that ingests rollout trajectories, computes advantage, and performs policy updates plus rollout-model weight synchronization for coding-agent post-training.
3. Implemented asynchronous rollout, replay-queue mini-batching, and router-assisted batching between rollout and training workers to overlap trajectory generation with policy optimization and improve distributed training throughput.
4. Developed trajectory/data transforms, token-level loss masks, stepwise-vs-trajectory advantage handling, rejection sampling, and batch balancing to improve GRPO signal quality and training stability.
5. Scaled long-context training recipes to 16K–32K contexts using Ulysses sequence parallelism, remove-padding, chunked prefill, and per-GPU token-budget tuning for DeepCoder and DeepScaleR-style runs.
6. Implemented sequence-parallel (SP) compatibility across the training stack so long-context post-training paths worked correctly with distributed attention, packed sequences, and rollout-to-training data flow.
7. Built SP-compatible MoE-LoRA kernel paths to support efficient distributed post-training for expert models without breaking sequence-parallel execution.
8. Integrated QuACK fused kernels into XoRL to improve kernel efficiency and support higher-throughput post-training recipes.
9. Added Qwen3.5 support and completed model bring-up across configs, training paths, and distributed recipes for reliable experimentation.
10. Diagnosed and fixed multi-node training failures, including `position_ids`, `cu_seqlens`, attention-mask, and MoE dispatch issues that destabilized distributed recipes across evolving model families.
11. Integrated long-context attention (Ulysses, Ring Attention) into Axolotl and supported SFT data flow from successful trajectories to extend supervised post-training to larger context windows.

* **Inference System: Pulsar & SGLang**
Motivation: High-throughput serving requires lower KV overhead and more stable speculative decoding across cache-hit patterns, batch sizes, and multi-node deployments.
Contributions:
1. Applied a Swift-KV caching strategy to accelerate prefill by reducing KV memory overhead and improving end-to-end latency.
2. Designed and implemented KV-cache prompt caching for the Phoenix speculator in Pulsar, stabilizing acceptance rates and reducing end-to-end latency.
3. Resolved tokenizer chat-template issues and Docker deployment bugs for reliable multi-node operation, then benchmarked cache behavior across batch sizes and cache-hit scenarios to explain acceptance-rate variability and optimize cache-hit logic.
4. Integrated and implemented Llama 4 support for sliding window attention

* **AgentGo**
Motivation: Tool-using agents alternate between long-context reasoning and external actions, but request-centric runtimes either evict useful KV state too early or waste memory by pinning it too long.
Contributions:
1. Co-developed the core idea of treating multi-turn agent workflows as first-class programs rather than isolated requests.
2. Helped build the staged system path from telemetry and shadow prediction to offline replay, observability, and config-gated runtime integration for prediction-aware scheduling.

* **LCFS**
Motivation: Multi-tenant LLM serving needs hierarchical fairness and performance isolation across shared instances and clusters without sacrificing throughput.
Contributions:
1. Contributed to design discussions around hierarchical fairness, vruntime-style accounting, and weight partitioning across distributed serving instances.
2. Participated in experiments evaluating performance isolation and fairness under multi-tenant LLM serving workloads.

– **Model Related**

* **CoderForge**
BLOG LINK, CODE LINK
Motivation: High-quality coding agents require strong trajectory data, stable post-training pipelines, and task-aligned optimization objectives for code generation.
Contributions:
1. Led the training pipeline for OpenHands R2E-Gym & SWE-Bench-scale data: curated high-signal SWE-smith / Rebench examples and fixed attention-mask plus position-ID issues in XoRL.
2. Distilled Qwen3-480B trajectories into a 30B coding model via supervised fine-tuning and activation distillation, then initiated MoE / RL scaling for Qwen3-30B to improve SWE-Bench solve rates.
3. Designed per-token loss formulations for coding-trajectory distillation and model-quality improvement.

• **Dolby** Mar. 2024 - Sep.2024
Research Intern *Sydney Australia*
Advisor: *Yucheng Liu* (Research Scientist, Dolby), *Shuaiwen Song* (Vice President of Research, Together.AI)
Industry Projects:

– **Extrem Efficient Video Coding System**
Motivation: Traditional codecs (H.264/H.265/AV1) lack content adaptivity and incur high compute/memory costs. Existing neural compressors are too heavy for real-time GPU and mobile streaming. A need for a low-latency, domain-aware solution that tailors compression to video content.
Contributions:
1. Invented and spearheaded $E^2ND - VC$ (Extreme Efficient Neural Domain Video Compression), a pioneering neural video compression framework that leverages content-aware quantization to deliver low-latency, high-quality streaming on both standard GPUs and mobile devices.
2. Designed Optimal Brain Stride-wise Quantization (OBSQ), a domain-specific quantization methodology that selectively compresses neural network weights based on content type (e.g., video conferencing, gaming), enabling real-time 1080p performance with minimal quality loss.
3. Engineered a multi-kernel, sensitivity-based quantization pipeline with mixed-bit precision assignments, dynamically allocating bit depths across convolutional kernels to preserve critical visual features while maximizing compression ratios.
4. Collaborated closely with cross-functional teams to implement PoC streaming pipelines, demonstrating significant reductions in power consumption and bandwidth usage without compromising visual fidelity.

• **DeepSpeed Team, Microsoft** Mar. 2023 - Feb. 2024
Research Intern *Sydney Australia*
Advisor: *Xiaoxia Wu* (Research Scientist, Microsoft), *Zhewei Yao* (Senior Researcher, Microsoft), *Shuaiwen Song* (Senior Principle Scientist, Microsoft)
Industry Projects:

– **DeepSpeed4Science**
PAPER LINK, CODE LINK
Motivation: To build unique capabilities through AI system technology innovations to help domain experts to unlock today' s biggest science mysteries.
Contributions:
1. Developed deepspeed4science's blog website through **Azure MySQL, Wordpress, Virtual Server Hosting, JavaScript HTML, CSS, AJAX, Azure Migration**. Website Link
2. Revised the blog content, font size, technical research architecture, and code related to GenSLMs-'Megatron-DeepSpeed for Large-Scale AI4Science Model Training'.

– **DeepSpeed Chat: Easy, Fast, and Affordable RLHF Training of ChatGPT-like Models at All Scales**
PAPER LINK, CODE LINK
Motivation: ChatGPT-like models have revolutionized the AI world, but an accessible end-to-end RLHF pipeline for training powerful ChatGPT-like models is still lacking within the AI community.
Contributions:
1. Apply INT4 and INT8 quantization to the RLHF pipeline, increase the batch size and improve the speed of the training and generation phases of RLHF without significantly compromising accuracy.
2. Investigated ColossalAI's pipeline, learned how to use ColossalAI's Zero-2, 3, and GeminiDDP, and adapted them for our RLHF algorithm.
3. Ran **400**+ benchmark experiments for DeepSpeed Chat, ColossalAI, and HuggingFace powered by native PyTorch. Summarized the results and conclusions in the DeepSpeed blog.
4. Revised DeepSpeed GitHub Landing Page, DeepSpeed Chat Blog, and produced DeepSpeed Chat video.

• **Future System Architecture (FSA) Lab, The University of Sydney (USYD)** Mar. 2022 - Present
Visiting Scholar, Ph.D. student *Sydney Australia*
Advisor: *Shuaiwen Song* (Associate Professor, USYD), *Chang Xu* (Associate Professor, USYD), *Yibo Yang* (Research Scientist in JD Explore Academy)
Research Projects:

– **RenAIssance: A survey into AI text to image generation in the era of large models**
PAPER LINK

Motivation: Text-to-image synthesis has become increasingly popular in the AI and computer graphics world (AIGC). However, there is no comprehensive survey paper that systematically introduces the frameworks and ideas behind text-to-image techniques. We aim to fill this gap in the literature.
Contributions:
1. Read over 100 papers, providing a literature review for each.
2. Collaborated with lab classmates to write the comprehensive survey paper.

- **Optimization of Diffusion Model Denoising Process**
  Motivation: Diffusion models currently require a large number of denoising steps, which we aim to reduce. One reason for the lengthy process is the lack of a clear relationship between the noise and the trained image. Our goal is to explore additional methods to establish a connection between noise and the denoised image, beyond guidance techniques, such as incorporating text embeddings into the raw noise.
  Contributions:
  1. Develop innovative ideas, implement them, and conduct comparative experiments to evaluate their performance.
- **Exploring Neural Collapse Phenomenon in Reinforcement Learning**
  PAPER LINK, CODE LINK
  Motivation: In reinforcement learning, agents may exhibit biased action selection in the environment due to incomplete understanding of the state and action distribution spaces. This research investigates whether the neural collapse phenomenon occurs in policy gradient networks as agents train with sufficient examples and examines its implications for balancing action selection in reinforcement learning agents.
  Contributions:
  1. Conducted experiments applying ETF classifiers to 5+ neural networks in 10+ discrete-action reinforcement learning environments (e.g., Atari, Gym Classic)
  2. Derived and proved the formula and geometric properties of policy gradient loss function
  3. Authored paper drafts and submitted the work to ICML.
- **Sparse Kernel Design in GPU TensorCore**
  PAPER LINK, CODE LINK
  Motivation: With the application of pruning methods, neural network weight matrices become increasingly sparse, but there is no implementation for sparse kernels in GPU TensorCore.
  Contributions:
  1. Conducted comparative experiments between our sparse kernel and Google's Sputnik.
  2. Summarized experiment results and figures in the paper.
- **DeepSpeed I/O Framework Support for AI4Science**
  CODE LINK
  Motivation: AI4Science models have revolutionized the AI world. DeepSpeed can support AI4Science models deployed across multiple nodes but lacks an I/O management framework for handling large amounts of training data efficiently.
  Contributions:
  1. Investigated DeepSpeed I/O support in Argonne HDF5 Luster System, analyzed data shuffling and fetching patterns for AI4Science models powered by DeepSpeed, and implemented algorithms to accelerate I/O.
  2. Implemented a ViT model for weather prediction.
- **CorDA: Context-Oriented Decomposition Adaptation of Large Language Models for Task-Aware Parameter-Efficient Fine-tuning**
  PAPER LINK, CODE LINK
  Motivation: Existing low-rank fine-tuning methods (e.g., LoRA) adapt LLMs without understanding which layers encode core knowledge vs. task-specific behavior, causing forgetting; we want a parameter-efficient method that adapts to the new task while preserving what the model already knows.
  Contributions:
  1. Designed the experimental methodology for evaluating task-aware parameter-efficient fine-tuning (dataset selection, baselines, and metrics across math/code/instruction-following).
  2. Implemented and executed large-scale experiments to compare CorDA against PEFT baselines, and helped collect and analyze empirical results used in the paper.
- **Survey of LLM Agents**
  Motivation: Rapid progress in agentic LLM systems makes it difficult to track design patterns spanning planning, memory, tool use, multi-agent coordination, and systems support.
  Contributions:
  1. Conducted a structured survey of recent LLM agent systems, synthesizing advances in tool use, memory, planning, multi-agent coordination, and serving/runtime design into an evolving research reference for follow-on projects.

- **School of Computer Science and Engineering, SYSU** Sep. 2018 - Mar. 2022
  Research Associate *Guangzhou, China*
  Advisor: *Dan Huang* (Professor, SYSU), *Yutong Lu* (Professor, SYSU)
  Research Projects:
  - **Pre-Expedite: Use Hierarchical Structure Space for Improving the Performance of Accessing Small Files in Parallel File System - Undergraduate Thesis**
    CODE LINK
    Motivation: Implemented an approach to reduce clients' I/O communication with MDS, leveraging minimal additional client-side resources. Ensured high usability without modifying POSIX standards.
    Contributions:
    1. Investigated the I/O bottleneck in parallel/distributed file systems for Big Data and Artificial Intelligence applications, identifying intensive metadata communication with the metadata server as a primary issue.
    2. Utilized POSIX to create ZERO file blocks (Loop Device). Established a VFS within the ZERO file blocks, allowing each user to store small files in their designated ZERO file blocks.
  - **HybridShare: Universal Resource Scheduling for Hybrid Jobs**
    CODE LINK
    Motivation: CPU- and GPU-centric applications allocate resources exclusively, leading to inefficient utilization of heterogeneous resources.
    Contributions:
    1. Analyzed the possibility of co-locating modern workflow - application in the same physical machine to share resources.
    2. Proposed HybridShare algorithms that can enable different resources-prefer jobs to be co-located in the same node and share hardware resources (e.g., GPU-concentric, CPU-concentric, Mem-intensive) through Slurm, Mesos, Kubernetes.
  - **MAEM - Multiple Applications co-Execution time Estimation**
    CODE LINK
    Motivation: There are few works to accurately estimate the slowdown of CPU/GPU applications based on the characteristic of applications & hardware architecture
    Contribution:
    1. Conducted a literature review on application profiling, interference and slowdown estimation, and interference-aware scheduling.
    2. Gathered resource consumption data for various benchmarks and analyzed their behavior.
- **Institute of Advanced Networks and Computing Systems, SYSU** Oct. 2018 - Mar. 2019
  Research Intern *Guangzhou, China*
  Advisor: *Hejun Wu* (Professor, SYSU)
  Research Projects:
  - **EmReal: A Digital Twin Framework of Emulated and Real Components for Robots with Reinforcement Learning**
    CODE LINK
    Motivation: Pioneered a digital twin framework for robots utilizing reinforcement learning (RL), bridging the gap between simulations and real-world deployments. Developed solutions to effectively transition RL algorithms from simulators to actual robots, advancing the field beyond its nascent stage.
    Contributions:
    1. Conducted a survey on robotics simulator systems and reinforcement learning algorithms.
    2. Designed and implemented a one-legged robot, integrating real and emulated components using XLM, Python, ROS, and Arduino C programming.
    3. Created a digital twin framework for robotic systems, employing reinforcement learning (RL) and seamlessly blending emulation, pre-training, connectivity, and hardware adaptation using ROS and PyBullet.
  - **Co-authored a book on deep learning in reinforcement learning, awaiting publication.**
- **Weixin Group, Tencent Holdings Ltd. & Dep. of CS, UIUC** Jul. 2018 - Jul. 2020
  Research Intern, Testing, Technical-Architecture Department *Champaign, IL, US & Guangzhou, China*
  Advisor: *Tao Xie* (Professor and Willett Faculty Scholar, UIUC), *Yuetang Deng* (Director)
  Industry Projects:
  - **JSidentify: A Hybrid Framework for Detecting Plagiarism Among JavaScript Code in Online Mini Games**
    TALK LINK, PAPER LINK
    Motivation: In cases of plagiarism for mini-games, deeply obfuscated code cloned from the original code often embodies malicious code segments and copyright infringements, posing great challenges for existing plagiarism detection tools. To address these challenges, we design and implement JSidentify, a hybrid framework to detect plagiarism among online mini games.
    Contributions:
    1. Worked under the guidance of Prof. Tao Xie, focusing on intermediate representation analysis in V8 & Node.js's Interpreter.
    2. Conducted literature review on code plagiarism detection methods and evaluations of clone detection tools.

3. **Developed an edit distance estimation and network flow algorithm** to measure similarity in bytecode generated by Ignition, TurboFan Interpreter.
4. **Designed a priority-queue-based framework** to consolidate multiple plagiarism detection algorithms.

- **Microsoft(China) Co.,Ltd. Guangzhou Branch** Sep. 2018 - Feb. 2019
  Project Assistant to Senior Cloud Acrchitect *Guangzhou, China*

  Advisor: *Zhen Guan* (Sr.Partner Technology Strategist, Microsoft)

  – **Textile-Focused Q&A System**
  Motivation: The textile industry in China lacked an accessible domain-specific intelligent Q&A service, while relevant information was scattered across heterogeneous web sources and difficult for users to query efficiently. We aimed to build a practical NLP system that could organize textile knowledge and provide question-answering support through a cloud-deployed service.
  Contributions:
  1. Learned Azure cloud architecture and model-serving workflows to support production-oriented deployment of machine-learning systems.
  2. Collected textile-domain Q&A data by crawling major industry websites and constructed a cleaned, serialized, and tokenized corpus.
  3. Implemented a pre-trained BERT model for the Q&A system and adapted it to the domain-specific dataset.
  4. Deployed the BERT-based Q&A model on Azure as an online service for demonstration and practical use.

- **SYSU-CMU Joint Institute of Engineering (JIE)** Feb. 2017 - Aug. 2017
  Research & Software Engineer Intern *Guangzhou, China*

  Advisor: *Xiaoyin Tang* (Professor, Southern University of Science and Technology)

  – Created a front-end website to integrate with a back-end deep learning model for efficient analysis of numerous fundus photographs.
  – Enabled detection of diabetic retinopathy (DR) and diabetic macular edema (DME) through seamless collaboration between the front-end and back-end systems.

- **Computational Medical Imaging Laboratory, SYSU** Jul. 2016 - Aug. 2017
  Research Intern *Guangzhou, China*

  Advisor: *Yao Lu* (Professor, SYSU)

  – **OHIF Viewer Web Project - Intelligent Medical Media Platform**
  PROJECT LINK
  Motivation: Medical imaging workflows often rely on fragmented tooling and cumbersome access to image data, making it difficult for clinicians and researchers to browse, manage, and analyze large collections of breast-cancer images efficiently. We aimed to build a web-based medical media platform that streamlined image visualization and supported practical clinical research usage.
  Contributions:
  1. Collected and organized breast-cancer data through web crawling with Scrapy to support platform development and evaluation.
  2. Developed an OHIF-based web viewer for medical-image browsing, visualization, and interactive review, and helped deploy the project online.
  3. Contributed to the associated **SIT (College Students' Innovative Entrepreneurial Training Plan)**, ID: 201502059, helping integrate project components into a usable prototype.
  4. Implemented traditional image-processing algorithms on mobile platforms to extend accessibility of medical-image analysis workflows.

Publications

Book

- ***C Language Programming* in Chinese**
  Xuemao Zhou, Wei Yi, **Zhongzhu Zhou**
  – **Tianjin University Press**
  – *ISBN: 9787561847251*
  – SALE LINK

Conference Paper

- ***JSidentify: A Hybrid Framework for Detecting Plagiarism Among JavaScript Code in Online Mini Games***
  Qun Xia, **Zhongzhu Zhou**, Zhihao Li, Bin Xu, Wei Zou, Zishun Chen, Huafeng Ma, Gangqiang Liang, Haochuan Lu, Shiyu Guo, Ting Xiong, Yuetang Deng, Tao Xie
  – **ICSE (International Conference on Software Engineering) 2020**
  – *Track: Software Engineering in Practice*
  – TALK LINK, PAPER LINK

- ***Flash-LLM: Enabling Cost-Effective and Highly-Efficient Large Generative Model Inference with Unstructured Sparsity***
  Haojun Xia, Zhen Zheng, Yuchao Li, Donglin Zhuang, **Zhongzhu Zhou**, Xiafei Qiu, Yong Li, Wei Lin, Shuaiwen Leon Song
  – **VLDB (International Conference on Very Large Databases) 2024**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***Imitate Optimal Policy: Prevail and Induce Action Collapse in Policy Gradient***
  Zhongzhu Zhou, Yibo Yang, Ziyan Chen, Fengxiang Bie, Haojun Xia, Xiaoxia Wu, Robert Wu, Ben Athiwaratkun, Bernard Ghanem, Shuaiwen Leon Song
  – **Submitted to ICML 2026**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***CorDA: Context-Oriented Decomposition Adaptation of Large Language Models for Task-Aware Parameter-Efficient Fine-tuning***
  Yibo Yang, Xiaojie Li, **Zhongzhu Zhou**, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, Bernard Ghanem
  – **NeurIPS (Annual Conference on Neural Information Processing Systems) 2024**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***Quant-LLM: Accelerating the Serving of Large Language Models via FP6-Centric Algorithm-System Co-Design on Modern GPU***
  Haojun Xia, Zhen Zheng, Xiaoxia Wu, Shiyang Chen, Zhewei Yao, Stephen Youn, Arash Bakhtiari, Michael Wyatt, Donglin Zhuang, **Zhongzhu Zhou**, Olatunji Ruwase, Yuxiong He, Shuaiwen Leon Song
  – **ATC (USENIX Annual Technical Conference) 2024**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***Ladder-Residual: Parallelism-Aware Architecture for Accelerating Large Model Inference with Communication Overlapping***
  Muru Zhang, Mayank Mishra, **Zhongzhu Zhou**, William Brandon, Jue WANG, Yoon Kim, Jonathan Ragan-Kelley, Shuaiwen Leon Song, Ben Athiwaratkun, Tri Dao
  – **ICML Forty-second International Conference on Machine Learning 2025**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***Understanding and Steering the Cognitive Behaviors of Reasoning Models at Test-Time***
  Zhenyu Zhang, Xiaoxia Wu, **Zhongzhu Zhou**, Qingyang Wu, Yineng Zhang, Pragaash Ponnusamy, Harikaran Subbaraj, Jue WANG, Shuaiwen Leon Song, Ben Athiwaratkun
  – **Submitted to ICML 2026**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***CARE: Covariance-Aware and Rank-Enhanced Decomposition for Enabling Multi-Head Latent Attention***
  Zhongzhu Zhou , Fengxiang Bie, Ziyan Chen, Zhenyu Zhang, Yibo Yang, Junxiong Wang, Ben Athiwaratkun, Xiaoxia Wu, Shuaiwen Leon Song
  – **ICLR (The Fourteenth International Conference on Learning Representations) 2026**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***KITTY: ACCURATE AND EFFICIENT 2-BIT KV CACHE QUANTIZATH CHANNEL-WISE PRECISION BOOST***
  Haojun Xia, Xiaoxia Wu, Jisen Li, Robert Wu, Junxiong Wang, Jue Wang, Chenxi Li, Aman Singhal, Alay Dilipbhai Shah, Alpay Ariyak, Donglin Zhuang, **Zhongzhu Zhou**, Ben Athiwaratkun, Zhen Zheng, Shuaiwen Leon Song
  – **MLSys (Ninth Annual Conference on Machine Learning and Systems) 2026**
  – *Track: Research*
  – PAPER LINK, CODE LINK

- ***When RL Meets Adaptive Speculative Training: A Unified Training-Serving System***
  Junxiong Wang, Fengxiang Bie, Jisen Li, **Zhongzhu Zhou**, Zelei Shao, Yubo Wang, Yinghui Liu, Qingyang Wu, Avner May, Sri Yanamandra, Yineng Zhang, Ce Zhang, Tri Dao, Percy Liang, Ben Athiwaratkun, Shuaiwen Leon Song, Chenfeng Xu, Xiaoxia Wu
  – **Submitted to ICML 2026**
  – *Track: Research*
  – PAPER LINK, PROJECT LINK

- ***CoderForge-Preview: SOTA open dataset for training efficient coding agents***
  Alpay Ariyak*, Junda Zhang, Junxiong Wang, Shang Zhu, Federico Bianchi, Sanjana Srivastava, Ashwinee Panda, Siddhant Bharti, Chenfeng Xu, John Heo, Xiaoxia Shirley Wu, James Zou, Percy Liang, Leon Song, Ce Zhang, Ben Athiwaratkun, **Zhongzhu Zhou***, Qingyang Wu* *Project Core Leads

  - **Together AI Blog**
  - *Track: Research*
  - BLOG LINK, CODE LINK
- ***Efficient Compression Algorithm-System Co-Design for Large-Scale Model Training and Inference***
  Zhongzhu Zhou
  - **Thesis of The University of Sydney**
  - *Track: Doctor of Philosophy (Engineering)*
  - UNDER REVIEW
- ***LCFS: Hierarchical Performance Isolation for Distributed LLM Serving with LLM Completely Fair Scheduler***
  - RELEASE SOON
  - *Track: Research*
- ***SQUEEZE THINK: Multi-Model Orchestration for Efficient Recursive Self-Aggregation***
  - RELEASE SOON
  - *Track: Research*
- ***Bio-Inspired LLM-Based Multiagent Systems***
  - RELEASE SOON
  - *Track: Research*
- ***AgentGo: Agent Self-Guided Optimized Program Scheduling for Tool-Using Large Language Models***
  - RELEASE SOON
  - *Track: Research*

Journal Paper

- ***Binary Neural Network for Automated Visual Surface Defect Detection***
  Wenzhe Liu, Jiehua Zhang, Zhou Su, **Zhongzhu Zhou**, Li Liu
  - **Sensors MDPI (Multidisciplinary Digital Publishing Institute)**
  - Track: *Special Issue Intelligent Sensing and Monitoring for Industrial Process*
  - PAPER LINK
- ***RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model***
  Fengxiang Bie, Yibo Yang, **Zhongzhu Zhou**, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, Shuaiwen Leon Song
  - **TPAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence)**
  - Track: *Survey Papers*
  - PAPER LINK

Preprint Paper

- ***DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales***
  Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, **Zhongzhu Zhou**, Michael Wyatt, Molly Smith, Lev Kurilenko, Heyang Qin, Masahiro Tanaka, Shuai Che, Shuaiwen Leon Song, Yuxiong He
  - PAPER LINK, CODE LINK
- ***DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery through Sophisticated AI System Technologies***
  Shuaiwen Leon Song, Bonnie Kruft, Minjia Zhang, Conglong Li, Shiyang Chen, Chengming Zhang, Masahiro Tanaka, Xiaoxia Wu, Jeff Rasley, Ammar Ahmad Awan, Connor Holmes, Martin Cai, Adam Ghanem, **Zhongzhu Zhou**, et al.
  - PAPER LINK

Patent

- ***KUBERNETES* 用户态应用中基于虚拟文件系统的小文件存储优化系统**
  Liang Du, Guixin Guo, Kangyou Zhong, Yunfei Du, Yutong Lu, **Zhongzhu Zhou**
  - **Chinese Patent**
  - 申请号: CN202010195318.5, 公开号: CN111475469A/B
  - CHINESE DOCUMENT LINK 1, CHINESE DOCUMENT LINK 2

Talks

- ***Panel Discussion: AI in Cross-Border Digital Technologies Infrastructure Platforms & Global Launchpads***
  - AI in the Cross-Border Digital Technologies Ecosystem: Infrastructure, Platforms & Global Launchpads, Nov, 20, 2025
- ***JSidentify: A Hybrid Framework for Detecting Plagiarism Among JavaScript Code in Online Mini Games***
  - ICSE (International Conference on Software Engineering), Jul, 11, 2020

Professional Service

- Institute of Electrical and Electronics Engineers (IEEE) Member ID: 97841404
- Association for Computing Machinery (ACM) Member ID: 6708618
- China Computer Federation (CCF) Member ID: B8293G
- Reviewer for Conferences
  - Thirty-ninth Conference on Neural Information Processing Systems (NeurIPS 2025)
  - Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)
- Reviewer for Journal
  - IEEE Transactions on Computers
- Program Committee
  - Computer Science Research Methods (CSRM 2023) (INFO5993/ INFO4990 in University of Sydney)
  - ACM International Conference on Architectural Support for Programming Languages and Operating Systems Artifact Evaluation Committee (ASPLOS'24 AEC)
- Web Chair
  - 32nd IEEE International Symposium on High-Performance Computer Architecture (HPCA 2026)

Conference Participation

- ***IEEE/ACM International Symposium on Code Generation and Optimization (CGO) 2026*** In-person Attendance; Jan 31 - Feb 4, 2026
- ***ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP) 2026*** In-person Attendance; Jan 31 - Feb 4, 2026
- ***IEEE International Symposium on High-Performance Computer Architecture (HPCA) 2026*** In-person Attendance; Jan 31 - Feb 4, 2026
- ***ACM SIGPLAN International Conference on Compiler Construction (CC) 2026*** In-person Attendance; Jan 31 - Feb 4, 2026
- ***International Conference on Learning Representations (ICLR 2025)*** In-person Attendance; April 24 - 28, 2025
- ***NVIDIA GPU Technology Conference (GTC 2025)*** In-person Attendance; March 16–21, 2025
- ***USENIX Annual Technical Conference (ATC 2024)*** Online Attendance; July 10 - 12, 2024
- ***ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2023)*** Online Attendance; March 25 - 29, 2023
- ***International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2022)*** Online Attendance; November 13 - 18, 2021
- ***China National Computer Congress (CNCC 2021)*** Online Attendance; December 16 - 18, 2021
- ***International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2021)*** Online Attendance; November 14 - 19, 2021
- ***ACM International Conference on Supercomputing (ISC 2021)*** Online Attendance; June 24 - July 02, 2021
- ***International Symposium on Computer Architecture (ISCA 2021)*** Online Attendance; June 14 - 19, 2021
- ***International Conference on High Performance Big Data and Intelligent Systems (HPBC&IS 2020)*** Online Attendance; May 23 - 23, 2020
- ***International Conference on Software Engineering (ICSE 2020)*** Online Attendance; Invited Talk; Jul 11, 2020

Teaching

- **COMP3520: Operating Systems Internals**
  Fall 2023, Tutor, The University of Sydney

Certification

- ***Pearson Test of English: Listening: 61; Reading: 71; Writing: 64; Speaking: 75; Overall: 67*** March, 21, 2024
- ***Course Certificate:Sample-based Learning Methods*** an online non-credit course authorized by University of Alberta, Alberta Machine Intelligence Institute and offered through Coursera; November, 14, 2021
- ***Course Certificate: Fundamentals of Reinforcement Learning*** an online non-credit course authorized by University of Alberta, Alberta Machine Intelligence Institute and offered through Coursera; Aug, 20, 2021
- ***IELTS Test Scores: Listening: 6.5; Reading: 7.0; Writing: 6.5; Speaking: 6.5; Overall: 6.5*** October, 08, 2020

Other Projects

- **LeetCode Record** Jun. 2017 - Present
  Honing Programming Skills Daily
  CODE LINK
  - Utilized languages such as C, CPP, Python3, Java, and Go to solve LeetCode algorithm questions based on my preference.
  - Maintained a repository containing my code and insights for each LeetCode problem.
- **System Related Conference Papers Crawler** Jun. 2021 - Present
  Web Scraper and Timeline for Top-tier Systems Conference
  CODE LINK
  - Leveraged Python, BeautifulSoup4, and Requests to scrape papers and crucial deadlines for major computer system conferences.
  - Employed Pandas and Matplotlib to create a timeline representing significant computer system paper submission deadlines.
- **DDLs** Dec. 2017 - May. 2018
  Course Project: Design and Development of Android Applications *Guangzhou, China*
  BACKEND CODE LINK FRONTEND CODE LINK
  - Developed DDLs, an Android application for personal deadline management, using Java and Android Studio for the front-end, incorporating MVC architecture, and NodeJS with Express.js for the back-end RESTful API.
  - Implemented features such as deadline administration with CRUD operations, adding, completing, and deleting deadlines in a timeline using SQLite for local storage, marking completed deadlines as unfinished, receiving server notifications through WebSocket, sharing timeline screenshots using Android's native sharing capabilities, and user authentication with JSON Web Tokens (JWT) for registration and login functionality.
- **ChainLoveHelp** May. 2018 - May. 2018
  South China Microsoft Hackathon Competition *Guangzhou, China*
  CODE LINK
  - ChainLoveHelp is dedicated to providing a peer-to-peer platform for university task posting and processing based on blockchain technology.
  - For the chain-end, employed Ethereum-based Parity to construct a consortium blockchain, operating two nodes on the chain for transaction processing, accounting, and consensus.
  - For the front-end, implemented a robust technology stack using PHP for server-side scripting, Apache as the web server, and MySQL for database management.
- **Guang Tu** Apr. 2017 - May. 2017
  South China Microsoft Hackathon Competition *Guangzhou, China*
  CODE LINK
  - Guangtu is a Windows-based map planning software that utilizes gesture recognition technology for enhanced user interaction.
  - The application was developed using Python for programming, Leap Motion for gesture recognition, PyQt5 for creating the graphical user interface, and Django for building the web framework and backend functionality.
- **Seven Seconds** Apr. 2017 - May. 2017
  SYSU Student Software Creative Design and Innovation Development Competition *Guangzhou, China*
  CODE LINK
  - Designed and developed an Android App to organize and record memories, leveraging the capabilities of Android Studio and Java. Successfully published the app on the 360 Mobile App Market.
  - Implemented a robust mobile App architecture, encompassing a user-friendly sidebar, homepage, memory management, as well as secure login and registration modules. Employed advanced data handling techniques, RESTful APIs, and seamless integration with a Node.js backend for efficient data processing and storage.
- **PVmedtech** Jul. 2016 - Aug. 2017
  Advisor: Yao Lu (Professor, SYSU) *Guangzhou, China*
  CODE LINK
  - Collected breast cancer data through web crawling Scrapy.
  - Developed an OHIF Viewer web project, available at LINK.
  - Hosted a **SIT (College Students' Innovative Entrepreneurial Training Plan)**, ID: 201502059.
  - Implemented traditional image processing algorithms on mobile platforms.

Skills

- *Programming Languages:* **Pascal** (11 yrs), **C** (11 yrs), **C++** (11 yrs), **Python** (6 yrs), **HTML, CSS, JavaScript** (6 yrs), **Java** (6 yrs), **SQL** (6 yrs), **Bash** (6 yrs), **LaTeX** (5 yrs), **Matlab** (5 yrs), **CUDA** (5 yrs), **R** (4 yrs), **Go** (4 yrs), **Triton** (2 yrs)
- *Systems and Infrastructure: MPI/OpenMPI* (6 yrs), **Linux Kernel** (5 yrs), **Distributed/Parallel File Systems - e.g., Lustre, HDFS** (5 yrs), **Kubernetes, Kubernetes Scheduler, Kubernetes SR-IOV** (5 yrs), **Docker** (5 yrs), **Hadoop** (4 yrs), **Spark** (4 yrs), **YARN** (4 yrs), **Mesos** (4 yrs), **NVLink** (2 yrs), **NVshMem** (2 yrs), **TensorCore, CudaCore Programming** (2 yrs)
- *Machine Learning and AI:* **TensorFlow** (5 yrs), **PyTorch** (4 yrs), **TorchServe** (4 yrs), **TensorBoard** (4 yrs), **Ray** (4 yrs), **JAX** (2 yrs), **Triton** (2 yrs), **DeepSpeed** (1 yr), **HuggingFace** (1 yr), **Reinforcement Learning** (4 yrs), **CNN, RNN, ResNet, Attention Block, UNet, Transformer, ViT** (5 yrs), **Neural Architecture Search** (3 yrs), **Diffusion Models** (1 yr), **GPT-2,3,4** (1 yr), **Reinforcement learning from human feedback** (1 yr), **VeOmni** (1 yr), **TorchTitan** (1 yr), **FSDP** (1 yr), **Zero 1,2,3** (1 yr)
- *Databases and Storage:* **MySQL** (6 yrs), **Oracle SQL** (6 yrs), **MongoDB** (6 yrs), **PostgreSQL** (6 yrs), **Redis** (4 yrs), **Hive SQL** (3 yrs)
- *Front-end Development:* **PHP** (6 yrs), **Vue.js** (6 yrs), **ReactJS** (6 yrs), **ASP.NET** (6 yrs), **jQuery** (6 yrs), **AngularJS** (6 yrs), **Apache** (6 yrs), **MeteorJS** (6 yrs)
- *Back-end Development:* **Spring Boot** (6 yrs), **Django** (6 yrs), **Flask** (6 yrs), **Node.js** (6 yrs), **Express** (6 yrs), **REST API Design** (6 yrs), **CI/CD**
- *Mobile Development:* **Android Studio** (6 yrs, Java, Kotlin), **XCode** (6 yrs, Swift, Objective-C), **React Native** (6 yrs, Cross-platform), **Flutter** (6 yrs, Cross-platform)
- *Web Crawling & Testing:* **Urllib** (6 yrs), **BeautifulSoup** (6 yrs), **Scrapy** (6 yrs), **Requests** (6 yrs), **JSON** (6 yrs), **Selenium** (6 yrs), **Pytest** (6 yrs), **JUnit** (6 yrs)
- *Version Control & Build Systems:* **Git** (8 yrs), **Gradle** (6 yrs, Android, Java), **Maven** (6 yrs, Java), **npm** (6 yrs, JavaScript), **pip** (6 yrs, Python)
- *Development Tools & Libraries:* **Airflow** (2 yrs), **Kafka** (2 yrs), **Elasticsearch** (2 yrs), **OpenCV** (5 yrs), **Pandas** (5 yrs), **NumPy** (5 yrs), **SciPy** (5 yrs), **NLTK** (5 yrs), **Matplotlib** (5 yrs), **Seaborn** (5 yrs), **Azure Data Factory** (2 yrs), **AWS** (2 yrs), **Google Cloud Platform** (2 yrs)

Extracurricular

- **Fitness: Fencing**(6 yrs), **Jogging** (7 yrs), **Bodybuilding**(6 yrs) (Hongxing Fitness Club Outstanding Students), **Table Tennis**(11 yrs), **Badminton**(11 yrs)
- **Leisure: Web & Mobile Application Development**(5 yrs), **Saxophone**(9 yrs), **Magic** (1 yrs), **Video Games** (more than **500+** PS5 game collections)
- **Volunteer:**
  - ***HPCA 2026 Conference, Volunteer (Registration & On-site Operations)*, Feb, 2026**
    * Supported on-site conference operations including registration check-in, badge distribution, and attendee guidance; coordinated with organizers to ensure smooth session flow and timely room transitions.
    * Assisted speakers and session chairs with logistics (A/V setup, timekeeping, and last-minute schedule updates), and helped handle ad-hoc issues to maintain a professional and welcoming conference experience.
  - ***Sun Yat-sen University, School of Computer Science and Engineering, Student Union, Vice President*, Jul, 2016 - Jul, 2017**
    * Mentored incoming freshmen, helped them acclimate to the university environment, and promoted a sense of belonging through inclusive campus activities and events.
  - ***Changjun High School Volunteer*, Jul, 2013 - Jul, 2014**
    * Enhanced the nursing home experience by engaging in meaningful conversations with elderly residents, preparing and serving fresh fruit, and maintaining a clean and sanitary environment for their well-being.