Zili Zhang

Ph.D. student · Peking University

Yanyuan Building 818, No.5 Yiheyuan Road, Haidian District, Beijing, Republic of China

zzlcs@pku.edu.cn | www.zilizhang.site | Gold-Sea | Zili Zhang

Education

Peking University
Beijing, China

Ph.D. in Computer Science and Engineering — *Sep. 2023 - Present*

- Advisor: Prof. Xin Jin

Peking University
Beijing, China

B.E. in Computer Science and Engineering — *Sep. 2019 - Jun. 2023*

- Overall GPA: 3.71/4.0 (top 9%)

Experience

Software Engineering Institute, PKU.
Beijing, China

Research Assistant — *Oct. 2020 - Present*

- Topic: Machine Learning System and Cloud Computing
- Advisor: Prof. Xin Jin

StepFun.
Beijing, China

Research Intern of LLM Training — *Apr. 2024 - Present*

- Optimization for Alibaba Serverless Computing Platform

Alibaba.
Beijing, China

Research Intern of Serverless Computing — *Jun. 2023 - Apr. 2024*

- Optimization for Alibaba Serverless Computing Platform

Moqi.
Beijing, China

Research Intern of Vector Database — *Oct. 2021 - Feb. 2023*

- Optimization for Faiss runtime

ByteDance Inc.
Beijing, China

Research Intern of DL compiler — *Jun. 2021 - Sep. 2021*

- Optimization for TVM compiler runtime

Teaching Experience

Introduction to Computer System (Honor Track)
Peking University

Teaching Assistant — *Sep. 2022 - Jan. 2023*

- Corrected students' homework and assisted teachers in the examination paper.
- Organized the seminar and prepared topics for discussion.

Introduction to Computer System
Peking University

Teaching Assistant — *Sep. 2021 - Jan. 2022*

- Corrected students' homework and assisted teachers in the examination paper.
- Organized the seminar and prepared topics for discussion.

Publications

DistTrain: Addressing Model and Data Heterogeneity with Disaggregated Training for Multimodal Large Language Models

In Preprint — *Aug 2024*

- **Zili Zhang**, Yinmin Zhong, Ranchen Ming, Hanpeng Hu, Jianjian Sun, Zheng Ge, Yibo Zhu, Xin Jin

RLHFuse: Efficient RLHF Training for Large Language Models with Inter- and Intra-Stage Fusion

In Preprint — *Sep 2024*

- Yinmin Zhong, **Zili Zhang**, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, Xin Jin

RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation

In Preprint *Apr 2024*

- Chao Jin, **Zili Zhang**, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, Xin Jin

Fast Distributed Inference Serving for Large Language Models

In Preprint *May 2023*

- Bingyang Wu*, Yinmin Zhong*, **Zili Zhang***, Gang Huang, Xuanzhe Liu, Xin Jin (* indicates equal contribution)

Fast Vector Query Processing for Large Datasets Beyond GPU Memory with Reordered Pipelining *Santa Clara, U.S.*

NSDI'24 *Apr. 2024*

- **Zili Zhang**, Fangyue Liu, Gang Huang, Xuanzhe Liu, Xin Jin

Jolteon: Unleashing the Promise of Serverless for Serverless Workflows *Santa Clara, U.S.*

NSDI'24 *Apr. 2024*

- **Zili Zhang**, Chao Jin, Xin Jin

dLoRA: Dynamically Orchestrating Requests and Adapters for LoRA LLM Serving *Santa Clara, U.S.*

OSDI'24 *Jul. 2024*

- Bingyang Wu, Ruidong Zhu, **Zili Zhang**, Peng Sun, Xuanzhe Liu, Xin Jin

Fast, Approximate Vector Queries on Very Large Unstructured Datasets *Boston, U.S.*

NSDI'23 *Apr. 2023*

- **Zili Zhang**, Chao Jin, Linpeng Tang, Xuanzhe Liu, Xin Jin

Ditto: Efficient Serverless Analytics with Elastic Parallelism *New York City, U.S.*

SIGCOMM'23 *Apr. 2023*

- Chao Jin, **Zili Zhang**, Xingyu Xiang, Songyun Zou, Gang Huang, Xuanzhe Liu, Xin Jin

Transparent GPU Sharing in Container Clouds for Deep Learning Training *Boston, U.S.*

NSDI'23 *Apr. 2023*

- Bingyang Wu, **Zili Zhang**, Zhihao Bai, Xuanzhe Liu, Xin Jin

Rise of Distributed Deep Learning Training in the Big Model Era: From A Software Engineering Perspective

TOSEM'23 *May. 2023*

- Xuanzhe Liu , Diandian Gu, Zhenpeng Chen, Jinfeng Wen, **Zili Zhang**, Yun Ma, Haoyu Wang, Xin Jin

Optimizing Half Precision Winograd Convolution on ARM Many-Core Processors *Online*

APSYS'22 *Aug. 2022*

- Dedong Xie, Zhen Jia, **Zili Zhang**, Xin Jin

Honors & Awards

University Awards

2024	**Presidential Scholarship of Peking University (4/200)**, Peking University Award	*Peking University*
2023	**Top 10 Bachelor Thesis of Peking University, EECS (10/408)**, Graduation Ceremony	*Peking University*
2023	**Outstanding Graduation Thesis of Peking University (33/4239)**, Graduation Ceremony	*Peking University*
2023	**Representor of Excellent Graduates (14/4239)**, Graduation Ceremony	*Peking University*
2023	**Excellent Graduates (613/4239)**, Graduation Ceremony	*Peking University*
2022	**Exceptional Award for Academic Innovation (5/408)**, Peking University Award	*Peking University*
2022	**Award for Scientific Research**, Peking University Award	*Peking University*
2022	**Lee Wai Wing Scholarship**, Peking University Award	*Peking University*
2021	**Award for Scientific Research**, Peking University Award	*Peking University*
2020	**Award for Academic Excellents**, Peking University Award	*Peking University*
2020	**Third Prize**, Peking University Award	*Peking University*

Skills & Interests

Programming	AWS, Docker, Kubernetes, C++, Python, CUDA, Java, Node.js, Latex
Languages	Chinese (native), English
Interests	Running (www.itra.run/RunnerSpace/ZHANG.Zili/4938114), Climbing, Traveling, Video Games