

# 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

PH.D. IN COMPUTER SCIENCE AND ENGINEERING

- Advisor: Prof. Xin Jin

Beijing, China

Sep. 2023 - Present

### Peking University

B.E. IN COMPUTER SCIENCE AND ENGINEERING

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

Beijing, China

Sep. 2019 - Jun. 2023

## Service

2025 **Shadow PC**, The 20th edition of EuroSys (EuroSys 2025)

Rotterdam

## 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

### 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)

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

Philadelphia, PA, U.S.

NSDI'25

Apr. 2025

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

### 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

APSYS'22

Online

Aug. 2022

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

## Internship

---

### StepFun.

RESEARCH INTERN OF LLM TRAINING

Beijing, China

Apr. 2024 - Present

- Optimization for Multimodal and RL Training

### Alibaba.

RESEARCH INTERN OF SERVERLESS COMPUTING

Beijing, China

Jun. 2023 - Apr. 2024

- Optimization for Alibaba Serverless Computing Platform

### Moqi.

RESEARCH INTERN OF VECTOR DATABASE

Beijing, China

Oct. 2021 - Feb. 2023

- Optimization for Faiss runtime

### ByteDance Inc.

RESEARCH INTERN OF DL COMPILER

Beijing, China

Jun. 2021 - Sep. 2021

- Optimization for TVM compiler runtime

## Teaching

---

### Operating System (Honor Track)

TEACHING ASSISTANT

Peking University

Feb. 2024 - July. 2024

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

### Introduction to Computer System (Honor Track)

TEACHING ASSISTANT

Peking University

Sep. 2022 - Jan. 2023

- Correcting students' homework and assisted teachers in the examination paper.
- Organizing the seminar and prepared topics for discussion.
- Coding for the course project.

### Introduction to Computer System

TEACHING ASSISTANT

Peking University

Sep. 2021 - Jan. 2022

- Correcting students' homework and assisted teachers in the examination paper.
- Organizing the seminar and prepared topics for discussion.
- Coding for the course project.

## Honors & Awards

---

### UNIVERSITY AWARDS

2024	<b>Merit Student of Peking University (1/56)</b> , PKU Award	Peking University
2024	<b>National Scholarship (Top 0.4% nationally)</b> , PKU Award	Peking University
2024	<b>Presidential Scholarship of Peking University (highest honor for Ph.D. students)</b> , PKU Award	Peking University
2023	<b>Top 10 Bachelor Thesis of Peking University, EECS (10/408)</b> , Graduation Ceremony	Peking University
2023	<b>Outstanding Graduation Thesis of Peking University (33/4239)</b> , Graduation Ceremony	Peking University
2023	<b>Representor of Excellent Graduates (14/4239)</b> , Graduation Ceremony	Peking University
2023	<b>Excellent Graduates (613/4239)</b> , Graduation Ceremony	Peking University
2022	<b>Exceptional Award for Academic Innovation (5/408)</b> , PKU Award	Peking University
2022	<b>Award for Scientific Research</b> , PKU Award	Peking University
2022	<b>Lee Wai Wing Scholarship</b> , PKU Award	Peking University
2021	<b>Award for Scientific Research</b> , PKU Award	Peking University
2020	<b>Award for Academic Excellents</b> , PKU Award	Peking University
2020	<b>Third Prize</b> , PKU Award	Peking University

## Skills & Interests

<b>Programming</b>	AWS, Docker, Kubernetes, C++, Python, CUDA, Java, Golang, Node.js, Latex
<b>Languages</b>	Chinese (native), English
<b>Interests</b>	Running ( <a href="http://www.itra.run/RunnerSpace/ZHANG.Zili/4938114">www.itra.run/RunnerSpace/ZHANG.Zili/4938114</a> ), Climbing, Traveling, Video Games