ZIHANG LIU

☐ +1 3106259919 | % Homepage | @ zihang.liu@berkeley.edu | 🔗 LinkedIn | ⚙ GitHub | 📍 Berkeley, CA

EDUCATION

University of California, Berkeley Berkeley, CA
Master in Electrical Engineering and Computer Science *Aug 2024 – Present*

Beijing University of Posts and Telecommunications Beijing, China
B.Eng. in Intelligence Science and Technology *GPA: 90%, Rank: 4/68* *Sep 2020 – Jun 2024*

Core courses: *Linear Algebra 98%, Discrete Mathematics 93%, Probability Theory and Statistics 93%, Neural Network and Deep Learning 97%, Design and Analysis of Algorithms 93%, Operating System 90%, Introduction to Reinforcement Learning 95%*

Research Interests: (LLM) Model Diagnosis, Robust Machine Learning, Efficient Model Training, Scientific ML.

PUBLICATIONS

1. **Model Balancing Helps Low-data Training and Fine-tuning** EMNLP 2024 Oral
   Zihang Liu, Yuanzhe Hu, Tianyu Pang, Yefan Zhou, Pu Ren, Yaoqing Yang

2. **EnsembleMOT: A Step Towards Ensemble Learning of Multiple Object Tracking** arXiv preprint
   Yunhao Du, Zihang Liu, Fei Su

RESEARCH EXPERIENCE

Reinventing Numerical Algorithms with Large Language Model Berkeley, CA
*Graduate Researcher, Advised by **Prof. Michael Mahoney** at **UC Berkeley*** *Aug 2024 – Present*

- Designed and implemented a transformer-based linear system solver that utilizes embedding and self-attention to learn preconditioning through in-context learning. Reaches comparable performance to conjugate gradient and steepest descent method.
- Designed an algorithm discovery pipeline with LLMs such as LLaMA-3.1-70B and GPT-4o. Introducing MCTS decoding algorithm to numerical algorithm discovery for the first time. Conducted empirical analysis of LLM's function discovery ability with level-1 and level-2 reasoning.

Model Diagnosis with Weight Analysis Hanover, NH
*Undergraduate Researcher, Advised by **Prof. Yaoqing Yang** at **Dartmouth College*** *Jul 2023 – Jul 2024*

- Proposed a layer-wise learning rate scheduler based on heavy-tailed self-regularization theory(HT-SR), that balances temperature parameters of neural network models.
- Modeled the heavy-tail behavior of optimizers (SAM) and model architectures. Rescheduling the learning rate to optimize the regularization effects, significantly improving performance on Image Classification and Language Modeling (NeurIPS 2023 Spotlight)
- Diagnosing the limitations of low-data training using Heavy-Tail metrics, and propose layer-wise model balancing to achieve model alignment, achieving up to 10% improvement in LLM fine-tuning. (EMNLP 2024 Oral)

Ensemble Methods in Multiple Object Tracking Beijing, China
*Undergraduate Research Assistant, Advised by **Prof. Fei Su** at **BUPT*** *Jun 2022 – Jan 2023*

- Proposed a model-independent ensemble method that integrates results from various MOT trackers to achieve higher overall performance, which we named EnsembleMOT.
- Proposed to use both spatial and temporal IoU(Intersection over Union) to merge and prune trajectories, achieving 3% improvement in MOTA and IDF1, alleviating ID-switch and abnormal bounding box problems.
- Co-authored a paper with an MCPRL lab member and submitted our work to the International Conference on Acoustics, Speech and Signal Processing (Submitted to ICASSP 2023).

PROJECTS AND INTERNSHIPS

Interpretable Multimodal Question Answering

- Implemented the interpretable visual question answering pipeline through modular code generation. Leverages code-generation LLMs such as CodeLLaMA, GPT-4o, Gemini-pro, etc, and Multimodal LLMs such as BLIP, BLIP-v2.
- Designed visual-augmented code generation that proposes visual information when generating executable code. Reduced syntax error and improve interpretability of reasoning through code.
- Brought improvement on GQA and OK-VQA datasets compared with baseline such as BLIP-v2, GPT-4o, and sViperGPT.

Reliable dialogue system trained with continuous learning and parameter-efficient fine-tuning

- Designed a dialogue system with improved consistency and reliability of responses from learning new knowledge while retaining existing knowledge.
- Constructed a GPT-2-based dialogue system with an interactive interface, fine-tuned with open-source dialogue datasets and parameter-efficient tuning.
- Proposed a dynamic learning strategy that combines Elastic Weight Consolidation (EWC) and data replay to improve OOD generalization while alleviating the catastrophic forgetting problem.

Backend Software Engineering Internship   Leadingtek Corp, Beijing

- Implemented efficient data structures such as hash tables and binary search trees for rapid data access. Utilized algorithms like quicksort for data organization, ensuring efficient data integration, storage, and retrieval.
- Designed optimized and structured SQL commands, focusing on complex joins, indices, and stored procedures to enhance the speed and accuracy of information retrieval.
- Collaborated in an agile setting with developers, QA, and product managers, contributing to sprint activities and code quality.

AWARDS & SERVICES

Reviwer: NeurIPS 2024 Workshop on Foundation Models for Science

Services: EMNLP 2024 Student Volunteer Coordinator

Grants: Berkeley Conference Travel Grant (2024)

Scholarships: First-class Scholarship (**Ranked 1/68**, 2023), Second-class Scholarship (**Ranked 4/68**, 2021, 2022)

National Mathematics Competition for College Students: Second Place (Top 10%), 2022

"Internet+" Innovation and Entrepreneurship Competition: Third Place (Top 10%), 2022

TALKS

EMNLP 2024	Model Balancing Helps Low-data Training and Fine-tuning (Recording)	Nov 12, 2024

SKILLS

English Proficiency: TOEFL IBT **114** (reading 30 listening 29, speaking 27, writing 28), GRE **328** (Quant 170)

Programming Languages: Python, C/C++, SQL, Rust, VHDL

Frameworks: Linux, Pytorch, Git, Slurm

Sports: I was a tennis athlete representing BUPT and have entered quarterfinals in regional and national championships as a doubles player.