Alan Zhu

Berkeley, California, USA
aczhu@berkeley.edu
https://www.linkedin.com/in/az1326/ | https://github.com/az1326

Education

University of California, Berkeley

Ph.D. in Computer Science | August 2024 - Present

Carnegie Mellon University

B.S. in Computer Science, Additional Major in Statistics | August 2020 - May 2024

- GPA: 4.0/4.0
- University Honors; Dean's List, High Honors; Senior Leadership Recognition Award

Awards

NSF Graduate Research Fellowship Recipient

- Awarded 2024

Peer-Reviewed Publications

- Zhang, Z., **Zhu, A.**, Yang, L, Xu, Y., Li, L., Phothilimthana, P. M., Jia, Z. 2024. Accelerating Retrieval-augmented Language Model Serving with Speculation. To appear at ICML 2024.
- Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Wong, R. Y. Y., **Zhu, A.**, Yang, L., Shi, X., Shi, C., Chen, Z., Arfeen, D., Abhyankar, R., Jia, Z. 2024. SpecInfer: Accelerating Generative Large Language Model Serving with Speculative Inference and Token Tree Verification. ASPLOS 2024.
- Cui, Z., Wang, S., Han V. Y., Rae-Gran, T., Yang, W. Y., **Zhu, A**., Hudson, S. E., Ion, A. 2024. Robotic Metamaterials. CHI 2024.

Research Experience

University of California, Berkeley; Department of Electrical Engineering and Computer Science; Sky Computing Lab

Graduate Student Researcher | August 2024 - Present

- Advised by Prof. Joseph Gonzalez.
- Working on developing frameworks and methods for training, serving, and controlling agentic AI systems, with a focus on agentic LLM systems.

Carnegie Mellon University, School of Computer Science, Catalyst Lab

Undergraduate Research Assistant | August 2022 - May 2024

- Worked under the supervision of Prof. Zhihao Jia on various ML Systems projects, including systems for speculative decoding and retrieval augmented generation. Work resulted in publications at ASPLOS and ICML.

University of Michigan, School of Information

Undergraduate Research Assistant | May 2021 - May 2024

- Worked under the supervision of Prof. Qiaozhu Mei on a project related to shortest path distances to samples in a graph, with implications on fairness in graph neural networks. First author manuscript in preparation for WWW 2025.

Carnegie Mellon University, School of Computer Science, Interactive Structures Lab

Undergraduate Research Assistant | May 2021 - May 2022

- Worked under the supervision of Prof. Alexandra Ion on a project developing robotic metamaterials, scalable and robust grids of cells able to produce motions by controlling actuated cells. Work resulted in a publication at CHI.

Teaching Experience

Carnegie Mellon University | Fall 2021 - Spring 2024

15-251 Great Ideas in Theoretical Computer Science

Fall 2023 - Spring 2024: Head Teaching Assistant

Fall 2023: Associate Head Teaching Assistant

Fall 2021 - Spring 2023: Teaching Assistant

- Led recitations and office hours, graded and provided feedback on student work, managed course logistics, and trained new course staff. Contributed towards development of new course material and infrastructure.

Work Experience

Apple | June 2022 - August 2022

AI/ML Intern

- Worked as part of Siri's Search Data Analytics team and developed a tool to detect heterogeneous effects in A/B tests regression-based variance reduction methods.

Skills

- Programming Languages: Python, R, C, C++, Java
- Scientific Packages: NumPy, SciPy, PyTorch, Tensorflow, NetworkX, Pyplot
- Tools: Google Sheets, Google Docs, LaTeX, Git, Conda, Docker
- Languages: English (native), Mandarin (fluent), French (intermediate)