Alan Zhu

3079 Cedarbrook Rd
Ann Arbor, Michigan, USA
(734) 272-7011
aczhu@andrew.cmu.edu | alanzhu2002@gmail.com
https://az1326.github.io
https://www.linkedin.com/in/az1326/ | https://github.com/az1326

Education

September 2020 - May 2024 (expected)

Carnegie Mellon University

*Bachelor of Science in **Computer Science**, Additional Major in **Mathematical Statistics***

- GPA: 4.0/4.0
- Dean's List, High Honors
- Relevant Coursework:
    - Deep Reinforcement Learning
    - Intermediate Deep Learning
    - Machine Learning Systems
    - Probability and Mathematical Statistics (Graduate Level)
    - The ABCDE of Statistical Machine Learning (Graduate Level)
    - Advanced Data Analysis
    - Real Analysis I and II

Manuscripts

- **Zhu, A.**, Ma, J., Mei, Q. Examining Shortest-Path Distances to Samples in Graphs. Submitted to ICWSM 2024.
- Zhang, Z., **Zhu, A.**, Yang, L., Xu, Y., Li L., Phothilimthana, P.M. Jia, Z. Accelerating Retrieval-augmented Language Model Serving with Speculation. Submitted to 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. SpecInfer: Accelerating Generative Large Language Model Serving with Speculative Inference and Token Tree Verification. Revise and Resubmit at ASPLOS 2024.
- Cui, Z., Wang, S., Han V. Y., Rae-Gran, T., Yang, W. Y., **Zhu, A**., Hudson, S. E., Ion, A. Robotic Metamaterials. To appear at CHI 2024.

Research Experience

Carnegie Mellon University, School of Computer Science

Undergraduate Research Assistant, Catalyst Lab | August 2022 - Present

- Working under the supervision of Prof. Zhihao Jia
- ***Accelerating Retrieval-augmented Language Model Serving with Speculation***
    - Developed system for latency improvements to retrieval-augmented language model pipelines by speculative retrieving documents from a cache
    - Parallel batch verification of cache-retrieved documents is faster than sequential retrievals from full database
    - Designed prefetching and adaptive speculation stride scheduler to improve cache hit rate
    - Built additional datasets for evaluation
    - Extended technique to k-nearest-neighbor language models
    - Second author on manuscript submitted to ICLR 2024
- ***SpecInfer: Accelerating Generative Large Language Model Serving with Speculative Inference and Token Tree Verification***
    - Speculative inference uses fast small models to predict the output of a slow large model, and verifies predictions in parallel, improving latency
    - Wrote proof demonstrating the superiority of our verification method over naive methods
    - Manuscript received Revise and Resubmit for ASPLOS 2024
- ***Model Performance Inference (MPI) on Language Models***
    - Created search space consisting of compressed language models upon which MPI metrics could be evaluated
    - Designed and tested new MPI metrics on the search space
    - Determined that existing MPI metrics (designed for computer vision settings) do not work well on language model settings

Undergraduate Research Assistant, Interactive Structures Lab | May 2021 - May 2022

- Worked under the supervision of Prof. Alexandra Ion
- ***Robotic Metamaterials***
    - Robotic metamaterials are grids of deformable or rigid square cells which can produce motions by actuating some cells able to control their own deformation
    - Implemented algorithm to design robotic metamaterials given desired mechanism motions, optimized for stability, build cost, and motion accuracy
    - Assisted in physical prototyping and manuscript writing
    - Manuscript received Revise and Resubmit for CHI 2024

University of Michigan, School of Information

Undergraduate Research Assistant | May 2021 - Present

- Working under the supervision of Prof. Qiaozhu Mei
- ***Investigating Fairness of Graph Sampling through Shortest Path Distances***
    - Ran simulations of various graph sampling techniques on various graphs to obtain shortest path distances
    - Adapted frameworks for theoretically estimating the distribution of shortest path distances found in literature to our specific case where the sample consists of multiple nodes and are not necessarily selected at random
    - Proposed heuristic for comparing sampling methods based on my framework and demonstrated its accuracy through simulations
    - Wrote manuscript as primary author
    - First author on manuscript submitted to ICWSM 2024

Teaching Experience

September 2021 - Present

Carnegie Mellon University

Great Ideas in Theoretical Computer Science

Head Teaching Assistant | September 2023 - Present

- **Manage course logistics**
- **Train new teaching assistants**
- **Maintain timely completion of responsibilities**
- **Inform professors of course staff and students' morale and feedback**
- Common responsibilities
    - Lead recitations and office hours
    - Grade student work and provide feedback
    - Create material for and lead small group reviews
    - Respond to student questions online

Associate Head Teaching Assistant | January 2023 - September 2023

- Assisted Head Teaching Assistant with bolded responsibilities above
- Performed common responsibilities listed above

Teaching Assistant | September 2021 - January 2023

- Served as Infrastructure team lead
    - Managed course gradebook, including importing and exporting scores and making scores available for students to view
- Performed common responsibilities listed above

Work Experience

June 2022 - August 2022

Apple

AI/ML Intern

- Worked as part of Siri's Search Data Analytics team
- Developed tool to detect heterogeneous effects in A/B tests
- Presented tool to senior management

Skills and Miscellaneous

- Programming Languages: Python, C, C++, Java, R
- Scientific Packages: NumPy, SciPy, PyTorch, Tensorflow, NetworkX, Pyplot, Pandas
- Tools: Google Sheets, Google Docs, LaTeX, Git, Conda, Docker
- Languages: English (native), Mandarin (fluent), French (intermediate)
- USAJMO (2016) and USAMO (2019) qualifier