Zain Sarwar

309-703-5234 | zsarwar@uchicago.edu | Linkedin | Webpage | Google Scholar | Github

RESEARCH INTERESTS

Designing sparse architectures for LLMs, Mixture of Experts, Inference optimization for LLMs, RAG, Data Markets

EDUCATION

University of Chicago Chicago, IL
PhD in Computer Science Aug 2025 (Expected)
GPA: 3.91/4.0
Relevant Coursework: Deep Learning, Machine Learning, Operating Systems, NLP, Algorithms & Distributed systems

PROFESSIONAL EXPERIENCE

Capital One New York, NY
Applied research intern - LLM Pretraining *June 2024 – Sept 2024*

Project : Improving Mixture of Experts with routed structured matrices

- Developed a method for augmenting experts in MoEs with structured matrices using hierarchical routing
- This method provides an alternative scaling strategy for MoE based LLMs which has a more efficient inference profile and outperforms a parameter and FLOP matched standard MoE in large scale pretraining.
- Evaluated method by pretraining 125M, 410M and 1.2B parameter models on multiple pretraining datasets, routing schemes, total activated parameters and learning rate schedules

University of Chicago Chicago, IL
Research Assistant *September 2021 – Present*

Ongoing projects

- Developing an algorithm for improving image classification models which finds gaps in the training data and optimally fills them by finding the most useful data in external datasets using supervised and unsupervised algorithms. This algorithm can be used to value large-scale private datasets in data markets

Past projects

- Invented a first-of-its-kind safety analysis tool for LLMs which can detect the model's tendency to hallucinate sensitive information by crafting semantically meaningful prompts using a retrieval mechanism
- Engineered state-of-the-art techniques for detecting text generated from Large Language models using graph neural networks that can be used to detect LLM generated fake news and hate speech
- Designed and developed the first ever generalizable video and VR based keystroke inference attack which uses a transformer based hand tracking and keystroke classification model trained on pseudo labels generated from hidden markov models to infer a user's typed content
- Created a voice privacy protection tool which prevents deep learning models from cloning an individual's voice to protect against identity theft using a voice anonymizing neural model

SELECTED PUBLICATIONS

Deepfake Text Detection: Limitations and Opportunities
Jiameng Pu, **Zain Sarwar**, Sifat Muhammad Abdullah, Abdullah Rehman, Mobin Javed, and Bimal Viswanath
IEEE S&P (Oakland) 2023

Can Virtual Reality Protect Users from Keystroke Inference Attacks?
Zhuolin Yang, **Zain Sarwar**, Iris Hwang, Ronik Bhaskar, Ben Y. Zhao, Haitao Zheng
USENIX Security Philadelphia, PA, August 2024.

TECHNICAL SKILLS

Languages: Python, Triton, C, C++, Java, JavaScript, Go, Haskell, Matlab, SQL
Libraries: PyTorch, DeepSpeed, Megatron-LM, TensorFlow, OpenCV, scikit-learn, pandas, NumPy, Selenium
Frameworks: Angular, React, Git, Docker, Flask, Node.js, Vue