Quan Mai

✉ quanmai@uark.edu
https://www.linkedin.com/in/quanmai-rzb https://quanmai.github.io

Profile

Ph.D. Candidate in Computer Engineering with expertise in deep learning, NLP, and high-performance computing (HPC). Experienced in implementing, training, and fine-tuning large language models (LLMs). Published researcher specializing in LLMs and neural retrieval. Previously worked for three years in the semiconductor industry, designing high-speed SRAMs. Proficient in Python, with experience in C++, CUDA, and DPC++.

Education

University of Arkansas, Fayetteville **Jan 2020 – Jun 2025**
PhD in Computer Engineering, GPA: 4.00/4.00

Danang University of Science & Technology, Vietnam **Aug 2011 – Jun 2016**
Bachelor of Engineering in Electrical and Electronics, GPA: 3.44/4.00, top 5% of Department

Skills

Languages: Python, C++, CUDA, DPC++, R

Frameworks and Libraries: PyTorch, Tensorflow, Hugging Face, Scikit-Learn, NLTK, Spicy

Experience

Research Assistant **Jan 2023– Present**
NLP Lab *University of Arkansas*

- Conducted advanced research in NLP, LLMs and information retrieval.
- Published research on enhanced retrieval performance as well as language understanding.
- Mainly focused on: LLMs, Mixture-of-Expert, Parameter-Efficient Finetuning, RAGs, Memory in LLMs.

Graduate Intern **Jan 2022– May 2022**
HPC Solution Architect *Intel, Oregon*

- Designed and implemented a molecular dynamics simulation using Intel OneAPI DPC++, achieving a 10x improvement in performance over standard C++ implementations.

Research Assistant **Jan 2020– May 2021**
Computer System Lab *University of Arkansas*

- Developed optimized solutions on HPC environments for computation and data-intensive simulations.
- Co-author of **CuSMC** - CUDA Sequence Monte Carlo package (CUDA, C++, R).

IP Design Engineer **Aug 2016– Oct 2019**
Circuit Design Team *eSilicon Vietnam (now Synopsys Inc.)*

- Specialized in developing high-speed and ultra-high-speed Pseudo Two-Port (P2P) SRAMs using cutting-edge semiconductor technologies, including 28nm, 14nm, 10nm, 7nm, and 5nm processes.

Selected Publication

Boolean-Aware Attention for Dense Retrieval
Q. Mai, S. Gauch, D. Adams
Submitted to ACL and currently under review

SetBERT: Enhancing Retrieval Performance for Boolean Logic and Set Operation Queries
Q. Mai, S. Gauch, D. Adams
2024 Eighth International Conference on Natural Language Processing and Information Retrieval

Sequence Graph Network for Online Debate Analysis
Q. Mai, S. Gauch, D. Adams, M. Huang
2024 International Conference on Information, Process, and Knowledge Management

BrainVGAE: end-to-end graph neural networks for noisy fMRI dataset
Q. Mai, U. Nakarmi, M. Huang
2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)