Yifan Qin

yqin3@nd.edu | (574) 401-9049 | Website | Google Scholar | LinkedIn

Summary

Ph.D. candidate working at the intersection of machine learning and computer architecture, focusing on inference-efficient LLMs and hardware-aware optimization under realistic system and device constraints. Research spans algorithm–system co–design, robustness/uncertainty-aware training, and runtime–oriented evaluation.

Education

University of Notre Dame **Notre Dame, IN**
Ph.D. candidate in Computer Science and Engineering *2022–present*

Huazhong University of Science and Technology **China**
B.S. in Electronic Science and Technology; M.S. in Software Engineering *2013–2021*
M.S. Efficient ML track

Experience

Research Assistant **Notre Dame, USA**
University of Notre Dame *2022–present*

- Developed hardware aware training and inference strategies for deep learning models on resource and device constrained accelerators (e.g., compute-in-memory), introducing a negative feedback based optimization framework (NeFT) to improve robustness and accuracy–efficiency tradeoffs under device variation and runtime constraints.
- Proposed a lightweight parameter sharing mechanism (TSB) for parameter efficient fine tuning and inference under memory and runtime constraints, reducing inference overhead while preserving model performance in hardware constrained settings.
- Developed the software and uncertainty aware inference algorithms for an end-to-end arrhythmia detection engine, including uncertainty analysis to characterize prediction reliability under noise and device variation, and supported an ISSCC 2025 silicon prototype demo achieving 1.75 μJ per inference under practical latency and energy constraints.
- Contributed to a tail robustness training approach using right censored noise to improve realistic worst case performance under hardware noise and variation, and the resulting paper received the ICCAD Best Paper Award.

Research Internship **Hong Kong**
AI Chip Center for Emerging Smart Systems(ACCESS) *2024.05–2024.07*

- Co-designed an end to end arrhythmia detection system with a CNN accelerator under strict latency and energy constraints, translating model requirements into deployment ready algorithm and system decisions.
- Optimized the model for hardware constraints, including quantization and pruning, reducing inference cost while preserving detection accuracy in an accelerator based pipeline.
- Implemented and integrated the software stack for real time inference and monitoring, supporting an ASP-DAC system demo on hardware prototypes at 10.60 μW and 150 GOPS under practical runtime constraints.

Research Assistant **China**
Huazhong University of Science and Technology *2018–2022*

- Studied hardware aware low precision learning and inference for CNN accelerators, improving robustness and accuracy–efficiency tradeoffs under non ideal device behavior in memristive and crossbar based computing.
- Published two papers on robust and energy efficient accelerator oriented inference, including a journal back cover feature.

Selected Publications

- **Yifan Qin**, Zheyu Yan, et al., "NeFT: Negative Feedback Training to Improve Robustness of Compute In Memory DNN Accelerators", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD'25).
  — *Hardware-aware training that improves robustness and accuracy–efficiency under variation.*
- **Yifan Qin**, Zheyu Yan, et al., "Sustainable Deployment of Deep Neural Networks on Non-Volatile Compute-in-Memory Accelerators", International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS'24).
- Likai Pei*, **Yifan Qin***, et al., "Towards Uncertainty Quantifiable Biomedical Intelligence: Mixed signal Compute in Entropy for Bayesian Neural Networks", IEEE/ACM International Conference on Computer-Aided Design (ICCAD'24) (acceptance rate 24%) (**Best Paper Award Candidate, 10 out of 802 submissions**).
  — *Bayesian inference design enabling uncertainty quantification under hardware constraints.*
- **Yifan Qin**, Zheyu Yan, et al., "TSB: Tiny Shared Block for Efficient DNN Deployment on NVCIM Accelerators", IEEE/ACM International Conference on Computer-Aided Design (ICCAD'24) (acceptance rate 24%).
  — *Lightweight parameter sharing for parameter efficient fine tuning under memory and runtime constraints.*
- Zheyu Yan, **Yifan Qin**, et al., "Improving Realistic Worst Case Performance of NVCIM DNN Accelerators Through Training With Right Censored Gaussian Noise", IEEE/ACM International Conference on Computer-Aided Design (ICCAD'23) (acceptance rate 22.9%) (**Best Paper Award, 2 out of 750 submissions**).
  — *Tail robustness training with right censored noise for improved worst case performance.*
- **Yifan Qin**, Zhenge Jia, et al., "A 10.60 μW 150 GOPS Mixed Bit Width Sparse CNN Accelerator for Life Threatening Ventricular Arrhythmia Detection", Asia and South Pacific Design Automation Conference (ASP-DAC'25) (demo on a silicon prototype).
- Jianbo Liu, Zephan Enciso, Boyang Cheng, Likai Pei, Steven Davis, **Yifan Qin**, et al., "A 65 nm Uncertainty Quantifiable Ventricular Arrhythmia Detection Engine With 1.75 μJ per Inference", IEEE International Solid-State Circuits Conference (ISSCC'25) (acceptance rate 27%) (demo on a silicon prototype).

Selected Demos

- **Real-time Ventricular Arrhythmia Detection: Efficient ML Inference System** (ASP-DAC'25). Video.
- **Uncertainty-aware ML Inference for Real-time Ventricular Arrhythmia Detection** (ISSCC'25). Video.

Selected Award

- IEEE/ACM ICCAD William J. McCalla Best Paper Award (2023)
- IEEE/ACM ICCAD William J. McCalla Best Paper Award Candidate (2024)
- DAC Young Fellow (2023–2025)
- Outstanding Graduate, Huazhong University of Science and Technology (2021)
- National Second Prize, China Undergraduate Mathematical Contest in Modeling (CUMCM; Ministry of Education, China) (2015)

Patents

- CN202011251999.9, A hardware neural network batch normalization system (Issued 5/20/2022).
- CN201910792384.8, A matrix vector multiplication circuit and calculation method (Issued 10/8/2021).

Skills

Technical: Python (PyTorch, NumPy, Pandas), C++, ONNX; Linux, Git, LaTeX

ML & Systems: Hardware aware ML and accelerator constrained inference; parameter efficient fine tuning; robustness and uncertainty evaluation under noise and variation; quantization and low precision deployment; runtime profiling and inference optimization