Yifan Qin

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

Summary

Ph.D. candidate in ML systems specializing in post-training optimization for efficient inference under realistic system constraints, with experience in reliability/uncertainty evaluation and prototype-to-demo integration.

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*

Experience

Research Assistant — **Notre Dame, IN**
University of Notre Dame — *2022–present*

- **Training and post-training for efficient inference:** Developed hardware-aware training and post-training strategies, including negative feedback training (NeFT) and lightweight parameter sharing fine-tuning (TSB), to improve inference robustness under noise while controlling runtime and compute overhead.
- **Uncertainty-aware inference engine (demo on silicon):** Implemented the software and uncertainty aware inference pipeline for an end-to-end arrhythmia detection engine; performed reliability analysis under noise/variation and supported a silicon-prototype demo achieving 1.75 μJ per inference.
- **Tail reliability under variation:** Contributed to a right-censored noise training approach that improves realistic worst-case performance under hardware noise and variation.

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

- **End-to-end inference pipeline:** Co-designed an arrhythmia detection system under strict latency and energy constraints, translating model requirements into deployment-ready algorithm and system choices.
- Implemented and integrated the software stack for real-time inference and monitoring on hardware prototypes; applied quantization and structured pruning, supporting a system demo that achieved 10.60 μW and 150 GOPS while preserving over 99% detection accuracy.

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

- Developed quantization-aware and low-bit CNN methods for efficient inference under non-ideal hardware behavior (e.g., device noise and variability), with results published in journals (including a back-cover feature).

Selected Projects & Demos

ISSCC'25: **Uncertainty-aware arrhythmia detection engine (demo on a silicon prototype).** Achieved 1.75 μJ per inference with uncertainty-aware inference and reliability evaluation under practical constraints. Demo video

ASP-DAC'25: **Real-time arrhythmia detection pipeline on ML accelerator (demo on a silicon prototype).** Integrated real-time inference and monitoring stack; demo achieved 10.60 μW and 150 GOPS with quantization and structured pruning. Demo video

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).
  — *Training-time optimization that improves robust inference under noise/variation with no extra inference overhead.*
- **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 weight sharing for parameter-efficient fine-tuning and inference under memory/runtime constraints.*
- 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**).
  — *Uncertainty-aware inference enabling reliability evaluation and deployment under system 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-robust training method targeting realistic worst-case inference behavior under noise and variation.*
- **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) (ML system demo).
- 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) (ML system demo).

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).

Awards

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

Skills

ML & Inference: PyTorch; post-training optimization (parameter-efficient fine-tuning, quantization, pruning); robustness and uncertainty evaluation; benchmarking under system constraints

Systems: C++; Linux; Git; performance analysis and profiling

Tools: NumPy, Pandas; LaTeX