Haomin (Harmin) Qi

harminchee.github.io

EDUCATION

University of California San Diego Sep. 2025 – Jul. 2027
M.S. in Electrical and Computer Engineering La Jolla, CA
GPA: 3.8 /4.0

The Chinese University of Hong Kong Sep. 2021 – Jul. 2025
B.S. in Mathematics and Information Engineering Hong Kong SAR
Double Major Graduation | Elite Stream Class

University of Leeds Jan. 2024 – Jul. 2024
Abroad Exchange in Computer Science Leeds, UK
GPA: 4.00 /4.00 | First Class Honour

PUBLICATIONS

TopoEdge: An Edge-assisted LLM Framework for Automated SDN Configuration Generation
Haomin Qi, Yuyang Du, Ziheng Kang, Yue Zhan, Soung Chang Liew
Under review at IEEE Network Operations and Management Symposium ***(IEEE NOMS 26')***

VeriRAG: A Retrieval-Augmented Framework for Automated RTL Testability Repair
Haomin Qi, Yuyang Du, Lihao Zhang, Soung Chang Liew, Kexin Chen, Yining Du
Under review at The International Symposium on Quality Electronic Design ***(ISQED 26')***

Governance-Aware Hybrid Fine-Tuning for Multilingual Large Language Models
Haomin Qi, Chengbo Huang, Zihan Dai, Yunkai Gao
IEEE International Conference on Big Data 2025 Workshop LLM4All ***(IEEE BigData 25'LLM4All)***

GraphCue for SDN Configuration Code Synthesis
Haomin Qi, Fengfei Yu, Chengbo Huang
IEEE Consumer Communications & Networking Conference 2026 ***(IEEE CCNC 26')***

Hybrid and Unitary PEFT for Resource-Efficient Large Language Models
Haomin Qi, Zihan Dai, Chengbo Huang
American Journal of Computer Science and Technology ***(AJCST)***

Transforming ABA Therapy: An IoT-Guided, Retrieval-Augmented LLM Framework
Haomin Qi, Chung-Ho Sin, Rosanna Yuen-Yan Chan, Victor Chun-Man Wong
IEEE Access

EXPERIENCE

Shang Data Lab, UC San Diego Sep. 2025 – Present
Research Assistant | Supervisor: Jingbo Shang La Jolla, CA

- Designed the **BenchInject** framework by linking execution traces with structured function indices, enabling automatic retrieval of target code regions and controlled insertion of fault patterns. Established a unified pipeline covering parsing, trace mapping, candidate extraction, and code rewriting
- Developed an end-to-end verification workflow that integrates LLM-guided modification with automated compilation and test execution. Demonstrated reliable bug activation and failure detection across large Java projects, providing a reproducible platform for evaluating LLM behavior in real software environments

Advanced Wireless Systems Group, CUHK Apr. 2024 – Sep. 2025
Research Assistant | Supervisor: Soung Chang Liew Hong Kong SAR

- Led the **VeriRAG** program, designing a retrieval-augmented generation (RAG) framework that integrates LLMs with Verilog compilation workflows to automatically detect and repair DFT-related errors, significantly improving accuracy in clock-domain crossing and scan-chain validation
- Developed **TopoEdge**, a topology-aware SDN configuration framework leveraging GNN-based contrastive learning and distributed LLM inference across edge devices, enabling efficient configuration repair and automated validation inside FRRouting's Topotest environment
- Led the long-term **full-stack development** and maintenance of the laboratory website, utilizing HTML, CSS, JavaScript, and backend integration to ensure continuous updates, professional presentation of research outcomes, and reliable access to resources

Deloitte
Sep. 2024 – Dec. 2024

Machine Learning Application Intern — Hong Kong SAR

- Developed the IoT-guided **ABA-RAG** framework, integrating multimodal sensor data (BVP, GSR, temperature, acceleration) with structured ABA task repositories. Achieved 73% classification accuracy and 0.90 recall in detecting key behavioral states, enabling more adaptive and context-aware task generation
- Deployed the system as a web-based platform with task retrieval, IoT feedback integration, and performance analytics dashboards. Supported 10 learners in pilot trials, with expert evaluation scores averaging 9.63/10, confirming effectiveness on par with traditional practitioner-led ABA interventions

Wireless Ad-Hoc & Sensor Networks Lab, NCU
Jun. 2024 – Sep. 2024

Research Intern | Supervisor: Min-Te Sun — Taoyuan, TW

- Proposed and implemented a Hybrid Fine-Tuning framework that dynamically integrates LoRA-GA and BOFT updates per layer, achieving near full fine-tuning accuracy while reducing training time by 2.1× and GPU memory usage by 50% on Llama3 models
- Extended unitary recurrent neural network (uRNN) principles into transformer-based LLMs, embedding structured unitary matrices into attention and feedforward layers to enhance gradient stability and convergence

R-Guardian
May. 2023 – Oct. 2023

Machine Learning Engineer Intern — Hong Kong SAR

- Developed AI-powered trademark search engine integrating image feature extraction, template matching, and reverse image search capabilities to enable accurate similarity analysis across global trademark databases
- Engineered scalable database system and cloud computing framework to process massive trademark data from multiple national IP offices, optimizing for real-time search and analysis capabilities

Artificial Intelligence & Computer Vision Lab, NYCU
Jun. 2023 – Aug. 2023

Research Intern | Supervisor: Jun-Wei Hsieh — Taipei, TW

- Contributed to research group developing DeepMAD framework, formulating mathematical programming approach to optimize CNN architecture design through entropy maximization and effectiveness constraints

Embedded AI & IoT Lab, CUHK
May. 2022 – Aug. 2022

Software Development Intern | Supervisor: Guoliang Xing — Hong Kong SAR

- Developed and tested data acquisition software for Smart Mobile Health Systems project (SMHS), implementing multi-threaded sensor data collection and real-time signal processing modules with 97.1% data transmission reliability across 60+ deployment sites

PROJECTS

MWRCNN: Wavelet-Based Dynamic CNN for Image Restoration
Fall 2025

- Proposed a multi-stage image restoration network combining dynamic convolution and wavelet-based frequency decomposition, enabling adaptive feature aggregation for denoising, deblurring, JPEG artifact removal, and super-resolution.
- Validated the model on synthetic and real-world benchmarks, achieving consistent gains over strong baselines, including over 3.5 dB PSNR improvement in heavy denoising (σ=50), $>$ 0.5 dB PSNR on real-world deblurring (GoPro), and up to 50% LPIPS reduction under complex degradations.

Adaptive Test-Time Scaling for LLM Safety Guards
Fall 2025

- Designed a confidence-aware test-time scaling framework for LLM safety guards, enabling dynamic allocation of reasoning compute through parallel sampling and confidence-weighted aggregation.
- Conducted systematic robustness–cost analysis on large-scale jailbreak benchmarks, demonstrating that confidence-triggered scaling recovers near-maximal safety performance while significantly reducing average inference overhead.

Intell-Pro Global Startup - Operations Director
Spring 2025

- Founded and served as CEO of Intell-Pro Global Limited, developing AI-powered trademark search engine serving law firms and IP agencies across US, Europe, and Asia markets.
- Led company strategy and financing initiatives, securing TSSSU funding (HK$675,000) and HK Tech300 Entrepreneurship Award, while establishing partnerships with major IP law firms and trademark agencies for market expansion

SKILLS

Languages: Python, C, Java, JavaScript, SQL, R, P4, Shell Script, HTML

Frameworks: PyTorch, TensorFlow, Hugging Face, OpenCV, FastAPI, MLflow, Git

Cloud & Tools: Azure ML, AWS, CUDA, Docker, OpenAI API, LangChain

ML/DL: Transformer, BERT, LLaMA, RAG, LoRA, PEFT, CNN/RNN, Self/Semi-Supervised Learning