Haomin (Harmin) Qi

(852) 9574-1340 | h5qi@ucsd.edu | harminchee.github.io

EDUCATION

University of California, San Diego
Sep. 2025 – Jul. 2027

M.S. in Electrical and Computer Engineering
La Jolla, CA

Track: Signal and Image Processing

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 Consumer Communications & Networking Conference 2026 (IEEE CCNC 26' Poster)

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 40th Annual AAAI Conference on Artificial Intelligence (AAAI 26')

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 Journal (2025-30591)

Hybrid and Unitary Fine-Tuning of LLMs: Methods and Benchmarking under Resource Constraints
Haomin Qi, Zihan Dai, Chengbo Huang

The Advances in Distributed Computing and Artificial Intelligence Journal (ADCAIJ)

EXPERIENCE

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

- Independently developed and documented step-by-step fine-tuning scripts for Llama3 model, creating accessible tutorials for model training and optimization processes in lab 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

- Implemented and compared fine-tuning methods (LoRA, BOFT, LoRA-GA) within ABA-RAG, showing that LoRA-GA reduced response latency to 498 ms while improving BLEU scores by 10%

- 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 $2.1\times$ 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, particularly for long-range reasoning tasks

- Benchmarked the hybrid method across diverse NLP and reasoning benchmarks (GLUE, GSM8K, MT-Bench, HumanEval) and multiple LLM scales, demonstrating consistent improvements in accuracy, code execution, and mathematical reasoning over baseline fine-tuning methods

R-Guardian

May. 2023 – Oct. 2023

Machine Learning Engineer

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

- Implemented machine learning pipeline combining string matching, text classification, and pattern recognition algorithms to automate trademark conflict detection with distributed cloud computing architecture

- 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

- Enhanced theoretical analysis of network expressiveness metrics, implementing novel evaluation methods that achieved 82.8% ImageNet accuracy while reducing computational costs by 50% compared to conventional architectures

Embedded AI & IoT Lab, CUHK

Hong Kong SAR

Software Development Intern | Supervisor: Guoliang Xing

May. 2022 – Aug. 2022

- 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

- Architected edge computing framework for behavioral monitoring, optimizing system performance through efficient memory management and parallel processing to achieve real-time analysis on resource-constrained IoT devices

PROJECTS

ArtTouch: Visual Art Recognition System

Spring 2025

- Developed multi-stage image processing pipeline integrating CLAHE contrast enhancement, Gabor filtering, and ESRGAN upscaling to optimize artwork digitization, achieving edge detection and texture preservation for 3D model generation

- Implemented adaptive bilateral filtering system with Sobel and Canny edge detection algorithms to enhance tactile feature recognition, enabling precise texture pattern reproduction for visually impaired art appreciation

Intell-Pro Global Startup - Operations Director

Fall 2024

- 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

Marine Heatwaves and Seagrass Resilience: A Machine Learning Analysis

Winter 2023

- Developed ML-based framework to analyze controlled experimental datasets of Halophila beccarii and Halophila ovalis, applying supervised models to identify key physiological and ecological indicators of thermal stress resistance

- Revealed species-specific resilience patterns under marine heatwaves, providing insights into ecosystem conservation strategies and advancing predictive modeling of climate-change impacts on seagrass communities

SKILLS

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

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

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

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