Haochen Li

harryli12321@gmail.com | 919-685-5555 | Durham, NC, USA | www.linkedin.com/in/haochen-harry-li | https://github.com/Harrisous

EDUCATION

Duke University

Master of Engineering, Artificial Intelligence **(GPA 3.9/4.0**) Aug. 2024 – Dec. 2025

Related Courses: Deep Learning Applications, Large Language Model for Generative AI, AI on Edge Devices, Explainable AI, Modeling Process & Algorithms, ML/AI Ops, Sourcing Data for Analytics

The University of Hong Kong

Bachelor of Science, Computer Science & Actuarial Science Sep. 2019 – Jul. 2024

Related Courses: Natural Language Processing, Machine Learning, Database Management Systems (RDBMS), Data Structure and Algorithms

GENERATIVE AI & LLM PROJECTS

AI Agent for Campus Information May. 2025

- Built a conversational AI agent for Duke University using the LangGraph framework to handle complex, multi-step user queries.
- Integrated a suite of tools including web scraping for dynamic content, Tavily for advanced web search, and API calls for accessing structured campus data.
- Result: Achieved high user satisfaction in evaluations, with human judges rating accuracy 4.5/5 and relevance 4.9/5.

RAG System for Financial Analysis Apr. 2025

- Developed a Retrieval-Augmented Generation (RAG) system enabling real-time querying of fresh financial data, overcoming the knowledge cutoff limitations of standard LLMs.
- Integrated Pinecone as a vector database for efficient semantic search, utilized OpenAI APIs for text generation, and built an interactive user interface with Streamlit.

LLM Fine-Tuning for Advanced Text-to-SQL Mar. 2025

- Fine-tuned a DeepSeek model on the BIRD dataset using Low-Rank Adaptation (LoRA) to significantly enhance its SQL generation capabilities for complex queries.
- Leveraged Unsloth to optimize memory efficiency during training, enabling faster iteration and model development.
- Result: Improved execution accuracy by 50% on simple queries while maintaining performance on medium-difficulty queries compared to the base model.

Flashback: AI Memory Augmentation System | Education Track Winner Dec. 2024

- Engineered an AI system on Raspberry Pi using Large Vision Language Models (VLMs) to capture, process, and retrieve user experiences through natural language queries.
- Architected a full-stack solution integrating hardware data capture, a cloud-based processing pipeline, and a multimodal LLM interface for contextual memory recall.

INTERN EXPERIENCE

Blue Insurance, **Machine Learning Engineer Intern**, Hong Kong Dec. 2022 – Jun. 2023

- Engineered a Python web scraping pipeline to collect and clean bond yield data, creating a reliable dataset for time-series forecasting models and reducing data collection time by 80%.
- Developed a custom Python tool using a binomial-tree algorithm to translate and validate complex financial models, enabling 50% faster parallel replication of a core risk algorithm.
- Built data automation scripts in Python to preprocess and structure raw data for machine learning inputs, improving the efficiency of valuation and reporting workflows.

Manulife Financial, **Machine Learning Engineer Intern**, Hong Kong Jul. 2021 – Jan. 2022

- Designed and deployed a Python automation pipeline for tracking COVID-19 severity, processing web data to generate structured datasets for predictive analysis and improving data availability tenfold.
- Analyzed product performance and risk data across multiple business units to identify key features and trends, presenting data-driven insights to inform risk management strategies.

SKILLS

- **Programming & Tools**: Python, SQL, Java, C++, Git, Docker, Streamlit, Raspberry Pi, CI/CD
- **Generative AI & LLM**: RAG, LLM Fine-Tuning (LoRA, QLoRA), AI Agents, LangGraph, NLP, Vector Databases (Pinecone), Transformers, Hugging Face, Unsloth, OpenAI API, Open Source LLMs (Llama, Mistral, Qwen models), Vision Language Models
- **Machine Learning**: PyTorch, Scikit-learn, NumPy, Pandas, Computer Vision (YOLOv8), Time-Series Forecasting
- **Development & Cloud**: AWS, GCP, Amazon SageMaker, Terraform, Kubernetes, Apache Spark, Hadoop, Docker, HPC, Flask