

# QIU Runheng

+86-17702766166 | [meloiseqrh@gmail.com](mailto:meloiseqrh@gmail.com) | [pragues.github.io](https://pragues.github.io) |  [pragues](#)

Longgang, Shenzhen, Guangdong, China, 518172

## OBJECTIVE

---

Seeking a master degree in computer science and engineering to leverage my expertise in coding and analytics. Aiming to further explore the world of programming, especially in llm inference and data mining. Will proactively seek industry job positions of machine learning engineer. Also remain open to research opportunities and, if possible, the pursuit of a PhD position.

## EDUCATION

---

- **The Chinese University of Hong Kong, Shenzhen** Aug. 2022 - Jun. 2026 expected  
*Bachelor of Science in Computer Science and Engineering* Shenzhen, China
  - Half-tuition and Shaw Bowen Scholarship
  - Under Graduate Teaching Fellowship: CSC1004 Computational Laboratory Using Java (2024 Spring)
- **Nanyang Technological University** Jan. 2025 - May. 2025  
*Exchange Program to College of Computing and Data Science* Singapore, Singapore

## INTERNSHIP EXPERIENCE

---

- **Huawei**  May. 2025 - Aug. 2025  
*AI Engineer Intern* Shanghai, China
  - Applied the quantization technique RaBitQ to BGE-Reranker. Use VNNI AVX512 and oneDNN engine to accelerate matrix multiplication block in forward inference.
  - Benchmark the MemOS with and without vector database and graph database. Integrated Membench datasets into the framework. Wrote comprehensive report on the performance of MemOS, mem0 on locomo, Longmemeval, Membench datasets.
  - Research on qwen omni vision lanaguage models, and the multi-vector indexing of maxsim and fast-plaid appeared in colbert and colpali.
- **SenseTime**  May. 2024 - Aug. 2024  
*Model Evaluation Intern* Shenzhen, China
  - Assisted in internal evaluations of the Sensechat and Cantonese multimodal large language model series, automating competitive end-to-end model evaluations via Python scripts.
  - Enhanced SenseBench for model benchmarking by integrating GPT-turbo as a judge for automated scoring, refining prompts for specific ToB business cases to improve scoring accuracy.
  - Evaluated model performance on test sets, provided feedback based on RLHF.
  - Created and maintained subjective test datasets.
- **FXKN, CUHK (Shenzhen)** May. 2023 - Jul. 2023  
*Wechat Miniprogram Developer* Shenzhen, China
  - Developed a WeChat Mini Program using WXML, WXSS, and WXS within the WeChat Developer Tools framework. Implemented basic features that direct users to result pages based on abnormal health indicators in their medical reports, providing potential disease risk assessments.

## PROJECTS

---

- **Agentic Memory for LLM Agents** Jul. 2025  
*Milvus, RAG* 
  - Implmentation of "A-MEM: Agentic Memory for LLM Agents"
- **Parallel Computing Course Project** Sept. 2024 - Oct. 2024  
*MPI, OpenMP, SSE SIMD, CUDA, C++*
  - Accelerated matrix multiplication using MPI, OpenMP, SSE SIMD, and CUDA.
  - Implemented Bilateral Filtering and optimized neural network computation.
- **Java Multi-User Chatroom Course Project** Feb. 2023 - Apr. 2023  
*Java, JDBC, JavaFX, TCP* 
  - Implemented user registration and login functionalities using JDBC and MySQL.
  - Developed a graphical user interface with JavaFX.
  - Enabled multi-user text, image, and voice messaging with TCP sockets and multithreading.

## HONORS AND AWARDS

---

- **Third Prize, Guangdong Province** Oct. 2023  
*2023 China Undergraduate Mathematical Contest in Modeling*
- **Grand Finalist** Sep. 2022  
*CUHK(SZ) Freshman English Debate Tournament*
- **Novice Finalist** Mar. 2023  
*NSDO5 English Debate Tournament 2022*

## SKILLS

---

- **Programming Languages:** Python, SQL, Java, C++
- **Web Technologies:** SpringBoot, React, Redis
- **Data Science & Machine Learning:** PyTorch, Pandas, NumPy
- **DevOps & Version Control:** Git, Docker, Kubernetes
- **Mathematical & Statistical Tools:** Numerical Linear Algebra, Ordinary Differential Equation
- **Other Tools & Technologies:** Linux, LaTeX