Zihan Qi

- **Mobile: (852) 61948792**
- **Email: normalman743@gmail.com**
- **GitHub: github.com/normalman743**
- **LinkedIn: zihan-qi-115b15301**

Education

The Chinese University of Hong Kong

BSc in Statistics, Minor in Computer Science | ***2022 - Present (Expected: July 2026)***

Academic Performance:

- Cumulative GPA: 3.616 / 4.0
- Major GPA (Statistics): 3.787 / 4.0

Honors:

- Dean's List, Faculty of Science, 2023–2024 & 2024–2025
- Head's List, Chung Chi College, 2024–2025

Awards:

- Alex Chung Shun Chan Science Faculty Scholarship (Chung Chi College Departmental Prize, HKD 10,000), 2025/26 – competitive merit-based scholarship.
- Dept. of Statistics Scholarship： 2023/24 & 2024/25
- Dept. of Statistics & Data Science Scholarship： 2025/26

Relevant Coursework:

1. **Computer Science & Machine Learning:**
    - Data Structures (CSCI 2100, A), Computer Organization (CSCI 2510, A), Intro to Operating Systems (CSCI 3150, A), Building Web Applications (CSCI 2720, A).
    - Data Mining & Statistical Learning (STAT 4001, A-), Introduction to Data Science (SEEM 2460, A).
    - Current Enrollment: Bayesian Learning (STAT 4010).
2. **Statistical & Mathematical Foundation:**
    - Statistical Computing (STAT 3006, A-), Intro to Stochastic Processes (STAT 3007, A-), Programming Lang for Stat (STAT 2005, A), Linear Algebra I (MATH 1030, A-).

Research Interests:

- Multimodal alignment and interpretability, with current work focusing on understanding representation behavior in large multimodal models. Also open to broader directions in AI, large language models, and statistical/data-science methodologies, with strong interest in developing further across these areas.

Technical Skills

Programming: Python, R, C/C++, SQL, JavaScript, HTML/CSS
ML & DL Frameworks: TensorFlow, scikit-learn, PyTorch
Data Analysis & Visualization: pandas, NumPy, statsmodels, ggplot2, seaborn, Excel, Tableau
Databases: MySQL, MongoDB
Tools & Platforms: Git, Linux, Jupyter Notebook, VS Code, Terminal
LLM Tools: OpenAI (GPT), Anthropic Claude, Google Gemini; FAISS, Chroma, basic RAG pipelines

Research & Project Experience

Research Preparation – Large Models & Multimodal Alignment **CUHK · Sep 2024 – Present**

- Conducted preliminary alignment experiments on OneLLM using a curated subset of the Flickr30k_images dataset (100 images × 5 prompts + 95 hard negatives per image) to probe image–text matching behavior.
- Modified the OneLLM codebase to hook and export hidden states from all 32 multimodal transformer layers (after the 32 multimodal tokens are formed), and wrote Python analysis scripts for layer-wise cosine similarity and positive–negative margin curves.
- Early results show overall cosine similarity rising with depth, with discrimination peaking in mid layers (≈15–22) before flattening again and occasional deep-layer outliers. The same pipeline is now being extended to 8 modality pairs (currently image–text and image–instruction) to compare alignment patterns across modalities.

Backend Developer – ICU (Intelligent CUHK) Campus Q&A Assistant | CUHK · Jun 2025 – Dec 2025

- Designed and implemented a modular FastAPI backend (≈30 REST endpoints) for a CUHK-specific study assistant, using a layered architecture (routers → services → models/schemas) with token-based auth, invite codes and audit logging.
- Built a course–semester–folder–file hierarchy on top of local file storage; used Celery workers and hashed physical-file storage to extract, chunk and embed PDFs/HTML into per-course and global Chroma vector stores (~1 GB of source files, ~12k RAG chunks).
- Deployed RAG and chat services that have so far served 25+ CUHK students, answered 600+ questions across 40+ courses and supported both course-specific and global study queries over lecture slides, school documents and curated "tree-hole" posts.
- Ran and analyzed a small randomized study using linear mixed-effects models ($F(1,42)=24.372$, $p<0.001$): not using ICU increased log-time by 1.108, implying roughly 67% longer problem-solving time; documented the experiment as evidence of learning impact.

RAG Backend Developer (Intern) | OPTISM ASD Support Platform **Remote · Apr 2025 – Sep 2025**

- Upgraded an autism-support Q&A platform to a Flask-based RAG backend using OpenAI embeddings, FAISS and MySQL, powering a bilingual knowledge base of 695 curated articles, 2,349 segments and 11,889 Chinese/English question templates.
- Implemented a streaming chat API (SSE) with conversation logging and API-key based access control, and designed multi-threaded FAISS index initialization plus monthly hot-update jobs that ingest new Drupal content and regenerate question–paragraph links.
- Built an internal admin console (dashboards, article management, auto question generation, Q&A log analysis, data sync) so non-technical staff can preview data, schedule hourly synchronization and export JSON; most CRUD and scheduling flows were implemented end-to-end by me.
- Co-designed and ran a 2,400-call latency/stability benchmark (Hong Kong + US, English + Chinese) showing the new SSE API cut average response time by ~32–34% and worst-case latency by up to ~58% versus the legacy HTTP endpoint, while maintaining ~99.9% success (2398/2400 calls).

Other Projects

- **Housing Price Forecasting (2024):** Led ARMA, regression and exponential-smoothing modelling on 20-year housing data from Shanghai, Sichuan, Guizhou and national markets; coordinated team workflows and validation, and designed smoothing models that improved forecasting accuracy by up to **42.04%** over baselines (e.g., Sichuan $\alpha=0.92$, $\beta=0.21$; national $\alpha=0.66$, $\beta=0.06$), with exploratory LSTM tests for regional comparison.
- **Auto Essay Grading (CUHK · 2023–2025)**: Ran a zero-shot LLM grading workflow on several hundred submissions across multiple UG/PG courses, with 93/104 students preferring "solution + AI feedback."
- **Multi-Disease Prediction (2024):** Processed a 4 GB clinical dataset and built deep-learning baselines, comparing their held-out performance against classical statistical models across patient subgroups.
- **Full-Stack Cultural Event Platform (2023–2024):** Implemented REST APIs (Node.js/Express, MongoDB) and a React/TypeScript frontend to support authentication, event CRUD and geo-search for 100+ cultural events.