LIHENG HUANG

liheng.huang@outlook.com | Hong Kong | Portfolio

Education

City University of Hong Kong — *MSc in Biostatistics*, GPA: 3.83/4.3 Sep 2025 – Oct 2026

East China Jiao Tong University — *B.Eng. in Data Science*, GPA: 84.3/100 Sep 2020 – Jun 2024

- Outstanding Graduation Thesis Award (Top 1% in School)

Work Experience

Hogenton Limited — *Intern (Business Systems Development & Reporting)* May 2026 – Aug 2026

- Customized an open-source clinic bookkeeping system linking patients, services, transactions, and prepaid session accounts; built daily, monthly, yearly, and service-level reports for reconciliation.
- Designed transaction rules for session issuance, redemption, and manual adjustments; synchronized balances after edits, deletions, and Excel import reversals, handling duplicate imports and same-name customers.
- Implemented pre-import backups, adjustment reason logs, and transaction history queries for data recovery and traceable accounting changes.

Shenzhen Qianhai Xinxin Digital Technology — *Data Analyst Intern* Jan 2024 – Jun 2024

- Used logistic regression, XGBoost, and WOE/IV methods for customer feature analysis and prioritization to support SMS audience selection.
- Wrote PostgreSQL joins across customer, transaction, and credit tables to prepare financial business data and produce statistical reports.

Projects

Text-to-SQL Fine-Tuning and Constrained Decoding Sep 2026 – Oct 2026
[GitHub]

- Fine-tuned Qwen3-4B with QLoRA on one 16 GB V100. Extracted table, column, and join labels from SQL syntax trees for auxiliary supervision and regularized training with a full-vocabulary KL penalty against the frozen base model.
- Enforced SQL grammar and alias scope during decoding, then used static checks to trigger one repair attempt. On 100 BIRD development questions, execution accuracy rose from 36% to 44% and executable queries from 71% to 77% relative to the base model.

Cox Filtration for Multivariate Functional Survival Data May 2026 – Jul 2026
[GitHub]

- Extended multiscale shared representation learning to right-censored outcomes; used Cox partial-likelihood score and information geometry to group ECG leads, extract shared and specific risk components, and learn subsequent directions through residualization.
- Implemented Fusion and Directional grouping based on agreement in lead-level risk contributions and standardized risk directions, respectively; recursively reconstructed predictions as lead-by-time loadings to interpret contributions across leads and time.
- Evaluated 154,744 twelve-lead ECG records across multiple data splits. At full training size, Directional reduced the mean component count from 72 to 36.5, with a mean C-index of 0.7460.

Predicting Survival from Longitudinal Clinical Records with a Transformer Apr 2025 – Aug 2025
[GitHub]

- Integrated multisource MSK-CHORD longitudinal clinical events in SQLite to construct landmark sequences; implemented training-side categorical encoding, numerical scaling, and variable-length sequence batching to control cross-patient and future-information leakage.
- Implemented categorical embeddings, a Transformer encoder, masked pooling, and a custom Cox loss in PyTorch, handling padding, tied event times, and numerical stability with extreme risk scores.

More projects available in my online portfolio.

Skills

Programming & Data: Python, SQL, R, PostgreSQL, Excel

Machine Learning & Tools: PyTorch, scikit-learn, XGBoost, Transformer, Feature Engineering, Git, Linux

Statistics & Experimentation: Regression Analysis, Time Series Analysis, Hypothesis Testing, Sampling and Interval Estimation, Model Evaluation

Languages: Mandarin, Cantonese, English (IELTS 7.0)