Qingyang Liu

+1 7348818747 qliu3@andrew.cmu.edu github.com/qingyangliu0065 linkedin.com/in/qingyangliu0605

Education

Carnegie Mellon University Pittsburgh, PA
Master's in Computational Data Science | GPA: 4.0/4.33 *Aug 2023 – May 2025 (expected)*
Courses: Machine Learning (PhD), Large Language Models, Advanced Cloud Computing, Multimodal ML (current)

University of Michigan - Ann Arbor Ann Arbor, MI
B.S. in Data Science (honor) and Economics | GPA: 3.99/4.0 *Sep 2020 – Apr 2023*
Courses: Deep Learning, NLP, Data Structures, Web System, DBMS, Regression Analysis, Theoretical Statistics

Professional Experience

New York Life Insurance New York, NY
Data Scientist Intern *Jun 2024 – Aug 2024*

- Initiated a **deep-Q-learning**-based underwriting framework to sequentially acquire costly health data and customize mortality risk prediction at individual case level, reduced cost by **47%** with only **0.04** drop in f1 score
- Performed in-depth **EDA** and **feature engineering** on tabular/text data with NLTK and scikit-learn to differentiate feature importance across demographic segments; enhanced **survival analysis** in R for data labeling
- Optimized **ETL** pipeline for querying **1M+** underwriting data on Redshift with **30%** improved efficiency

NeuLab | CMU Pittsburgh, PA
Research Engineer *Feb 2024 – Current*

- Build a unified and scalable Python framework for **LLM** bias and reliability analysis, with composable operations e.g. bias quantification, hallucination analysis, prompting, retrieval-augmented generation (**RAG**)
- Developed example testbeds for toxicity analysis, adapting and integrating 4 detoxification approaches into library

Accenture Shanghai, CN
Strategic Data Science Intern *Jun 2021 – Sep 2021*

- Provided Proof of Concept (**PoC**) of enhanced ML system design for an energy company, projected **20%** boost in energy distribution efficiency via client segmentation and usage forecasting using XGBoost and LSTM
- Developed a modular Python codebase for interactive **Tableau** dashboards integrated with **Streamlit** to visualize impacts of digitalization on key business metrics, simplifying decision-making for clients and cross-functional teams

Selected Projects

Kaggle: Predictive Modeling of Energy Behavior *Jan 2024 – May 2024*

- Developed a **time-series** LGBM+DeepTables ensemble model to forecast electricity consumption and production
- Conducted EDA and feature engineering for lag features, ranked top **6%** among 2700+ teams with **bronze** medal

LLMs for Long Dialogue Summarization *Sep 2023 – Dec 2023*

- Enhanced parameter-efficient finetuning on Llama-2 with **QLoRA** and shifted attention window to extend input context length from 4k to **16k**, integrated DeepSpeed ZeRO for efficient distributed training
- Outperformed baselines including zero-shot prompt engineering on GPTs/Llama-2 and task-oriented models like DialogLM and Pegasus, with on average 25% increase in BLEU, 5% in Meteor, and **10%** in BertScore

Group-Robust Text Toxicity Detection *Feb 2023 – Apr 2023*

- Developed a BERT+BiLSTM classifier for toxicity detection on **450k** real online comments, with a two-stage training regime to mitigate spurious correlation with 8 sensitive demographic groups
- Innovated a data balancing technique of up-sampling the error set and down-sampling the non-error set as the input to the second training stage, with a **25**% reduced training time and **4**% increase in worst group accuracy

Algorithmic Fairness for Health Assessment *Jan 2022 – Jul 2022*

- Enhanced machine learning fairness by mitigating racial bias in a widely-used health risk prediction algorithm
- Designed an iterative importance re-weighting algorithm based on gradient boosting to reduce fairness violation
- Achieved a **6-8%** boost in AUROC score, and **20%** reduction in demographic parity violations on **100k**+ records

Technical Skills

Languages: Python, C/C++ (OOP), R, Java, JavaScript, SQL (Oracle, MySQL, PostgreSQL)
Frameworks: PyTorch, TensorFlow, scikit-learn, Transformers, Pandas, Matplotlib, Seaborn, NLTK, spaCy, PySpark
Tools: AWS, Azure, Agile, Git, Shell, Docker, Kubernetes, CI/CD, NoSQL, Tableau, HTML/CSS, Jupyter Notebook