Haonan(Eric) Gao

Email: eric.gao@yale.edu
Phone: (+1) 203-675-5865
Github | Google Scholar | LinkedIn

EDUCATION

Yale University **2024.09 - 2026.05**
Master of Science in Biostatistics (Data Science track)
- Current cumulative GPA: 4.0/4.0

University of Toronto **2019.09 - 2024.05**
Honor Bachelor of Science (with high distinction)
- Cumulative GPA: 3.9/4.0
- Major in **Computer Science** and **Statistics**; Minor in **Mathematics**
- Relevant Courses: Neural Networks and Deep Learning, Software Tools and Systems Programming, Computation Theory, Data Structure and Analysis, Computer Organization, Operating Systems, Database, Time Series Analysis

HONORS / AWARDS

- **Dean's List Scholar Awards** **2019 - 2024**
  University College, University of Toronto
- **Dr. James A. & Connie P. Dickson Scholarship In Science & Mathematics** **2020.10**
  University College, University of Toronto, Award amount: 500 CAD
- **Covid-19 Emergency Student Bursary Scholar** **2020.05**
  The University of Toronto Scholars Program, University of Toronto, Award amount: 2,000 CAD
- **J. S. Mclean Scholarship** **2019.10**
  University College, University of Toronto, Award amount: 5,000 CAD
- **University Of Toronto Scholar** **2019.09**
  The University of Toronto Scholars Program, University of Toronto, Award amount: 7,500 CAD

RESEARCH / PROJECT EXPERIENCES

RAG in Finance **2024.09 - Present**
Supervisor: Prof. Song Ma, Yale School of Management, Yale University
- Developed a Retrieval-Augmented Generation (RAG) GenAI system to predict upcoming projects of publicly traded companies, estimating both implementation costs and market values for economic and financial analysis.
- Designed and implemented a Python pipeline to parse, sanitize, and embed over 2TB of diverse data—including SEC filings (10-K, 8-K), patents, Wall Street Journal articles, and press releases etc.
- Integrated OpenAI embeddings with LangChain and Pinecone, utilizing isolated memory instances per company (gvkey) to prevent context accumulation and reduce token usage by 30%. Automated the extraction and computation of key quantitative metrics (top three Q-values).
- Fine-tuned the RAG model to enhance prediction accuracy, achieving an 85% success rate in forecasting strategic projects and enabling scalable processing of large datasets.

User Aware MultiLingual Text Simplification Project **2022.09 - 2024.03**
Supervisor: Prof. En-Shiun Annie Lee, Department of Computer Science, University of Toronto
- Led an annotation team in gathering, choosing, and labeling data, and supplied automated evaluation tools in Python, created a multilingual dataset factoring in users' education level for text simplification across 11 languages, generating corresponding styles for different education backgrounds.

- Designed and applied NLP sentence segmentation and cleaning algorithm upon over 10 million chosen sentences and selected 2500 target complex sentences.
- Utilized and fine-tuned the zero-shot text simplification capability of the pre-trained seq-seq model mT5, BART and Pegasus.

Research topic on Penalized Regression without Cross-validation **2023.05 - 2023.09**

Supervisor: Prof. Jun Young Park, Department of Statistics, University of Toronto

- Utilize reduced summary data sourced from public databases to delve into the neurobiological mechanisms of behavior using univariate penalized regression models such as Ridge, LASSO, or Elastic-Net. Actively engaged in determining the threshold for permissible data limitation/incompleteness.
- Leveraged both direct methods and the Coordinate Descent Algorithm for Beta estimation. Introduced the Covariance Regression Model to facilitate the selection of potential models, effectively determining the optimal tuning parameters through careful analysis.

ParaMath Lab (Compilers and LLVM Project) **2023.01 - 2023.05**

Supervisor: Prof. Maryam Mehri Dehnavi, Department of Computer Science, University of Toronto

- Utilized mainstream compilers such as GCC, LLVM, and Clang to generate and analyze their optimization reports; By changing the optimization level, identified non-optimizable sparse computation code patterns and pinpoint corresponding Domain-Specific Languages(DSLs) for those patterns.
- Concentrated on evaluating popular benchmarks such as NAS, HPCG, Parboil, and Intel MKL by C/C++. Executed tests on both CPU and GPU cores (CUDA) to identify and diagnose bottleneck codes limiting compiler optimization.

Lee Lab (A Multilingual Dataset and Benchmark for Math Word Problems) **2022.11 - 2023.01**

Supervisor: Prof. En-Shiun Annie Lee, Department of Computer Science, University of Toronto

- Created an extensive multilingual parallel corpus of Math Word Problems (MWPs) in nine languages including English, Chinese, Hindi, Urdu, etc. This is the largest multilingual MWP dataset to date.
- Utilized the dataset and show the viability using pre-trained multilingual seq-seq languages models prMSLMs for auto-regression MWP generation in both monolingual and multilingual setups.
- Presented the multilingual constraint-based MWP generator implemented on mBART50 as the most promising for low-resource languages.
- The paper has been accepted by *IEEE/ACM Transactions on Audio, Speech, and Language processing (TASLP)*.

Diagnostic Question Analysis Model **2021.09 - 2022.03**

Supervisor: Prof. Shlomo Ta'asan, Department of Computer Science, Carnegie Mellon University

- Implemented a modified Item Response Theory (IRT) model to analyze extensive data derived from student interactions with diagnostic questions, thereby facilitating a precise assessment of individual learning statuses and automating curriculum recommendations.
- Compared to the baseline model, the fine-tuned IRT model shows a better performance after 40 iterations, and with approximately 5% increase in accuracy after 100 iterations.

WORK EXPERIENCE

Cloud Data Engineer; Full-time **2022.04 - 2023.05**

Huawei Technologies Canada, Toronto

- Participated in developing a distributed, low-latency, reliable data engine and developed a connector for the time-series Database InfluxDB based on the query engine, to analysis and migrate across data centers and hybrid clouds with over 40 different Databases.
- The engine is now deployed and used in the top 5 banks in China with over 3 million users, the InfluxDB connector provided efficiently usage for operations monitoring, LoT sensor data and real-time analytics, such as stocks.
- Carried out a solution to push down SQL query's predicates under cloud cluster environment, the overall improvements that ultimately increased the processing speed of TPC-H 1000GB benchmark queries by an average of 20%.

Software Developer Summer Intern; Full-time **2021.04 - 2021.08**

Hebei Shengteng information technology co. Ltd., Hebei, China

- Participated in developing an Integrated Agricultural Management System for Hebei provincial government, where the system allows for comprehensive information display, including information on law enforcement agencies, agricultural sales maps, planting, farming, etc.
- Supported data collection for functionality tests by using Excel VBA, LaTeX, and other tools to write analysis reports for future improvement of the management system.
- Visualized raw data using R and other tools in preparation for the follow-up audition.

PUBLICATIONS

[1] David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, **Haonan Gao**, Annie En-Shiun Lee. "FTC-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects" arXiv:2309.07445. Presented at *The 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024)*

[2] Gamage O. Ishendra, Surangika Dayani Ranathunga, Annie Lee, Mehreen Alam, **Haonan Gao**, et al. "A Multilingual Dataset (MultiMWP) and Benchmark for Math Word Problem Generation." Accepted by *IEEE/ACM Transactions on Audio, Speech and Language Processing, VOL. 31, 2023*

[3] Gao, **Haonan**. "A Diagnostic Question Analysis Model based on a Modified Item Response Theory." *2022 6th International Seminar on Education, Management and Social Sciences (ISEMSS 2022)*. Atlantis Press, 2022

Skills

- **Language Proficiency:** English(Fluent), Mandarin(Native), Python (Pandas, TensorFlow, PyTorch, Numpy), Java, C/C++, R, SQL, Linux, InfluxDB, MongoDB, LaTex
- **Development Tools/Tech Stack:** Docker, Git, Trino, Huggingface, AWS(EC2/SageMaker), Azure, GCP, Kubernate, Langchain, Pinecone