LI YAO

322 - 26 Everett St, Cambridge, MA, 02138
⋄ +1 (617) 763-4249 ⋄ li_yao@fas.harvard.edu

EDUCATION

M.S., Data Science Sept 2022 - May 2024
Harvard University, Cambridge, USA GPA: 4.0/4.0

B.S., Statistics and Mathematics Sept 2018 - Aug 2021
Simon Fraser University, Burnaby, Canada GPA: 3.86/4.0

SKILLS

Computing: Python, R, SQL, MATLAB, SAS, C/C++, GCP, AWS, Docker, Tableau
Tools: Pandas, Numpy, Matplotlib, sklearn, PyTorch, TensorFlow, Keras, JAX, Genism, NLTK, WandB, Git

PROJECT EXPERIENCES

TeamBirth: Teamwork for patient-centered childbirth, Ariadne Labs-Harvard Spring 2024

- Developed a fully functioning end-to-end chatbot using **LangChain** and **MLOps** to analyze postpartum surveys from American Women's Hospitals, generating actionable insights to improve care processes across 15 states
- Deployed the Llama-7b-chat LLM on **GCP**, ensuring robust privacy safeguards for hospital data

MLOps: Sign Language Translation, Harvard Fall 2023

- Finetuned a transformer-based sign language translation model by **Vertex AI** and **WandB** and incorporated workflow orchestration into our project with the use of **Kubernetes**
- Built **RESTful APIs** to serve our models and designed user interfaces for seamless user interactions

Language Models Represent Space and Time, MIT Fall 2023

- Extracted the activation map of each layer of the LMs such as BERT and GPT-Neo using **PyTorch** and analyzed if there were any correlations between activation and the geographic coordinates associated with the respective tokens
- Examined whether LM's rich embeddings contain world knowledge by employing a linear probe

Predictive Analytics for Breast Cancer Risk, Harvard Fall 2022

- Performed data preprocessing, feature engineering, and feature selection on the breast cancer dataset using **Pandas** to enhance predictive modeling for assessing cancer risk
- Built and evaluated machine learning models such as Logistic Regression, Random Forest, and Gradient Boosting, using **sklearn**, and employed metrics such as F1 score, recall, and precision to ensure robust model performance

RESEARCH AND WORK EXPERIENCES

Keyword-Assisted Embedded Topic Modeling, Harvard July 2023 - May 2024

- Proposed a semi-supervised approach to improve the VAE-based embedded topic model, and compared our methodology with SOTA methods such as vONTSS, ETM, BERTopic on different corpus like 20Newsgroups and bbc-news
- Wrote an open-source **Python package** and prepared paper for publication

Data Scientist, Statistics Canada May 2021 - Aug 2022

- Conducted data cleaning and analysis on the Canadian Vital Statistics death database, and utilized generalized linear models (quasi-Poisson) and time series models in **R** to produce the excess mortality estimates caused by COVID-19 on a monthly basis
- Used Canada Census database to produce as of July 1st, 2021 centenarian population estimates using **SQL**, **Pandas**, **Numpy**, **Matplotlib**, etc.

Goodness-of-Fit Based on Empirical Distribution Function, SFU May 2021 - Dec 2021

- Wrote programs for the calculation of goodness-of-fit test statistics such as Cramér–von Mises statistic (W^2), Anderson–Darling statistic (A^2), and Watson statistic (U^2) and their P-values
- Built an open-source **R package** and submitted it to CRAN (`https://CRAN.R-project.org/package=EDFtest`)

TEACHING EXPERIENCES

Teaching Fellow, Harvard University
AC 215, Advanced Practical Data Science, MLOps Fall 2024
CS 109b, Advanced Topics in Data Science Spring 2024
AM 207, Stochastic Methods for Data Analysis, Inference and Optimization Fall 2023