Zhengyi Shan

(1) 672-380-5580 | zhengyishan.ds@gmail.com

Personal Website | GitHub | LinkedIn

PROFESSIONAL EXPERIENCE

Data Science Capstone Project | Boeing Digital Solutions Jeppesen Ongoing

- Designed an automated system to parse and categorize daily NFDD PDF updates, reducing manual processing time by 80% and error rates by 15%.
- Streamlined workflows using data chunking and classification, cutting 90% irrelevant data and boosting database accuracy/retrieval efficiency.

Research Assistant | University of British Columbia Ongoing

- Processed 500-hour multilingual audio datasets with Pandas and automated workflows, cutting preprocessing time by 50% for linguistic analysis in Praat.

AI Speech Testing Intern | NetEase Youdao 01/2024-04/2024

- Developed AI-based IELTS speaking score evaluation using ASR, optimizing model accuracy with feature analysis.
- Created high-quality testing datasets through audio segmentation and manual scoring, conducting error analysis to improve ASR performance and offering actionable recommendations for model optimization.
- Assisted with data annotation, including labeling and proofreading English-language datasets, and contributed to technical documentation using LaTeX to ensure clear and professional project reporting.

EDUCATION

Master of Data Science (Computational Linguistics) | University of British Columbia 08/2024-06/2025

- GPA: A-

Exchange in Quantitative Social Analysis | Hong Kong University of Science and Technology 01/2023-06/2023

BA in English Language and Literature (Global Studies Honors Program) | Shantou University 09/2020-07/2024

- GPA: 4.13 (91.3%) | Ranked 2/72 | Graduated with Distinction

SKILLS

Programming Languages: Python, R, SQL (PostgreSQL, MongoDB)

Data Processing & Analysis: Data Wrangling (Pandas, NumPy), Feature Engineering, Missing Data Imputation, Statistical Analysis (SciPy), Visualization (Matplotlib, ggplot2, Altair)

Machine & Deep Learning: Supervised/Unsupervised Learning (scikit-learn), Model Evaluation, Neural Networks (PyTorch, TensorFlow – CNNs, RNNs, Transformers), LLM fine-tuning, LangChain

PROJECTS

LLM-Based Multilingual NLP Workflow for Toxicity Detection and Detoxification

- Designed an agentic workflow using LangChain and Granite-3.0-2B to perform toxicity detection (toxic/non-toxic) and toxic-to-non-toxic style transfer on multilingual text.
- Used Helsinki-NLP translation model, achieving 85% detection accuracy (5,000+ samples) and 80% toxicity reduction (manual score: 8/10).

House Prices: Advanced Regression Techniques – Kaggle Competition

- Designed a machine learning pipeline to predict house prices, including data preprocessing (handling missing values, encoding, scaling), feature engineering (e.g., total square footage, house age, total bathrooms), and feature selection using Random Forest (59 key features).
- Trained and tuned a Gradient Boosting Regressor with RandomizedSearchCV after model selection, achieving an RMSE of 0.1322 (log scale) on the dev set.