Linan Wu

Mobile: +86 181 3805 5392 | Email: linan_huilaioffer@163.com

EDUCATION

Beijing Normal – Hong Kong Baptist University (BNBU)

*Bachelor of Science (Honors) in **Data Science*** *09/2022-06/2026*

*Minor in **Economics***

- Weighted Average Score: 3.7 / 4.0

Awards & Achievements:

- Guangdong Provincial Third Prize in 2024 China Undergraduate Mathematical Contest in Modeling (CUMCM)
- BNBU Second-tier Scholarship (2022-23 & 2023-24)

COMPUTER TECHNIQUES

- Programming languages: Python, R, C, Java
- Data processing, analysis & visualization: Pandas, NumPy, SQL, Matplotlib, Seaborn, Plotly, WordCloud
- Machine Learning and Deep Learning: LSTM, BERT, RoBERTa, Transformer
- Big Data and Distributed Computing: Hadoop, Spark, Docker
- Web Development and Database: Django, MySQL

PROJECT EXPERIENCES

Construction of Large Language Model Inference System *02-06/2025*

Aimed to build a high-accuracy large language model inference system to answer complex natural science multiple-choice questions generated by GPT-3.5

- Built a multi-model architecture based on open-source models such as Mistral-7B, Yi-34B, and LLaMA 2, using Zero-shot, Few-shot, and SFT fine-tuning strategies, combined with H2O LLM Studio, to optimize the model, achieving an accuracy improvement of over 10% in specific fields such as physics/biology
- Used LoRA fine-tuning and cached past_key_values, increasing inference speed by 60%
- Dynamically enhanced context by integrating RAG and LangChain, building a corpus containing 60million paragraphs based on multi-source Wikipedia data, developing a FAISS vector retrieval system, and entering pre-designed multi-round Prompt templates
- Implemented confidence integration based on results of multiple models, combined with methods such as TF-IDF re-ranking and Embedding similarity, to increase the Top3 hit rate to 93%; and the integrated model was 15% more accurate than the single model

Identification of Offensive Language in Chinese Social Media Environment *10-12/2024*

Aimed to build an automated language detection and recognition system, based on deep learning technology, to help platforms manage implicit and explicit offensive remarks in the Chinese social media environment

- Conducted the collection, cleaning, and annotation of Chinese social media data; built and optimized multiple dedicated datasets including COLD, ToxiCN, and ToxiCloakCN

- Built and trained LSTM neural networks and pre-trained language models (BERT and RoBERTa) to achieve accurate identification and classification of aggressive language
- Evaluated and optimized the model through performance evaluation indicators (Accuracy, ROC-AUC, F1-score); eventually decided on the RoBERTa model, which achieved an accuracy of 78.54% and performed outstandingly in complex language recognition

Predictive Analysis of Diamond Prices *05-06/2024*

Used Python and R languages to model and analyze 5000 pieces of diamond transaction data, attempting to build a high-precision price prediction model.

- Implemented data cleaning, feature engineering, and multivariate transformation (Box-Cox) to optimize model performance
- Explored data features through data visualization tools, identified multicollinearity between variables, and improved model prediction accuracy
- Adopted stepwise regression and cross-validation methods; the final model achieved an R^2 of 0.9853 and a cross-validation RMSE of 0.1308, showing superior and stable performance
- The model successfully passed regression assumption tests such as residual independence, normal distribution, and homoscedasticity to ensure the robustness of predictions

Processing and Analyzing of Earthquake Data from 1900 to 2023 *03-05/2023*

Used big data and distributed computing technologies to carry out in-depth mining and analysis of global earthquake data from 1900 to 2023 to cast light on the spatiotemporal patterns of seismic activity and potential risk areas

- Utilized big data frameworks such as Hadoop and Spark for large-scale data storage and parallel computing, completing efficient processing and analysis of global long-time series earthquake data
- Performed data cleaning and time format standardization, used reverse geocoding technology to obtain accurate geographical information, and conducted in-depth analysis of the spatiotemporal distribution characteristics of earthquakes
- Adopted visualization tools such as Plotly and WordCloud to clearly display the spatial hotspots and key event characteristics of earthquake data, providing strong data support for subsequent risk prediction and disaster prevention and mitigation

ADDITIONAL EXPERIENCES

BNBU Dream Weavers Club *10/2022-Now*

- Have served in the club, one of the most renowned student organizations at BNBU, from initially an ordinary member to currently a vice president
- Help plan, organize, and coordinate various student activities including fund-raising charity campaigns, guest speeches, volunteer programs, and so forth