Hao-xuan (Horace) Wang

+86 18622468042

billweasley20092@gmail.com

Github: https://github.com/billweasley

Linkedin: https://www.linkedin.com/in/horace-haoxuan-wang

Personal Website: http://shellcottage.me

Work Experience

2025.03 - Now • **Senior Data Scientist, Integrity and customer experience** @ Grab

- Initial project is about producing an enhanced on-device keyword spotting model and combining it with service-side multimodality LLM to detect safety issue during taxi hailing.

- 2022.03 - 2025.02 **Machine Learning Engineer, ASR and Language Tech** @ Zoom

  - Designed and implemented a **multilingual ASR service** supporting **36 languages** and **5+ workflows**. The architecture is based on **AsyncMQ**/**Kafka**, deployed via **Kubernetes (K8s)** and **Istio**, supporting dynamic CPU-based scaling to serve global Zoom customers.
  - Conducted **ASR model performance optimization**, fine-tuning inference parameters, memory usage, and CPU allocation for efficient large-scale deployment.
  - Developed **LLM-based ASR error correction workflows**, achieving a **Rare Word WER reduction from 37.8% to 17.5%** on medical datasets using closed-source LLMs and bias word lists.
  - Independently implemented and optimized **Whisper inference**, leveraging **in-house VAD (Voice Activity Detection)** and **WhisperX**, achieving higher throughput and lower WER than OpenAI's reference implementation.
  - Built **Multi-Head Attention (MHA)-based time alignment** on **LAS**/**Seq2Seq models**, providing accurate **word-level timestamps** for multilingual transcription. * [Filed a US patent for this] *
  - Independently experimented with LLM for ASR in multimodality setting, aiming to improve the consistency of ASR decoding results. The orthographic (fully-format) WER (word error rate), and the rare word WER achieved a better result compared with the production model.
  - Created **LLM-based data augmentation pipelines** using **Mistral MoE 8x7B**, generating diverse textual datasets for model training.

- 2019.08 - 2021.12 **Data Scientist and Software Engineer** @ Barclays

  - Initially a Java Software Engineer supporting Barclaycard, supporting backend cache for bouns.
  - Company address matching and entity matching without internal GPU and labeled data available. Solve using an active learning method. Start from constructing some small datasets only with external data and training an XGBoost tree, then label samples in the boundary and fine-turn BERT models in an iterative way. Finish the inference on 6 million internal pair-wised samples with this model on a CPU cluster, using a DJL based pipeline built from scratch on my own. It achieved a very satisfying result of **94% F1 score on a noisy testing dataset from 89% where we started**. The model does inference offline on our Spark cluster in a distributed way. For 6 million pair-wised samples, the running time is under 1 hour (on a cluster with 80 CPUs).
  - Built **time-series forecasting models** using the **Informer architecture**, predicting aggregated user transaction volume and value. Constructed **counterfactual analyses** to assess financial losses during system downtimes.

Education

2018 - 2019	MSc Web Science and Big Data Analytics @ **University College London,** Distinction
2016 - 2018	BSc Internet Computing @ **University of Liverpool *,** First class
2014 - 2016	BSc Information and Computing Science @ **Xi'an Jiaotong-Liverpool University** *

***Note:** 2+2 pathway routine (first 2 years in Suzhou, China and final 2 years in Liverpool, UK), dual degree.

Personal Project

- 2024.06 - **Fine-tuning and evaluation of medical record data on Large Language Models (LLMs)**

  Fine-tuned various **LLMs (Llama3-instruct, Llama3 Chinese-chat, Qwen2)** on **Chinese medical records datasets**, focusing on tasks such as **department classification, record summarization, and discharge certification**.

  - **Consultation Summary**/**Discharge Summary:** BLEU improved from **0%-30%** to **49%-55%**, ROUGE-L from **20%-30%** to **60%-64%**.
  - **Department Classification:** Accuracy improved from **0%-36%** to **69%-71%**.

  Future plans include open-sourcing the dataset.

Technical Article

- **"Accelerating Deep Learning on the JVM with Apache Spark and NVIDIA GPUs"**

  Author: Haoxuan Wang, Qin Lan [AWS], Carol McDonald [Nvidia]; Link: https://www.infoq.com/articles/deep-learning-apache-spark-nvidia-gpu/?itm_source=articles_about_ai-ml-data-eng&itm_medium=link&itm_campaign=ai-ml-data-eng

Early Stage Project

- 2019.06 - 2019.09 **Project Internship (Master Degree Thesis)** @ Astroscreen

  Built data crawlers for **Gab.com** and pre-processed datasets using **Regular Expressions**. Fine-tuned **BERT and XLNet** for classification tasks, achieving an **86% F1 score** on hashtag-balanced datasets. Visualized results using **t-SNE**, performed **cross-validation**, and evaluated metrics including **Accuracy, F1 score, and Matthews Correlation Coefficient**.

- 2025.02 - 2025.02 **Work trail** @ Finalround.ai

  In a one-week project focused on intent detection from intermediate ASR results, I independently implemented a complete detection pipeline and achieved an F1 score of 87% on a validation meeting. Notably, half of the intents were detected ahead of the ASR final utterance. This work enables me to receiving a job offer from them. The complete pipeline included: 1) Rule-based handling of greeting utterances.2) Evaluating sentence completeness using segment-any-text, syntactic parsing, and perplexity scoring. 3)Detecting confirmation-type questions (e.g., "Can you hear me?") using Sentence-BERT embeddings. 4) Classifying final question intents with a small language model. Also, I developed prompts for extracting resume information, which improved the personalization and quality of LLM-generated responses.