Yuxuan Zhang (Wayne)

yxzwayne@gmail.com | https://github.com/yxzwayne

Experience

Machine Learning Intern Sep. 2023 – Current

Salk Institute for Biological Studies *San Diego, CA*

- Training autoregressive **Transformers** and **variational autoencoders** using **PyTorch** and **Tensorflow** on over 10 million viral protein sequences to recover hierarchical information and interpret drug-resistant epistasis.
- Developed a **PCA**-based tool for filtering cryogenic electron microscopy image data that uncovered new insights in protein structures by surpassing traditional K-means and GMM clustering limitations.

Machine Learning Intern Mar. 2024 – Jun. 2024

trufflepig.ai *Remote*

- Developed a complete document ingestion and retrieval pipeline to PDFs across various domains with **FAISS** vector search and BGE cross-encoder reranking.
- Evaluated hit rates and mean reciprocal ranks for BGE-m3 and MiniLM cross-encoders, leveraging GPT-4 and Hermes-2-Pro-Llama3 as synthetic data generator and judges, generating over 10,000 synthetic queries.
- Deployed an open-source cross-encoder model to Azure, boosting retrieval accuracy by over 10%.

Software Engineer Dec. 2021 – Jul. 2022

Awesung Inc. *Cranbury, NJ*

- Developed an internal logistics management application using **React.js**, **Django**, and **MySQL**, centralizing operation data previously scattered across multiple platforms.
- Optimized backend queries, reducing data serialization time by **20%** and **RestAPI** response times by **50%**.
- Optimized logistics operations using previously unincorporated third-party data, devised operational plans by analyzing data with **Pandas** and **Matplotlib**, reducing Item-Not-Received complaints by **15%** in two months.

Projects

Medilora — Large language models, Fine-tuning Oct. 2023 - Dec. 2023

- Fine-tuned OpenHermes-2.5-Mistral-7B with Q-LoRA using **Axolotl** on 300 million medical text tokens.
- Improved **PubMedQA** and **MedQA** evaluation scores by over **20%**, matching the state-of-the-art 70B **Meditron** on **MMLU-Medical** with **0.05%** of the data size.

Open-ended Document Classification — Large language models, Classification May 2023 - Jun. 2023

- Implemented the recursive summarization method in 2109.10862 by generating summaries for ArXiv papers using ChatGPT API and generating open-ended class labels with **RoBERTa**.

CompassX Platform — Software Engineering Jan. 2023 - Jun. 2023

- Founding developer of CompassX at UCSD, led full-stack development using **Flask**, **JavaScript**, and **PostgreSQL**, scaled data schema and backend to support multi-college expansion.

Education

University of California San Diego La Jolla, CA

M.S. Data Science. GPA: 3.5 *Sep. 2022 - Present*

- **Courses and Involvements** Attention/Diffusion Reading Group, RL/LLM seminar, Search and Optimization, Learning Algorithms, Maths of Deep Learning, Scalable Data Systems, Text Mining, Statistical Modeling
- **Teaching Assistant** DSC 30: Data Structures & Algorithms

University of Washington Seattle Seattle, WA

B.S. Informatics. GPA: 3.5 *Sep. 2017 - Jun. 2021*

Technical Skills

Programming: Python, Java, SQL, JavaScript, R
Computing Libraries: Keras, PyTorch, XGBoost, Pandas, NumPy, Matplotlib
Frameworks: React, Django, Node.js, Flask, JUnit
Tools: Git, Docker, Linux, AWS, Vim, Jupyter