Zhihao (Johnson) Du

zhihaodu@berkeley.edu | GitHub | Personal Website

Experiences

Apple Inc. Seattle, WA
Machine Learning Engineer 09/2024 - present
- AIML - SII (Siri and Information Intelligence)

Toby AI Company San Francisco, CA
Machine Learning Engineer Intern 05/2024 - 08/2024
- Developed real-time speech-to-speech translation system with latency <2.0s and ASR-BLEU >40 supporting 10+ languages
- Scaled development to handle 400 concurrent users with batched parallel inference and autoscaling instances using Kubernetes
- Leveraged Parameter Efficient Finetuning on Whisper using personalized glossaries to achieve 18% decrease on WER
- Replaced the end-to-end model with highly flexible modularized text-cascading system improving latency by 47.8%
- Hand-crafted full evaluation pipeline and performed extensive gridsearch on hyperparameters for translation policies
- Launched on Product Hunt authored as 1 of the 4 core makers of the product, achieveing 3rd product of the day

Sky9 Capital San Francisco, CA
Software Engineer Intern 06/2023 - 06/2024
- Designed LLM-powered Chatbots hosting creative text-based party games catering to 500K users on Discord using Python
- Engineered product data feedback pipelines enabling reinforcement learning cycles based on past user interactions
- Optimized dialogue generation algorithms using efficient prompt chaining and RAG, decreasing token count by 40%
- Addressed scalability challenges using Docker containers and automatic cross-server deployment scripts
- Collaborations with incubator: LoRA training for pet art generation; AIGC for product marketing via SEO

University of California, Berkeley Berkeley, CA
Head Teaching Assistant 01/2023 - 05/2024
- Courses taught: INFO 259 Natural Language Processing; EECS 182 Introduction to Deep Learning
- Led 10 member course staff on course logistics, content planning, HW and Exam creation, and grading for 300 students
- Taught sections and held office hours covering topics on CNN, Transformers, Meta-Learning and Generative Models

Software Researcher, Project AEI (Artificial Emotional Intelligence) 01/2022 - 05/2023
- Leveraged Pytorch and MS Azure platform to build CNN-Transformer model for emotion classification from real-time speech
- Installed real-time model evaluation system on Raspberry Pi while overcoming the platform's low-resource constraints
- Improved model performance by 3% with innovative speech augmentation techniques using librosa and torchaudio

DataCVG Co Ltd Shanghai, China
Data Engineer Intern 05/2021 - 08/2021
- Performed extract-transform-load (ETL) on client FosunPharma's two major relational database, resolved conflicts with hand-designed schemas, and consolidated all data into a single database
- Programmed 100+ complex SQL query scripts and engineered data merger pipelines using the DBeaver ETL software

Education

University of California, Berkeley Berkeley, CA
M.Anlytx (Analytics), Artificial Intelligence Concentration - GPA: 3.88/4.00 08/2023 - 05/2024
B.A. Computer Science, Statistics - GPA: 3.82/4.00 08/2019 - 05/2023

Skills

Languages – Python, Java, SQL, C, HTML+CSS

Tools/Frameworks – Python Frameworks: [Numpy/Pandas, Matplotlib, Scikit-learn, Pytorch/Jax, Flask, LangChain], AWS/MS Azure, Docker, PostgreSQL, MongoDB, Git, Unix System

Skills – Machine Learning/Artificial Intelligence, Natural Language Processing, Prompt Engineering, Full Stack Development, Object Oriented Programming, Data Structures,