Daniel Yu

Software Engineer | Machine Learning Engineer

talldanielyu@gmail.com | (650) 278-0824 | danielyu.app | linkedin.com/in/itsyangyu | github.com/kalig0

Experience

Software Development Engineer, eGain – Sunnyvale, CA Jan 2025 – present

- Built end-to-end AI agent pipelines integrating embeddings, OpenSearch, rerankers, and LLM prompts with system/user instruction separation reducing response generation time by 25%
- Led development of reusable chat widget supporting authentication, portal selection, and multi-tenant embedding
- Migrated and standardized search APIs and OpenSearch schemas to support keyword and semantic search, multilingual responses, and consistent response contracts
- Designed and maintained cloud-native AI services using Docker, AWS, OpenSearch, DynamoDB, SQS, and S3

Machine Learning Engineer Intern, CVTE – Guangzhou, China May 2024 – July 2024

- Developed and implemented logic enhancements for a machine learning model utilized in over 5 million classrooms, improving the generated classroom analysis by 20%
- Used Python, Matplotlib, and Pyplot to process classroom data to analyze and identify dips in classroom engagement

Software Engineer, ReadMKT – San Francisco, CA Aug 2022 – Aug 2023

- Led development of a rich text editor using JavaScript, allowing users to intuitively edit content, embed media, and formal text in a WYSIWYG environment
- Implemented infinite scrolling with lazy loading, enhancing the user experience and improving page performance.

Full Stack Developer Intern, 7G BioVentures – San Francisco, CA May 2022 – Aug 2022

- Built a responsive internal dashboard using React, enabling employees to track and manage investment projects
- Created modular front-end components with reusable design patterns for scalable UI architecture

Education

Worcester Polytechnic Institute, MS in Computer Science – Worcester, MA Aug 2023 – May 2025

- Fine-tuned LLMs and built RAG pipeline for educational feedback, improving teacher response quality by 30%

University of California, Santa Barbara, BS in Mathematical Sciences – Santa Barbara, CA Aug 2018 – June 2022

Projects

ASSISTments Jan 2023 – present

- Fine-tuned large language models using QLoRA to enhance the accuracy of feedback for student responses to math questions
- Trained models using TensorFlow on millions of student response data to develop and implement effective improvement strategies for student feedback
- Implemented vector database using FAISS to generate feedback and score for studend responses with retrieval augmented generation, improving large language model generated content accuracy by 30%

Skills

Languages: Python, Typescript, C++, HTML, CSS

Frameworks: PyTorch, TensorFlow, React, Node.js

Infrastructure: AWS