Yu Xiu

(434) 422-2707 | xiuyutiffany@gmail.com | LinkedIn | Github | Portfolio

Experience

Stealth AI Startup Bay Area, CA | **Demo**
AI Software Engineer (Independent Project) *Aug 2025 – Present*

- Architected and deployed full-stack AI tutoring platform with FastAPI backend, React frontend, and RESTful APIs for real-time grading and multi-turn conversational feedback
- Fine-tuned BERT intent classifier (94% macro F1) for 6-class query routing, with stratified sampling and class balancing to address data imbalance
- Designed RAG-enhanced grading system and raised grading accuracy from 49% to 74%, systematically evaluating 5 prompt strategies (zero-shot, few-shot, chain-of-thought, self-consistency, self-consistency + CoT) across GPT-4o-mini and GPT-5.2
- Fine-tuned BERT/RoBERTa as cost-efficient alternatives to GPT, reducing inference cost by 95% and latency by 10x while maintaining classification accuracy
- Built multilingual conversation support with Unicode-based language detection and dynamic prompt engineering for consistent response language matching
- Implemented production features: OpenAI Moderation API for content safety, ChromaDB vector store for retrieval, streaming responses, and confidence-based fallback routing
- Developed user progress tracking with answer history and auto-generated wrong-answer workbooks for personalized practice and knowledge gap remediation

NASA Ames Research Center Mountain View, CA
Systems and Software Engineering Intern *Oct 2023 – Apr 2024*

- Optimized test pipeline by analyzing MySQL telemetry data to identify faulty components and obsolete equipment, reducing average test execution time by 20%
- Developed fault management simulation models in MATLAB/Simulink for EPIC rocket booster system, implementing and validating control logic for 4 propulsion components
- Implemented backend test algorithms from sequence diagram specifications, collaborating with scientists and IT to debug and resolve system design issues
- Processed experimental test data and created visualizations (pie charts, tables) to communicate analysis findings and support engineering decisions

Project

AutoGrader - Code Assessment Platform *Nov 2025* | **Demo**

- Full-stack auto-grader: FastAPI backend and React (Vite) frontend with RESTful APIs; real-time Python code grading via pytest in isolated subprocesses across multiple problem sets
- Integrated OpenAI (GPT-4o-mini) for code analysis and bilingual (EN/ZH) feedback
- Deployed frontend on Vercel and backend on Render with CORS and env-based configuration

Education

San José State University San José, CA
Master of Science in Computer Science; GPA: 3.6 *2022 – Dec 2024*

Stanford University Stanford, CA
Visiting Graduate Student, Computer Science (Summer Session) *2024*

San José State University San José, CA
Bachelor of Science in Computer Science; GPA: 3.7 *Dec 2020*

Technical Skills

Languages: Python, TypeScript, JavaScript, Java, SQL, MATLAB
AI/ML: HuggingFace Transformers, BERT, RoBERTa, Fine-tuning, OpenAI API, LangChain, RAG, Prompt Engineering
Backend: FastAPI, REST APIs
Frontend: React, TypeScript, HTML/CSS
Data: ChromaDB, MySQL, Vector Databases, Data Analysis
Tools: Git, Docker, Simulink, Render