Jiajun Huang

(530) 908-2643 eddyislearning.ai
recruit@jiajunh.me linkedin.com/in/jiajunh/

Education

Northeastern University (Silicon Valley) San Jose, CA
M.S. in Artificial Intelligence *Jan. 2026 – Dec. 2027*

University of California, Davis Davis, CA
B.S. in Computer Science *Sep. 2021 – Aug. 2025*

Technical Skills

Languages: Python, SQL, Golang, C/C++, JavaScript, Bash, HTML/CSS
Frameworks & Libraries: PyTorch, TensorFlow, scikit-learn, XGBoost, LangChain, FastAPI, React, NumPy, Pandas
Infrastructure & Tools: AWS (EC2, S3), Docker, Kafka, Redis, Spark, Airflow, MLflow, PostgreSQL, FAISS, Git

Work Experience

NextTier Sacramento, CA
Machine Learning Engineer Intern *Aug. 2025 – Dec. 2025*

- Built a client risk scoring pipeline that served 3,000+ daily predictions, engineering 40+ features from transaction, credit, and demographic data with **Pandas** and **Spark**, improving accuracy by 12% over the prior rule-based system.
- Trained and tuned **XGBoost** and **LightGBM** models on 2M+ records using **Optuna** for hyperparameter search, achieving an AUC of 0.91 and a 15% lower false positive rate compared to the previous production model.
- Deployed model serving via **FastAPI** and **Docker** with RESTful APIs, achieving <200ms P95 latency and automated versioning through **MLflow** for experiment tracking and reproducibility.
- Built a monitoring dashboard tracking model drift and feature distribution shifts, enabling proactive retraining and maintaining <2% accuracy degradation over a 3-month deployment cycle.

Eth Tech Newark, CA
Software Engineer Intern *May. 2024 – Aug. 2024*

- Migrated legacy monolithic services to a microservices architecture using **Golang** and **gRPC**, reducing API response latency by 40% and supporting 5,000+ RPS under high-concurrency traffic.
- Implemented a write-through caching layer with **Redis** for the user segmentation module, reducing **PostgreSQL** read load by 35% and cutting average query time from 120ms to 45ms.
- Engineered real-time data ingestion pipelines with **Apache Kafka**, processing 50K+ user activity events per minute and ensuring 99.9% data availability for downstream analytics dashboards.
- Increased code coverage from 60% to 85% by building unit and integration test suites with **Testify**, reducing production bug rate by 30% across weekly release cycles.

Project Experience

DocLens Individual Project
AI-Powered Document Q&A System *Nov. 2025 – Jan. 2026*

- Built an end-to-end **Retrieval-Augmented Generation (RAG)** system enabling natural language queries over 500+ page document collections, achieving <3s average response latency using **LangChain**, **OpenAI API**, and **FAISS** for vector similarity search.
- Implemented a chunking and embedding pipeline supporting PDF, Markdown, and HTML ingestion with recursive text splitting and **text-embedding-3-small**, achieving 92%+ retrieval relevance on benchmark queries.
- Developed a **FastAPI** backend with streaming responses and a **React** frontend, featuring conversation history, source citation display, and document upload for a production-ready user experience.
- Optimized retrieval accuracy through hybrid search (dense vector + **BM25**) and **re-ranking** with **cross-encoder** models, improving answer accuracy by 18% on a custom evaluation set of 200+ Q&A pairs.