Kayvan Zahiri

kzahiri@dons.usfca.edu https://www.linkedin.com/in/kayvan-zahiri/ https://github.com/Kayvan-Zahiri/

AWS and GCP-certified AI Engineer and Software Developer with full-stack, GenAI, and LLM fine-tuning experience.

Experience

AI Engineer **Remote**

Asurion *January 2026 - Present*

- Built a voice turn detection system, improving auto-labeling accuracy from 50% to 88%+ (90%+ on a 400 sample validation set).
- Designed and scaled an automated labeling pipeline to auto-label 400k+ audio samples, enabling large-scale fine-tuning of a Whisper-based PyTorch model.
- Deployed the model in a real-time voice AI pipeline (LiveKit), enabling end-to-end conversational agent testing with low-latency inference.

AI Engineer **Remote**

Spotly Jobs *October 2025 - January 2026*

- Engineered ATS data pipelines using dlt and dbt, expanding job posting coverage by 5%+ and scaling automated processing to 500+ postings per pipeline run.
- Built 7 Dagster assets and sensors to automate ETL workflows, cutting manual engineering intervention by 50%+ and saving 10+ engineer-hours per week.
- Improved data ingestion throughput by 30–40% through optimized scraping logic and REST API handling, processing 150–200 additional postings per hour at no added infrastructure cost.

AI Engineer **Remote**

Outlier AI *May 2024 - September 2025*

- Trained LLMs using Reinforcement Learning from Human Feedback, improving instruction-following accuracy by 20%.
- Evaluated 1M+ tokens across 5+ datasets, identifying 300+ instruction-following errors that directly informed targeted finetuning.
- Refactored 3+ AI/ML models to generate more precise and succinct output.

Software Developer **San Francisco, CA**

University of San Francisco *September 2024 - May 2025*

- Lead developer on IDEA team software development, delivering 3 full-cycle applications.
- Created 3 business/entrepreneurship GenAI Python apps, adopted by 50+ students and professors in coursework.
- Conducted research on 10+ existing AI tools and technologies.

Software Developer **San Francisco, CA**

Founders Network *August 2024 - December 2024*

- Built Slack site integrations into an existing codebase of over 10,000 lines of code.
- Delivered 8 sprint cycles in an Agile environment with a 95% on-time delivery rate.
- Collaborated directly with the CEO and senior engineering team, gaining executive-level product visibility as an intern.

Software Developer **San Francisco, CA**

University of San Francisco *May 2024 - August 2024*

- Created and deployed 2 generative AI Thunkable apps, one of which is featured on PeopleCode.ai's front page.
- Lead developer on a full-stack PeopleMuseum project app and built a codebase of over 5,000 lines of code.
- Used React JS to create front-end UI, achieving load times under 2 seconds with optimized API connections.
- Worked on a team with 5+ graduate students and professors, delivering 100% of milestones on schedule.

Projects

Founder of ResumeAI

- AI-Powered ATS Resume Optimizer (Next.js, TypeScript, React, Supabase/Postgres, Anthropic Claude API, Stripe, Chrome Extension) — https://withresumeai.com/
- Founded and built (solo) a full-stack AI resume platform with ATS compatibility scoring, Claude-powered bullet rewriting, AI cover letter generation, PDF/DOCX export, and a companion Chrome extension; reached active users across 7+ countries with 80%+ week-over-week new-user growth.

ParkCast SF — End-to-End MLOps System for Parking Occupancy Prediction

- Shipped a production parking-occupancy API (FastAPI, Docker, GCP Cloud Run) serving predictions for 12.7K SF blocks at 8.98 MAE / 0.73 R^2 from a 33-feature LightGBM residual model.
- Built a fully automated MLOps retraining pipeline in GitHub Actions — DataSF ingestion, MLflow tracking, and a champion-challenger gate rescored on a shared test set to block silent regressions before GCS upload and Cloud Run rollout.
- Engineered leakage-safe temporal features (lag windows, recency-weighted training, KNN-inferred block aggregates) to extend metered-only data to citywide coverage and improve event-hour accuracy.

Creator of PantrIQ

- Leveraged AI to develop a full-stack mobile application.
- Built with React Native, TypeScript, TailwindCSS, Supabase, Google/Apple Authentication, OpenAI, and more.
- Scan, voice, and text input options, AI recipe suggestions, automated meal planning, automatic shopping list generation, and pantry tracking.

AI/ML Industry Data Dashboard

- Live data dashboard aggregating AI/ML research from arXiv and Medium with automated GCS scraping, trend visualizations, and Anthropic-powered topic clustering — deployed on Google Cloud Run.

Creator of the PeopleCodeOpenAI Python library

- Allows for easy access and use of OpenAI's tools and API access: https://www.peoplecode.ai/
- Contains text-to-speech, voice recognition, RAG, and more.
- Being used in intro courses at the University of San Francisco.

Developed a full-stack search engine

- Full-stack search engine with inverted index (Java, 1910 lines Javadoc, 1746 SLOC).
- Backend: Java 21, Apache OpenNLP, TF-IDF ranking, Log4j2, multithreading (locks + queues).
- Web crawler: sockets, HTTP, regex, reentrant locks.
- Web server: Eclipse Jetty + servlets; Frontend: HTML + Bootstrap.
- Collaboration: GitHub PRs, Eclipse IDE, JUnit testing, Maven.

Technical Skills

Languages: Python (PySpark, SparkSQL, PyTorch, Pandas, NumPy, Scikit-Learn, Matplotlib, Plotly), Java, C, SQL, NoSQL
AI/ML: Generative AI, Large Language Models (LLM), LLM fine-tuning, Prompt Engineering, Model Evaluation, RAG, Voice AI, Regression, Classification, Unsupervised Learning, Deep Learning
Tools: Git, GitHub, Bash, Docker, ETL Pipelines, Distributed Computing, AWS, GCP, MongoDB, Apache Airflow, LiveKit

Certifications

AWS Certified Cloud Practitioner
Google Cloud Digital Leader

Education

University of San Francisco — **San Francisco, CA**
M.S. in Data Science and Artificial Intelligence — *July 2025 - June 2026*

University of San Francisco — **San Francisco, CA**
B.S. in Computer Science — *August 2021 - May 2025*