ARTHUR WEI

+1 (510) 945-8276 ◇ Fremont, CA
arthur.wei50@gmail.com ◇ linkedin.com/in/haichuanwei ◇ github.com/Arthur-Systems

EDUCATION

M.S. Computer Science – AI Focus — University of California, Irvine
Expected June 2026 — GPA: 4.0/4.0
Focus on modern AI systems, deep learning, and user-centric ML applications.
Courses: Deep Learning, NLP, Reinforcement Learning, Computer Vision, Systems Design

B.S. Computer Science — University of California, Santa Cruz
June 2025 — GPA: 3.8/4.0
Trained in systems, software engineering, and applied AI development.
Courses: Data Structures, Operating Systems, AI, Software Engineering, Algorithms

PROFESSIONAL SUMMARY

AI-focused Software Engineer passionate about building intuitive, intelligent systems that empower users to create, adapt, and solve problems. Experienced across the AI lifecycle from prototyping LLM and vision-based features to deploying performant backend systems at scale. Committed to designing thoughtful interfaces, writing maintainable code, and collaborating across disciplines. Skilled in Python, TypeScript, and FastAPI, with hands-on experience in AI tooling, cloud-native development, and cross-functional team environments.

SKILLS & TOOLS

Languages: Python, TypeScript, SQL, C++
Frameworks/Libraries: FastAPI, React, Flask, PyTorch, Hugging Face, OpenCV
AI/ML: LLM Integration, Embeddings, Prompt Engineering, CV Pipelines, Model Serving
Cloud/DevOps: GCP (Vertex AI, Cloud Functions), AWS (Lambda, S3), Docker, Linux
Systems: PostgreSQL, Redis, WebSockets, GraphQL, REST APIs
Core Strengths: Product-focused AI Development, Full-Stack Collaboration, User-Centered Engineering

WORK EXPERIENCE

Cloud and AI Engineering Intern — Jan 2024 – Apr 2024
MetafoodX — *San Jose, CA*

- Developed AI-enhanced backend services using FastAPI and GCP Vertex AI to classify food types in real time, improving user labeling efficiency by 30%.
- Designed scalable retraining pipelines with GCP Cloud Functions and Pub/Sub to adapt models to evolving datasets.
- Refined containerized inference services, reducing model response latency by 20%.
- Collaborated with UX and ML teams to monitor and iterate on end-to-end user-facing AI features.

Backend Software Engineering Intern — May 2023 – Sept 2023
ArteraAI — *Los Altos, CA*

- Engineered async WebSocket-based data pipelines supporting AI model workflows, improving throughput by 40%.
- Built developer tooling for ML observability and issue tracking, cutting debug turnaround by 50%.
- Tuned database queries and endpoints for performant data delivery to model consumers.
- Contributed to cross-functional product sprint teams focused on improving AI-powered diagnostics.

PROJECTS

Grading Automation Tool — July 2024
Baskin Engineering — Python, Redis, Canvas API

- Built an AI-adjacent autograding platform to reduce manual grading by 50%, serving over 150 students per term.
- Designed intelligent name resolution logic to handle messy student data, increasing accuracy and saving instructors hours per week.

- Implemented Redis-based caching to reduce redundant API calls and boost backend processing speed by 30%.

Discord AI Collaboration Activity – Polynomial Apr 2024
CSE115A — React, TypeScript, Node.js, PartyKit

- Collaborated in a team of 6 to build a live Discord app where users could plan and track tasks in voice channels.
- Engineered real-time sync features with WebSockets; added authentication and notification logic to keep users engaged.
- Used agile methods and Scrum ceremonies to drive iterative releases and improve team coordination.

AI-Assisted Smart Home Controller Mar 2024
LA Hacks Finalist — Python, ChatGPT API, Firebase, React

- Developed a context-aware home automation system that adapted to users' habits using lightweight AI.
- Integrated OpenAI LLMs for natural chat-driven controls; trained behavior models for temperature and lighting prediction.
- Delivered a polished web + mobile dashboard and earned top finalist honors out of 200+ teams.

Autonomous Driving Simulation (Waymo Dataset) Feb 2024
w/ Prof. Leilani Gilpin, UCSC

- Adapted CV models to simulate safe, autonomous behavior in uncertain edge scenarios.
- Improved object detection precision using active learning and fine-tuning in PyTorch.
- Focused on model explainability and visualization to support future research.

Pan-Audit AI Toolkit (Internal) May 2023
ArteraAI

- Prototyped real-time object recognition system with YOLOv8 and CV pipelines.
- Developed human-in-the-loop interface for model feedback and tuning.

Real-Time Communication System May 2023
ArteraAI Internship — Python, WebSocket, Redis, PostgreSQL

- Built a production-ready async backend for real-time AI model communication between users and services.
- Implemented JWT-secured data access and reduced round-trip latency by 25% via Redis caching and database indexing.