Jianxiao Cai

La Jolla, CA | shawn.jx.cai@gmail.com | (657)627-6520

linkedin.com/in/jianxiao-shawn-cai | github.com/ShawnCai223

Education

University of California, San Diego — La Jolla, CA
M.S. in Computer Engineering — Sep 2024 – Expected Mar 2026

Shanghai Normal University — Shanghai, China
B.Eng. in Electronic Information Engineering — Sep 2020 – Jun 2024

Work Experience

ivector — Sacramento, CA (Remote)
Software Engineering Intern — Jul 2025 – Expected Sep 2025

- Designed and implemented the phone terminal interaction pages in Unity, integrating backend logic for user navigation.
- Developed C# login functionality with Firebase authentication, enabling secure user sign-in and session management.
- Set up Firebase database and authorization rules, ensuring consistent data flow and reducing access errors by 20%.
- Mapped Unity page assets with backend code, improving app responsiveness and cutting loading times by 15%.

Sentari — New York City, NY (Remote)
Software Engineering Intern — Jun 2025 – Jul 2025

- Developed modular service using cosine similarity and topic matching to track emotional continuity in user transcripts.
- Built a TypeScript parser extracting user intents from 200+ diary entries with over 90% pattern-based NLP accuracy.
- Used 768-dim MiniLM embeddings with storage to simulate full NLP pipeline for cold-start and long-term users.
- Implemented clean architecture with utility separation, typed interfaces, and comprehensive pipeline logging.
- Proposed workflow enhancements: clearer task breakdown, structured ownership, and standardized code reviews.

FORVIA HELLA — Shanghai, China
Advanced Engineering Intern — Jan 2024 – May 2024

- Worked on BLDC motor control system development using Model-Based Design (MBD) and embedded integration.
- Built motor control logic in Simulink and auto-generated code for Arduino, reducing low-level coding time by 40%.
- Designed and validated motor driver circuits, ensuring firmware-hardware compatibility across control states.
- Developed Simulink models for speed, rotational direction, and fault behavior, targeting real-time deployment.
- Established a reusable MBD workflow adopted by team, improving consistency and reducing test iteration cycles.

Projects

AI Blog Generator – LLM Integration & Backend Automation — May 2025 - Jun 2025

- Developed full-stack web app for keyword-based AI-generated blog posts tailored to specific topics.
- Integrated OpenAI GPT-3.5 API for prompt-engineered, SEO-friendly content generation with dynamic API handling.
- Automated scheduled content creation using Python's schedule library for periodic blog post generation and updates.
- Designed responsive interface with HTML templating, delivering real-time feedback and seamless backend integration.

Offline Edge AI Health Assistant on Jetson Nano – Edge AI Inference — Apr 2025 - Jun 2025

- Developed offline health system on Jetson Nano with five sensors and voice interaction, ensuring data privacy.
- Deployed TinyLLM under 4 GB RAM, achieving 90% voice recognition accuracy in low-resource environments.
- Compressed language model by 60% and replaced three libraries for stable real-time on-device inference.
- Built Python pipeline synchronizing sensor input and voice processing for cohesive real-time health data integration.

Skills

- **Programming Language:** Python, C/C++, Java, Shell, SQL, Go, Matlab, C#
- **Framework and Tool:** Git, Linux, Docker, MySQL, Flask, REST APIs, Spring Boot, HTML