Joshua Ko

Santa Clara, CA | joshua.ko.jko@gmail.com | in/joshua-ko-jko | github.com/joshuakojko | joshuako.dev

Education

San Jose State University San Jose, CA
B.S. Computer Science *Expected Graduation: Dec 2026*

- Intro to Computer Systems, Data Structures and Algorithms, Object-Oriented Design, Computer Architecture
- Organizations: ACM@SJSU Backend Dev, ML@SJSU Project Member, SCE Dev

Skills

Languages: Python, Java, JavaScript, TypeScript, SQL, MIPS Assembly
Frameworks: React.js, Node.js, Express.js, Next.js, Hono, Flask, FastAPI
Developer Tools: Git, Docker, SQLite, MongoDB, Pinecone, Firebase, Prometheus, Grafana, Postman, Cloudflare, Vercel

Experience

Software Engineering Fellow July 2024 – Sept. 2024
Headstarter AI

- **Top 3 finalist in Headstarter Hackathon** - Pam Track against 500+ competitors. In team of 3, jointly developed voice-based AI car dealership agent (WebSocket, OpenAI, Deepgram, Twilio) with real-time analytics dashboard (Next.js, MongoDB), enabling automated customer inquiries, appointment scheduling, and call data visualization
- Built and deployed 5 Full-Stack AI applications (RAG pipelines, LangChain, OpenAI) within 5 weeks in a team of 3

Software Engineering Intern June 2024 – Aug. 2024
San Jose State University - Software & Computer Engineering Society

- Optimized **video caching** for **Dockerized streaming server** in **Python**, implementing preemptive cache downsizing to prevent cache overflow and ensure efficient cache management before streaming to **RTMP media server**
- Added custom **Prometheus metrics** to monitor **API data rate**, **cache performance**, and **HTTP requests** for remote streaming server; utilized **PromQL** for metric queries and added **Grafana panels** for **real-time data visualization**
- Refactored **Express.js endpoint** to secure print requests, implemented **server-side validation** to enforce print limits and update print count, centralized print request logic to backend, preventing **unauthorized access**
- Designed and implemented comprehensive **unit tests** for the updated API endpoint using **Mocha**, **Chai**, and **Sinon**

Undergraduate Research Assistant Sep. 2023 – Aug. 2024
San José State University - Dr. Robert Chun

- Mentored high school students in research project on potentials of a **College Major Assessment Chatbot** via **OpenAI API's chat completion models** compared to traditional questionnaire interest profilers
- Participated in a collaborative research project investigating current application and limitations of **AR/VR/XR technology** in university education through the **Meta Quest 3**

Projects

Patent Mate - Stanford LLM x Law Hackathon 4 | *Next.js, React, TypeScript, LlamaIndex, Pinecone, OpenAI, Groq*

- Co-developed Patent Mate, a **full-stack AI-driven platform**, to streamline legal patent prefiling service by offering patentability search, NDA, invention disclosure analysis, and patent attorney/law firm recommendations features
- Leveraged **Groq's fast inference** for patentability criteria analysis and comprehensive patenteability report generation
- Utilized **LlamaParse** to parse public patent PDFs, **OpenAI model for embeddings**, and Pinecone for vector storage to enable **advanced RAG and similarity searches** for prior art and attorney recommendations

Ginder | *Next.js, React, TypeScript, GitHub OAuth, Flowise, Drizzle, Cloudflare (Workers, Vectorize, D1, KV, AI)*

- Led frontend development and key integrations for Ginder, a **personalized chatbot** enabling developers to discover **Open Source repositories** based on their GitHub repository metrics, skills, and interests.
- Collaborated on **LLM orchestration** flow with Flowise for personalized recommendations and integrated **WebSocket protocol** for low-latency, real-time response streaming, enhancing recommendation accuracy and user experience.
- Configured **GitHub OAuth** to authenticate Octokit fetching repositories for analysis, summarization, and **vector embeddings** with **Cloudflare Vectorize**, enabling accurate similarity search against Open Source projects
- Worked in 3-person team to deploy app on **Vercel** with **Cloudflare serverless functions**, ensuring accurate real-time recommendations