

Joshua Yi

Santa Clara, CA | jryi.work@gmail.com | linkedin.com/in/jryi | github.com/j-r-yi | U.S. Citizen · No Sponsorship

EDUCATION

Santa Clara University

Santa Clara, CA

B.S. in Computer Science, GPA: 3.97/4.0

Expected June 2027

Relevant Coursework: Machine Learning, AI, Computer Networks, Data Structure & Algorithms

SKILLS

Languages: Python, JavaScript, Java, C++, SQL (MySQL, PostgreSQL), HTML/CSS

Frameworks & Libraries: FastAPI, React, Next.js, Node.js, Express, Spring Boot, SQLAlchemy, LangChain, Redis

Tools & Technologies: AWS (EC2, Lambda, ECR), Azure (AI Foundry & Blob Storage), Docker, Git, GitHub

Actions CI/CD, Pytest, Postman, VS Code, RAG, REST APIs, Elasticsearch, Kafka

PROFESSIONAL EXPERIENCE

CrossRealms International

Chicago, IL (Remote)

AI Junior Engineer Intern

Jun 2026 – Aug 2026

- Contributed to an internal AI-powered knowledge and onboarding platform featuring **RAG**, **PII guardrails**, and automated developer tools that accelerates development workflows.
- Co-developed an **LLM**-driven unit-test generation pipeline using **Python**, **LangChain**, and **Azure AI Foundry**, generating 100+ tests across 25 modules and reducing test-development time by 40%.
- Implemented Git utilities and Python AST / **Pytest** checks for automated code change detection, LLM context building, and validation with retries for failed test generations.
- Hardened RAG pipeline by centralizing **PII guardrails** across **FastAPI** modules to ensure fail-closed protection.
- Provisioned 12 Azure resources using **Terraform (IaC)**, enabling reproducible environments and reducing onboarding time from 6 hours to 15 minutes.

Capgemini

Shanghai, China

Software Engineer Intern

Jun 2025 – Aug 2025

- Contributed to a real-time analytics platform to detect EV event risks and KOL insights, driving data-informed decisions and enabling faster issue detection.
- Implemented **Kafka**-based streaming pipelines to analyze sentiment and compute risks, storing results in **MySQL**.
- Optimized **FastAPI** endpoints and queries with **Redis** caching, reducing API latency from ~200ms to under 40ms.
- Built interactive **React** dashboards for real-time event tracking, risk assessment, and KOL analysis.
- Dockerized** the insights module to streamline workflows and deployment consistency.

PROJECTS

NextAI | Conversational Document AI Agent

Jun 2025 – Jan 2026

- Built an AI-powered PDF chat assistant supporting voice and text querying using **React**, **Express**, and **LangChain**.
- Engineered a custom AI agent with **GPT-5** to facilitate multi-document retrieval and long-form content analysis.
- Optimized document retrieval efficiency with in-memory vector caching to reduce latency.
- Integrated speech-to-text (**STT**) and text-to-speech (**TTS**) to improve accessibility and enable hands-free usage.

Poli Shop | Full-Stack E-Commerce Web Application

Jul 2024 – Oct 2024

- Built a full-stack e-commerce platform with **React**, **Next.js**, and **Spring Boot**, then deployed it on **AWS EC2**, featuring user authentication, checkout, order processing, and inventory management.
- Designed a responsive, accessible user interface with **TailwindCSS** and **Chakra UI**, leveraging **Redux** for efficient global state management.
- Implemented **Redis** caching to improve API responsiveness, **PostgreSQL** persistence, and **AWS S3** storage.
- Containerized the application with **Docker** to ensure consistent development and deployment environments.