Ankush Rai

Santa Clara, CA | +1 (408) 334-2219 | arai4@scu.edu | LinkedIn | GitHub | Portfolio

Summary

Software Engineer specializing in scalable backend systems, distributed systems, and applied AI. Experienced shipping LLM/RAG/MCP-powered products and multi-tenant data integrations on AWS for Fortune 100 enterprise clients. Currently building event-driven ingestion infrastructure at Allyvia, taking systems from prototype to production.

Education

Santa Clara University — Santa Clara, CA
Master of Science in Computer Science and Engineering — *Expected December 2026*

SRM Institute of Science and Technology — India
Bachelor of Technology in Computer Science and Engineering — *June 2020*

Experience

Allyvia — March 2026 – Present
Software Engineer Intern — *San Francisco*

- Built an event-driven client data ingestion pipeline using AWS SQS to decouple third-party sync jobs from the API, enabling reliable retries and back-pressure handling for high-volume QuickBooks/Square webhook traffic.
- Engineered a multi-tenant Python/Django REST API that integrates with QuickBooks and Square, unifying finance, inventory, and operations data into a single AI-driven ERP hub for SMB customers.

ZS Associates — March 2022 – January 2025
Software Engineer — *India*

- Led migration from a Tableau dashboard to a custom React + FastAPI web app, shipping a platform that reduced load time by **40%** and eliminated **$100K+** in licensing costs.
- Scaled FastAPI backend services with Redis caching and PostgreSQL pooling, instrumenting GitLab CI/CD, Pytest, and Prometheus for production monitoring, cutting P95 latency by **45%**.
- Designed and deployed an LLM-powered workflow on AWS for Pfizer (Fortune 100 pharma client), evaluating models and prompts that automated extraction from **10K+** reports, saving **20+** analyst-hours per week.

EY — October 2020 – February 2022
Software Engineer I — *India*

- Eliminated race conditions in concurrent Python services using threading locks, reducing transaction errors by **15%** on the General Motors OnStar production platform.
- Optimized response efficiency across Flask and Express.js applications using pagination and in-memory caching, improving API response time by **35%** for **10K+** monthly requests.

Projects

PaperTrail: AI Document Intelligence | *Python, FastAPI, LangChain, Docker, AWS*

- Engineered an LLM-powered platform that processes unstructured PDFs into structured data with OCR, semantic chunking, and vector embeddings, automating manual document analysis. Won NVIDIA's "Agents for Impact" hackathon.

Argus: MCP Prompt-Injection Firewall | *Python, FastAPI, MCP, JWT, Claude-as-Judge, WebSockets*

- Designed a bidirectional MCP proxy with layered defense (regex + LLM-as-judge) and JWT-based per-agent scope policy that intercepts tool calls/responses, blocks prompt injection and credential exfiltration, and streams events to a live dashboard.

FormWhisper: Voice-Driven AI Tool | *Python, FastAPI, React 19, TypeScript, LangGraph*

- Shipped a full-stack AI tool that reads PDF forms using a Vision-Language Model (Qwen2.5-VL) and fills them through voice conversation. Won Best Future Unicorn at Hack for Humanity 2026 from **80+** teams.

Technical Skills

Languages: Python, JavaScript, TypeScript, Java, SQL
Frontend: React, Next.js, TypeScript, HTML/CSS, Responsive UI
Backend: Django, FastAPI, Flask, Express.js, REST APIs, Microservices
AI/ML: LLMs, RAG, AI Agents, LangChain, LangGraph, Prompt Engineering, scikit-learn, PyTorch
Databases: PostgreSQL, MySQL, Redis, Vector Databases
Cloud & DevOps: AWS (SQS, EC2, S3), Docker, Kubernetes, Kafka, GitLab CI/CD, Git, Prometheus, Pytest
AI Coding Tools: Claude Code (advanced), Cursor, GitHub Copilot

Awards

Spot Award (ZS Associates) – Recognized for shipping a Python LLM-powered workflow for Pfizer that automated manual pharma report analysis.
Best Future Unicorn – Hack for Humanity 2026, selected from **80+** teams for FormWhisper, an AI voice-driven form-filling tool.