

Manan Shah

San Francisco, CA (Open to Relocate) | +1 (872) 279-4692 | mananshah03691@gmail.com | [Linkedin](#) | [Github](#)

Summary

AI Engineer with 3+ years of software engineering experience building production applications across fintech and enterprise domains. Hands-on experience with LLM-powered solutions, generative AI, embeddings, semantic search, Amazon Bedrock, and Pinecone. Strong background in Python, Java, REST APIs, microservices, AWS, Docker, and Kubernetes, with a focus on building reliable and scalable AI-enabled systems.

Skills

Programming Languages: Python, Java 17, TypeScript, JavaScript, SQL

Generative AI & LLMs: LLMs, RAG, Prompt Engineering, Embeddings, Semantic Search, AI Agents, Tool Calling, MCP

AI Frameworks & Platforms: LangChain, LangGraph, Amazon Bedrock, OpenAI API, Anthropic Claude, Hugging Face

Databases & Vector Search: Pinecone, ChromaDB, PostgreSQL, MongoDB, Redis

Backend & APIs: FastAPI, Spring Boot, REST APIs, Microservices, Apache Kafka

Cloud & MLOps: AWS (EC2, S3, RDS, EKS, Lambda, CloudWatch), Docker, Kubernetes, GitHub Actions, CI/CD

Tools & Practices: Git, Postman, JIRA, Agile Scrum, System Design, LLM Evaluation

Professional Experience

PayPal | *Software Engineer*

Apr 2025 – Present

- Built a Python and FastAPI based AI service to help support teams analyze merchant disputes and retrieve relevant resolutions from previously resolved cases.
- Created a Retrieval-Augmented Generation (RAG) pipeline using Amazon Bedrock embeddings and Pinecone for semantic search across historical dispute cases and supporting documents.
- Developed document processing workflows to clean dispute data, chunk large case records, generate embeddings, and store metadata for filtered retrieval.
- Integrated Amazon Bedrock language models with retrieved case context to generate concise resolution summaries and suggested next steps for support agents.
- Added prompt templates, structured outputs, retrieval filters, and validation checks to keep generated recommendations relevant and grounded in dispute data.
- Used Apache Kafka to process new dispute events asynchronously and trigger background indexing workflows without slowing down customer-facing APIs.
- Containerized the application with Docker and deployed it on AWS using Kubernetes, GitHub Actions, and CloudWatch for CI/CD, monitoring, and production troubleshooting.

Duck Creek Technologies | *Software Engineer*

Mar 2021 – July 2023

- Developed Java, Spring Boot, and Hibernate services for policy administration, billing, and claims workflows on an enterprise P&C insurance platform.
- Designed REST APIs for quote, policy, billing, and claims workflows and integrated backend services with payment processors and third-party insurance systems.
- Worked with PostgreSQL and MongoDB to store policy and transaction data, optimizing queries and indexes for high-volume application workflows.
- Built asynchronous processing workflows using Apache Kafka to exchange policy and claims events between distributed backend services.
- Helped break down parts of a legacy Java application into independently deployable Spring Boot microservices, improving maintainability and release flexibility.
- Investigated production issues using application logs, API traces, and database queries, identifying root causes across distributed services and third-party integrations.

Projects

TrailMail | *React, TypeScript, Spring Boot, PostgreSQL, REST APIs*

trailmails.com

- Built and deployed a Chrome extension that converts LinkedIn profile information into personalized outreach emails, with a React and TypeScript interface for reviewing and editing generated drafts.
- Developed Spring Boot REST APIs for contact lookup, template management, and outreach history, with PostgreSQL for persistent user and campaign data.
- Integrated Hunter.io and Gmail APIs using OAuth2, including token management, retries, and error handling, and published the product on the Chrome Web Store.

DialDo | *TypeScript, Electron, Twilio, REST APIs, SQLite*

dialdo.com

- Built a local-first macOS voice agent where users call a Twilio number to trigger tasks on their computer through a cloud gateway and Electron desktop client.
- Designed a tool execution architecture with a plugin-based registry, allowing the agent to discover and invoke new capabilities without modifying the core routing logic.
- Implemented persistent task state and execution history with SQLite and packaged, signed, and notarized the application for macOS distribution.

Education

Illinois Institute of Technology

Master of Science in Computer Science

Aug 2023 – May 2025

Chicago, IL