Affan Arif Khamse

New York City, NY · (516) 853-4972 · khamseaffan@gmail.com · github.com/khamseaffan · linkedin.com/in/affan-khamse

Education

Master of Science in Computer Science (GPA: 3.8/4.0) — New York, USA
New York University — *May 2025*

Bachelor of Engineering in Computer Engineering (GPA: 9.00/10.0) — Mumbai, India
University of Mumbai — *June 2023*

Technical Skills

Programming and Scripting Languages: Java, Python, TypeScript, C#, JavaScript, C++
Frameworks and Libraries: Spring Boot, FastAPI, Flask, .Net MAUI, Django, React, LangChain, Node.js
Cloud & Infrastructure: AWS (EC2, DynamoDB, S3, Lambda, SQS, SES), Azure, Docker, Kubernetes
Databases: SQL (MySQL, PostgreSQL), NoSQL (MongoDB, DynamoDB), Redis, Firebase
Testing & Monitoring: JUnit, PyTest, Django Test, CloudWatch, Chrome DevTools
DevOps & CI/CD: GitHub Actions, Travis CI, Docker Compose, AWS SAM, Git, Swagger/OpenAPI
System Design: Microservices, Event-Driven, Distributed Systems, RESTful APIs, WebSockets

Work Experience

Software Engineer — New York, USA
Novum AI — Jun 2025 – Present

- Architected an AI-powered call center assistant that provides **real-time** smart suggestions to representatives based on live transcription analysis, helping agents close deals faster and improve conversion rates
- Improved end-to-end suggestion latency from 3–4 seconds to **under 1.5 seconds** by parallelizing context retrieval, optimizing chunking strategy, and tuning AWS Lambda memory to reduce cold-start overhead
- Redesigned the retrieval pipeline using **layout-aware hierarchical chunking** with sparse search, dense vector search, and **cross-encoder re-ranking**, improving context relevance during live calls
- Built **event-driven backend services** with AWS SAM and CI/CD pipelines, implementing session-based authentication with auto-refresh and **multi-tenant RBAC** for 50+ concurrent sessions
- Reworked authentication architecture by **separating authentication and authorization** into dedicated middleware, reducing request complexity and improving maintainability using Redis-backed sessions
- Implemented **production security hardening** including webhook signature verification, CORS enforcement, and authorization middleware to prevent injection and scripting attacks

Software Engineer — New York, USA
InquisAI (Startup Project) — Jun 2024 – Mar 2025

- Developed an AI assistant platform using **LangChain**, **OpenAI Embeddings**, and **Chroma vector store**, achieving accurate document retrieval on 500+ internal test queries with semantic search
- Redesigned backend by migrating Flask to FastAPI with asynchronous processing and optimized database queries, reducing API latency by 30% through performance testing
- Led **Agile development** in a 3-person team, using Azure DevOps to manage sprints and implementing **CI/CD pipelines** for automated testing and deployment to AWS

Projects

Home Store – Multi-Tenant E-Commerce Platform — Jan 2025 - Present
Spring Boot, Spring Cloud, Eureka, Docker, PostgreSQL, React Router, Firebase, SwaggerUI — GitHub

- Designed a microservices architecture using Spring Boot, Eureka service discovery, and API Gateway, with modular services enabling independent deployment and **horizontal scaling** across tenants
- Containerized services with Docker and Docker Compose and exposed REST APIs with SwaggerUI documentation, enabling **consistent testing and rapid iteration**

FlashBids - Anti-Sniping Auction Platform — Sep 2024 - Dec 2024
AWS EC2, DynamoDB, S3, Redis, WebSockets, CloudWatch, Flask, Python, REST APIs, DevOps — Demo | GitHub

- Created a **real-time auction platform** using Flask and WebSockets to prevent last-second sniping, implementing a fair bidding mechanism with automatic time extensions
- Constructed a **scalable backend** using AWS EC2 with Auto Scaling and DynamoDB, designed to handle 1000+ of concurrent users through elastic infrastructure
- Engineered **non-blocking APIs** and Redis pub/sub event pipeline, achieving **30% improved latency** and sub-second response times verified with 25-30 concurrent users

Leadership & Technical Contributions

- **Mentored 50+ NYU students** in C++ and Java as Teaching Assistant; several secured internships post-course
- Introduced company-wide **Ruff / ESLint / Prettier** setup at Novum AI, improving code consistency and reducing review cycles