Affan Arif Khamse

New York City, NY · +1 5168534972 · khamseaffan@gmail.com · linkedin.com/in/affan-khamse · github.com/khamseaffan

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), Supabase, Redis, Pinecone
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, RAG Systems

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
- Reduced median suggestion latency from 3 seconds to ∼**1.2s** by parallelizing context retrieval, restructuring LLM prompts for 72% token reduction (4,900 → 1,950), and tuning AWS Lambda memory to the full-vCPU threshold
- Redesigned the retrieval pipeline using **layout-aware hierarchical chunking** (parent-child architecture) and **hybrid search** combining dense embeddings with sparse BM25 vectors via Pinecone, replacing flat chunking that destroyed table structure and other structure data
- Built **event-driven backend services** with AWS SAM and CI/CD pipelines, implementing Redis-backed session authentication and **multi-tenant RBAC** with role, permission, and company-scoped access control
- 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

Stoca – AI-Native Local Commerce Platform — April 2026 - Present
Next.js, TypeScript, FastAPI, Python, PostgreSQL, Supabase, pgvector, Claude API, GitHub Actions — GitHub

- Built an AI-native commerce platform where store owners manage their entire business through conversation with a Claude-powered assistant supporting **19 tool calls** via Vercel AI SDK streaming, covering inventory, pricing, orders, and promotions
- Engineered an autonomous product enrichment pipeline combining **Claude Vision** for shelf-photo identification, **pgvector semantic search** for catalogue matching, and Pexels API for image sourcing

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 with automatic time extensions to prevent last-second bid sniping
- Profiled Redis connection bottleneck (50-80ms per new connection during bid floods) and implemented connection pooling, dropping **p95 latency from 800ms to 200ms**
- Built scalable backend on AWS EC2 with Auto Scaling and DynamoDB, designed for 1,000+ concurrent users through elastic infrastructure

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