Affan Arif Khamse

khamseaffan@gmail.com · +1 (516) 853-4972 · linkedin.com/in/affan-khamse · github.com/khamseaffan

Technical Skills

Programming and Scripting Languages: Java, Python, TypeScript, JavaScript, C++
Frameworks and Libraries: Spring Boot, FastAPI, Flask, Django, Next.js, React, LangChain, Node.js
Cloud & Infrastructure: AWS (EC2, DynamoDB, S3, Lambda, SQS, SES), GCP, Docker, Kubernetes
Databases: SQL (MySQL, PostgreSQL), NoSQL (MongoDB, DynamoDB), Supabase, Redis
Testing & Monitoring: PostHog, JUnit, PyTest, Django Test, CloudWatch, Chrome DevTools
DevOps & CI/CD: GitHub Actions, Travis CI, Docker Compose, AWS SAM, Git, Swagger/OpenAPI
System Design: Agentic Systems, Microservices, Event-Driven, Distributed Systems, RESTful APIs
AI & Retrieval: LLM Tool Calling, RAG, Vector Databases (pgvector, Pinecone), Vercel AI SDK

Work Experience

AI Engineer — New York, USA
Centh AI (South Park Commons - backed) — May 2026 - Present

- Forward-deployed across enterprise clients, owning system design end-to-end and turning live client requirements into production build decisions
- Built an agentic credentialing workflow: **Google Document AI** for OCR, LLM-based classification and extraction, requirement checks, **primary source verification** across **50+ different sources**, and sanction searches
- Cut each primary-source verification from about **5 minutes** of manual work (navigate, type, read, save as PDF, upload to the ATS, approve) to under **30 seconds**, running them in parallel instead of one at a time
- Built a user-facing **extension** that puts the agent inside the credentialer's and onboarder's existing workflow, **adopted by 12 pilot users**, with their feedback driving the fixes
- Implemented security controls: RBAC, session activity access control, infrastructure firewall, webhook verification, injection prevention, encryption

AI Software Engineer — New York, USA
Novum AI — Jun 2025 - Apr 2026

- Architected an AI call assistant end-to-end, reducing suggestion latency from **3.4s to under 1.5s** via parallelized retrieval and 60% prompt token reduction (4,900 to 1,950)
- Redesigned the retrieval architecture with **layout-aware hierarchical chunking** and **hybrid search** combining dense embeddings with sparse BM25 vectors, replacing flat chunking that destroyed document structure
- Deployed **event-driven backend services** with AWS SAM, Redis-backed session auth, **multi-tenant RBAC**, and production security hardening (webhook verification, CORS enforcement, injection prevention)

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 business through conversation with a Claude-powered assistant with **55 tools** via Vercel AI SDK streaming, covering products and inventory, orders, promotions, analytics, and store settings
- Integrated **Claude Vision**, **pgvector semantic search**, and Pexels API into an autonomous product enrichment pipeline for shelf-photo identification and catalogue matching

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

- Co-built a **real-time auction platform** in a team project using Flask and WebSockets, with automatic time extensions to prevent last-second bid sniping, deployed on AWS EC2 with Auto Scaling and DynamoDB
- Profiled a Redis connection bottleneck (50-80ms per new connection under load testing) and implemented connection pooling, dropping **p95 latency from 800ms to 200ms**

Leadership & Technical Contributions

- **Mentored 50+ NYU students** as Teaching Assistant across 4 semesters, teaching C++ and Java along with UML diagrams, system design, and design principles
- Introduced company-wide **Ruff / ESLint / Prettier** setup, improving code consistency and reducing review cycles

Education

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

- Relevant coursework: Cloud Computing, Information Security and Privacy, Software Engineering, Algorithms

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

- Technical Member, Google Developer Student Club (GDSC)