Affan Arif Khamse

Backend Software Engineer | LLM Systems | Scalable Architectures | NYU CS Grad 2025
New York, NY · (516) 853-4972 · khamseaffan@gmail.com · github.com/khamseaffan · linkedin.com/in/affan-khamse

Education

Master of Science in Computer Science (GPA: 3.78/4) — New York, USA
New York University — *Expected May 2025*
Courses: Software Engineering, Cloud Computing, Machine Learning, DSA, Databases

Bachelor of Engineering in Computer Engineering (GPA: 9.00/10) — Mumbai, India
University of Mumbai — *June 2023*

Technical Skills

Programming and Scripting Languages: C++, Java, Python, Node.js, JavaScript
Frameworks and Libraries: Flask, Spring Boot, FastAPI (RESTful), Django, ReactJS
Cloud and DevOps: Amazon Web Services / AWS, Azure, Docker, Kubernetes, GitHub Actions, Travis CI
Databases & Query Languages: SQL(MySQL, PostgreSQL), NoSQL(MongoDB, Firebase)
Testing & Debugging: JUnit, PyTest, Django Test, Chrome DevTools
Tools and Methodologies: Git, GitHub, Figma, UML, Agile (Scrum), Object-Oriented Design
Specialized Expertise: API Design, Microservices Architecture, LLM Applications, Scalable Systems

Work Experience

Software Engineer — New York, USA
InquisAI (NYU R&D Project) | inquis-ai.com — Jun 2024 – Present

- Spearheaded development of an AI assistant builder that leveraged **LangChain** and **OpenAI** Embeddings to vectorize uploaded documents to generate domain-specific responses via GPT-4o
- Redesigned backend architecture, migrating from Flask to FastAPI, improving request throughput and **cutting API latency by 30%** with asynchronous processing
- Architected and deployed high-performance, scalable RESTful APIs, optimizing data operations to enhance efficiency and scalability, demonstrating the ability to support **1K+ concurrent users**
- Led Agile sprint management for a 3-person tech team using **Azure DevOps**, driving architecture decisions, infrastructure scaling, and AI model integration to reduce project delivery timelines by 25%

Software Teaching Assistant – Object-Oriented Programming — New York, USA
Courant Institute of Mathematical Sciences @ New York University — Jan 2024 – Present

- Mentored 50+ students in Object-Oriented Programming (**C++** and **Java**), applying software engineering principles to real-world scenarios and debugging complex issues
- Conducted code reviews and provided constructive feedback on programming assignments, enhancing code quality, maintainability, and adherence to best practices

Projects

Home Store – E-Commerce Platform (Backend) — Jan 2025 - Present
Spring Boot, PostgreSQL, React Router, Firebase — GitHub

- Developed a scalable inventory and store management platform for small-to-midsize retailers, with modular **RESTful APIs** created in Spring Boot
- Connected Firebase Cloud Storage to manage image uploads, with PostgreSQL used for schema-based data storage
- Designed the system with scalability in mind, planning a transition to **Dockerised microservices** using Spring Cloud, with well-defined service boundaries

Live Flash Auctioning System (Team Academic Project) - Full Stack Developer — Sep 2024 - Dec 2024
AWS EC2, DynamoDB, S3, Redis, WebSockets, CloudWatch, Flask, Python, REST APIs, DevOps — Demo | GitHub

- Constructed a real-time bidding platform using AWS EC2 with Auto Scaling, DynamoDB, and Redis, handling over **20K+ concurrent user sessions**
- Developed **non-blocking APIs** for auction creation and bidding, reducing response latency by 30%
- Built a low-latency WebSocket communication layer to broadcast live bids, **reducing client-server latency by 40%** and improving system reliability
- Orchestrated an **event-driven** bidding pipeline using Amazon SQS/SES (bid placement → live updates → auction close → email notifications), **achieving 99% bid placement success**

VibeCheck (Full-Stack Project) — Sep 2023 - Dec 2023
Django, Python, Bootstrap, AWS Elastic Beanstalk, PostgreSQL, Redis, TravisCI — Demo | GitHub

- Created a social platform that paired users based on real-time Spotify listening data and music preference similarity
- Established a reactive messaging framework using Redis Pub/Sub, reducing end-to-end **chat latency by 30–40%**
- Deployed via AWS Elastic Beanstalk with **CI/CD (Travis CI)**, a**chieving 87% test coverage** ensuring reliability