Sami Haroon Khan

Principal Engineer · AI/ML and Fintech Systems · Engineering Leadership

iam@samiharoon.com · +92 313 301 5179 · +1 650 262 6220 · linkedin.com/in/samihk · github.com/samihk · Karachi, PK

Principal Engineer with 8+ years building AI/ML systems and fintech payment infrastructure across seed-stage startups and Fortune 500 client environments. Took the co-founder path twice, which means shipping under pressure, owning architecture decisions end to end, and leading teams without the luxury of ambiguity. Looking to apply that depth inside a high-impact engineering organization where the problems are hard and the standards are high.

Core Stack: FastAPI, NestJS · LangChain, vLLM · AWS, GCP · PostgreSQL, Redis · Next.js, React.js · TypeScript, Python

PROFESSIONAL EXPERIENCE

Co-Founder and CTO | **Vupechat Inc.**, Remote USA *May 2024 – Present*

- Built a niche VTubers social media platform from zero to one. Developed WebRTC motion tracking plugins and shipped MVP in 3 months. Hired SWEs/ML engineers and cultivated a culture of high agency, ownership and grit.
- Wrote a computer vision classifier for content moderation where the cost of a false positive equalled a false negative; held 99.2% production accuracy on a platform where user trust was the only retention mechanism
- Built the WebRTC communication layer from scratch, holding sub-100ms latency under 1,000+ concurrent sessions and improving user retention by 80%
- Architected the full distributed systems stack for 18,000+ VTubers: real-time avatar animation, spatial audio processing, and synchronized virtual environments running at scale
- Built a user-matching system using embedding models and classifiers to segment users by persona and engagement pattern, measurably increasing community interaction depth

Staff Software Engineer | **Blossom.team**, Remote USA (concurrent contractor) *Dec 2022 – Apr 2024*

- Built an ML content moderation pipeline processing 100,000 daily interactions where the accuracy requirement did not soften as volume scaled; cut manual review by 90% and saved $200,000 annually
- Rebuilt the search system using preference modeling; relevance improved by 300% and user satisfaction ratings moved from 3.2 to 4.8 out of 5
- Hired and onboarded 8 senior engineers into a structured team org that supported 5x product growth within the following 12 months

Technical Lead | **Remotebase**, Karachi Pakistan *Feb 2022 – Jun 2023*

- Brought in as Technical Lead to own system design, architecture, and delivery across Fortune 500 accounts, coordinating teams of up to 25 engineers across concurrent client engagements
- Placed at Arrow Payments for 7 months as Engineering Manager; ran a 12-person team through two delivery cycles while contributing directly to core architecture decisions
- Migrated infrastructure to AWS CDK, cutting cloud costs by 35%; reliability improved in the process, which meant the cost reduction came at no expense to system stability
- Built session transcription and analytics tooling deployed across 5 enterprise clients, improving user engagement metrics by 60%

Senior Software Engineer | **Afiniti Inc.**, Karachi Pakistan *Dec 2020 – Apr 2022*

- Redesigned call routing and queue management logic, hitting 60% benchmark improvement and cutting memory footprint by 30%, without disturbing the contractual uptime SLA the platform was held to across Fortune 500 clients
- Wrote core shared-memory APIs in C/C++ for an AI-driven Omnichannel routing platform; the architecture supported high-availability requirements across enterprise deployments
- Established the company's first structured API documentation system; developer onboarding dropped from several weeks to a few days and cross-team productivity improved by 40%

Software Engineer | **Kepler Analytics**, Karachi Pakistan *Jan 2020 – Dec 2020*

- Built the analytics reporting layer for 25+ global retail chains, processing 500+ business metrics per client and contributing to an average 20% improvement in client revenue
- Implemented POS and ETL integrations across retail accounts, reducing data processing time by 70% and enabling real-time reporting at scale

Software Engineer | **Oraan**, Karachi Pakistan *Nov 2018 – Dec 2019*

- Took a fintech product from concept to production, shipping the complete PWA and mobile suite backed by a Java Spring-Boot microservices architecture
- Owned the technical roadmap for GCP deployment; designed and implemented role-based data management and secure API layer using KeyCloak and PostgreSQL

Software Engineer | **Techlogix**, Karachi Pakistan *Oct 2017 – Oct 2018*

- Built payment processing modules for Bank of Punjab handling over $1M in daily transactions; zero security incidents across the full deployment period
- Optimized multi-cluster deployments for the New England Journal of Medicine, cutting infrastructure costs by 25% and improving system performance by 35%

EDUCATION

Bachelor of Engineering in Software Engineering *Dec 2013 – Nov 2017*

NED University of Engineering and Technology, Karachi, Pakistan

Grade A · *Awarded Fully Funded Scholarship*

TECHNICAL EXPERTISE

Backend: FastAPI, NestJS, ElysiaJS, Spring Boot

AI and ML: LangChain, vLLM, Google A2A, PyTorch, Claude Code, Ollama, Docling, PaddleOCR

DevOps: AWS, GCP, Docker, Terraform, Kafka, Nginx, Jenkins

Databases: PostgreSQL, Redis, MongoDB, ChromaDB, FAISS

Web: Next.js, React.js, WebRTC, Node.js, Bun

Languages: TypeScript, Python, Java, Bash