QALAB HASSNAIN AGHA

CTO | AI Systems Architect | Real-Time Systems | Cloud Infrastructure

+92 323 7586006 | aghaqalabhassnain@gmail.com | Islamabad, Pakistan | LinkedIn | GitHub | Portfolio & Writing

PROFESSIONAL SUMMARY

CTO at Quickgen Technologies with 4+ years building production-grade AI systems, scalable backend architectures, and IoT-integrated platforms. Deep expertise in real-time data pipelines, LLM-powered automation, computer vision, MLOps, and cloud-native deployment on AWS and Azure. Hands-on technical leader who has taken multiple products from zero to deployment — architecting microservices, leading cross-functional teams, and shipping AI-driven solutions across healthcare, hospitality, fintech, and consumer tech. Technical author with 8 published engineering case studies on production AI deployment. Proven ability to reduce latency, cut infrastructure costs, and deliver measurable business outcomes through intelligent systems engineering.

CORE SKILLS

Languages & Frameworks	Python (FastAPI, Flask, Django), .NET Core, Node.js, REST APIs, WebSockets, gRPC, MQTT, UDP
AI / ML & MLOps	LLMs (Gemini, GPT-4, Claude), Whisper, DeepGram, YOLOv8, Computer Vision, NLP, RAG, Prompt Engineering, Model Fine-tuning, Model Quantization, Gait & Kinematics ML, ML Pipelines
Data Science	TensorFlow, Keras, scikit-learn, Pandas, NumPy, Matplotlib, Seaborn, Jupyter, LSTM, Data Visualisation
Real-Time & IoT	PCM Audio Ingestion, High-Frequency Sensor Streaming (200Hz+), BLE 5.0, ESP32, Edge AI, Anomaly Detection, Signal Processing
Cloud & DevOps	AWS (EC2, S3, Lambda, SQS), Microsoft Azure, GCP, Docker, Kubernetes, CI/CD, Nginx, Load Balancing, Auto-scaling, Vercel
Databases & Caching	PostgreSQL, SQL Server, MongoDB, Redis, Firebase, Supabase, Vector Databases (ChromaDB), Query Optimisation, Database Sharding
Architecture	Microservices, Event-Driven Architecture, Monolith-to-Micro Migration, API Gateway, Message Queues, System Design, Distributed Systems
Frontend / Mobile	React, Next.js, Flutter, React Native, TypeScript, Tailwind CSS, WebSocket Clients, Real-Time Dashboards
Security & Monitoring	JWT, OAuth2, RBAC, Encrypted Channels, Sentry, Grafana, Prometheus, System Monitoring, Alerting

PROFESSIONAL EXPERIENCE

Chief Technology Officer (CTO) | **QuickComm AE** • Islamabad, PK — 2026 – Present

QuickComm — Hospitality Real-Time AI Communication Platform (AWS)

- Architected a real-time audio system replacing walkie-talkies in hotel & security environments — cut staff response time by ~45% and reduced per-property operational cost to ~$3/month
- Built PCM audio ingestion pipeline with Whisper & DeepGram achieving 94%+ transcription accuracy; intent classification via Gemini LLM at ~88% precision for automated task routing
- Designed and executed a full monolith-to-microservices migration on AWS, improving throughput 3× and reducing deployment downtime to near-zero via blue-green deployments
- Implemented real-time monitoring & alerting; reduced critical incident response time by 70% through automated anomaly detection in service health metrics

IoT & Full Stack Developer | **upLYFT** • London (Remote), UK — 2024 – Present

Physical Rehabilitation SaaS Platform (Azure)

- Designed and deployed the complete backend & web-app architecture for a two-sided rehab platform connecting clinicians and patients, reducing onboarding setup time by ~60%

- Integrated an ML-based kinematics pipeline — gait analysis and kinetics models processing real-time body motion from IoT wearable sensors with 92%+ movement classification accuracy
- Built high-frequency sensor data ingestion at 200Hz+ over WebSockets and UDP; sub-100ms end-to-end latency on Azure with auto-scaling supporting 500+ concurrent sessions
- Led a 5-person cross-functional team through 3 major releases, delivered on schedule

Chief Technology Officer (CTO) | **Quickgen Technologies** • Islamabad, PK 2022 – Present

Leading technical strategy, architecture, and end-to-end delivery across a portfolio of AI, IoT, and SaaS products in healthcare, hospitality, consumer tech, and fintech. Directing 5+ engineers across the full SDLC — from system design through deployment, monitoring, and post-launch iteration.

CCTV Anomaly Detection — Multi-Phase Computer Vision System

- Evolved a surveillance system across 5 development phases to production using YOLOv8 for crowd surge, loitering, intrusion, and fight detection across 8+ simultaneous camera feeds
- Achieved 91% detection accuracy with configurable per-zone alert thresholds; reduced false-positive security alerts by ~35%

PaperIntel — RAG Research Intelligence System

- Built an end-to-end Retrieval-Augmented Generation pipeline: PDF ingestion, section-aware chunking, BGE-M3 embeddings, ChromaDB vector store, cross-encoder reranking, and citation-aware generation
- Implemented hybrid BM25 + dense vector retrieval with query expansion and multi-hop decomposition, achieving higher precision vs single-mode baselines

Apollo Golf Companion App

- Developed complete backend for a GPS-based golf app serving 1,000+ active users — ball tracking, scorecard management, shot analytics with <80ms query responses and ±2m GPS accuracy

Real-Time Multilingual Audio Translation Platform

- Built real-time translation backend capturing live PCM audio, transcribing and translating into 10+ languages in <800ms end-to-end; 96%+ accuracy via custom fine-tuned Whisper

The Giving Cube — Smart IoT Donation Ecosystem

- Designed complete platform (smart donation box + mobile app + web dashboard) — processed £50K+ in donations; 99.9% transaction reliability via Stripe; BLE/WiFi dual-mode sync

Additional Projects at Quickgen

- BadarAI Voice Assistant (Deepgram STT → Gemini LLM → ElevenLabs TTS, <800ms latency, live tool-calling), MRO Digital Twin (7 role-based dashboards, 34+ DB tables, full vehicle lifecycle), AI Pendant (ADHD wearable, 160+ API endpoints), Smart Boxing Gloves AI (<50ms data-to-insight, 200Hz+ sensor), Scentix (TVOC smart scent system), Ergomatics (18-module BLE cushion)

AI Engineer & Backend Developer | **Freelance (Upwork & Direct)** • Remote 2021 – 2022

- Delivered 15+ AI and backend solutions for international clients — computer vision (OCR, pose estimation, object detection), NLP automation pipelines, full-stack web applications
- Maintained 5-star Upwork client rating; average project delivery 20% ahead of schedule; stack: Python, FastAPI, React, Docker, AWS

IT Intern | **CareCloud** • Islamabad May – Aug 2021

- Built health services REST APIs in .NET Core C# and wrote SQL stored procedures for a live production healthcare system

Software Engineer Intern | **PTCL** • Islamabad Aug – Sep 2019

- Developed an Employee Record Search desktop application for HR using Python and deep learning

EDUCATION

MS Computer Science | **Sir Syed CASE Institute of Technology** • Islamabad *2023 – 2025*

B.E. Computer Engineering | **National University of Sciences & Technology (NUST)** • Islamabad *2017 – 2021*

TECHNICAL WRITING & THOUGHT LEADERSHIP

8 long-form engineering case studies documenting production AI deployments — qalab-e-hassnain.github.io/blog

→Deploy a CV Model to Production (Zero to 91% Accuracy)

→RAG Architecture in Production: ChromaDB + BM25 Hybrid Retrieval

→Real-Time IoT Platform: BLE, WebSockets & 200Hz Sensor Streaming

→Model Quantization in Production: 60% Cost Reduction

→Building an LLM Real-Time Audio Pipeline at Scale

→Monolith to Microservices: 3× Throughput on AWS

→Production AI Deployment Checklist: 23 Gates Before You Ship

→YOLOv8 in Production: Multi-Camera CCTV Anomaly Detection

NOTABLE PERSONAL PROJECTS

- **Fake Ad Detection** — NLP fraud classification for 50,000+ online ads; 94% accuracy using LSTM, SVM, Doc2Vec & TF-IDF ensemble; 40% faster inference vs baseline
- **AI Motion Tracking System** — real-time wearable sensor processing (gyro, accelerometer, BPM at 200Hz+) with movement classification at 91% accuracy
- **Book Cover Recognition App** — computer vision + OCR desktop app; 100% accuracy on Goodreads review retrieval across 5,000+ trained book covers
- **Demand Forecasting System** — ML model for SME inventory prediction improving stock decision accuracy by ~30%; E-Commerce Price Intelligence Platform scraping 10+ brands in real-time

CERTIFICATIONS

Neural Networks and Deep Learning — Coursera (5HAQEMXEHZ6J)

Applied Machine Learning in Python — Coursera (3YUUQ4ZWT8LJ)

Applied Text Mining in Python — Coursera

Microsoft Office Specialist Word 2013 — Microsoft

Applied Data Science with Python — Coursera (TBX3K474H74E)

AI for Medical Diagnosis — Coursera (NWJ6EAVUDZNC)

Applied Social Network Analysis in Python — Coursera (F6RA6Z9R4C47)

Python 3.6 Complete Course — Udemy