

Awais Khan

Senior Software Engineer (Platform & AI Infrastructure)

Kabirwala, Pakistan | Born 25/10/2001 | +92 306 1675704

awais.push@gmail.com | [LinkedIn](#) | [GitHub](#)

I can provide my own visa/work permit for Germany within 5–8 weeks.

My notice period is 15 days.



Technical Skills

Python (6 years of project experience) | TypeScript (5) | Node.js (5) | PostgreSQL (5) | AWS (5) | Kubernetes (5) | Rust (3) | LLM & Voice AI (3) | NestJS | FastAPI | Django | gRPC | GraphQL | REST | Redis | MongoDB | Cassandra | Docker | Terraform | ArgoCD | CI/CD (GitHub Actions, Jenkins) | Prometheus | Grafana | LangChain | LangGraph | Twilio | LiveKit | WebRTC/SIP

Work Experience

01/2025 – today

Senior Software Engineer — Platform & AI Infrastructure

CodiuX, Lahore, Pakistan (Onsite)

Own a shared multi-region voice-AI architecture powering three products: real-time conversational pipeline telephony (Twilio / SIP trunking / LiveKit / WebRTC-SDP) → STT → LLM orchestration (Claude, GPT-4) → TTS — with voice-activity detection (VAD) and turn-taking for natural, low-latency conversation; on-call and incident response under HIPAA, GDPR and APP audits.

Project: [CallJess](#) — AI voice receptionist for veterinary clinics that answers calls, books appointments and handles client intake autonomously.

- Scaled the call engine to 2,000–5,000 live calls per hour across the EU, US and Australia.
- Built the LLM eval harness gating every release of the conversational pipeline.

Technologies used: Python | Rust | PostgreSQL | Twilio | SIP | LiveKit | WebRTC/SDP | Claude & GPT-4 | AWS

Project: [Simz](#) — US-based, HIPAA-compliant conversational voice-AI assistant for health-care, legal, hospitality, restaurant and auto verticals.

- Delivered real-time intake, appointment scheduling and live transfer with integrations into Salesforce, Twilio, Five9 and HubSpot.

Technologies used: TypeScript | NestJS | PostgreSQL | LangGraph | Twilio | LiveKit | Claude & GPT-4 | AWS

Project: [SimpleTalk](#) — US-based no-code voice-AI sales platform with 70,000+ active users and \$1M+ ARR.

- Ran AI agents that make and answer calls, qualify leads and book appointments 24/7 up to 1,800 simultaneous conversations in 12 languages.
- Integrated live CRM sync with Salesforce, HubSpot, Pipedrive, Zoho and GoHighLevel on GKE.

Technologies used: Node.js | MongoDB | LangChain | LangGraph | Twilio | LiveKit | Claude & GPT-4 | AWS | Kubernetes (GKE)

02/2023 – 01/2025

Full-Stack Engineer — Infrastructure & Security

Xref, Sydney, Australia (Remote)

Project: [Xref](#) — Australia-based, ASX-listed (ASX:XF1) HR-tech SaaS for automated reference & background checking and engagement surveys with thousands of customers across APAC, Europe and North America, 7M+ references across 195 countries.

- Owned the refactor of microservices on AWS Lambda, SQS, SNS, S3 and PostgreSQL for the multi-tenant platform, cutting per-request cost 20–30% at peak.
- Ran penetration testing on customer-facing APIs and closed several critical vulnerabilities.
- Tuned Python/Django endpoints to raise throughput and cut p99 latency.

Technologies used: Python | Django | AWS Lambda | SQS | SNS | S3 | PostgreSQL | REST | API & SQL-injection penetration testing

08/2021 – 02/2023

Software Engineer

StudioX (formerly Secomind, a SECO S.p.A. company), San Jose, USA (Remote)

Project: [Clea](#) — European industrial-grade IoT / Edge-AI framework by SECO S.p.A. (Italy, Euronext STAR Milan-listed; German office in Stuttgart) for managing large device fleets — vending machines, EV chargers, coffee machines and trusted by Hitachi Energy, Google, La Marzocco, Nayax, SACMI and Riello UPS.

- Owned IoT SaaS platform development, making real-time processing reliable across sensor networks on the Astarte data-orchestration platform under tight SLAs.
- Built REST and GraphQL inference APIs on AWS that cut enterprise integration from weeks to days.

Technologies used: Python | Rust | PostgreSQL | Cassandra | MQTT | Astarte IoT | REST | GraphQL | AWS

Open-Source Projects

QuicMQ (Rust)

QUIC message broker giving each subscription its own stream to avoid the head-of-line blocking Kafka and NATS have. WAL durability, replay, Prometheus metrics, and a CLI. github.com/Awaishkhan404/quicmq

Sesame-csm (Rust)

Streaming inference runtime (Tokio, gRPC) that turns batch-only voice models into sub-second conversational backends. github.com/Awaishkhan404/sesame-csm

Education

10/2018 – 11/2022

Bachelor of Science in Information Technology

University of Education, Lahore, Pakistan

03/2016 – 07/2018

FSc Pre-Engineering (HSSC)

Superior College, Khanewal, Pakistan

Achievements / Certifications

OpenAI Academy (Agents and Workflows) | HackerRank (Software Engineer)

Language Skills

English (Fluent) | Urdu (Mother tongue) | German (Basic, currently learning)

Hobbies

Padel | Travel

Awaishkhan