

Talha Mukhtar

SENIOR MOBILE & APPLIED AI ENGINEER

Multan, Pakistan | +92 306 7447900 | talha.sumra11@gmail.com | linkedin.com/in/talhasumra | Open to Remote

PROFESSIONAL SUMMARY

Senior Mobile and Applied AI Engineer with 8+ years of experience delivering production Android and Flutter applications across EV charging, IoT, communication, monitoring, queue management, and business operations. Builds AI agents and intelligent products using local and hosted LLMs, deterministic guardrails, structured actions, streaming interfaces, on-device inference, and evaluation workflows. Built and maintained 12+ mobile applications, including systems handling 1M+ records, plus conversational, voice, coaching, and computer-vision AI products.

CORE COMPETENCIES

Mobile Engineering: Flutter, Dart, Android, Kotlin, Java, Jetpack Compose, Android SDK, XML, responsive UI, tablet support

AI Agent Engineering: Conversational agents, domain-specific agents, local and hosted LLMs, Qwen3 8B, Ollama, prompt design, structured actions, stateful workflows, domain guards, human-in-the-loop review

Multimodal & On-Device AI: Sherpa-ONNX, WhisperKit Large-v3, Silero VAD, Kokoro TTS, PP-OCRv5, ONNX Runtime, streaming speech, computer vision

Backend & Cloud: Python, FastAPI, Next.js, TypeScript, REST/SOAP APIs, JSON, WebSockets, SQLite, Prisma, Firebase, AWS, Google Cloud Platform, Cloudflare

Architecture & Realtime: MVVM, clean architecture, repository pattern, WebSockets, Pusher, BLE, TCP/IP, UDP, Wi-Fi, IoT, offline-first workflows

Quality & Delivery: Automated testing, prompt evaluations, Phoenix LLM evals, profiling, debugging, Sentry, GitHub, Bitbucket, Jira, Basecamp, Slack

PROFESSIONAL EXPERIENCE

App Technologies - Multan, Pakistan

Senior Mobile App Developer | January 2018 - Present

Career progression: Senior Flutter Developer, Senior Android Developer, Mid-Senior Android Developer, Mobile Developer

- Developed and maintained 12+ production mobile applications across EV charging, IoT monitoring, communication, reporting, queue management, and business workflow domains.
- Built and optimized applications handling 1M+ records, improving data loading, responsiveness, low-latency access, and stability for large operational datasets.
- Delivered native Android products using Kotlin, Java, Jetpack Compose, Android SDK, XML, Firebase, REST/SOAP APIs, and third-party SDK integrations.
- Built Flutter applications with reusable widgets, responsive layouts, API integrations, multiple product flavors, and reliable state-driven user flows.
- Implemented realtime and device communication using WebSockets, Pusher, BLE, TCP/IP, UDP, Wi-Fi, and offline workflows for connected products.
- Expanded delivery into applied AI with LLM agent workflows, local model orchestration, deterministic guardrails, streaming interfaces, on-device OCR, and automated evaluation datasets.
- Improved release quality through profiling, crash monitoring, Sentry diagnostics, production issue resolution, and mentoring 10+ trainees and junior developers.

AI AGENT & APPLIED INTELLIGENCE PROJECTS

TREX - Real-Time Restaurant Voice Ordering Agent | Python, FastAPI, WebSockets, Qwen3 8B, Sherpa-ONNX, WhisperKit, Silero VAD, Kokoro TTS

Built a local-first English voice agent for TFC with streaming audio, live captions, natural endpointing, barge-in, and low-latency responses. The agent answers menu questions, validates inventory, manages pickup or delivery orders, updates customer details mid-conversation, and persists orders in SQLite. Added structured SOP/test datasets and Phoenix-compatible LLM evaluations.

MedList - Private On-Device Medicine OCR | Flutter, Dart, PP-OCrv5, ONNX Runtime, SQLite

Created an iOS and Android app that turns medicine-order photos and screenshots into editable lists without an OCR API or runtime network dependency. Implemented text detection, recognition, confidence scoring, medicine-field parsing, audit storage, and mandatory human review before saving.

Nuvio - Local-First AI Chat Platform | Next.js, TypeScript, React, Ollama, SQLite, Prisma, Cloudflare Tunnel

Developed a responsive local AI chat MVP with streaming model responses, persistent conversation history, installed-model selection, passcode sessions, prompt limits, per-IP rate controls, health checks, and secure public demo access without exposing the Ollama service.

Snooker Coach AI | Next.js, TypeScript, React, OpenAI/Ollama, Vitest

Built a level-aware coaching agent for snooker rules, technique, cue-ball control, drills, practice plans, strategy, and match-note review. Added provider abstraction, deterministic domain and safety guards before model calls, typed drill content, unit tests, and prompt-evaluation fixtures.

SELECTED MOBILE & CONNECTED PRODUCT PROJECTS

EOSVOLT | Flutter, iOS, Android, Kotlin, Jetpack Compose

EV charging product focused on quick and reliable charging experiences across cross-platform and native mobile applications, including multi-flavor build and release workflows.

AeroGuard Air Monitor | Android, Kotlin, Jetpack Compose, IoT

Real-time air-quality monitoring application that surfaces live environmental information from connected devices.

PowerFuel | Android, Java, APIs

Mobile application supporting intelligent charging stations and related operational and administration workflows.

Queue Management Suite | Android, Kotlin/Java, Offline Workflows

Worked across customer feedback, counter display, tablet, appointment, and offline queue-management applications for service environments.

EquiLoco | Android, Java, Kotlin

Mobile-first platform connecting horse owners, riders, and experts through communication and service workflows.

Kurira | Flutter, Dart, Realtime Communication

International calling application designed for accessible low-rate calls, especially to destinations in Africa.

EDUCATION

Master of Computer Science - Virtual University of Pakistan

LANGUAGES

English (Advanced) | Urdu (Native)