

Sheharyar Ahmed

sheryahmedme1@gmail.com | sheharyar.dev | github.com/sheharyar-ahmed | linkedin.com/in/sheharyar-ahmed

Islamabad, Pakistan (UTC+5) · 4 hrs daily overlap with US Eastern · Remote

Summary

Senior software engineer with four years of end-to-end delivery at a senior-led studio, building web, mobile and AI systems for fintech, logistics and SaaS clients. Provide technical leadership across concurrent engagements, owning architecture and reviewing the work of 4 to 6 engineers. Build production LLM systems where correctness is enforced structurally rather than prompted for, and verified by automated testing and offline evaluation harnesses rather than asserted.

Experience

Devkross

Sep 2022 – Present · Remote

Senior-led software studio delivering custom web, mobile and AI products across fintech, healthcare, logistics and SaaS.

Senior Software Engineer

Jan 2026 – Present

- Own system architecture across concurrent client engagements and review the work of **4 to 6 full-stack and mobile engineers**, defining service boundaries, API schemas, database isolation models and the failure modes each has to survive before sprint kickoff.
- Lead AI feature delivery on client products, covering LLM integration, retrieval pipelines and evaluation, so model behaviour is measurable rather than anecdotal.
- Delivered a smart-glasses-to-cloud video transport and device-provisioning system for a hardware client, splitting control plane from data plane so recordings upload direct to cloud with **zero media bytes transiting the phone**: BLE discovery, AES-encrypted serial verification and a time-boxed Wi-Fi Direct window in a React Native control app, and an on-glasses Kotlin agent with a **deduplicated WorkManager upload queue surviving process death**.
- Engineered that transport for remote diagnosis under **zero physical hardware access**, with all verification running through a remote tester loop, and gated signed QA builds on **64 Jest cases across 7 suites** plus Kotlin JUnit tests, lint, APK signature parity checks and version lockstep across two repositories.

Software Engineer

Oct 2024 – Dec 2025

- Owned feature delivery on concurrent client engagements, taking systems from architecture through production support.
- Owned database design and access control across client projects: schema modeling, migrations, PostgreSQL row-level security and multi-tenant data boundaries.
- Ran PR review and CI release gates for the delivery pod, and handled production incident triage on live client systems.

Associate Software Engineer

Sep 2022 – Sep 2024

- Shipped the customer onboarding and document-verification flow on a fintech client portal, covering multi-step form state, validation and error recovery, and authenticated file upload, from schema design through production release.
- Integrated third-party services and internal APIs behind consistent REST contracts, with automated tests wired into CI release gates.
- Grew from assigned tickets to owning modules end to end: data modeling, implementation, code review and deployment.

Featured Projects

Autonomous Support-Triage Agent and Evaluation Harness

2026 · github.com/sheharyar-ahmed/inbox-triage-agent

TypeScript, Claude Managed Agents, MCP, Zod, Vercel Serverless, Vitest

- Engineered an 849-line SSE consumer that survives the failure modes vendor examples get wrong: a three-condition terminal gate immune to a decoy idle event arriving microseconds early, stash-and-dispatch resolution of discriminator-less blocking signals, and reconnect-with-history-replay whose dedupe never gates the terminal check. Passed a **31/31 ship gate across ten grader-evaluated tickets**, with an embedded prompt-injection payload refused, quoted as evidence and escalated rather than obeyed.
- Verified the system offline: **230 hermetic tests (401 assertions)** replay committed production traces from a clean clone with no API key at **\$0 spend**, a generated-JSON-Schema drift test pins the published tool contract to its Zod source, and a memory-write tripwire separates integrity halts from format failures. 71 numbered architecture decisions recorded with rejected alternatives.

Local-First Retrieval-Augmented Document Q&A Engine

2025 · github.com/sheharyar-ahmed/airlock

Python, FastAPI, Next.js 15, React 19, sentence-transformers, NumPy, typed SSE

- Built a fully offline RAG pipeline with no framework, no vector database and no cloud API, for documents that cannot legally be pasted into a hosted model. Page-scoped chunking (800-character windows, 150 overlap, never crossing a page boundary) makes every citation map to **exactly one page by construction**; retrieval runs as L2-normalized cosine over an (n, 384) embedding matrix with np.argmax partition O(n) top-k, so documents produce **zero network egress at runtime**.
- Hardened a typed contract across a racy streaming boundary: 4 Pydantic-typed SSE event classes, a hand-written client frame parser handling partial TCP reads and split multi-byte UTF-8, a generation-counter fence dropping callbacks from superseded streams, and a finally-guaranteed settle converging four distinct failure paths onto **one terminal UI transition**.

Cross-Platform AI Field-Quoting and Evidence-Cited Estimation Engine 2025 · github.com/sheharyarr-ahmed/quotelens

React Native (Expo), TypeScript, LangGraph, FastAPI, faster-whisper, Supabase Postgres RLS

- Architected a seven-node LangGraph pipeline (parallel per-photo vision fan-out, join, bounded retry edge) turning site photos and spoken narration into itemized quotes, where every line item is cross-checked against a code-stamped observation set and priced only from a fixed catalog, so invented catalog IDs drop to unpriced and the model **structurally cannot fabricate a price**. Secured tenancy with **34 RLS policies across 9 tables** and a JWT verifier with pinned per-algorithm key selection.
- Implemented an event-sourced realtime mobile UI over Postgres CDC: a framework-free reducer with idempotent event folding and a monotonic event-id watermark that makes history replay a no-op across regenerates, proven **three independent ways** (75 offline tests, a committed production-run fixture, a 70-assertion live harness). Cached-context regeneration re-runs in **3 text-model calls** with zero re-paid vision inference.

Database-Enforced Multi-Role Isolation and Proof Suite 2026 · github.com/sheharyarr-ahmed/load-lane

PostgreSQL 16 RLS, plpgsql, TypeScript, Drizzle, Vitest, Docker

- Architected four-role marketplace data isolation as **10 row-level-security policies** under ENABLE + FORCE ROW LEVEL SECURITY with a dedicated non-owner login so policies cannot be bypassed, a NULLIF(current_setting(...)) idiom making unset actor context **deny-by-default**, and a BEFORE UPDATE plpgsql guard enforcing a column-level old-versus-new rule that RLS structurally cannot express.
- Verified the boundary with a **13-case isolation matrix asserting the denial mechanism** rather than mere failure (SQLSTATE 42501 versus P0001 with message match versus zero-row filtering), each case in a rolled-back transaction as the non-owner role.

Additional Projects

Asynchronous ML Audio Source-Separation Platform 2025 · github.com/sheharyarr-ahmed/stemfoundry

Python, FastAPI, Celery, Redis, PyTorch (MPS), Demucs, Next.js 15, Stripe

- Diagnosed a PyTorch MPS convolution kernel ceiling from stride arithmetic and clamped the model's internal segment window rather than enabling silent per-op CPU fallback, benchmarking **0.28-0.58x realtime GPU separation** against a 0.89x CPU baseline; a compare-and-swap job claim was race-tested to **exactly one winner** across a **25/25 legal-transition matrix**, with 130 network-free tests.

Real-Time Voice-Agent Scheduling and Observability Platform 2026 · github.com/sheharyarr-ahmed/voxdesk

Next.js 15, TypeScript, ElevenLabs Agents, Cal.com API v2, Drizzle, PostgreSQL, HMAC-SHA256

- Engineered synchronous mid-conversation tool contracts for a production voice agent booking real calendar events, with per-tool time budgets derived from measured latency (**111ms p50** availability lookups under a 2800ms budget) and idempotent CTE ingest whose deliberate no-op ON CONFLICT DO UPDATE defeats an MVCC snapshot race; a build-time adapter seam swaps the metered voice SDK for a scripted fake, so the **120-test suite runs at \$0 platform spend**.

Citation-Enforced LLM Compliance Audit Engine 2026 · github.com/sheharyarr-ahmed/auditdoc-app

Python, FastAPI, Anthropic forced tool-use, PyMuPDF, Next.js BFF, TypeScript

- Made hallucinated citations **structurally impossible** rather than prompt-discouraged: forced tool-use restricts the model's citation vocabulary to integer indexes into the extractor's own chunk array, every returned index is type- and bounds-checked, uncitable FAIL verdicts auto-downgrade to lower-confidence PARTIAL, and a Pydantic validator backstops the invariant, yielding **4 independent enforcement layers** each attacked by its own adversarial test.

Native iOS Focus-Session Engine with Lock-Screen Live Activities 2025 · github.com/sheharyarr-ahmed/focusframe

Swift 6, SwiftUI, SwiftData, ActivityKit, WidgetKit, Keychain Services

- Built a two-target SwiftUI app with **zero third-party dependencies** under Swift 6 strict concurrency, and designed inter-process Live Activity state under a denied App Group entitlement: the timer epoch rides in static ActivityAttributes so the system render process ticks the countdown at **zero IPC per second**, with ContentState pushes fired only on semantic events to preserve the ActivityKit update budget.

Technical Skills

Languages: TypeScript, JavaScript (ES6+), Python, Kotlin, Swift 6, SQL, HTML/CSS

Frontend: React 19, Next.js 15 (App Router, RSC, Server Actions), Tailwind CSS, TanStack Query, styled-components

Mobile: React Native (Expo), Android (Kotlin, BLE, Wi-Fi Direct P2P, WorkManager, OkHttp), Rokid CXR-M smart-glasses SDK, iOS (Swift 6, SwiftUI, SwiftData, ActivityKit, WidgetKit)

Backend and APIs: Node.js, Express, FastAPI, REST, Server-Sent Events, WebSockets, Celery, Drizzle, Prisma, Zod, Pydantic

Data: PostgreSQL (row-level security, CDC, pgvector), Supabase, Redis, SQLite, schema design and migrations

Generative AI and Agents: Generative AI, AI agents and agentic orchestration (Claude Managed Agents, MCP, LangGraph),

Anthropic Claude API, forced tool-use, RAG pipelines, LLM evaluation harnesses, prompt-injection defense

ML and Modeling: PyTorch, Hugging Face Transformers, sentence-transformers, LoRA fine-tuning, Diffusers, PEFT, ControlNet, pgvector embeddings

Cloud, DevOps and Quality: AWS (S3, ECS), Microsoft Azure, Docker, Vercel, GitHub Actions CI/CD, Vitest, Jest, pytest, contract and property-based testing, observability and structured logging

Education

MS Software Engineering, National University of Sciences and Technology (NUST) 2024 – 2026 (expected)

BS Software Engineering, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) 2020 – 2024