

HASSAN ABBAS

AI Engineer — Generative AI · LLM Systems · RAG · Agentic Automation

Karachi, Pakistan hassanabbas2975@gmail.com github.com/hassan-abbas-arain linkedin.com/in/hassan-abbas-arain

AI Engineer building production-grade LLM, RAG, and agentic automation systems. 2+ years shipping scalable enterprise AI at Liquid Technologies and Teknoloje Solutions — from LLM orchestration and governed agent execution to voice automation, computer vision, and backend architecture.

SKILLS

- Languages:** TypeScript, Python, JavaScript, Rust, SQL
- Backend:** Node.js, NestJS, Express, FastAPI, Flask, TypeORM, Drizzle ORM, BullMQ, Socket.IO, SSE, REST/OpenAPI, Jest
- Frontend:** React, Next.js, Vite, Redux Toolkit / RTK Query, TanStack Query, Ant Design, shadcn/ui, TailwindCSS, Zod
- AI / LLM:** LLM tool calling and agent loops, human-in-the-loop approvals, Model Context Protocol (client and server), RAG (pgvector, hybrid search, reciprocal rank fusion, LLM re-ranking), embeddings, structured outputs, AI usage metering, Anthropic Claude, OpenAI, Google Gemini, vLLM, LangChain, Whisper, Parakeet, ONNX Runtime, YOLOv8, ArcFace, PaddleOCR, BioBERT
- Data & Security:** PostgreSQL (pgvector, full-text search, advisory locks, triggers), Redis, SQLite, FAISS, AWS S3, JWT, OAuth 2.0/2.1 (PKCE), HMAC-signed webhooks, AES-256-GCM, Argon2, SSRF defence, multi-tenant isolation, append-only audit logs
- Infrastructure:** Docker, Kubernetes, NGINX, GitHub Actions CI/CD, Tauri 2, Turborepo, pnpm, Playwright, n8n, Stripe, Slack Bolt

EXPERIENCE

AI Engineer · Liquid Technologies · Karachi, Pakistan April 2026 — Present

Sally — Multi-Tenant AI SaaS Platform

- NestJS 11, PostgreSQL + pgvector, Redis/BullMQ, Socket.IO, React 18 — principal contributor across backend and frontend.
- Designed governed agent execution: every agent action is risk-classified fail-closed, high-risk writes wait for owner approval and replay from the stored payload, and the audit log is append-only at the database, keeping every agent write accountable.
 - Built a bidirectional Model Context Protocol integration: an MCP server exposing 44 tenant-scoped tools to Claude Desktop behind an OAuth 2.1 authorization server, plus a multi-transport MCP client governed by the same gate.
 - Closed the tenant-isolation gaps the product is sold on: per-person feature switches on every route, a positions-based access model rolled out in shadow mode with zero regressions, and an audit that fixed cross-tenant reads.
 - Cut the cost of adding integrations and AI providers: a four-provider calendar connector platform with zero frontend code per provider, per-tenant encrypted OAuth apps, and bring-your-own-key LLM routing across OpenAI, Anthropic, Gemini and vLLM.
 - Shipped seven business domains end to end, from a PMO cockpit over Google Sheets, Odoo and Trello to a passwordless client portal, plus a meeting-intelligence pipeline turning Recall.ai and Meetstream.ai transcripts into Jira and Asana action items.

Sally Desktop

Offline-first meeting recorder for Windows, macOS and Linux (Tauri 2, Rust, Next.js) with on-device Whisper/Parakeet transcription on ONNX Runtime, GPU acceleration and meeting sync to Sally — meetings are transcribed without audio leaving the machine.

LeadGenPro

Lead-generation SaaS (NestJS, PostgreSQL, React) that turns keyword searches on LinkedIn, Reddit, Twitter/X and Hacker News into CRM-ready leads via a Playwright worker, a BullMQ LLM-qualification pipeline and GoHighLevel sync.

Upwork Analytics Platform

Solo-built read-only multi-account BI platform (Next.js, NestJS, SQLite/Drizzle) with incremental OAuth sync, AES-256-GCM token encryption and a CI-enforced read-only boundary that fails the build on any Upwork mutation.

AI Engineer · Teknoloje Solutions · Karachi, Pakistan · 18 months October 2024 — April 2026

- Built TekInsight — an enterprise AI platform with LLM-powered chat, FAISS + LangChain RAG document intelligence, and JWT multi-tenant authentication deployed on Flask/IIS, reducing retrieval time by 60% and query latency by 40%.
- Engineered TekConдор — a hybrid YOLOv8 + ResNet Siamese network signature verification pipeline at 94%+ accuracy with sub-300ms PaddleOCR cheque digitisation inference, eliminating manual review for 75% of submissions.
- Built a real-time multi-camera AI surveillance system using YOLOv8 + ArcFace embeddings with cosine similarity matching and polygon zone tracking, achieving 95%+ face recognition accuracy across distributed cameras.
- Designed an AI Voice Agent Platform using Twilio + Retell AI + n8n that monitors inbound emails, extracts leads via LLM parsing, and triggers AI outbound calls — cutting lead response time from hours to under 2 minutes.
- Built a LLaMA 3.2 + LangChain FastAPI medical summarisation service converting doctor-patient conversations to structured clinical JSON at sub-3s latency, reducing documentation effort by 70%.

SELECTED PROJECTS

- Enterprise AI Recruitment & ATS Platform** — Next.js, NestJS, PostgreSQL, pgvector, OpenAI, LangChain, Docker, Kubernetes · 2026
AI recruitment platform automating the hiring funnel: resume parsing, skill extraction, interview scorecards, semantic candidate matching via pgvector, and GDPR / EU AI Act compliance with adverse-impact monitoring.
- Enterprise Conversational AI Platform** — Next.js, NestJS, PostgreSQL, OpenAI, Google Gemini, Stripe · 2026
Conversational AI platform with scalable CRM-integrated agents, shared inbox, campaign broadcasting, WhatsApp Commerce, and Stripe billing on Redis/BullMQ background processing.
- CureBridge — AI Telehealth Platform** — BioBERT, XLNet, PyTorch, LLM, Socket.IO, JWT, Docker · FAST NUCES FYP · 2025
Telehealth platform with LLM diagnostic assistance, BioBERT medical NER, XLNet sentiment analysis, and real-time video consultation (87%+ disease

prediction accuracy).

EDUCATION

FAST NUCES — BS Computer Science

Final Year Project: CureBridge — AI Telehealth Platform (2025)

Karachi, Pakistan

CERTIFICATIONS

Supervised Machine Learning — Coursera · Andrew Ng · 2024–2025

Advanced Learning Algorithms — Coursera · Deep Learning, TensorFlow · 2024–2025

Data Cleaning & Processing — DataCamp · Python · 2024

Astera Bootcamp: Data Integration — 2023