

MIRZA MOHIBULLAH BAIG

Lead AI-Product Engineer · Full-Stack & AI/LLM Systems

Lahore, PK · +92 327 0765525 · mirzamohib572@gmail.com

[GitHub](#) · [LinkedIn](#) · [Portfolio](#)

SUMMARY

Full-stack engineer who ships production AI systems end to end — RAG pipelines, LLM orchestration, payments, and the infrastructure underneath them. Primary developer on multiple live AI products across FastAPI, Node.js, and React, with independently launched applications on the App Store and Google Play. Seeking mid-level full-stack or AI/LLM engineering roles — remote, or open to relocation.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript (ES6+), SQL

Backend: FastAPI, Node.js, Express.js, Celery, REST APIs, Server-Sent Events

Frontend: React.js, Next.js, Tailwind CSS (v4), HTML5 / CSS3

Databases: PostgreSQL, MongoDB (Atlas), Redis, Beanie ODM, SQLAlchemy, Weaviate (vector search)

AI & ML: RAG architecture, LangChain, LangGraph, vector search, Jina multimodal embeddings, OpenAI / Anthropic APIs, hybrid BM25 + vector retrieval, tiktoken token budgeting, PyTorch, TensorFlow / Keras

DevOps & Cloud: Docker, Docker Compose, GitHub Actions, CI/CD, AWS (EC2, S3, EBS), DigitalOcean, Nginx, Linux

Integrations: Stripe, RevenueCat, AWS S3, third-party LLM and OCR services

PROFESSIONAL EXPERIENCE

Cognara Training — *Lead AI-Product Engineer*

Lahore, PK · Jul 2025 – Present

- Lead team operations — code reviews, architecture direction, and coding standards — and work directly with the client stakeholder and designer to turn ambiguous product requirements into scoped, estimated tickets, owning the delivery board that gives non-technical stakeholders progress visibility.
- Primary developer on four production AI-powered platforms, owning architecture, data modeling, API design, deployment, and maintenance for each.
- Build retrieval-augmented generation systems end to end for dense policy and compliance documents — hi-res multimodal extraction, structure-aware chunking that groups elements into logical sections on heading boundaries rather than fixed-size windows to preserve clause integrity, multimodal embeddings via the Jina API, vector retrieval, and grounded generation.
- Cut time-to-first-token on RAG queries from roughly 30 seconds to 5–10 seconds by introducing hybrid retrieval — BM25 lexical filtering to narrow the candidate set ahead of vector similarity search.
- Orchestrate multi-step LLM workflows with LangChain and LangGraph, including multi-turn conversation state and tool-calling sequences.
- Engineer real-time streaming LLM chat over Server-Sent Events, delivering token-by-token responses to a React frontend alongside custom video-handling functionality.

- Design asynchronous processing with Celery and Redis for long-running document, audio-transcription, and OCR pipelines, with artifacts persisted to AWS S3.
- Implement Stripe subscription billing across multiple products — checkout, webhooks, plan management, and access control — and model both document and relational data layers (MongoDB/Beanie, PostgreSQL/SQLAlchemy).
- Own deployment infrastructure: multi-container Docker Compose stacks on AWS EC2 with dedicated EBS storage, automated GitHub Actions CI/CD, and separated staging and production environments.
- Design token-budgeted pipelines with tiktoken-precise accounting and staged caps at the embedding, retrieval-context, and conversation-history layers, bounding worst-case cost per request instead of letting it scale with document size.
- Integrate AI-assisted development tooling (Claude Code with custom review subagents and MCP integrations) into the team workflow to accelerate delivery without sacrificing code quality.

INDEPENDENT PROJECTS

Independent / freelance development — 2024 – Present

Pictofy AI — *AI Pet Portrait Application · pictofy.ai*

12/2025 – 4/2026

Live on the Apple App Store and Google Play

- Designed, built, and shipped a full-stack AI PWA end to end — generation pipeline, user accounts, mobile-first frontend — and took it through review to publication on both app stores.
- Implemented dual monetization with Stripe for web checkout and RevenueCat for in-app subscriptions across iOS and Android; deployed and maintained on DigitalOcean.

VECTA — *Voice-Enabled Crypto Trading Assistant*

08/2025 – 06/2026

- Built a real-time financial forecasting assistant using a Temporal Fusion Transformer with structural modifications, benchmarked against PatchTST and trained on historical Binance market data.
- Applied self-attention and quantile-loss objectives to improve forecast accuracy and quantify prediction uncertainty rather than emitting point estimates alone.
- Wrapped the model in a full product — a Next.js application with subscription billing and a companion browser extension — and authored a research paper documenting the architecture and evaluation methodology.

Impacting Life Productions — *Media Platform (Client Project)*

02/2025 – 04/2025

- Delivered a MERN-stack media platform with a custom admin panel for a UK-based film production company, and established the DigitalOcean deployment pipeline for secure production releases.

PhoneFixer POS — *Point-of-Sale System*

05/2024 – 06/2024

- Built and deployed a Next.js point-of-sale system in daily production use at a phone repair business, handling inventory management, repair ticketing, and sales reporting.

Additional Machine Learning Work

- Multi-task age and gender prediction model (hybrid CNN + ANN); image caption generator built on a MobileNetV2 encoder with a GRU decoder; lesion classifier, a CNN trained from scratch.

EDUCATION

Bachelor of Science in Computer Science — *Superior University, Lahore, PK*

2022 – 2026