

SHOAIB RASOOL

Full Stack AI Engineer

Islamabad / Wah Cantonment, Pakistan | github.com/shoaibrasool | shooaib6@gmail.com | +923119599700 | <https://www.linkedin.com/in/shoaib-rasool-193150221/>

PROFESSIONAL SUMMARY

Full Stack AI Engineer with hands-on industry experience spanning frontend, backend, and applied AI/LLM engineering since late 2022. Currently AI Lead on Sentrío, a production multi-agent security risk platform, architecting agentic workflows with LangGraph, FastAPI, and multi-model LLM orchestration. Proven track record building production RAG systems, agentic pipelines, and full-stack applications end to end, from ERD and system design through deployment. Strong foundation in React/Next.js, Node.js/FastAPI, and modern AI infrastructure including LangGraph, Qdrant, vector search, and RAG architecture.

WORK EXPERIENCE

Backend AI Engineer — AI Lead, Sentrío | Anemoia

Wah Cantonment, Pakistan (On-site) | January 2026 – Present

- Serve as AI Lead on Sentrío, a production multi-agent security risk platform; core Backend + AI engineer building agentic workflows with FastAPI and LangGraph multi-agent orchestration.
- Designed a FAIR-based Monte Carlo risk simulation engine for quantitative security risk analysis.
- Built a Gemini-based batch alert classification pipeline for high-volume security alert triage.
- Implemented prompt injection guardrails covering 14 distinct attack categories to secure agentic pipelines.
- Built a Qdrant-backed, RBAC-aware RAG system with semantic caching, model tiering, and token cost optimization.
- Designed a source-aware alert normalizer unifying Wazuh, GuardDuty, and Suricata alert formats into a single schema.
- Diagnosed and resolved Gemini API 429 quota exhaustion in production by implementing deduplication, priority queuing, and model fallback strategies.
- Integrated human-in-the-loop review mechanisms and MCP tools/skills into agentic workflows.
- Authored technical documentation and a client-facing cost/token usage guide for the platform.

Full Stack AI Engineer | Cofitech.ai

Islamabad, Pakistan | June 2025 – March 2026

- Led the Rebot project end to end, managing a team of two junior developers.
- Built Rebot, a no-code chatbot creation platform, entirely from scratch, including ERD design and overall system architecture.
- Built the full stack with React (frontend), FastAPI (backend), and MongoDB (database).
- Built native integrations with Slack, websites, and embeddable widgets.
- Integrated Stripe to handle payment methods and billing.
- Implemented webhooks to stream LLM responses to clients in real time.
- Applied CORS management and security guardrails to prevent data breaches.
- Optimized the RAG architecture for hybrid search and reranking, and added model evaluation to ensure retrieval accuracy.
- Reduced hallucinations through strict context grounding, retrieval reranking, and confidence-based fallback responses.

Full Stack Developer | Booma Enterprises

United Kingdom (Remote) | March 2024 – May 2025

- Built the full ecosystem for NAVIGO Taxi, a taxi-hailing platform, from admin tooling to rider/driver mobile apps.
- Built a Next.js admin dashboard integrated with Firebase, including server-side functions and authentication.
- Worked with a senior developer to migrate the backend from Firebase to Node.js.
- Built two React Native apps (driver and rider) with Google Maps integration, webhooks, real-time tracking, theming, and in-app notifications and messaging.

React Developer | Enigma Software Solutions

Islamabad, Pakistan (On-site) | July 2022 – February 2024

- Served as the core frontend React developer on an attendance management system.

- Built the entire frontend using React Hooks (including useReducer and useRef), component-based architecture, and state management.
- Supported other teams with frontend development tasks across concurrent projects.

TECHNICAL SKILLS

- Languages & Frontend: JavaScript, TypeScript, React, Next.js, Vue, Redux, Context API, Tailwind CSS, Material UI, Bootstrap
- Backend: Node.js, Express.js, FastAPI, Python
- AI / LLM Engineering: LangGraph, LangChain, RAG system design, prompt engineering, multi-agent orchestration, model evaluation, fine-tuning (OpenAI SFT API, LoRA/QLoRA concepts)
- LLM APIs & Models: OpenAI, Anthropic Claude, Google Gemini, Groq, Meta Llama, Ollama (local inference)
- Vector Databases & Search: Qdrant, Pinecone, FAISS, BM25, hybrid search, Reciprocal Rank Fusion, semantic caching
- Databases: MongoDB, PostgreSQL, MySQL, Supabase, SQLite, Firebase
- Mobile: React Native, Expo
- Infrastructure & DevOps: Docker, Docker Compose, CI/CD, AWS EC2, Azure, DigitalOcean, Vercel, Railway, Nginx, systemd
- Integrations & APIs: Stripe, Slack API/webhooks, Cal.com API, Notion API, Tavily, Google Maps API, WebSockets/SSE
- Other Tools: Pydantic, tiktoken, Git/GitHub, OpenCode, SQL performance tuning (indexing, stored procedures)

FEATURED PROJECTS

Lead Qualification Agent — Embeddable AI Sales Widget

- Embeddable, Gemini-powered AI chat widget that qualifies inbound website leads against an Ideal Customer Profile through natural conversation.
- Built a deterministic lead scoring engine (0–100, zero LLM calls) using a weighted rubric across 6 dimensions: budget, authority, need, timeline, company fit, and engagement.
- Score-based routing: automatically books Cal.com demos and sends Slack notifications for qualified leads, flags mid-tier leads for human follow-up, and declines low-fit leads.
- Stack: FastAPI, LangGraph, Gemini 2.5 Flash Lite, MongoDB checkpoint, Cal.com v2 API, Slack webhooks, and a vanilla JS Shadow DOM widget frontend.
- Deployed on Railway and AWS EC2 with Nginx and systemd.
- Production-hardened with retry logic (exponential backoff/jitter) on outbound Gemini calls, fast-fail MongoDB timeouts with graceful degradation, per-node timeouts inside the LangGraph flow, structured logging, and a 16-conversation synthetic seed test suite validating scoring and routing behavior.

Support Deflector — RAG-Powered SaaS Support Tool

- RAG-powered support ticket deflection tool backed by a Notion knowledge base of 150+ help articles.
- Combines dense and BM25 sparse retrieval via Reciprocal Rank Fusion (70% dense / 30% sparse).
- Started with hosted Gemini Embedding 2 (3072-dim), then migrated to a locally hosted BGE-large-en-v1.5 embedding model to remove API rate-limit bottlenecks during bulk re-ingestion.
- Built idempotent re-ingestion using deterministic MD5-based point IDs to prevent duplicate vectors on re-runs.
- Added request rate limiting, Qdrant health checks, and Pydantic-validated ingestion; fully containerized with Docker Compose.
- Stack: FastAPI, Qdrant, Notion API, BGE-large-en-v1.5, Gemini Embedding 2, Docker Compose.

LinkedIn Content Automation Pipeline

- Nine-phase agentic content automation pipeline built with LangGraph to power Auralistic AI's content strategy.
- Four separate LangGraph workflows: draft generation, idea generation, angle brainstorming, and weekly theme clustering.
- Tavily-based signal detection and Qdrant-based deduplication.
- LLM-as-judge scoring using Groq LLaMA 3.3 70B.
- Two-layer authenticity gate combining a 110+ banned corporate-phrase filter with adversarial LLM critique.
- Human review dashboard; pipeline does not auto-publish. Covered by 47 automated tests; uses Supabase for storage.
- Stack: LangGraph, Tavily, Qdrant, Groq (LLaMA 3.3 70B), Supabase.

Voice Companion Agent

- Real-time voice AI companion built with LangGraph and FastAPI, designed to converse, consult, calm, and motivate the user.
- Dual-purpose build: personal daily-use tool and client-facing portfolio piece.
- FastAPI core with a streaming WebSocket connection to Gemini via an async client.
- LangGraph agent with SQLite persistence (SQLAlchemy) and full token streaming through the graph back over FastAPI WebSockets.
- Voice I/O pipeline: Moonshine (Hugging Face Transformers) for speech-to-text and Kokoro (kokoro-onnx) for text-to-speech, with sentence/clause-aware chunking, ordered async audio delivery, and thread-pool-offloaded synthesis.
- Frontend built with React, Vite, and TypeScript, featuring a WebSocket streaming chat UI.
- Entire codebase written by hand without AI code generation, to deepen backend and FastAPI fundamentals.
- Repositories: github.com/shoaibrasool/voice-companion-agent, github.com/shoaibrasool/companion

Personal Portfolio Site

- Personal portfolio site built around an AI chat widget as the primary hero and interaction element.

EDUCATION

BS, Artificial Intelligence / *Air University, Islamabad, Pakistan*

September 2021 – June 2025

- Ran industry AI/LLM engineering work in parallel with degree studies from late 2022 onward.