NABEEL HUSSAIN

Senior AI Engineer | Agentic AI, Voice AI & LLM Systems

+92 308 7023514 | nabeel73jb@gmail.com | nabeelhussain.com | linkedin.com/in/nabeel-hussain1018 | github.com/ItsNabeelHussain

Lahore, Pakistan | Open to Relocation: Gulf / Europe | Full EU Overlap, 4-6h US Eastern

PROFESSIONAL SUMMARY

Senior AI Engineer with **6+ years** building production **Large Language Model (LLM)** systems and scalable **Python** backends. Specialise in **multi-agent systems**, real-time **voice agents** (**Pipecat**, **Deepgram**, **Cartesia**, **Silero VAD**), **agentic AI** orchestration with **LangGraph**, and **Retrieval-Augmented Generation (RAG)** with **hybrid search** and **reranking**. End-to-end ownership across multiple shipped products, from architecture through deployment and evaluation, including **on-premise AI for regulated banks**. Experienced leading engineering teams and delivering **GDPR**-compliant AI for regulated industries.

CORE COMPETENCIES

Agentic AI | Multi-Agent Systems | Voice AI Engineering | Large Language Models (LLM) | Retrieval-Augmented Generation (RAG) | LLM Evaluation and Grounding | Prompt Engineering | Vector Search | Multi-Tenant SaaS | Microservices Architecture | REST API Design | Cloud Infrastructure (AWS, Azure) | Data Pipelines | Technical Leadership

TECHNICAL SKILLS

Programming Languages: Python, JavaScript, C++, SQL

LLM & Agentic AI: LangGraph, **LangChain**, **multi-agent orchestration**, **tool calling / function calling**, structured outputs, **LLM-as-judge evaluation**, **citation grounding and hallucination correction**, **multi-provider model routing** (Claude, GPT-4, Gemini, Grok, Mistral, Qwen), **semantic caching**, **token cost optimization**, **prompt engineering**

Voice AI & Real-Time: Pipecat, Daily.co **WebRTC**, **Deepgram** (Speech-to-Text / STT), **Cartesia** and **ElevenLabs** (Text-to-Speech / TTS), **Silero VAD** (Voice Activity Detection), turn detection, barge-in handling, latency budgets, Wav2Vec2

RAG & Vector Databases: Retrieval-Augmented Generation (RAG), **hybrid retrieval (BM25 + dense)**, **reranking**, chunking strategy, **BGE-M3 embeddings**, **Qdrant**, **pgvector**, Elasticsearch, retrieval evaluation

Machine Learning & CV: PyTorch, **TensorFlow**, **OpenCV**, **YOLOv8**, **OCR**, **Hugging Face**, natural language processing (NLP), anomaly detection

Backend & APIs: FastAPI, **Django**, **Django REST Framework**, Flask, **Celery**, SQLAlchemy, Pydantic, asyncio, **REST APIs**, GraphQL, **WebSockets**, SSE, **microservices**, HMAC-signed gateways

Databases: PostgreSQL, **Supabase** (row-level security), MySQL, MongoDB, **Redis**, Elasticsearch

Cloud & DevOps: AWS (EC2, S3, RDS, Lambda, CloudFront), **Azure** (Container Apps, Static Web Apps), **Docker**, Nginx, **CI/CD**, **GitHub Actions**, Git, Linux

Data & Payments: Apache Airflow, Scrapy, Playwright, ETL pipelines, **Stripe**, **Stripe Connect**, **double-entry ledger** design

Security & Compliance: Presidio PII redaction, GLiNER, **GDPR**, **on-premise deployment**, encryption at rest, **tamper-evident audit logging**, role-based access control (RBAC), **automated testing**

PROFESSIONAL EXPERIENCE

Senior AI Engineer **Aug 2023 - Present**

Conovo Technologies *Remote*

- Designed the retrieval layer behind production **RAG** systems using **hybrid BM25 and dense search** over **BGE-M3 embeddings** in **Qdrant** with **reranking**, making query responses **10× faster**.
- Built **multi-provider LLM routing** across **Claude**, **GPT-4** and **Gemini** with **semantic caching**, streaming and per-call cost logging, cutting LLM API costs by **35%** while improving reliability.
- Established evaluation rigour with **LLM-as-judge validation gates in CI** covering response quality, hallucination rate and retrieval precision.
- Engineered **GDPR**-compliant AI pipelines for regulated clients, including **Presidio** and GLiNER **PII redaction**, consent handling, **audit logging** and data-residency controls.
- Built an intelligent document assistant on **LangChain** and **vector search**, making document search **2× faster**.
- Orchestrated **20+ daily Apache Airflow pipelines** across **250+ production data sources**, synchronising terabytes of structured and unstructured data from **AWS S3** and the web and processing **1M+ records per run**.
- Led a distributed engineering team, running weekly **code reviews** and maintaining **CI/CD** pipelines with **Docker** and **GitHub Actions**.

Python Developer **Oct 2019 - Aug 2023**
ArhamSoft Technologies *Lahore, Pakistan*

- Led design and rollout of **backend platforms** for **finance, healthcare and e-commerce** clients, scaling to tens of thousands of active users.
- Improved application responsiveness by up to **60%** through **query optimization**, refactoring and in-memory **caching**.
- Built document-processing pipelines combining **OCR** and **NLP**, converting scanned legal and financial documents into structured data and cutting manual work by **70%**.
- Managed **AWS** infrastructure (EC2, RDS, S3) with **auto-scaling** and monitoring, sustaining **99.9% uptime** while cutting operational costs by **20%**.
- Integrated **Stripe** billing, Slack and Jira into enterprise client platforms, and mentored junior engineers through **code reviews** and pair programming.

KEY PROJECTS

PersonifAI - Procedure and Compliance Analysis (Financial Services) | Multi-agent LLM, on-premise | askpersonifai.com

- Built a multi-agent platform in which **10 specialist agents** audit bank operating procedures for compliance, controls, risk and automation readiness, producing audit-ready reports and process maps.
- Designed an **anti-hallucination layer** that verifies every regulatory citation live against **11 UK and US regulator sources**; provider-agnostic across **4 LLM providers**, shipped **on-premise** for regulated banks with **1,300+ automated tests**.

Global Agentic Market Exchange (G.A.M.E.) | Django, PostgreSQL, Celery, Redis, Stripe Connect, AWS | agenticmarket.exchange

- Built a marketplace where AI agents hire other agents through an **HMAC-signed gateway**, with per-call pricing, a **double-entry ledger** with hold-and-release settlement and **Stripe Connect** payouts.
- Built entry and **anti-fraud** systems for a **$10M prize challenge**: batch registration of **10,000 agents per request**, liveness and uniqueness checks, near-duplicate detection, designed to scale to **1M+ agents**.

ProofDial - Independent Voice Agent Audits | Pipecat, Deepgram, Cartesia, Silero VAD, Claude, Presidio, Supabase | proofdial.com

- Audits the calls a **voice agent** has taken and places automated test calls against production agents, scoring each with **Claude as LLM judge**, with **Presidio** PII redaction and a **multi-tenant** backend; maps 13 failure types to 38 named controls.

SharpSource ATS - AI-Native Applicant Tracking | sharpsource.global

- An applicant tracking system run by AI: it writes the job description, screens and ranks applicants with a **weighted fit score**, schedules interviews, drafts the offer and forecasts acceptance; every AI call can be overridden and lands in an **audit log**.

Webcurve - Counterfeit Seller Intelligence | Qwen2.5, Claude 3.5, Apache Airflow, Scrapy, PostgreSQL, AWS

- Led a team of three building a **fraud detection** platform that tracks sellers of counterfeit products, with **Airflow**-scheduled scraper DAGs and a custom-trained **LLM** automating scraper maintenance; query optimisation gave a **30%** performance gain.

LiquidDocs - Expert-Verified AI Due Diligence | liquiddocs.ai

- **AI agents** extract, structure and score every document in a data room, and a named expert verifies each finding into a source-linked, audit-trailed **Golden Record**.

21Swap - AI Face Transformation | Python, TensorFlow, OpenCV, Flask, AWS GPU

- Led development of a deep-learning face transformation system for image and video, with **GPU acceleration** cutting inference time by **40%**.

AssetCurve - Enterprise Asset Management | Django, DRF, PostgreSQL, AWS, Nginx, Stripe

- Asset and IT hardware management platform tracking equipment costs and recertifications, built on **Django REST Framework** with **Stripe** payments on AWS.

Balie - AI Receptionist | usebalie.nl

- Answers customers **24/7** on the website chat and **WhatsApp** in Dutch and English, books appointments into Google Calendar or Outlook, captures leads and hands the conversation to a person when needed.

NX Innoverse - Healthcare Innovation Platform | nx-innoverse.com

- A platform built around nurses: nurses pitch their innovations privately, investors back them, and employers post jobs and hire, with **role-based onboarding** for each.

EDUCATION & CERTIFICATIONS

Bachelor of Software Engineering | University of Agriculture, Faisalabad	2017 - 2021
Microsoft Certified: Azure AI Apps and Agents Developer Associate	Oct 2026
Microsoft Certified: Azure Fundamentals	Jun 2026
AWS Certified Cloud Practitioner	Jul 2024
Udemy: Machine Learning A-Z: AI, Python & R + ChatGPT Prize	Dec 2025