IRFAN ALI

Senior AI Engineer · LLMs, RAG, Agents, Voice AI · 7+ years experience
+91 99716 27567 · irfan.ali@datacortex.in · linkedin.com/in/irfanalidv · github.com/irfanalidv

Senior AI engineer with **7+ years** building production LLM systems: RAG pipelines, multi-agent orchestration, multi-provider LLM infrastructure, and voice AI on telephony. **Founding AI engineer** at Kuration AI (Hong Kong) and **sole AI hire reporting to the CTO** at Schneider Electric. Maintains **11 open-source Python libraries** on PyPI and **2 peer-reviewed IJAINN publications**. Open to full-time senior AI engineer roles: remote, India-based, global hours.

Experience

Senior AI Engineer | Applied LLM & Voice AI · Open-Source — Dec 2025 – June 2026 · Remote

- Built **Reflecta** (getreflecta.com), a voice-first AI wellness app: phone-based check-ins, post-call LLM analysis, and personalised recommendations. Stack: FastAPI, Bolna telephony, Groq, Neon Postgres + pgvector, JWT/OTP auth with dual-layer rate limiting, async webhook processing, multi-provider LLM fallback. Solo build, end-to-end.
- Built **Stacksift** (stacksift.in), a B2B domain product analyzer running a **5-stage LLM classification pipeline** (DSPy + GPT-4.1) over crawled web data, with structured extraction, deduplication, and verdict scoring at **~$0.03 per analysis (~84% margin)**. FastAPI backend, beta-key auth, IP rate limiting.
- Author and maintainer of **11 open-source Python libraries** on PyPI, including:
  - **RAGNav**: RAG routing & retrieval; **R@3 = 0.956 on SQuAD** via hybrid BM25 + dense retrieval with RRF fusion, **72% test coverage / 131 tests**.
  - **ragfallback**: resilient retrieval with query rewriting and confidence scoring; **AgentEnsemble**: multi-agent orchestration (ReAct, Swarm, Pipeline, Debate, WorkflowGraph); **scrapeflow-py**: Playwright-based structured web extraction.

Founding AI Engineer | Kuration AI — Jun 2024 – Nov 2025 · Hong Kong SAR · Remote

- **First AI hire** at an early-stage B2B startup. Designed and owned the **full AI architecture**: agent pipelines, RAG, structured extraction, multi-provider LLM orchestration, and supporting data pipelines.
- Built the LLM orchestration layer across **3 providers (GPT-4o, Claude, Gemini)** with automatic fallback, prompt management, and cost controls.
- Built the data enrichment system: agent pipelines pulling from **multiple contact / company / funding / people APIs**, with fallback when any single provider degraded.
- Shipped the surrounding ETL workflows: merging, deduplication, validation, and per-call cost tracking.

Senior Manager – Data & Analytics, R&D | Luminous Power Technologies (Schneider Electric) — Aug 2023 – Jun 2024 · India

- **Sole AI hire** reporting directly to the CTO. Built the R&D data and AI function from scratch, then **hired and led a team of 3** (data scientist, data engineer, data analyst).
- Shipped an **LLM-powered dealer assistant** that helped the sales network recommend products and run load and battery sizing calculations, cutting time-to-quote and standardising recommendations across regions.
- Built an internal R&D dashboard for geospatial data, real-time power-outage tracking across Indian states, and live solar monitoring; owned the underlying Azure data platform.

Data Analytics & Automation Associate | Lynk — Feb 2022 – Jul 2023 · India · Remote

- Built analytics pipelines and **NLP-powered search** for an expert-matchmaking platform, cutting expert-discovery time for research and investment teams.

Head of Data & Analytics | brainsfeed Ltd. — Aug 2018 – Feb 2022 · Hong Kong SAR · Remote

- First data hire at an early-stage research marketplace. Built the data function from scratch and grew a **distributed team of 10+** across multiple time zones. Shipped **Infosphere**, an internal intelligence platform with NLP enrichment and natural-language search that improved research discovery and supported new-client wins.

Skills

Core AI: LLMs, Retrieval-Augmented Generation (RAG), AI Agents, Multi-agent Systems, LLM Orchestration, Prompt Engineering, Voice AI, Fine-tuning, Evaluation
Frameworks: LangChain, LangGraph, DSPy, LlamaIndex, FastAPI, Pydantic, Playwright, Pandas, NumPy
LLM Providers: OpenAI (GPT-4, GPT-4o), Anthropic (Claude), Google (Gemini), Meta (Llama), Mistral, Groq, Ollama
Data & Backend: Python, PostgreSQL, MongoDB, pgvector, Pinecone, ChromaDB, REST APIs, WebSockets, Async Python, Docker
Voice & Infra: Bolna, Twilio, ASR output handling, async webhooks, JWT/OTP auth, rate limiting, multi-provider fallback patterns
Cloud & MLOps: AWS, Azure, GCP, Vercel, Supabase, Neon, GitHub Actions, CI/CD

Education

M.Sc. Data Science & AI | IISER Tirupati | 2025–2026 | **CGPA 8.9/10** | *Institute of National Importance, Govt. of India*
B.Tech, CSE | Alliance University, Bengaluru | 2013–2017 | *Exchange semester at ISEP, Paris*

Publications

Mental Health AI on MentalChat16K – IJAINN Dec 2025 [DOI] · Neural-Symbolic Topic Evolution – IJAINN Oct 2025 [DOI]