

MUHAMMAD UMER KHAN

AI Engineer | Agentic AI & RAG | Voice AI & Workflow Automation

Karachi, Pakistan · Open to Remote · +92 343 2187868 · muhammadumerk546@gmail.com

linkedin.com/in/muhammadumerkhan-ai · github.com/MuhammadUmerKhan · buildwithumer.site

SUMMARY

AI Engineer, 2+ years building production LLM systems: agentic RAG assistants, multi-agent backends, voice and WhatsApp agents, CRM automation. Delivered 12+ projects end to end. The work spans fine-tuning to evaluation-gated CI deploys, including white-label SaaS live in production, and every system is red-teamed against OWASP LLM Top 10 with findings published.

TECHNICAL SKILLS

Agentic AI: AI Agents, Multi-Agent Systems, Prompt Engineering, Tool Calling, LangGraph, LangChain, CrewAI

Fine-Tuning & Evaluation: LoRA/PEFT, Quantization, LLM-as-Judge, Faithfulness/Relevancy Metrics, DeepEval, RAGAS

RAG & Vector Search: Hybrid Search (BM25 + dense), Rank Fusion, Cross-Encoder Reranking, Qdrant, FAISS

LLM Safety: Guardrails, Prompt-Injection Defense, Red-Teaming, Threat Modeling, OWASP LLM Top 10, MITRE ATLAS

MLOps & Deployment: FastAPI, REST APIs, Docker, GitHub Actions CI/CD, Distributed Tracing, MLflow, LangSmith

ML & Data Science: Python, PyTorch, TensorFlow, Scikit-Learn, NLP, Computer Vision, Pandas, NumPy

Voice & Automation: Inbound/Outbound Voice Agents, Speech-to-Text, WhatsApp Automation, Vapi, Retell AI, n8n, Make

Databases & Cloud: PostgreSQL, MongoDB, SQL, Supabase, AWS, Azure, Git, React, Next.js, Expo (React Native)

Languages: English (Professional), Urdu (Native)

EXPERIENCE

AI Engineer — Revera Innovations

Karachi (Hybrid) | Feb 2025 – Present

- Deployed two vision products to production: a virtual try-on system, and an autonomous skin-analysis agent that detects skin concerns in a user's photo and reasons over live inventory to recommend products.
- Built inbound and outbound AI voice agents on Vapi and Retell AI. One is a dental-clinic receptionist: it books open slots through Cal.com, creates the lead and deal in HubSpot, and alerts the team.
- Building an AI WhatsApp lead-automation system that turns inquiries into database leads and relays the team's replies back to customers. Also automated CRM operations with n8n and Make across HubSpot, GoHighLevel, Airtable and Jotform.
- Shipped two white-label society-management SaaS platforms on React admin panels, Expo mobile apps and Supabase backends, with offline HMAC-signed QR gate security and multi-tenant isolation. One is live in production.

AI Developer — Bright Solutions

Karachi (Hybrid) | Dec 2024 – Feb 2025

- Launched AI voice agents and automated lead-generation systems, now part of daily production operations.

AI Engineer Intern — SaylaniTech

Karachi (On-site) | Jun 2024 – Dec 2024

- Developed LLM and NLP features and shipped them into applications behind production-grade APIs.

Data Science Intern — Digital Empowerment Network

Remote | May 2024 – Jun 2024

- Built predictive models on social-impact datasets, plus the data-cleaning pipelines and reports behind them, in Python.

PROJECTS

AskUmer — Hybrid-RAG Conversational Assistant

[GitHub](#) | 2026

- A LangGraph ReAct agent over hybrid retrieval: Qdrant dense and BM25 sparse search, rank fusion and FlashRank reranking across 898 indexed chunks. An in-memory knowledge graph handles multi-hop queries.
- Hardened with NeMo Guardrails input gating and Portkey failover (gpt-oss-120b → gpt-oss-20b, retries, caching). Deploys are gated by a DeepEval LLM-as-judge CI suite at ≥ 0.7 faithfulness and answer relevancy.

DineMate — Agentic AI Food-Ordering System

[Live Demo](#) · [Security Report](#) | 2025

- Built a four-agent LangGraph backend (ordering, order management, kitchen, admin) with voice ordering via Whisper ASR and TTS. Role-gated dashboards track revenue, peak hours and top items.
- Red-teamed it across 8 attack categories mapped to OWASP LLM Top 10 and MITRE ATLAS. A critical price-manipulation flaw was fixed with server-side recomputation, and a Llama Prompt Guard 2 classifier now screens input.

SupportGenie — Dual Fine-Tuned Banking Support Chatbot

[GitHub](#) | 2025

- LoRA fine-tuned Mistral-7B (17,000 instruction pairs, $r=32$) and LLaMA-3-8B (1,764 pairs) in 4-bit quantization on one free-tier T4. Final loss: 1.78 / 1.49. FAISS FAQ retrieval answers first, falling through to the models.

Diagnosisfy — LLM-Powered Medical Report Insights

[Live Demo](#) | 2025

- Top-10 finalist at the SMIT AI Hackathon, built in 12 hours: OCR + LLM lab-report analysis with a RAGAS-scored RAG chatbot.

EDUCATION

B.S. in Computer Science — Sindh Madressatul Islam University

Karachi | 2023 – 2027 (Expected)

CERTIFICATIONS

- DeepLearning.AI — Deep Learning + NLP (2025) · IBM — ML Professional Certificate (2024) · SMIT — AI & Data Science (2025)