UMAR JAVED

Sialkot, Punjab, Pakistan | +92 309 4780810 | ranaumarranaumar705@gmail.com
linkedin.com/in/ranaumar4880 | github.com/Umar4880 | ranaumar.tech

AI engineer with over a year of experience (nine months at GulzarSoft + freelancing) building RAG systems and multi-agent backends in Python. Strong in hybrid retrieval, tool-calling agents (LangGraph, MCP) and async backends with FastAPI, PostgreSQL and Qdrant. Looking for an AI engineer role where I can build reliable LLM features for real users.

EXPERIENCE

Junior AI Engineer, *GulzarSoft, Gujrat*

- Built, as the only developer, a multi-tenant platform to let businesses handle customer queries and orders on WhatsApp, Messenger, Instagram and web chat, using FastAPI, PostgreSQL and Qdrant.
- Develop a supervisor agent to route messages to a RAG pipeline or order database, using intent classification.
- Design a lazy-loading tool gateway to keep full tool schemas out of the prompt until a tool is called, cutting the fixed prompt from 1,060 to 580 tokens (45%).
- Added tenant isolation to stop one business seeing another's data, using RS256-signed JWTs, Qdrant metadata filters and tenant-scoped PostgreSQL queries.

AI/ML Intern, *GulzarSoft, Gujrat*

- Built a Q&A system over 30+ company documents to give cited answers, using RAG and Docling chunking.
- Added crash-safe chunk caching to avoid re-processing whole documents, using SHA-256 hashes so only changed chunks are re-embedded.
- Prototyped a LangGraph supervisor agent for intent routing, and a voice agent using LiveKit and Gemini.
- Built MCP servers (database, calculator, weather) and connected them to Claude.

SKILLS

Languages:	Python, SQL
AI / LLM:	LangGraph, LangChain, Tool Calling, MCP, RAG, Ragas, DeepEval
Retrieval/Data:	Qdrant, FAISS, Redis, PostgreSQL, BM25, Reranking
Backend:	FastAPI, asyncio, Pydantic, SQLAlchemy, Celery, Docker
Models/APIs:	OpenAI, Anthropic, Gemini, DeepSeek, Ollama

PROJECTS

Growmina *(LangChain, LangGraph, FastAPI, Celery, Redis, Qdrant, Docling, PostgreSQL, Docker, MinIO)*

- Built an agent to handle customer's scattered quries from omnichannel using LangGraph and LangChain.
- Design ingestion pipeline to survive crash and duplicate embeddings using hashing and caching, and reduce hallucination by grounding context via hybrid retrieval (sparse + dense) in Qdrant (under 640 ms).
- Reduce token costs up to 45% using prompt KV-caching and implementing tools gateway.

Bravous *(LangGraph, LangChain, FastAPI, PostgreSQL, MinIO, Redis Next.js)*

- Clone Claude and ChatGPT Projects-style workspace with projects, custom instructions, memory and file uploads, using FastAPI, Next.js.
- Built a deep-research agent to answer multi-search questions, using LangGraph agents that search, extract, find gaps and re-search, with PostgreSQL checkpoints and Tavily.
- Stream agent traces, thinking and steps to showcase live working of agent.

Comply-Analytics *(LangChain, Docling, Qdrant, FastAPI, DeepEval, Next.js)*

- Built a RAG app to answer questions on over 30+ compliance and financial documents with citations, using Docling, hybrid retrieval with reranking, and Ragas for faithfulness, precision and recall.
- Added SHA-256 incremental sync to skip unchanged files, cutting re-indexing compute by about 90%.
- Implement chunk caching to prevent re-embedding by storing temporary values in Redis with (24hr TTL).

async-social-bridge *(Python, Pydantic, PyPI v0.1.1)*

- Published a library to unify WhatsApp, Messenger and Instagram webhooks, using one typed async interface.

EDUCATION

University of Gujrat, BS Computer Science *(2021 – 2025)* CGPA: 3.45/4.00