

UMAR JAVED

Sialkot, Punjab | +92 309 4780810 | ranaumarranaumar705@gmail.com

[LinkedIn](#) | [Portfolio](#) | [GitHub](#)

PROFESSIONAL SUMMARY

Applied AI Engineer specializing in agentic AI, retrieval-augmented generation, and production-grade LLM backends with hands-on experience designing multi-tenant, multi-agent systems with LangChain, LangGraph, and FastAPI including tool-orchestration architectures, RAG pipelines, vector databases, and real-time voice agents.

PROFESSIONAL EXPERIENCE

Junior AI Engineer - GulzarSoft

(May 2026 – Present)

- Sole developer architecting a multi-tenant B2B SaaS platform enabling businesses to handle customer queries and orders through an AI agent embedded in WhatsApp, Messenger, and Instagram.
- Designed a hierarchical multi-agent system with a head orchestrator that classifies user intent and routes to specialized tools (RAG retrieval, live product/order database) or resolves the query directly.
- Designed a custom lazy-loading tool registry that exposes lightweight tool manifests and loads full schemas only on demand, reducing the always-present tool-definition prompt footprint by 45% (1,060 to 580 tokens), independent of the number of registered tools.
- Implemented multi-tenant data isolation via metadata filtering in Qdrant and tenant id scoping in a PostgreSQL database.

AI/ML Intern - GulzarSoft

(March 2026 - May 2026)

- Built a RAG-based Q&A system with citation tracking over 10 real company documents using heading-based chunking with incremental reingestion so document updates reprocess only changed sections not full document.
- Implemented crash-safe chunk-level caching, reducing recovery reprocess from 100% to only failed chunks.
- Independently prototyped a multi-hierarchical agent system with a supervisor routing queries to sub-agents by intent.
- Prototyped a real-time voice agent using LiveKit and Gemini real-time models for voice-driven task execution.
- Built and connected MCP (Model Context Protocol) servers (database, calculator, weather) to Claude, exploring tool-use integration outside the LangChain ecosystem.

TECHNICAL SKILLS

- **AI/LLM:** Agentic AI, RAG, Tool Calling, MCP, Context Engineering, LLM Evaluation
- **Frameworks:** LangGraph, LangChain, FastAPI, SQLAlchemy
- **Retrieval/Data:** Qdrant, FAISS, Redis, PostgreSQL, BM25, Reranking
- **Models/APIs:** OpenAI, Anthropic, Gemini, DeepSeek, Open-source LLMs
- **Infrastructure:** Docker, Celery, GitHub, Ollama

SELECTED PROJECTS

Bravous Multi-Agent System (*LangGraph, PostgreSQL, SQLAlchemy, Redis, FastAPI, Next.js*)

- Built a *ChatGPT/Claude* Projects-style workspace with scoped projects, custom instructions, and file uploads.
- Developed a deep-research agent tool that decomposes queries, iteratively explores via web search, critiques findings for gaps, re-loops if needed, then synthesizes an answer.
- Architected a supervisor agent that triggers the research tool on demand, using LangGraph state management and PostgreSQL checkpointing for long-running tasks.

AI-Powered Learning Assistant (*DeepSeek, Pegasus, Whisper, FAISS, Django*)

- Built an educational assistant that transcribes, summarizes, and indexes video content into concise study notes with recommendation-oriented retrieval.

Comply Analytics RAG System (*LangChain, Qdrant, FastAPI, Next.js, Redis*)

- Built a full-stack local RAG application for document Q&A with upload, chat sessions, and citation tracking, using hybrid retrieval (embeddings + BM25 + reranking) evaluated on faithfulness, precision, and recall.

AI Chat Assistant - RAG Chrome Extension (*LangChain, Redis, FastAPI, Gemini*)

- Built a Chrome extension replicating Gemini's in-browser side-panel feature, a split-view panel that lets users ask questions directly about the currently open tab's content, without switching to a separate chat window.

EDUCATION

University of Gujrat - Bachelor of Science in Computer Science (2021–2025)

CGPA: 3.45/4.00