MUHAMMAD HAZIQ IMRAN

AI SOFTWARE ENGINEER

Rahim Yar Khan, Pakistan | +92 325 1234666 | hazziq595959@gmail.com
linkedin.com/in/haziqimran59 | github.com/HaziqImran18 | haziqimran18.github.io/MY-PORTFOLIO

PROFESSIONAL SUMMARY

AI Software Engineer specializing in the development of LLM-powered applications, agentic AI systems, multi-agent architectures, and retrieval-augmented generation (RAG) pipelines. Skilled in AI agent orchestration, context and memory management, and reliability engineering, supported by a foundation in machine learning, deep learning, and model evaluation. Proficient in Python, backend development, and REST API design, with expertise in automated testing, evidence grounding, structured-output validation, and prompt-injection defense to improve AI system reliability, accuracy, latency, and cost efficiency.

TECHNICAL SKILLS

Languages	Python, SQL
LLMs & Agents	LangGraph, LangChain, Crew, MCP, OpenAI API, Groq, HuggingFace Transformers, Multi-Agent Systems, Agentic Workflows, Tool/Function Calling, Prompt Engineering, Structured Outputs
RAG & Vector Search	RAG Pipelines, Text Embeddings, Semantic Search, ChromaDB, Pinecone, sqlite-vec, FAISS, Document Chunking, PyMuPDF
Evaluation & Testing	Benchmark Design, Baseline and Ablation Studies, Claim Verification, Citation Validation, Prompt-Injection Defense, Pydantic, pytest, Mocked LLM Testing
Machine Learning	PyTorch, Scikit-learn, BERT Fine-tuning, Text Classification, Model Evaluation, Pandas, NumPy
Backend & Tools	FastAPI, Flask, Streamlit, SQLite, Firebase, MongoDB, PostgreSQL, REST APIs, Docker, Git, Vercal

PROJECTS

JobPilot: MCP-Powered Job Search Agent

Python, FastMCP, OpenAI API, SQLite, sqlite-vec, Pydantic, Streamlit, Docker

- Built a multi-tool AI agent with six custom MCP servers for job discovery, candidate profiling, memory, job matching, CV tailoring, and application tracking.
- Designed an agent harness with tiered tool permissions, human approval for external actions, schema validation, prompt-injection defenses for scraped content, and an LLM gateway with response caching and cost limits.
- Implemented token-budgeted context assembly, rolling-summary compaction, and semantic and episodic memory with confidence scoring, deduplication, and user-controlled deletion.
- Developed a claim-verification layer that checks generated CV bullets against candidate profile facts to reduce unsupported claims.
- Designed an evaluation framework comparing LLM-based job matching against a TF-IDF baseline using human-labeled job descriptions and ablation studies.

AutoDev-Triage: Evidence-Grounded Multi-Agent GitHub Issue Triage

Python, LangGraph, OpenAI API, Pydantic v2, Flask, pytest

- Built a multi-agent system of Diagnoser, Researcher, Verifier, and Integrator agents that analyzes GitHub issues, researches potential fixes, and generates root-cause analyses.
- Implemented deterministic abstention, independent claim verification, and code-enforced URL validation to flag unsupported conclusions.
- Achieved 92% root-cause accuracy on a 12-issue benchmark, with correct abstention on both no-evidence cases and zero ungrounded URLs.

- Developed an evaluation harness comparing the multi-agent workflow against a single-LLM baseline on accuracy, latency, and API cost.
- Wrote 55 offline tests using mocked LLM and network calls, and built a Flask interface that processes GitHub tokens per request without storing them.

EngiGraph: Multi-Agent RAG Assistant for Equipment Manuals

Python, LangGraph, LangChain, Groq API, Google Embeddings, ChromaDB, PyMuPDF, Pydantic, Flask, pytest

- Built a LangGraph workflow with investigator, planner, and verifier agents that generates equipment diagnostic reports grounded in PDF manuals, with page-level citations and a verification-revision loop.
- Implemented a RAG pipeline with page-aware PDF extraction, overlapping chunking, Google embeddings, persistent ChromaDB search, and resumable batch indexing.
- Added structured-output validation, retry handling for malformed JSON, citation validation against retrieved pages, and deterministic safety and uncertainty notes.
- Developed a Flask app with PDF upload, background processing, live pipeline progress, per-claim citations, and Markdown report download.
- Designed a 57-question manually verified benchmark (direct, multi-passage, and unanswerable questions) to compare accuracy, citation validity, latency, and token cost against a single-call RAG baseline.
- Wrote 30+ automated tests with mocked LLM and retrieval components for offline testing at zero API cost.

FINAL YEAR PROJECT

ShopGen AI: WhatsApp-Integrated Conversational Shopping Agent 2025 – 2026

Python, LangGraph, LangChain, OpenAI GPT-4o Mini, Whisper, Whapi, Pinecone, Firebase, FastAPI, SerpAPI

- Architected a multi-agent system on WhatsApp using LangGraph workflows for autonomous product discovery, real-time price comparison, personalized fashion recommendations, and conversational commerce.
- Implemented stateful agent memory with persistent session context via LangChain and LangGraph, enabling coherent multi-turn conversations and long-term preference tracking across 100+ interactions.
- Integrated Pinecone for semantic product search via dense embeddings, and Firebase Realtime Database for inventory and session management supporting concurrent users.
- Built a multimodal text and voice pipeline (OpenAI Whisper STT / TTS) over WhatsApp via the Whapi REST API, with query response latency under 3 seconds.

WORK EXPERIENCE

AI/ML Engineer Intern, Ezitech Institute Jun 2025 – Sep 2025

- Fine-tuned a BERT-based fake news detection model with HuggingFace Transformers, reaching ~93% accuracy on 5,000+ labeled news articles.
- Built a phishing website detection classifier with a Scikit-learn pipeline, feature extraction, and evaluation; attained ~91% accuracy, 0.90 F1-score, and 0.95 ROC-AUC.
- Constructed an end-to-end Comic Story Generator combining Groq LLM narrative generation with Coqui TTS voice narration, reducing manual content creation time by ~70%.

EDUCATION

BS Artificial Intelligence, Khwaja Fareed University of Engineering & IT, Rahim Yar Khan 2022 – 2026

Coursework: Machine Learning, Deep Learning, NLP, Computer Vision, Reinforcement Learning, Database Systems, Programming for Artificial Intelligence

CERTIFICATIONS

- Career Essentials in Generative AI, Microsoft & LinkedIn Learning
- AI Advanced Course, Xeven Solutions