ZULQARNAIN KHAN

AI Engineer — Agentic AI — RAG Systems — Cloud & MLOps

Email: zulqar4791@gmail.com — Phone: +92 3177896472

Address: Islamabad, Pakistan

LinkedIn: https://www.linkedin.com/in/zulqarnain-khan-b4a4aa212/ — GitHub: github.com/ENGRZULQARNAIN — HuggingFace: huggingface.co/engrzulqarnain

PROFESSIONAL SUMMARY

AI Engineer with 3+ years building production AI systems across fintech, edtech, and enterprise platforms. My core focus has been RAG pipelines, multi-agent systems, and AI orchestration using LangChain, LangGraph, and Model Context Protocol (MCP). At Developers Den, I built multi-modal RAG infrastructure for the US mortgage sector, processing 10,000+ financial PDFs from US lending institutions with 90+ percent retrieval accuracy. On the infrastructure side, I have shipped AI workloads on GCP, and AWS using Docker and Kubernetes, running zero-downtime rolling deployments through CI/CD pipelines. I work in Agile teams, take part in sprint planning, and collaborate directly with Data Scientists and Product Owners to get things into production.

TECHNICAL SKILLS

Programming Languages: Python (Expert), SQL (Advanced – Stored Procedures, Window Functions, Recursive Queries, Temp Tables), C++

Agentic AI & Orchestration: LangChain, LangGraph, Model Context Protocol (MCP), Multi-Agent Systems, Autonomous Workflows, ReAct, Tool Calling, Function Calling, Chain-of-Thought (CoT), LLM-as-a-Judge

LLM Fine-Tuning & Alignment: Open-source LLMs (LLaMA-3, Mistral, Gemma, Phi-3, Qwen), Hugging Face Transformers, PEFT (LoRA, QLoRA), RLHF, TRL, Unsloth, PyTorch, Model Distillation

RAG & Knowledge Systems: LangChain, LangGraph, DSPy, LiteLLM, Vector Databases (Pinecone, ChromaDB, FAISS, Weaviate), Multi-modal RAG, GraphRAG, Document Processing, Metadata Extraction, Semantic Search

Backend & API Development: FastAPI (REST + WebSockets), Microservices Architecture, Async Programming (asyncio), Celery, Redis, Rate Limiting, JWT Authentication

Cloud & Infrastructure: GCP (Vertex AI, GCS, GKE), AWS (EC2, S3, Lambda, ECR, ECS)

Containerization & Orchestration: Docker, Kubernetes (K8s), Zero-Downtime Rolling Deployments, Helm Charts, Container Registries

CI/CD & DevOps: GitHub Actions, CI/CD Pipelines, Git (GitHub/GitLab), Linux, Shell Scripting

Model Deployment & MLOps: vLLM, Ollama, Hugging Face Inference Endpoints, GPTQ/AWQ Quantization, MLflow, Model Versioning, TFX

Evaluation & Monitoring: Langfuse, LangSmith, LM-Eval-Harness, Accuracy/Latency/Cost Tracking

Deep Learning Frameworks: TensorFlow, PyTorch, RNNs, LSTMs, GRUs

PROFESSIONAL EXPERIENCE

AI Engineer

Wanclouds Inc — Remote, San Francisco Bay Area — May 2025 to Present

- Built multi-agent systems using MCP and LangGraph, connecting tool-calling agents to enterprise knowledge bases across autonomous multi-step workflows
- Maintained RAG pipelines backed by ChromaDB 1,000+ document corpus, handling semantic search and metadata enrichment
- Ran AI services on Kubernetes with zero-downtime rolling deployments; CI/CD via GitHub Actions covered build, test, and release stages
- Packaged all services as Docker containers, configured autoscaling policies and health probes across multi-node K8s clusters
- Wrote FastAPI microservices with WebSocket endpoints for real-time assistants; Celery and Redis handled the background task queue for content generation
- Cut time-to-relevant-content by 45 percent using chain-of-thought and ReAct-style prompting for curriculum personalization agents
- Took part in sprint planning and backlog grooming every two weeks, working directly with Product Owners

to scope and prioritize AI features
- Tracked latency, token spend, and user feedback in Langfuse across all production LLM calls

Associate NLP Engineer

Developers Den LLC — On-site, Islamabad — August 2024 to May 2025

- Built a multi-modal RAG pipeline for US mortgage document analysis – loan applications, appraisal reports, and underwriting PDFs from US lending institutions – using LangChain, FAISS, and LiteLLM across 10,000+ documents; applied chunking, metadata enrichment, and deduplication to keep retrieval clean
- Reduced document processing time by 50 percent by rewriting the ingestion pipeline for US mortgage data, replacing a sequential parse-and-embed loop with parallel async workers and smarter chunking
- Shipped the AI platform on Kubernetes using Docker, with zero-downtime rolling deployments managed through GitHub Actions; the same pipeline handled testing, image builds, and cluster releases
- FastAPI backend (REST and WebSockets) supported 500+ concurrent users in the US financial services sector, with Celery queues, Redis caching, rate limiting, and JWT auth throughout
- Wired up Apache Kafka to handle real-time document ingestion from multiple US data providers, decoupling the ingestion layer from the processing workers
- Used LangSmith and Langfuse to debug hallucinations and track accuracy, latency, and cost per request across production mortgage query workloads
- Implemented a GraphRAG layer with Neo4j that routed complex US mortgage regulatory queries through a knowledge graph before hitting the vector store

AI Developer

Maktek.ai — Remote, Dubai, UAE — November 2023 to July 2024

- Fine-tuned LLaMA-3, Mistral, and Gemma for code generation using LoRA and QLoRA (4-bit) with TRL and Unsloth on consumer GPUs; hit 95 percent on LM-Eval-Harness and published the resulting models to Hugging Face
- Served the fine-tuned models via vLLM on RunPod (AWS and GCP); GPTQ and AWQ quantization brought throughput up 3x and cut hosting costs by 50 percent
- Built a FastAPI and LangChain RAG system over Pinecone and ChromaDB, logging accuracy, latency, and token cost per query to MLflow
- Packaged model serving with Docker, deployed to Kubernetes, and wired up CI/CD so model updates went to production without downtime
- Ran an RLHF alignment loop using TRL and generated synthetic training data through model distillation, which dropped data labeling costs by 60 percent

AI Intern

Maktek.ai — Remote, Dubai, UAE — August 2023 to October 2023

- Shipped 3+ RAG applications using LangChain, OpenAI, and FastAPI into client production environments
- Built 5+ chatbots across different platforms to automate client business workflows, ranging from support triage to internal knowledge retrieval
- Ran large-scale web scraping jobs, cleaned and structured the output, and loaded it into knowledge bases backing the RAG systems

Research Assistant

National Center of AI, UET Peshawar — Peshawar, Pakistan — April 2023 to June 2023

- Researched multi-class text emotion classification using deep learning, reaching 85 percent average accuracy across multiple datasets
- Compared traditional classifiers (k-NN, Naive Bayes, SVM, stacking ensembles) against sequence models to understand where each approach broke down
- Wrote custom preprocessing pipelines for tokenization, normalization, and feature extraction that fed directly into model training
- Built and benchmarked RNN, LSTM, and GRU architectures in TensorFlow across the same datasets for a clean comparison

EDUCATION

Bachelor of Computer Software Engineering with Honours

University of Engineering and Technology, Mardan — Mardan, Pakistan — June 2020 to June 2024

CERTIFICATIONS

Generative AI with Large Language Models — AWS, Coursera — 2024
Deep Learning Specialization — Stanford University, Coursera — 2023
Machine Learning Specializations — Stanford University, Coursera — 2023
Database Design and SQL — University of Colorado, Coursera — 2023

LANGUAGES

English (Professional Working Proficiency) — Urdu (Professional Working Proficiency) — Pashto (Native Proficiency)