MUAAZ AHMAD SAEED

Senior AI/ML Engineer

Riyadh, Saudi Arabia | +966 53 526 3766 | muaazahmed93@gmail.com | Iqama: 2627296466

linkedin.com/in/muaazdev | github.com/muaazdev | muaazdev.com

PROFESSIONAL SUMMARY

AI/ML Engineer with 5+ years of production experience specializing in Generative AI, LLMs, and Computer Vision. Expert in building end-to-end AI systems including RAG pipelines, multi-agent architectures, and LLM fine-tuning (LoRA, QLoRA). Strong MLOps background with CI/CD pipelines, Docker containerization, and cloud deployment (AWS, GCP). Proven track record deploying high-throughput inference services with <200ms latency, optimizing ML pipelines, and building scalable production systems. Proficient with both open-source LLMs (Llama, Mistral) and closed-source APIs (OpenAI GPT, Claude, Gemini).

TECHNICAL EXPERTISE

LLMs & GenAI: RAG, Graph RAG, Multi-Agent Systems, LangChain, LangGraph, MCP, Prompt Engineering, LLM Fine-tuning (LoRA, QLoRA, PEFT), Open-source (Llama, Mistral, Gemma), Closed-source (GPT, Claude, Gemini)

MLOps & Cloud: AWS (ECS, EC2, S3, Lambda, SageMaker, CloudWatch), GCP, Docker, Kubernetes basics, CI/CD (GitHub Actions), Model Serving, Auto-scaling, TensorRT, vLLM, ONNX

NLP: Intent Recognition, Entity Extraction, Text Classification, Question Answering, Summarization, Conversational AI

Computer Vision: Object Detection & Tracking (YOLO, ByteTrack), Segmentation (SAM), Pose Estimation, Video Analytics, Stable Diffusion

Backend & Data: Python, FastAPI, Node.js, REST APIs, Async Programming, PostgreSQL, MongoDB, Redis, Vector DBs (Qdrant, Pinecone, ChromaDB)

PROFESSIONAL EXPERIENCE

AI Engineer / AI Consultant (Remote, Part-Time)

Accelerated Innovation Group | Georgia, USA | Sep 2025 - Present

- Provide AI/MLOps architecture guidance for enterprise teams deploying LLM-based systems to production
- Design CI/CD pipelines for ML model deployment with automated testing and evaluation gates
- Guide teams on RAG pipelines, multi-agent orchestration, prompt engineering, and infrastructure decisions

Machine Learning Engineer

Rapid Labs | Islamabad, Pakistan | Oct 2024 - Sep 2025

- Built No-Code Multi-Agent Platform on AWS ECS with Docker, auto-scaling, and load balancing; handles concurrent users with sub-second latency
- Architected high-throughput inference services using FastAPI async endpoints, request batching, and Redis caching for 10x latency reduction
- Implemented CI/CD pipelines with GitHub Actions: automated testing, linting, Docker builds, and deployment to AWS ECS
- Built evaluation harness for LLM agents: automated test generation, faithfulness scoring, and regression detection blocking bad merges
- Developed RAG pipelines with vector databases for document Q&A; optimized video analytics from 8 FPS to 25+ FPS via GPU quantization

AI Team Lead / Machine Learning Engineer

Metex Labz | Islamabad, Pakistan | Jun 2023 - Oct 2024

- Fine-tuned Llama 2 using LoRA/QLoRA on legal corpus; deployed with FastAPI and optimized inference using quantization
- Built multi-agent system with LangGraph for legal research, document generation, and question answering
- Deployed GPU-optimized inference services on GCP and on-premises; managed model versioning and rollback procedures
- Built production Stable Diffusion APIs: model serving, request queuing, horizontal scaling for face restoration and image editing
- Led team of 4 ML engineers; established code review practices, testing standards, and deployment workflows

Team Lead - AI Solutions

MetaFourt Pvt Ltd | Faisalabad, Pakistan | May 2022 - Jun 2023

- Led end-to-end AI model development for NLP and computer vision with focus on reproducibility and deployment automation
- Implemented Docker containerization for all ML services; built automated testing frameworks integrated with CI/CD

AI Engineer

Hive Technologies | Pakistan | Sep 2021 - Apr 2022

- Developed healthcare AI solutions including medical image analysis; optimized for edge inference using TFLite
- Built event-driven PubSub systems using Google Cloud, Node-RED, and Cloud Run services for real-time data processing
- Integrated Vertex AI for ML model deployment and inference pipelines on Google Cloud Platform

Python Developer

PhotonSofts | Faisalabad, Pakistan | Sep 2020 - Aug 2021

- Developed Python-based web applications with AI integration for keyword analysis and search functionality
- Built REST APIs and managed server-side operations ensuring optimal application performance and uptime

KEY PROJECTS

DocMind - Multi-Tenant RAG Platform

- Production-grade multi-tenant RAG platform with hybrid search (OpenAI embeddings + BM25), Cohere reranking, adaptive chunking, real-time SSE streaming, inline citations, and RAGAS evaluation pipeline
- Multi-tenant isolation with namespaced vector collections and tenant_id enforcement; deployed on Kubernetes (GKE) with HPA autoscaling and Prometheus monitoring

Tech: *Python, FastAPI, Next.js, PostgreSQL, Redis, ChromaDB, LangChain, Cohere, Docker | github.com/muaazdev/DocMind*

LegalMind - AI Legal Knowledge Assistant

- Production-grade modular RAG system for legal document Q&A with hybrid search (ChromaDB + BM25), Cohere reranking, mandatory citation validation, and 3 AI evaluation agents (Adversarial Lawyer, Compliance Auditor, Shepardizer)
- RAG Triad metrics with strict thresholds and CI/CD pipeline; deployed on Kubernetes (GKE) with HPA autoscaling

Tech: *Python, FastAPI, Llama 2, PEFT, LangGraph, ChromaDB, Pytest | github.com/muaazdev/legalmind*

No-Code Multi-Agent Builder Platform

- Platform enabling users to create and deploy custom AI agents without coding; production-deployed on AWS ECS with auto-scaling

Tech: *LangChain, LangGraph, Qdrant, PostgreSQL, Docker, AWS ECS, FastAPI, GitHub Actions CI/CD*

High-Throughput LLM Inference Service

- Inference service handling concurrent requests with <200ms p95 latency; auto-scaling based on metrics

Tech: *FastAPI async, Redis caching, Request batching, AWS ECS, CloudWatch monitoring*

Real-Time Sports Video Analytics

- Custom YOLO training for player/puck tracking; optimized from 8 FPS to 25+ FPS via TensorRT quantization

Tech: *PyTorch, YOLO, TensorRT, Gemini API, AWS GPU, Docker*

Pictro AI - AI Art Generator & Anime

- AI-powered image generation Android app with 100,000+ downloads; Stable Diffusion pipelines with GAN enhancements for anime-style visuals

Tech: *Stable Diffusion, GANs, FastAPI, Docker, GCP Cloud Run*

ToonMe - AI Cartoon & Anime Photo Editor

- Production-scale AI photo editing app with 50,000+ downloads; SDXL, GFPGAN, ESRGAN for style transfer and face restoration

Tech: *Stable Diffusion, GFPGAN, ESRGAN, FastAPI, Docker, GCP*

EDUCATION

Bachelor of Science in Computer Science | University of Agriculture Faisalabad

ADDITIONAL

Location: Based in Riyadh, Saudi Arabia (Transferable Iqama)
Languages: English (Professional), Urdu (Native) | **Leadership:** Team management, technical interviews, mentoring