MUAAZ AHMAD SAEED

Forward Deployed AI Engineer

Riyadh, Saudi Arabia | +966 53 526 3766 | muaazahmed93@gmail.com
linkedin.com/in/muaazdev | github.com/muaazdev | muaazdev.com

PROFESSIONAL SUMMARY

Forward Deployed AI Engineer with 5+ years taking AI from problem discovery to production. I work directly with business teams and enterprise clients to understand the real workflow, then design, build, and operate the GenAI and agentic systems that change it. Production experience in LangGraph multi-agent systems, RAG, Document AI, tool use/MCP, and LLM evaluation, with privacy-aware design (PII redaction, tenant isolation, encryption) for legal, healthcare, and financial documents. Solid engineering base: Git, automated testing, Docker, CI/CD, AWS and GCP, plus full-stack delivery with FastAPI and React/Next.js.

TECHNICAL EXPERTISE

LLMs & GenAI: RAG, Graph RAG, Multi-Agent Systems, LangChain, LangGraph, LlamaIndex, Haystack, MCP, Prompt Engineering, Structured Extraction (Pydantic / JSON-schema outputs), LLM Fine-tuning (LoRA, QLoRA, PEFT), Open-source (Llama, Mistral, Gemma), Closed-source (GPT, Claude, Gemini)

Evaluation & Observability: LLM evaluation harnesses, RAGAS, RAG Triad metrics, faithfulness scoring, regression gates in CI, Prometheus, CloudWatch

Document AI: OCR (AWS Textract, Azure Form Recognizer, Tesseract), Invoice/Contract Parsing, Table and Line-Item Extraction

Backend & Frontend: Python, FastAPI, Flask, Django, Node.js, REST APIs, Async Programming, React, Next.js, Angular, TypeScript

Data & ML: Pandas, NumPy, PyTorch, TensorFlow/TFLite, scikit-learn, Optimization (OR-Tools / PuLP)

Databases: PostgreSQL, MongoDB, Redis, Vector DBs (Pinecone, Milvus, Weaviate, Qdrant, ChromaDB)

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

Security: Encryption at rest/in transit (TLS, KMS), Secrets Management, RBAC, Tenant Isolation, PII Redaction

Customer Delivery: Stakeholder discovery and requirements gathering, POCs to production, demos, handover and documentation

NLP & Vision: Intent Recognition, Entity Extraction, Text Classification, Question Answering, Summarization, Conversational AI; YOLO, ByteTrack, SAM, Pose Estimation, Video Analytics, Stable Diffusion

PROFESSIONAL EXPERIENCE

Senior AI Engineer (Remote, Contract)

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

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

Machine Learning Engineer

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

- Built a No-Code Multi-Agent Platform that lets non-technical users create and deploy their own AI agents; on AWS ECS with Docker, auto-scaling, and load balancing at sub-second latency
- Built evaluation harness for LLM agents: automated test generation, faithfulness scoring, and regression detection blocking bad merges
- Implemented CI/CD pipelines with GitHub Actions: automated testing, linting, Docker builds, and deployment to AWS ECS
- Architected high-throughput inference services using FastAPI async endpoints, request batching, and Redis caching for 10x latency reduction
- 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

- Built multi-agent system with LangGraph for legal research, document generation, and question answering, including document ingestion and clause-extraction pipelines for legal contracts with citation validation against source text
- Fine-tuned Llama 2 using LoRA/QLoRA on legal corpus; deployed with FastAPI and optimized inference using quantization
- 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 on Google Cloud, Node-RED, and Cloud Run for real-time data processing; integrated Vertex AI for model deployment and inference pipelines

Python Developer

PhotonSofts | Faisalabad, Pakistan | Sep 2020 - Aug 2021

- Developed Python web applications with AI integration for keyword analysis and search; built REST APIs and managed server-side operations

KEY PROJECTS

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

DocMind - Multi-Tenant RAG Platform

- Hybrid search (OpenAI embeddings + BM25), Cohere reranking, adaptive chunking, SSE streaming, inline citations, and RAGAS evaluation pipeline
- Multi-tenant isolation with namespaced vector collections and tenant_id enforcement; on GKE with HPA autoscaling and Prometheus monitoring
  Tech: Python, FastAPI, Next.js, PostgreSQL, Redis, ChromaDB, LangChain, Cohere, Docker | github.com/muaazdev/DocMind

No-Code Multi-Agent Builder Platform

- Platform enabling users to create and deploy custom AI agents without coding; production-deployed on AWS ECS with autoscaling
  Tech: LangChain, LangGraph, Qdrant, PostgreSQL, Docker, AWS ECS, FastAPI, GitHub Actions CI/CD

Multi-Agent Platform with Subagent Routing

- Node.js agent-builder platform with parent agents and subagents; semantic routing via knowledge-base summarization, conversation history management, and MCP server integrations
  Tech: Node.js, PostgreSQL/pgvector, MCP, Railway

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 & ToonMe - Consumer GenAI Apps

- AI art/anime generator (100,000+ downloads) and AI photo editor (50,000+ downloads); Stable Diffusion/SDXL, GANs, GFPGAN, ESRGAN for style transfer and face restoration
  Tech: Stable Diffusion, GANs, FastAPI, Docker, GCP Cloud Run

EDUCATION

Bachelor of Science in Computer Science | University of Agriculture Faisalabad

ADDITIONAL

Languages: English (Professional), Urdu (Native) | **Leadership:** Team management, technical interviews, mentoring