



Muhammad Husnain Ali

AI Engineer

AI Solution Architect

ABOUT ME

AI Engineer and Solution Architect delivering enterprise-grade Generative AI for UAE energy, banking, and government sectors. 4+ years building agentic systems, RAG pipelines, and event-driven architectures, with direct delivery to Fertiglobe, RTA, TAQA, First Abu Dhabi Bank, and Dubai Islamic Bank.

EDUCATION

Bachelor of Science (B.Sc.), Computer Science

Superior University, Lahore

CGPA: 3.65/4.0

CERTIFICATIONS

Microsoft Copilot Studio & Azure AI
LangChain & LangGraph: Agentic AI
Azure AI: Cloud-Native Architecture
Prompt Engineering & LLM Optimization
AWS Certified Machine Learning Engineer Associate (MLA-C01), Expected Q3 2026

TECHNICAL SKILLS

Generative AI

Generative AI, Large Language Models (LLM), Agentic AI, Multi-Agent Orchestration, Sub-Agents & Skills, Harness Engineering, Context Engineering, Prompt Engineering, Retrieval-Augmented Generation (RAG), Guardrails, Fine-Tuning, LoRA/QLoRA, Embeddings, Semantic Search

Solution Architecture

Event-Driven Architecture, Microservices Architecture, Modern Monolith Architecture, Apache Kafka, Azure Event Hubs, Enterprise Solution Design, Distributed Systems, System Design, API Design

LLM Stack

LangChain, LangGraph, CrewAI, GPT-4o, Azure OpenAI, Claude, LLaMA, Hugging Face Transformers, BERT

Vector Databases

Azure AI Search, Pinecone, FAISS, Chroma, Weaviate, Neo4j, Knowledge Graphs, Vector Search, Hybrid Search

Cloud, Infra & MLOps

Microsoft Azure, Azure AI Foundry, Azure Machine Learning, Copilot Studio, Container Apps, Azure Functions, Cosmos DB, Azure Kubernetes Service (AKS), Helm, Kubernetes, Docker, AWS, Azure DevOps,

+92 300-0916-426

linkedin.com/in/muhammadhusnainali873

muhammad.husnain.ali.738@gmail.com

github.com/MuhammadHusnainAli

USD 50M+ Revenue Impact

73% RAG Latency Reduction (15s to 4s)

1000+ Hours Saved Annually

1500+ Connectors (NovaFloat Platform)

500+ DB Tables via Text-to-SQL

300K+ Legal Cases Indexed

95–98% Eval Accuracy (Enterprise)

65%+ Cross-Department Productivity

WORK EXPERIENCE

AI / ML Engineer & AI Solution Architect

DigiFloat
Lahore, Pakistan

Nov 2024 – Present

- Built 25+ agentic AI systems for enterprise Law, Data, and UAE/GCC clients using LangChain, LangGraph, Azure AI Services, and Copilot Studio.
- Pioneered production use of MCP (Model Context Protocol) and A2A (Agent-to-Agent) for secure, interoperable multi-agent systems.
- Set up MLOps on Azure DevOps (CI/CD, model versioning); deployed across Container Apps, Functions, AI Foundry. Integrated Langfuse for observability.
- Architect enterprise AI platforms end to end across event-driven, microservices, and modern-monolith designs, deployed on AKS with Helm charts for small to mid-size model serving and backend service rollout.

Key Projects:

- Data2Decisions (Enterprise Data Intelligence):** Multi-agent text-to-insight platform built around a central orchestration agent that coordinates 10+ specialised sub-agents, each with its own scoped skills. Unifies unstructured sources (SharePoint, S3, Google Drive) with 25+ structured connectors (MySQL, Oracle, MongoDB) over a unified knowledge-graph layer that powers cross-source reasoning. Data governance and access controls scope every sub-agent to its permitted domain. **USD 50M+ revenue impact**; 1000+ engineering hours saved annually.
- NovaFloat (Enterprise AI Productivity Platform):** AI platform unifying **1500+ prebuilt connectors**, custom **MCP servers** authored in Python and attached straight from the UI, and an n8n-style visual workflow builder for bespoke automations. On top of any connector, users compose knowledge bases and AI pipelines, while a **harness-engineered agent** reasons across the relevant connectors, MCP tools, knowledge bases, and live web search, forming the platform's core interactive layer. Backend engineered in **FastAPI** with **Microsoft Entra ID** OAuth, PostgreSQL for application state, **Azure AI Search** for knowledge bases, and Redis caching. Rolled out to HR, DevOps, infrastructure, and cross-functional teams, lifting productivity **65%+** across departments.
- Cheque Clearance Automation (First Abu Dhabi Bank):** Event-driven platform on **Kafka / Azure Event Hubs** where partitioned topics stream cheques through microservice consumers: image preprocessing, **YOLO** detection, OCR extraction, signature matching, then automated clearance. Tuned topic partitioning and consumer groups to remove stalling under load; deployed on **AKS** with Helm, private networking, and autoscaling for secure, bank-grade throughput.
- Plutus (Dubai Islamic Bank):** Conversational banking agent for balances, transfers, and beneficiaries. Wrapped core banking REST APIs into a custom **MCP server** exposed to CrewAI agents, applying a layered governance model of guardrails, both AI-based and hard-coded, enforced at tool and agent level (e.g. balance replies constrained to the single numeric value the tool returns), keeping every response within regulated, auditable banking boundaries.
- Fabric Extraction Pipeline (Fertiglobe UAE):** End-to-end Microsoft Fabric pipeline ingesting policy PDFs from SharePoint and extracting 40+ structured fields into Delta tables. Azure AI Foundry agents with JSON-schema response formats drive a multi-step per-PDF extraction plus a dedicated evaluation agent for sensitive financial figures, reaching **95% extraction accuracy and 98.75% evaluation accuracy**.
- CFO Digest (UAE Client):** Automated multi-stage AI + data pipeline, end-to-end from scrape to email. Serper API pulls date-scoped market/financial data into Fabric warehouse tables; downstream agents load and threshold-filter records, draft stakeholder narratives, then hand off to an evaluation agent. Runs weekly to generate a Power BI PDF report emailed per stakeholder group, orchestrated via an Azure AI Foundry agent.
- CFO Board Agent (Fertiglobe UAE):** Board-level Copilot Studio agent that distils source content into concise 3-line briefs from clear structured instructions, built on secure document grounding with a dedicated evaluation loop, reaching **87% evaluation accuracy**.
- RTA Mahboob (UAE Govt, 1000+ users):** Multilingual RAG (EN/AR) that routes each query by language

CI/CD, MLOps, LLMOps, MLflow

Data Engineering

Microsoft Fabric, ETL / ELT Pipelines, Delta Tables, Azure Data Factory, Lakehouse, Data Warehouse, SQL, Text-to-SQL, Power BI

Computer Vision

YOLO Object Detection, Deep Learning, OCR Pipelines, Image Preprocessing, Signature Matching, Document Extraction

Protocols & Standards

MCP (Model Context Protocol), A2A (Agent-to-Agent), REST APIs, GraphQL, OAuth 2.0, Microsoft Entra ID

Integrations & CRM

HubSpot CRM, SharePoint, MS Teams, Serper API, Webhooks

AI Governance

Guardrails, Access Control (RBAC), Evaluation & Grounding, Responsible AI, Data Governance, Auditability

Observability

Langfuse (Tracing, Logging, Monitoring), Evaluation Pipelines, Model Monitoring

Backend & ML

Python, FastAPI, PyTorch, TensorFlow, Hugging Face, Machine Learning, Deep Learning, Natural Language Processing (NLP), PostgreSQL, Redis, Git, Azure VNet, API Management Gateway, Managed Identity

INDUSTRIES

Banking / FinTech (DIB, FAB)
Government (RTA UAE, Data2Decisions)
Energy (Fertiglobe UAE: DOA, VAT)
Legal Tech (LawGPT)
Healthcare (MedConverse, HealChat)
HR Tech (Screening, Email Automation)
Enterprise Productivity (NovaFloat)

- into a dedicated English or Arabic pipeline for low latency, alongside text-to-SQL over 500+ tables via FastAPI. Cut response latency from 15s to 4s (**73%**).
- UAE VAT Agent (Fertiglobe UAE):** FTA-reference, amendment-aware tax-guidance agent grounded on curated FTA source documents for verifiable, citable answers. Deployed to Microsoft Teams. **98% accuracy** across 100+ eval cases.
 - DOA (Fertiglobe UAE):** Policy Q&A via Copilot Studio. Source PDFs encoded values as colour coding rather than literal numbers, so a preprocessing pipeline converts colour-coded documents into clean Markdown before the agent sees them, fixing the data problem ahead of the agent problem. SharePoint + Entra ID integrated; **95% accuracy** on 200+ eval cases.
 - LawGPT (Legal AI, Pakistan):** Platform for lawyers and judges spanning legal chat, document drafting, and document analysis, where a centralised orchestrator agent delegates to feature-specific sub-agents on demand. 300K+ legal cases indexed.
 - Tanzim (Driver Support Automation):** AI support platform where drivers raise tickets across 5 to 6 workflows: lost-and-found items recovered in vehicles, fare-dispute resolution, and more. Each inbound email triggers an AI workflow that attempts autonomous resolution by drafting and sending replies; when a case exceeds its confidence or scope, the agent escalates and auto-creates a **HubSpot CRM** ticket for a human agent to close the loop. Entire workflows engineered in FastAPI and orchestrated with LangChain and LangGraph, with layered guardrails and governance controls over every automated action, hand-off, and escalation. Hardened for production: an **Azure API Management** gateway fronts the AI app for secure, throttled access, Container Apps run inside a private VNet, and the **MCP server** is deployed under the hood of the same VNet and exposed only through the API gateway.

Data Scientist / ML Engineer

Swift Solver Jan 2024 – Jul 2024
Remote

- Built end-to-end ML pipelines across 50K+ records; fine-tuned SLMs with LoRA/QLoRA for a 25% accuracy gain.
- Built NLP classification and sentiment pipelines using Transformers and BERT; set up model versioning, experiment tracking, and automated retraining.

Python Developer / AI Engineer

Pure Logics Jun 2023 – Dec 2023
Lahore, Pakistan

- Fine-tuned GPT-3.5-turbo via OpenAI API with a 40% accuracy gain; built preprocessing pipelines across 100K+ documents.
- Shipped FastAPI and Streamlit apps in production with auth, rate limits, and logging; applied few-shot and chain-of-thought prompting.

KEY ACHIEVEMENTS

- Delivered 25+ production agentic AI systems across legal, banking, energy, and government domains for UAE/GCC enterprise clients.
- Drove **USD 50M+ revenue impact** and saved 1000+ engineering hours annually through the Data2Decisions multi-agent platform.
- Achieved **95% extraction and 98.75% evaluation accuracy** on the Fertiglobe Microsoft Fabric extraction pipeline via schema-governed evaluation agents.
- Cut multilingual RAG latency by **73% (15s to 4s)** for 1000+ government users on the RTA Mahboob platform.
- Pioneered production **MCP** and **A2A** adoption with tool- and agent-level governance guardrails for secure, auditable multi-agent systems.
- Architected event-driven, microservices, and modern-monolith systems for enterprise and banking workloads on **Kafka / Event Hubs**, shipped to **AKS** via Helm with private networking.