Muhammad Ghulam Jillani

LinkedIn | Hire Me (My Upwork Profile) | m.g.jillani123@gmail.com | Kaggle | GitHub | Medium

(5+) Years of Experience

Professional Summary

Lead AI & Cloud Data Scientist (*Generative AI | LLMOps | Agentic RAG | Autonomous AI Systems | Multi-Cloud Architect*)

I design and deploy production-grade Generative AI systems that automate complex business workflows, enhance decision intelligence, and deliver measurable ROI.

With 5+ years of experience and 44+ successful AI implementations, I specialize in architecting scalable LLM-powered platforms across AWS, Azure, and GCP. My work spans healthcare, finance, legal, retail, and SaaS, where I transform fragmented data and manual operations into intelligent, autonomous AI systems.

I operate at the intersection of AI architecture, cloud engineering, and business strategy, building systems that are not just innovative, but reliable, secure, and enterprise-ready.

Core Capabilities:

- Architecting LLM-driven applications using GPT-4/5, Claude 4, Gemini v3, DeepSeek v3, and LLaMA 3/4
- Designing multi-agent and Agentic RAG systems using LangChain, LangGraph, LlamaIndex, AutoGen, n8n ai agents and LangSmith
- Implementing advanced retrieval pipelines with Pinecone, Weaviate, FAISS, Knowledge Graphs, and multimodal data integration
- Deploying scalable AI systems with robust MLOps and LLMOps frameworks using MLflow, ZenML, Docker, CI/CD, and cloud-native orchestration
- Building AI Copilots, Compliance Automation Engines, Intelligent Document Systems, Predictive Analytics Platforms, and Autonomous Decision Systems

Impact Snapshot:

- 44+ AI/ML systems delivered End-to-End
- Up to 65% reduction in manual workflows
- Up to 75% improvement in retrieval accuracy
- Up to 40% increase in operational efficiency across enterprise clients

Recognition:

- 24× LinkedIn Top Voice in AI and Cloud ML | Top 100 Global Kaggle Contributor
- NVIDIA Developer Program Contributor | Engaged in AI research, experimentation, and community knowledge sharing within Google Developer Group and AWS AI ecosystems

I build AI systems that move from concept to production with precision, scalability, and real business impact.

Technical Skills

AI Engineering & LLM Systems:

- **Large Language Models:** GPT-4o/4.5, Claude 3/4, Gemini 3, LLaMA 3/4, DeepSeek v3
- **LLM Frameworks:** LangChain, LangGraph, LlamaIndex, AutoGen, LangSmith, PhiData, and n8n Ai agents
- **Architectures:** Agentic RAG, Multi-Agent Systems, Tool Calling, Function Orchestration, Prompt Engineering, Fine-Tuning, Evaluation Pipelines
- **Vector & Retrieval:** Pinecone, Weaviate, FAISS, ChromaDB, Qdrant, Knowledge Graph Integration
- **Applications:** AI Copilots, Enterprise Search, Compliance Automation, Intelligent Document Processing, Conversational AI

Machine Learning & Data Science:

- Machine Learning, Deep Learning, NLP, Time Series Forecasting, Predictive Analytics, Feature Engineering, Model Optimization, Statistical Modeling, Model Deployment
- **Frameworks & Tools:** Scikit-Learn, TensorFlow, Keras, PyTorch, Hugging Face Transformers
- **Data Processing:** Pandas, NumPy, PySpark
- **Visualization:** Plotly, Matplotlib

Backend & AI Application Development:

- Python, FastAPI, Flask, Streamlit
- REST APIs, GraphQL
- Microservices Architecture
- Docker, GitHub Actions, CI/CD Pipelines

Cloud, MLOps & LLMOps:

- **Cloud Platforms: AWS** (SageMaker, Bedrock, Lambda, EC2, Step Functions, **EC2 Virtual Machines**), **Azure** (Azure ML, Azure AI Studio, App Services, AKS, **Azure Virtual Machines**), **GCP** (Vertex AI, Cloud Functions, BigQuery ML, **Compute Engine VMs**), Heroku, Vercel, and Hugging Face.
- **Practices & Tools:** MLOps, LLMOps, AIOps, MLflow, ZenML, AutoML, Model Monitoring, Experiment Tracking, Cloud ML Pipelines, Scalable Deployment & Orchestration

Databases:

- SQL, NoSQL
- **Vector Databases:** Pinecone, FAISS, Weaviate, ChromaDB, Qdrant
- Neo4j (Knowledge Graphs)

Leadership & Strategy:

- Technical Leadership, AI Architecture Design, Product Thinking, Stakeholder Alignment, Agile Delivery, Business Intelligence

Communication & Collaboration:

- Technical Documentation, Executive Reporting, Client Communication, Cross-Functional Collaboration

Languages:

- English – Fluent

Projects

- **AI-Powered Blog Generation Platform – LangGraph + Parallel Orchestration:** Developed an AI-driven platform to automate end-to-end technical blog creation by intelligently routing topics, gathering real-time research from authoritative sources, and generating structured multi-section content in parallel. Delivered high-quality, citation-backed blog posts with AI-generated technical diagrams, enabling content teams to produce professional articles faster and more consistently.
  Impact: Achieved 5x faster content production through parallel section generation, reduced manual writing effort by 80%, and maintained consistent quality across 3–15 section blog posts.
  Technologies: LangGraph, GPT-4.1-mini (OpenAI), Tavily API, Google Gemini, RAG Pipeline, Python, Pydantic, Streamlit, Docker, AWS EC2, S3.
- **AI-Powered ESG Compliance Analyzer – Agentic RAG + LLMOps:** Developed an AI-driven platform to automate ESG (Environmental, Social, Governance) compliance reporting by analyzing corporate disclosures and detecting regulatory gaps. Delivered high-accuracy compliance checks tailored to EU CSRD requirements, enabling enterprises to meet sustainability reporting obligations faster and more efficiently.
  Impact: Achieved 90% accuracy in ESG compliance checks, reduced manual auditing time by 70%, and accelerated corporate sustainability assessments.
  Technologies: LangChain, GPT-4o (Azure OpenAI), Chroma DB, RAG Pipeline, Python, Azure Blob Storage, FastAPI, Streamlit.
- **Enterprise Knowledge Navigator – Agentic RAG for Corporate Wikis:** Built an AI-powered internal assistant for navigating vast enterprise documentation across departments. Combined multi-agent reasoning, metadata tagging, and vector-enhanced retrieval to help employees access policies, SOPs, HR guidelines, and technical manuals.
  Impact: Reduced internal search time by 70% and improved employee productivity across support and engineering teams.
  Technologies: LangGraph, Pinecone, GPT-4o, RAG Fusion, LangChain, FastAPI, AWS S3, Azure OpenAI, Streamlit, Role-based Access.
- **RetailGPT Copilot – AI Assistant for Product Recommendations & Inventory Planning:** Engineered a GenAI Copilot for retail enterprises that generates dynamic product recommendations using real-time user behavior and inventory forecasts. Integrated reinforcement learning for personalized optimization and used AWS Bedrock for model orchestration.
  Impact: Boosted conversion rates by 32% and improved inventory planning accuracy by 48%.
  Technologies: AWS Bedrock, RLHF, LangChain, Scikit-learn, Snowflake, Postgres + PgVector, Airflow, FastAPI.

- **MedIQ Insights – AI-Powered Clinical Decision Support System:** Developed a GenAI platform for hospitals to analyze patient history, lab reports, and symptoms using a hybrid agentic RAG pipeline. The system offers recommendations and real-time insight based on ICD-10 mappings and updated medical literature.
  Impact: Improved diagnostic accuracy by 25% and reduced patient onboarding time by 40%, empowering clinicians with intelligent decision support.
  Technologies: LangChain, LangGraph, RAGatouille, LLM (Claude 4), GCP Vertex AI, Weaviate, BigQuery, LangSmith, FastAPI.
- **FinServe GPT – Autonomous Financial Risk Advisor:** Built a secure cloud-native system that leverages LLMs and multi-agent architecture to monitor transactions, assess portfolio risk, and suggest investment strategies based on real-time market data and user goals.
  Impact: Enabled financial analysts to reduce manual workload by 60%, while clients received tailored recommendations backed by real-time AI analysis, boosting retention and ROI.
  Technologies: CrewAI, LangGraph, AutoGen, GPT-4o (via Azure), PostgreSQL + PgVector, Azure Functions, ScaNN, Kubernetes.
- **AI Compliance Copilot – Regulatory Document QA & Risk Detection:** Engineered an AI copilot that parses and validates compliance reports, contracts, and regulatory updates (e.g., GDPR, HIPAA, SOC 2). Incorporated OpenSearch for full-text audit, dynamic prompt routing, and AI-generated risk flags.
  Impact: Reduced compliance risk exposure by 45% and decreased manual review time for legal teams by 60%.
  Technologies: LangChain, LangSmith, OpenSearch, GPT-4o (Bedrock), AWS Lambda, DocTR, PII Redaction, FastAPI.
- **LegalDoc AI – Intelligent Document Retrieval & Summarization System:** A secure enterprise platform that uses Agentic RAG to analyze legal contracts and case files. Integrated an LLM-powered chatbot for dynamic querying and document-based reasoning across multi-format legal datasets.
  Impact: Reduced document review time by 70% and enabled legal teams to find relevant clauses and insights in seconds, improving operational efficiency and decision-making.
  Technologies: LangChain, FAISS, GPT-4o (via AWS Bedrock), Streamlit, Pinecone, Docker, OpenSearch, AWS S3, FastAPI.
- **AI-Powered Meeting Assistant (AIMA) – Generative AI for Intelligent Meeting Management:** Designed and deployed an AI- powered meeting assistant leveraging GPT-4o, LangChain, FastAPI, and Pinecone to automate meeting transcription, summarization, and knowledge retrieval. Integrated Google Sheets API for roadmap estimations and Gmail API for seamless email dispatch. The system enhances productivity by providing instant access to key meeting insights and AI-driven MoM (Minutes of Meeting) generation. Deployed on AWS for real-time scalability.
  Impact: Reduced manual meeting documentation by 80%, improving decision-making efficiency.
  Technologies: Python, GPT-4o, LangChain, FastAPI, Pinecone, Google Sheets API, AWS (Lambda, S3, EC2, Bedrock).
- **Conversational AI Chatbot for Customer Support – GPT-4o, AWS Bedrock:** Built an enterprise-grade conversational AI chatbot powered by GPT-4o, integrated with LangChain and AWS Bedrock for real-time customer interactions. Implemented NLP techniques such as intent recognition, tokenization, and dynamic response generation, ensuring context-aware and personalized support.
  Impact: Reduced customer query resolution time by 40%, increasing customer satisfaction.
  Technologies: Python, GPT-4o, AWS Bedrock, LangChain, Flask, PyTorch, NLP, AWS Lambda, API Gateway.
- **DocuWiz AI – AI-Powered Document Intelligence System:** Developed DocuWiz AI, a document intelligence tool leveraging BART, DistilBERT, and Retrieval-Augmented Generation (RAG) for text extraction, summarization, and sentiment analysis. Designed for legal, academic, and corporate sectors, enabling automated document processing and deep content insights.
  Impact: Reduced document review time by 60%, improving workflow automation.
  Technologies: Python, BART, DistilBERT, RAG, FastAPI, Streamlit, AWS Lambda, Azure Cognitive Services.
- **Intelligent Document Summarization Tool – NLP & Pegasus:** Built an AI-powered document summarization tool using transformer-based models (Pegasus) to generate concise yet informative summaries. Supports PDF and DOCX formats, streamlining document analysis in legal, financial, and research domains.
  Impact: Reduced document processing time by 60%, increasing productivity.
  Technologies: Python, TensorFlow, Pegasus, PyPDF2, Streamlit, AWS Lambda, GCP Vertex AI.
- **AI-Powered Plant Disease Detection – Deep Learning & Azure AI Studio:** Developed an AI-driven agricultural assistant for real-time plant disease detection using ResNet50 for image classification. Integrated Azure AI Studio for cloud-based model training and Azure App Services for web deployment. The system enables farmers to detect crop diseases early and receive automated treatment recommendations.
  Impact: Reduced disease detection time by 70%, improving crop yield predictions by 30%.
  Technologies: Python, ResNet50, TensorFlow, Flask, React, OpenCV, Azure AI Studio, Azure App Services.
- **Predictive Analytics for Sales Forecasting – LSTM & ARIMA:** Built an AI-powered sales forecasting model using LSTM and ARIMA to predict future sales trends based on historical data. Integrated with a Flask-based web dashboard for real-time insights, enabling data-driven decision-making for business growth and inventory optimization.
  Impact: Improved sales forecasting accuracy by 30%, optimizing business operations.
  Technologies: Python, TensorFlow, LSTM, ARIMA, Flask, Matplotlib, AWS SageMaker.
- **Real-Time Traffic Monitoring System – Computer Vision & Edge AI:** Designed an AI-driven traffic monitoring system leveraging YOLOv7 for real-time vehicle detection in live video feeds. The system integrates Docker for containerized deployment and FastAPI for real-time data streaming, aiding urban traffic authorities in congestion analysis and road safety improvements.
  Impact: Enhanced traffic flow optimization and improved road safety analytics.
  Technologies: Python, YOLOv7, OpenCV, Docker, FastAPI, AWS Lambda, Edge AI.
- **Real-Time Object Tracking for Security – DeepSORT & YOLO:** Built a real-time security surveillance system integrating DeepSORT with YOLO for multi-object tracking. Designed for crowd monitoring, traffic surveillance, and smart security systems. Integrated CI/CD pipelines using GitHub Actions for rapid deployment.
  Impact: Enhanced security efficiency and automated threat detection.
  Technologies: Python, YOLO, DeepSORT, Docker, CI/CD, MLflow, AWS S3, AWS Lambda.
- **AutoML-Studio – No-Code/Low-Code ML Platform:** Built a no-code AutoML platform supporting classification, regression, time-series forecasting, and clustering. Integrated MLflow for tracking, SHAP for explainability, and data drift detection for model reliability. Deployed with FastAPI and Streamlit, enabling non-technical users to create and deploy AI models seamlessly.
  Impact: Accelerated model development for non-technical users, democratizing AI adoption.
  Technologies: Python, MLflow, SHAP, FastAPI, Streamlit, AWS Lambda, Google Cloud Run.
- **AIOps-Driven Predictive Maintenance System – AWS SageMaker & Kafka:** Developed a predictive maintenance platform using AWS SageMaker and Kafka to detect anomalies in industrial equipment. Integrated real-time LSTM models for automated failure detection and predictive analytics.
  Impact: Reduced downtime by 40%, improving maintenance efficiency.
  Technologies: Python, AWS SageMaker, Kafka, Databricks, LSTMs, CI/CD, AIOps.
- **Customer Satisfaction Prediction System – MLOps with ZenML & MLflow:** Created a customer satisfaction prediction model using ZenML and MLflow, achieving 93% accuracy in forecasting customer feedback. Built within a MLOps framework for seamless pipeline management and model deployment.
  Impact: Enhanced customer experience optimization for enterprises.
  Technologies: Python, ZenML, MLflow, Scikit-learn, AWS SageMaker, Azure Machine Learning.

CERTIFICATIONS

- **IBM Machine Learning Specialization Professional Certificate,** IBM.
- **Microsoft Azure AI Fundamentals AI-900 Exam Prep Specialization,** Microsoft.
- **Machine Learning Engineering for Production (MLOps),** DeepLearning.ai.
- **Generative Adversarial Networks (GANs) Specialization,** DeepLearning.ai.
- **Preparing for Google Cloud Certification: Machine Learning Engineer,** Google.
- **AWS Cloud Solutions Architect Professional Certificate,** Amazon Web Services.
- **Data Science on the AWS Cloud Specialization**, DeepLearning.ai.
- **MLOps | Machine Learning Operations Specialization,** Duke University.
- **Advanced Machine Learning on Google Cloud Specialization,** Google Cloud.
- **Google Advanced Data Analytics Professional Certificate,** Google.
- **Deep Learning Specialization,** DeepLearning.ai.
- **Prompt Engineering for ChatGPT,** Vanderbilt University.
- **AI Product Management,** Duke University.
- **IBM Generative AI Product Managers,** IBM.
- **IBM AI Product Manager,** IBM.
- **Google Project Management: Professional Certificate,** Google. **Practical**
- **Machine Learning Specialization,** DeepLearning.AI.
- **Google Business Intelligence Professional Certificate,** Google.
- **Large Language Model Operations (LLMOps),** Duke University.