

HUSSNAIN ARSHAD

AI Engineer | Generative AI, LLMs, RAG, AI Agents

Lahore, Pakistan | hussnainarshad85@gmail.com | +92 334 0408039
linkedin.com/in/hussnain-arshad-4b1896211 | hussnainarshad.github.io

PROFESSIONAL SUMMARY

AI Engineer with 4+ years building and shipping production Generative AI systems: LLM chatbots, RAG pipelines, multi-agent workflows, text-to-SQL, and document intelligence. Hands-on with LangChain, LangGraph, CrewAI, MCP, OpenAI, Anthropic, and Amazon Bedrock. Strong in Python, FastAPI, Docker, AWS, and Azure. MS in Data Science; AWS Certified Machine Learning Engineer - Associate and AWS Certified AI Practitioner.

PROFESSIONAL EXPERIENCE

Senior Machine Learning Engineer | 7Vals

Apr 2025 – Present

Cloud products for equipment tracking, rental management, and IT Asset Management (ITAM).

- Designed and deployed production Generative AI / LLM features across four enterprise platforms, from POC to production.
- Built RAG chatbots, agentic workflows, text-to-SQL automation, AI reporting, and semantic search to automate complex enterprise workflows.
- Architected LLM orchestration and multi-agent systems with LangGraph, LangChain, OpenAI, Anthropic, and MCP for task routing and enterprise API interaction.
- Developed an authenticated MCP server for product APIs (asset create/retrieve/update), cutting manual operational effort by ~80%.
- Automated asset creation with OCR and Whisper speech-to-text, reducing manual data entry by ~90%.
- Shipped Smart Diagnosis, AI Summaries, Similar Tickets, Recommendations, and Smart Ticket Assignment to speed support workflows.
- Built Python/FastAPI microservices; deployed containerized AI workloads with Docker, AWS Lambda, and ECR.
- Improved AI response accuracy by ~20% and cut API latency by ~35% via evaluation, prompt engineering, semantic filtering, and caching.
- Led a 3-person AI innovation team delivering AI capabilities across four core products with product and engineering partners.

ML/AI Engineer | Premier NX

Feb 2024 – Mar 2025

Service company focused on finance and accounting solutions.

- Deployed production LLM apps with LangChain, Mixtral, Claude, and Azure OpenAI, improving response accuracy by ~20%.
- Architected hybrid RAG pipelines with PGVector and PostgreSQL, reducing manual information retrieval effort by ~60%.
- Orchestrated multi-agent workflows with CrewAI, cutting manual intervention in repetitive processes by ~70%.
- Built document intelligence pipelines with Azure Document Intelligence, reducing registration-card processing time by ~90%.
- Integrated Azure Speech Services for speech-enabled AI experiences; shipped secure FastAPI microservices with authenticated REST APIs.
- Containerized and deployed AI services on Microsoft Azure with Docker for scalable production execution.

Machine Learning Engineer | One Machine Software

Jun 2022 – Jan 2024

Healthcare management solutions.

- Built a multilingual medical AI chatbot with LangChain, Pinecone, and Whisper covering 800+ medications.
- Created a domain medical knowledge base from government and clinical sources; evaluated OpenAI, Llama, Mistral, and AWS Bedrock for accuracy, latency, and cost.
- Automated insurance document verification with PaddleOCR and Tesseract (~50% less manual entry); implemented text-to-SQL for natural-language healthcare queries.
- Deployed containerized AI services on Azure with Docker (99.9% uptime); used CrewAI agentic workflows to cut insurance claim turnaround by ~40%.
- Built NLP pipelines for sentiment analysis and intent detection; ran internal workshops on prompt engineering and LLM evaluation.

TECHNICAL SKILLS

Languages & Frameworks: Python, FastAPI, REST APIs, TensorFlow, Keras

Generative AI & LLMs: LangChain, LangGraph, CrewAI, MCP, OpenAI, Anthropic, Amazon Bedrock, Ollama, RAG, text-to-SQL, fine-tuning, prompt engineering, multi-agent systems
Data & Retrieval: Pinecone, PGVector, PostgreSQL, OCR (PaddleOCR, Tesseract), Whisper, document parsing, semantic search
Cloud & DevOps: AWS (Lambda, ECR), Microsoft Azure, Docker, Git

EDUCATION

MS, Data Science | University of Management and Technology, Pakistan 2022 – 2024
GPA: 3.75 / 4.00 | Focus: Machine Learning, Deep Learning, NLP

BS, Electrical Engineering | University of Lahore 2017 – 2022
GPA: 3.1 / 4.00 | Coursework: Programming, Data Structures, Artificial Intelligence

CERTIFICATIONS

AWS Certified Machine Learning Engineer - Associate | Feb 2026 – Feb 2029
credly.com/badges/efcda18f-b478-4ae5-9dc1-4b2cd4020976

AWS Certified AI Practitioner | Aug 2025 – Aug 2028
credly.com/badges/66ccb7df-b106-4b56-af2c-4b56b76d83af

PUBLICATIONS

PhageVir: Assessing Computational Intelligence Models for Accurate Identification of Phage Virion Proteins. International Journal of Advanced and Applied Sciences (IJAAS), Vol. 12, Issue 5, 2025.
science-gate.com/IJAAS/2025/V12I5/1021833ijaas202505013a.html

LANGUAGES

English, Urdu