MAMOON UR RASHEED

AI/ML ENGINEER

+923181393178 | mamoon.aiwork@gmail.com | Lahore, Pakistan

SUMMARY

AI/ML Engineer with 4+ years of experience in designing, fine-tuning, and deploying Large Language Models (LLMs) and Generative AI solutions for enterprise applications. Expertise in developing Retrieval-Augmented Generation (RAG) pipelines, transformer-based NLP systems, and AI-powered chatbots using both open-source (LLaMA-2, Mistral, Falcon, BERT) and commercial LLMs (GPT-4, Gemini, Claude). Skilled in optimizing model performance—reducing latency by 40% via quantization and ONNX runtime—and deploying scalable MLOps pipelines on AWS/GCP/Azure using Docker, Kubernetes, and FastAPI. Successfully automated document processing (70% efficiency gain) and built multi-LLM hybrid systems to balance cost and accuracy. Passionate about delivering low-latency, production-ready AI with robust monitoring (Prometheus/Grafana) and seamless integration.

PROFESSIONAL SKILLS

Skilled in Python, Flask, and FastAPI for robust API development. Deeply proficient in Generative AI, transformer architectures, MLOps, and deep learning techniques. Strong foundation in Natural Language Processing (NLP), including text classification, summarization, embeddings, and semantic search. Experienced in deploying AI solutions using cloud platforms like AWS, Azure, and GCP, with a focus on containerization and scalable infrastructure. Hands-on expertise in Speech Recognition and Text-to-Speech (TTS) technologies for building voice-enabled intelligent applications. Adept at designing, fine-tuning, and serving Large Language Models (LLMs) across a wide range of industry use cases.

In 2025, I developed n8n automation skills, building AI-powered workflows for data processing, chatbots, and task automation. I mastered node configuration, API integrations, and workflow optimization to streamline business processes efficiently

WORK EXPERIENCE

AI/ML ENGINEER

(2021 - 2025)

ALGONLP | Onsite

- Developed and deployed LLM-based solutions (GPT, BERT, LLaMA, T5) for chatbots, document automation, and virtual assistants.
- Built RAG pipelines using FAISS, Pinecone, and LangChain for domain-specific applications.
- Created APIs with FastAPI and Flask, integrating models into production systems with low latency.
- Applied MLOps practices for scalable deployments on AWS, GCP, and Azure.
- Worked with speech technologies (ASR, TTS) to enhance AI-powered user interfaces.

EDUCATION

Bachelor of Commerce (B.Com)

Self-Taught AI/ML Engineer

Pursued independent learning through online resources, research papers, open-source projects, and hands-on experimentation.
Focused areas include:

- Large Language Models (LLMs)
- NLP & Deep Learning
- MLOps & Model Deployment
- RAG Pipelines & Vector Databases