

Muhammad Ahmad Mujtaba

AI Engineer | Deep Learning | NLP | MLOps

m.ahmad50519@gmail.com | +92 349 084 7790 (WhatsApp) | +92 329 586 3484 (Calls) | ahmad-mujtaba.vercel.app |

[linkedin.com/in/muhammad-ahmad-mujtaba](https://www.linkedin.com/in/muhammad-ahmad-mujtaba) | github.com/Ahmad-Mujtaba-09

PROFILE

AI engineer with three years of experience building production AI systems: LLM fine-tuning, RAG pipelines, and full-stack ML backends with complex optimization layers. Have worked as the sole engineer on end-to-end projects for international clients, handling everything from raw data to deployed services. Holds a B.Sc. in Artificial Intelligence.

TECHNICAL SKILLS

AI / ML: PyTorch | Hugging Face Transformers | LangChain | LightGBM | TensorFlow | YOLO | Stable Diffusion | Agno

Techniques: LLM Fine-tuning | LoRA / QLoRA | RAG | NLP | Computer Vision | Price Optimization | Time-Series (STL)

LLM APIs: llama.cpp | Ollama | OpenAI | Anthropic | Gemini | Azure AI

Backend & MLOps: Python | Java | JavaScript | FastAPI | Docker | MongoDB | Qdrant | Supabase | Git | DVC | GCP | DigitalOcean | Linux | n8n | Prefect

EXPERIENCE

Contract AI Engineer | HKS TechLabs | *Remote* | Feb 2026 - Present

- Act as the sole ML engineer building an AI pricing recommender from the ground up, managing the full pipeline from Supabase ingestion to SKU classification and STL decomposition.
- Developed demand forecasting models using LightGBM (Tweedie + quantile regression), incorporating elasticity simulations and a two-phase price optimizer (LP + SLSQP) with per-SKU explainability.

Contract AI Engineer | SaaS Startup (*Client name withheld under NDA*) | *Remote* | Jun 2025 - Jan 2026

- Built a taxonomy-driven RAG recommendation system using Qdrant; designed embeddings around semantic taxonomies instead of raw objects, cutting query times from 15-20s to 5-7s on short queries and 50-60s to 10-13s on larger ones.
- Migrated the same system from a monolithic setup (4-6 concurrent users) to a microservice architecture with async request handling, scaling to 40-50 concurrent users.
- Built end-to-end data pipelines, including web scrapers, LLM-based JSON cleaning/structuring, and entity extraction (skills, education, tasks) before embedding.

Junior AI Engineer | EnDevSols LLC | May 2024 - Jun 2025

- Designed and maintained automated training pipelines on DigitalOcean and physical GPU servers.
- Trained and deployed computer-vision models across object detection, classification, and medical imaging.
- Built and deployed NLP and generative AI models behind APIs, including LoRA/QLoRA fine-tuning workflows.

AI Engineer Intern | EnDevSols LLC | Jun 2023 - Apr 2024

- Trained and deployed models, handled data preprocessing, integrated vector databases, and built APIs with FastAPI.

EDUCATION

B.Sc. Artificial Intelligence, COMSATS University Islamabad

September 2022 - July 2026 | CGPA: 3.61 / 4.00