Islamabad, Pakistan
+92-3045146995
mohsinmahmood675@gmail.com

Mohsin Mehmood

linkedin.com/in/mohsin
github.com/mohsinm-dev

Machine Learning Engineer with 3+ years of experience designing and deploying **scalable ML and LLM-based systems**. Strong background in **model optimization, distributed inference, and ML infrastructure**, with production impact in healthcare and consumer applications. Active contributor to **Flax, CPython, and PyTorch** ecosystems.

Work Experience

Forward Deployed ML Engineer, Kodamai **Nov 2025 – Present**
Remote, UK

- Strengthened the core RAG pipeline for a **multilingual (Arabic + English)** platform, improving end-to-end reliability and query performance across document ingestion, retrieval, and inference.
- Designed and optimized a document ingestion pipeline handling both text-based and scanned PDFs, integrating PaddleOCR **(PP-OCRv5)** and multilingual preprocessing for robust **Arabic** and English OCR.
- Improved document structure extraction by combining Docling with VLM-based parsing approaches, reducing OCR-only dependency and increasing retrieval quality.
- Optimized **embedding** and **reranking** workflows to reduce latency and improve relevance, while operating under on-prem deployment constraints with careful GPU utilization and memory control.
- Identified and removed system bottlenecks across ingestion, vector search, and inference, resulting in faster end-to-end query times and more stable demos.
- Collaborated closely with backend and infra teams to align FastAPI services, Celery workers, queues, and databases, improving scalability and fault isolation.

Lead Data Scientist, Exora AI **July 2024 – Nov 2025**
Remote, Singapore

- Led the architecture of a **healthcare** GenAI platform, reducing LLM latency by **35%** and infrastructure cost by **25%** via context caching, token optimization, and SLM offloading (FastAPI, Redis, Qdrant, Docker, AWS).
- Built production RAG and triage pipelines used in 500+ patient sessions by 80+ doctors, improving diagnostic reliability with hybrid vector search, safety-filtered prompts, and SSE streaming.
- Delivered multimodal AI capabilities (**SigLIP + BLIP embeddings**), enabling visual diagnosis and multi-stage consultations, cutting LLM token usage by **40%.**
- Standardized PHI-compliant observability across 7 microservices, implementing structured logging, latency tracing, and automated PII/PHI redaction.

Senior Machine Learning Engineer, The Quell App **Jan 2024 – May 2024**
Remote, Canada

- Built a Dockerized GenAI MVP combining chat and personalized routines, directly contributing to $250K pre-seed funding.
- Evaluated and deployed open-source LLMs, selecting **Mistral-7B** for optimal quality–cost tradeoff and designing a self-hosted inference workflow with safety guards.
- Reduced prompt failure rate by ~30% through structured **A/B evaluations** and policy-enforced chat flows across FastAPI and Celery services.
- Partnered with product and design teams to translate user journeys into clear data flows and API contracts.

Machine Learning Engineer, DiveDeepAI **April 2023 – Jan 2024**
Islamabad, Pakistan

- Delivered an end-to-end e-commerce analytics pipeline processing ~5k products/week across 5 retailers with over 90% scraping accuracy (**Scrapy, Playwright**).
- Built price recommendation models (**XGBoost, scikit-learn**), achieving ~12% MAPE using competitor, demand, and attribute data.
- Deployed ML services on AWS EC2 with FastAPI, Redis, and Celery, maintaining ~99% uptime and introducing CI, testing, and model versioning.
- Designed indexed PostgreSQL schemas powering BI dashboards and cost-tracking workflows.

Machine Learning Engineer [Profile], **Upwork** **Jan 2022 – Present**
Remote, Pakistan

- Delivered **20+** ML and GenAI projects with a **100%** client success rate across NLP, computer vision, and generative modeling.
- Fine-tuned LLMs (GPT-3.5, BERT, T5) using LoRA, prompt tuning, and mixed-precision training, improving quality while reducing inference cost.
- Trained and deployed CV models (CNNs, YOLOv5), achieving >90% mAP on real-world datasets.
- Built end-to-end ML pipelines containerized with Docker and served via **FastAPI/Flask** for real-time inference.

Recent Open Source & Projects [Github]

- **Google / Flax** — **Top 5** contributor (last 12 months)
  Fixed core issues in NNX transforms, variable hooks, tabulation crashes, and documentation.
- **CPython** — Authored fixes across the interpreter and standard library (TextIOWrapper, Unicode handling, asyncio, threading docs).
- Contributor to PyTorch, CausalML, Astropy, and scientific ML libraries.

Education and Certifications

- **B.Sc. Computer Science,** Capital University of Science and Technology, Islamabad. **2019 – 2023**

Technologies and Languages

- **Languages:** Python, C++
- **ML & GenAI:** PyTorch, JAX, Flax, RAG, LoRA, Vector Search, LangChain, LangGraph
- **Backend & Systems:** FastAPI, Flask, Docker, Redis, RabbitMQ, PostgreSQL
- **Cloud & DevOps:** AWS, Azure, CI/CD, Model Versioning, Observability
- **Data & Infra:** Qdrant, SQL, Streaming APIs