# Noor Aftab

**Lead AI / MLOps Engineer | AI Platform Architect**
📧nooraftab779@gmail.com | 🌐https://www.nooraftab.info/
🔗LinkedIn: linkedin.com/in/noor-aftab-6a5535219 | 💻GitHub: github.com/nooraftab779

---

## Professional Summary

Lead AI / MLOps Engineer with over **4+ years of professional experience** designing, deploying, and operating **enterprise-grade AI platforms** in production environments. Strong background in **LLM systems, computer vision, OCR pipelines, and Kubernetes-based MLOps**, with hands-on responsibility for **architecture design, performance optimization, cost control, and production stability**. Proven ability to scale AI services to **100,000+ daily requests**, reduce infrastructure costs, and deliver reliable systems in **regulated and compliance-driven domains**. Experienced in working with distributed, remote-first engineering teams and complex stakeholder environments.

---

## Core Competencies

**Artificial Intelligence & Machine Learning**
Computer Vision (YOLO, Mask R-CNN), OCR (PaddleOCR, Tesseract), Vision-Language Models, Large Language Models (GPT, Gemini, Claude, Qwen, LLaMA), vLLM, ONNX Runtime, model evaluation on live traffic

**MLOps & Platform Engineering**
Kubernetes (multi-cluster environments), Docker, CI/CD pipelines, GPU & CPU mixed workloads, inference optimization, model lifecycle management, deployment automation

**Backend & Data Engineering**
Python, FastAPI, Cython, RESTful APIs, Elasticsearch, MongoDB, PostgreSQL, data pipelines

**Cloud & Infrastructure**
Azure Functions, AWS Lambda, on-premise deployments, hybrid cloud architectures, serverless systems

**Observability & Operations**
Elastic Stack, centralized logging, monitoring dashboards, alerting, performance analysis, cost optimization

---

# Professional Experience

### Lead AI / ML Engineer – MLOps & Platform Architecture

**ShuftiPro | Remote**
*November 2022 – Present*

- Lead the design and operation of a distributed AI platform for **KYC, identity verification, and document intelligence**, processing **100,000+ verification requests per day** across **60+ Kubernetes clusters**.
- Architected end-to-end AI pipelines integrating **OCR, computer vision models, and LLM-based semantic extraction**, designed with confidence-based fallback and validation mechanisms.
- Optimized end-to-end inference latency from **\~10 seconds to \~4 seconds** by refactoring processing pipelines, optimizing model execution paths, and tuning infrastructure.
- Re-architected LLM inference using **vLLM-based shared GPU services**, reducing GPU consumption from **60+ dedicated GPUs to 3–5 shared GPUs**.
- Achieved significant infrastructure cost reduction by lowering monthly operational expenses from **approximately USD 20,000 to USD 7,500**.
- Designed and maintained **100+ production microservices**, deployed using automated CI/CD pipelines with rolling and canary strategies to ensure **zero downtime**.
- Implemented centralized logging, request tracing, and encrypted data handling using Elasticsearch to meet **audit and compliance requirements**.
- Acted as technical owner for production systems, performing architecture reviews, code reviews, incident analysis, and cross-team coordination.

### Python / Machine Learning Engineer

**Machine Learning 1 | Lahore**
*January 2022 – October 2022*

- Developed and deployed OCR-based document processing pipelines with extensive preprocessing and post-processing logic to improve accuracy and robustness.
- Built ML-backed REST APIs using Flask and FastAPI to serve real-time inference workloads.
- Automated model training workflows, including dataset preparation, annotation coordination, and validation processes.
- Improved OCR accuracy, throughput, and system reliability across multiple document types used in production.

---

# Selected Technical Projects

### Enterprise AI & MLOps Platform – KYC & Document Intelligence

- Large-scale AI platform combining OCR, computer vision, and LLM inference under strict latency, security, and compliance constraints.
- Multi-region Kubernetes deployment with GPU-aware scheduling and encrypted data pipelines.
- Continuous monitoring and evaluation of models on live production traffic.

### AI-Powered Multimodal Product Categorization (Qubyk)

- Fine-tuned **IDEFICS-9B Vision-Language Model** using **LoRA and 4-bit quantization**, enabling training of large models on limited GPU resources.
- Designed event-driven training and inference pipelines using message queues and asynchronous processing.
- Automated hierarchical product categorization across **7 main categories and 40+ subcategories** for an e-commerce platform.

### ASX Announcement Intelligence Pipeline

- Designed and implemented a serverless ETL pipeline using **Azure Functions** for real-time processing of financial announcements.
- Integrated **Claude 3 Haiku** for intelligent PDF parsing and structured data extraction.
- Reduced manual processing effort by approximately **95%**.

### Shein Product Intelligence & Modesty Classification System

- Built a production-grade scraping and ETL pipeline using Selenium Stealth and rotating proxies.
- Developed a hybrid computer vision stack combining **NudeNet, Detectron2, and DensePose** for automated modesty classification.
- Automated approximately **95% of manual review workload** through AI-driven classification.

### Real Estate Crawling & EPC OCR System (UK)

- Implemented nationwide crawling across Rightmove, Zoopla, and OnTheMarket.
- Built YOLO-based detection and OCR pipelines for EPC rating extraction and validation.
- Enabled long-term price tracking and sustainability analytics for real-estate intelligence.

---

## Education

**Bachelor of Science in Computer Science (BSCS)**
COMSATS University, Lahore — 2021

---

## Certifications

- IBM Data Engineering Professional Certificate — Coursera
- Continuous Integration & Continuous Delivery (CI/CD) — Coursera
- Introduction to Containers with Docker & Kubernetes — Coursera

---

## Languages

- English — Professional proficiency
- Urdu — Native

---

## Availability

Open to **Germany-based remote or hybrid roles**
Senior / Lead AI Engineer • MLOps Engineer • Applied Machine Learning Engineer