Noor Aftab

nooraftab779@gmail.com | +92 331 4816207 | nooraftab.info | linkedin.com/in/noor-aftab-6a5535219 | github.com/nooraftab779

PROFESSIONAL SUMMARY

MLOps / ML Platform Engineer with 4+ years of experience building and operating production-grade AI platforms at scale. Proven track record in end-to-end MLOps pipeline design, GPU infrastructure management, and inference optimization (vLLM, TensorRT, NVIDIA Triton, ONNX Runtime). Orchestrated model serving across 60+ Kubernetes clusters processing 100,000+ daily requests. Achieved 60% infrastructure cost reduction through GPU consolidation and shared inference architectures. Deep expertise in containerized deployments with Terraform-managed infrastructure, automated CI/CD, agentic LLM tooling, and production observability using ELK and Grafana.

TECHNICAL SKILLS

MLOps & Platform: Kubernetes (60+ multi-cluster), Docker, CI/CD (rolling & canary), GPU-aware scheduling, model registry & versioning, model lifecycle management, reproducible ML pipelines

GPU & Inference: vLLM (PagedAttention, continuous batching), TensorRT, NVIDIA Triton Inference Server, ONNX Runtime, model quantization (FP16, 4-bit, LoRA), VRAM optimization, batch tuning

AI/ML: Computer Vision (YOLO, Mask R-CNN, Detectron2, DensePose), OCR (PaddleOCR, Tesseract), Vision-Language Models, LLMs (GPT, Gemini, Claude, Qwen, LLaMA), live model evaluation

Agentic AI: Multi-agent orchestration, LLM function/tool calling, Vertex AI, typed tool registries with allow-lists, NL-to-SQL pipelines, retrieval grounding, sandboxed code execution

Backend & Data: Python, FastAPI (REST/gRPC), Cython, Elasticsearch, MongoDB, PostgreSQL, message queues, event-driven architectures, async pipelines

Cloud & IaC: AWS (Lambda, EKS, ECR), Azure Functions, Terraform, hybrid cloud, serverless, on-premise deployments

Observability: Elastic Stack (Elasticsearch, Filebeat, Kibana), Grafana, Prometheus, centralized logging, request tracing, alerting, cost-per-inference tracking

PROFESSIONAL EXPERIENCE

Lead AI / ML Engineer — MLOps & Platform Architecture

Nov 2022 – Present

ShuftiPro | *Remote*

- Lead the design and operation of a distributed AI platform for KYC, identity verification, and document intelligence, processing 100,000+ requests/day across 60+ Kubernetes clusters with GPU-aware scheduling.
- Architected end-to-end MLOps pipelines — data ingestion, model inference (OCR, computer vision, LLM extraction), evaluation, and monitoring — with confidence-based fallback and validation mechanisms.
- Re-architected LLM inference using vLLM-based shared GPU services with continuous batching and optimized VRAM allocation, reducing GPU consumption from 60+ dedicated GPUs to 3–5 shared GPUs.
- Achieved 60% infrastructure cost reduction, lowering monthly GPU/compute expenses from USD 20,000 to USD 7,500 through GPU consolidation, resource right-sizing, and workload optimization.
- Optimized end-to-end inference latency from ~10s to ~4s (60% reduction) by refactoring pipelines, tuning model execution paths with TensorRT and ONNX Runtime, and optimizing GPU/CPU resource allocation.
- Designed and maintained 100+ production microservices using FastAPI, deployed with Docker on Kubernetes via Terraform-managed infrastructure, with automated CI/CD, rolling updates, and canary strategies.
- Built agentic LLM tooling on Vertex AI that automates deployment orchestration, merge-request code review, and production latency investigation across multi-environment clusters.
- Implemented production observability using Elastic Stack (Filebeat, Kibana) and Grafana dashboards — request tracing, structured logging, performance metrics, and alerting for audit and compliance.
- Deployed and tuned models via NVIDIA Triton Inference Server and vLLM for production-grade latency and throughput across vision and LLM workloads.
- Acted as technical owner for production systems — architecture reviews, code reviews, incident analysis, post-mortems, and cross-team coordination with ML engineers and data scientists.

Python / Machine Learning Engineer

Jan 2022 – Oct 2022

Machine Learning 1 | *Lahore*

- Developed and deployed OCR-based document processing pipelines with preprocessing and post-processing logic to improve accuracy and robustness in production.
- Built ML-backed REST APIs using Flask and FastAPI to serve real-time inference workloads with low-latency requirements.
- Automated model training workflows including dataset preparation, annotation coordination, experiment tracking, and validation processes.

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, Triton-based model serving, encrypted data pipelines, and automated model versioning.
- Continuous monitoring and evaluation of models on live production traffic with drift detection and automated alerting via Grafana and ELK.

Agentic ML Deployment Control Plane

- Architected an agentic MLOps control plane wrapping kubectl, AWS ECR, MongoDB, Rancher, and GitLab as a typed tool registry with audited dispatch and per-agent allow-lists.
- Built a conductor pipeline (plan → validate → execute → review → rollback) with autonomy levels, kill-switch enforcement, and canary chunked rollouts with inter-chunk health gates.
- Shipped AI GitLab merge-request review with large-file recovery and coverage reporting; cut MongoDB counter query load ~30x by retuning poll cadence and lookback windows.

Multi-Agent OCR & Document Verification Platform

- Built a Supervisor-led multi-agent system routing OCR triage, document-type classification, NER extraction, and request-ID investigations to specialist agent teams over typed tool interfaces.
- Enforced tool-grounded provenance rules that eliminated fabricated MRZ/OCR/NER fields, and reduced representative visual-compare sessions from ~385s to ~57s (~6x faster).
- Consolidated OCR, DocType, NER, and SPMS investigation onto one conversational surface covering 2,600+ document types across 200+ countries.

AI-Powered Multimodal Product Categorization (Qubyk)

- Fine-tuned IDEFICS-9B Vision-Language Model using LoRA and 4-bit quantization, reducing VRAM requirements by ~75% and enabling training on limited GPU resources.
- Designed event-driven inference pipelines using message queues and async processing for decoupled, scalable GPU workloads.

ASX Announcement Intelligence Pipeline

- Serverless ETL pipeline using Azure Functions for real-time financial announcement processing; integrated Claude 3 Haiku for PDF parsing. Reduced manual effort by ~95%.

Shein Product Intelligence & Modesty Classification

- Hybrid computer vision stack (NudeNet, Detectron2, DensePose) for automated modesty classification. Automated ~95% of manual review workload.

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