

Nadeem Yousaf

Senior ML Engineer | Agentic AI, LLM Orchestration & Production ML Systems | Computer Vision, NLP & Deep Learning

Lahore, Punjab, Pakistan • nadeem.yousaf.4444@gmail.com • 0332 4621172 • linkedin.com/in/nadeem-yousaf-

Open to senior / lead remote roles (US & EU) and consulting engagements — production AI, LLMops & AI strategy.

Last updated: June 2026

SUMMARY

Senior ML Engineer with 6+ years across Agentic AI, LLM orchestration, Computer Vision, NLP, and Deep Learning. Gold medalist (BS) and Silver medalist (MS in AI/ML) with published research in *Computers in Biology and Medicine* (IF 7.7). Currently sole ML engineer building safety-critical production AI for a mental health platform — 12 systems shipped from zero infrastructure on a single GPU. Expert in production systems including multi-agent architectures (LangGraph, LangChain, CrewAI, N8N), LLM serving & MLOps (vLLM, OpenAI-compatible APIs, evaluation harnesses), RAG optimization, fine-tuning, and scalable deployments using GPT, Llama, Claude, Mistral, and AWS Bedrock. Specialized in taking GenAI from demo to production — with reliability, cost control, and safety built in — across healthcare, security, finance, and marketing.

30+

AI/ML systems & projects shipped
(6 yrs)

~5.4×

cheaper inference via cost analysis

+183%

lead conversion (12% → 34%)

94%

faster reporting (12h → 45min)

92%

accuracy in enterprise legal Q&A

EXPERIENCE

Technology Digest, Inc. — Senior Machine Learning Engineer Feb 2026 – present · 5 mos
Maryland, United States · Remote

- Sole ML engineer on a safety-critical mental health AI platform; built **12 production AI systems from zero infrastructure** on a single GPU, spanning conversational AI, real-time safety, clinical documentation, and search.
- Designed and deployed an LLM serving + MLOps stack (vLLM, OpenAI-compatible API, token streaming, session memory & automatic context compaction) behind a stable contract layer that made model swaps invisible to downstream teams.
- Built a **multi-tier real-time content-safety & crisis-detection system** (regex + lightweight classifier + LLM cascade) resolving >99% of traffic in under 100 ms at ~99.7% accuracy, with multilingual crisis detection.
- Led adversarial safety evaluation (200+ scripted clinical-QA turns across scenarios) and built permanent regression harnesses; migrated intent routing from 40+ keyword rules to an LLM classifier covering 22+ intents in any language.
- Delivered structured clinical-note generation (JSON-schema output, WHO/NICE-aligned), offline multilingual semantic transcript search (no external vector DB), and an OCR document-intake pipeline.
- Authored a build-vs-buy cost analysis from real token budgets, keeping self-hosted inference **~5.4× cheaper** than an external API at scale; informed the infrastructure decision.

TransData — Senior Machine Learning Engineer Sept 2023 – Jan 2026 · 2 yrs 5 mos
Lahore, Pakistan

- Architected multi-agent systems with LangGraph and LangChain orchestrating GPT-4, Llama 2/3, Claude 3.5, and Mistral for enterprise automation, achieving **92% accuracy** in legal Q&A and reducing support tickets by **40%**.
- Developed a production LLM chatbot with OAuth2/Okta integrating Jira, Databricks, and SharePoint via RAG (ChromaDB, contextual compression), achieving ~8-second response time and **89% user satisfaction**.
- Implemented AI cold-outreach automation with N8N processing Apollo CSVs, intelligent lead scoring, and GPT-4-powered personalized emails, increasing conversion from **12% to 34%** and driving **2.3× deal closure**.
- Designed question-answering systems with document loaders (PDFPlumber, PDFMiner), optimized vector embeddings (ChromaDB, Faiss, Pinecone), and multi-stage verification agents.
- Created text-to-image APIs with Stable Diffusion (SSD-1B, OpenJourney) on Paperspace GPU infrastructure (FastAPI, Docker); built auto-labeling pipelines for YOLO training and parallel image-captioning APIs (BLIP, GIT, BLIP-2).
- Optimized LLM inference with Llama CPP 4-bit quantization, reducing memory by **75%**.

TechnoGenics SMC PVT LTD — Software Engineer (Data/ML) Jan 2023 – Oct 2023 · 10 mos
Lahore, Pakistan

- Developed production RESTful APIs with Flask and Django implementing authentication.
- Managed NoSQL infrastructure (Redis, Elasticsearch) and built real-time data streaming with Apache Kafka.
- Deployed applications on GCP (Compute Engine, Cloud Storage, GKE, BigQuery) with automated CI/CD pipelines.
- Contributed to cybersecurity pipelines processing **5M+ daily events** with ML threat detection (Random Forests, Gradient Boosting) and anomaly detection (Isolation Forests).
- Conducted statistical analysis with the Python stack (Pandas, NumPy, Scikit-learn), generating automated reports (Matplotlib, Plotly).

Information Technology University — Research Associate Apr 2021 – Jan 2023 · 1 yr 10 mos
Lahore, Pakistan

- Implemented predictive modeling for crime-location analysis using LSTM networks, geospatial analysis, and graph-based models.
- Engineered smart surveillance (Nazra project): real-time human counting, face recognition, and behavioral analysis on edge devices.
- Applied object detection (YOLOV5-7, YOLOX, Detectron2) and tracking (ByteTrack, FairMOT, CenterTrack) achieving **30+ FPS**.
- Developed alerting for boundary crossing, abandoned objects, loitering, and person re-identification with deep metric learning.
- Implemented knowledge graphs for disease-outbreak prediction (DeepWalk, Node2Vec, MetaPath2Vec) and a face-recognition pipeline (YOLO v5-v7, DSFD low-light, fine-tuned FaceNet).

Information Technology University — Teaching Assistant & Graduate Research Assistant Sept 2019 – June 2021
Lahore, Pakistan

- Assisted graduate Advanced Algorithms Analysis and undergraduate Theory of Automata courses, supporting 150+ students.

- Developed face-parsing modules with semantic segmentation (BiseNet); modified VGGFace/FaceNet with custom attention achieving **22.4% improvement** on VIP-attribute.
- Engineered custom ranking-based loss functions for BMI estimation and satellite house counting, reducing outliers by **40%** vs MSE loss; implemented Grad-CAM saliency and t-SNE analysis.

EDUCATION

Master of Science, Computer Science — Information Technology University

Sept 2018 – June 2020

- CGPA 3.83/4.00, Silver Medalist (Summa Cum Laude), AI & Machine Learning specialization; Graduate Student Fellowship.
- Thesis: *Face-to-BMI Prediction using Semantic Segmentation based Region-Aware Pooling* (published in *Computers in Biology and Medicine*, IF 7.7).

Bachelor of Science, Computer Science — University of Sargodha

Sept 2014 – June 2018

- CGPA 3.6/4.00, Gold Medalist & Valedictorian. Thesis: *Artificially Intelligent Digital Assistant (AIDA)*.

SELECTED HONORS & CERTIFICATIONS

- Silver Medalist, MS CS (3.83 CGPA, Summa Cum Laude, 2020), ITU.
 - Gold Medalist & Valedictorian, BS CS (3.6 CGPA, 2018), University of Sargodha.
 - Published in *Computers in Biology and Medicine* (IF 7.7, 2021).
 - Graduate Student Fellowship (2019–2020), ITU.
- Punjab Educational Endowment Fund Scholarship; Fauji Foundation Scholarship.
 - Neural Networks & Deep Learning; Convolutional Neural Networks — DeepLearning.AI.
 - Deep Learning with TensorFlow; Deep Learning with Keras — IBM.

SELECTED PROJECTS

Production AI Platform for a Mental Health Service (Safety-Critical)

Technology Digest · 2026 – present

Sole ML engineer; 12 production systems on one GPU. LLM serving (vLLM, OpenAI-compatible API, streaming, context compaction); multi-tier real-time safety & multilingual crisis detection (>99% resolved <100 ms, ~99.7% accuracy); structured clinical-note generation; offline multilingual semantic search; adversarial safety evaluation with permanent regression harnesses; LLM-classifier intent routing (22+ intents, any language); build-vs-buy cost analysis (~5.4× cheaper self-hosted).

LinkedIn Content Automation Platform with Multi-Agent AI

TransData · Mar 2024 – present

Django REST backend + Android frontend with persona-based LLM fine-tuning; N8N multi-agent workflow for idea generation, drafting, quality verification, and scheduling (50+ posts/week); trend-aware tone/style transfer via TF-IDF + embedding similarity (95% brand consistency); PostgreSQL + Redis with monitoring dashboards.

AI Cold-Outreach Automation System

TransData · Jan 2024 – June 2024

N8N workflow with Apollo CSV processing, dedup/validation, custom ICP scoring, and GPT-4 few-shot personalized emails; React + Material-UI admin dashboard for 1000+ leads with CRM integration. Lifted qualified-lead conversion 12% → 34% (+183%), cut manual qualification 90%, drove 2.3× deal closure.

Multi-Agent Legal Q&A with Verification Workflow

TransData · Jan 2024 – Mar 2024

CrewAI + LangGraph workflow with generation (GPT-4), cross-verification (Claude 3.5), and source-attribution agents; legal-optimized chunking with Faiss; Dockerized on AWS EC2 with user-configurable LLM backends and SOC 2 compliance.

Agent-Based BI Report Generation (Marketing & Finance)

TransData · Nov 2024 – present

LangGraph workflow with specialized extraction/analysis/insight/compilation agents across CSV/PDF/Word/Excel; dynamic charts (Matplotlib, Plotly); Dockerized on AWS EC2 with CloudWatch and email delivery. Cut report generation 12 hrs → 45 min (94%), enabling 3× capacity growth.

TECHNICAL SKILLS

Agentic AI & LLM Orchestration: LangGraph, LangChain, CrewAI, multi-agent systems, GPT-4/3.5, Claude 3.5 Sonnet/Opus, Llama 2/3 (7B–70B), Mistral 7B, Mixtral 8x7B, Qwen, AWS Bedrock, prompt engineering, few-shot, chain-of-thought.

LLM Serving & MLOps: vLLM, OpenAI-compatible APIs, token streaming (SSE), context compaction, evaluation harnesses, groundedness/faithfulness metrics, observability & quality-drift monitoring, p95 latency & cost-per-task, build-vs-buy cost modeling.

Machine Learning & Deep Learning: PyTorch, TensorFlow, Keras, Scikit-learn, Random Forests, Gradient Boosting, SVM, CNN, RNN, LSTM, GRU, Transformers, attention, transfer learning, fine-tuning, quantization (4/8-bit), custom loss functions.

NLP & Retrieval: RAG, ChromaDB, Faiss, Pinecone, Qdrant, Weaviate, embeddings (OpenAI, HuggingFace, BGE/Sentence Transformers), contextual compression, MMR, BM25, TF-IDF, NER, BERT, RoBERTa, multilingual NLP.

Enterprise LLM Chatbot with Jira & Databricks Integration

TransData · Apr 2024 – July 2024

OAuth2 + Okta SSO with RBAC, session management, and audit logging; integrated Jira REST API, Databricks SQL, SharePoint, and document loaders; RAG pipeline (ChromaDB, OpenAI embeddings, contextual compression) with conversational memory.

Automated Form-Filling with AI Data Extraction

TransData · Feb 2024 – Mar 2024

Users upload documents with many questions; the system extracts questions, answers them via vector storage + user-chosen LLMs, generates charts, and stitches answers/charts back into the original form order. Dockerized on AWS EC2.

Text-to-Image Generation APIs for Product Imagery

TransData · Apr 2024 – May 2024

FastAPI microservices for product imagery with Stable Diffusion (SSD-1B, OpenJourney); Dockerized on Paperspace GPU (A100), ~3–5s generation for 1024×1024 images. Plus parallel image-captioning APIs (BLIP, GIT, BLIP-2) cutting processing 8s → 2.5s/image.

Nazra — AI Smart-Surveillance for Crime Prevention

Information Technology University · Apr 2021 – Nov 2022

Real-time detection/tracking (YOLO v3-v7, Detectron2, ByteTrack, FairMOT) at 30+ FPS on edge devices; face recognition (YOLO + DSFD low-light + fine-tuned FaceNet); parking management (YOLO v7, ~99% accuracy); boundary/loitering detection; anomaly detection (Anomalib); night-to-day (EnlightenGAN); OCR; super-resolution.

Face-to-BMI Prediction via Region-Aware Pooling (Publication)

ITU · Computers in Biology and Medicine, 2021

Semantic-segmentation-based region-aware pooling (Reg-GAP) on FaceNet/VGG-Face features for BMI estimation; 22.4% improvement on VIP-attribute. arxiv.org/pdf/2104.04733.pdf

Computer Vision: YOLO (v3-v8), Detectron2, Faster/Mask R-CNN, tracking (ByteTrack, FairMOT, DeepSORT, CenterTrack), face recognition (FaceNet, ArcFace, DSFD), segmentation (BiseNet), BLIP/GIT/BLIP-2, OCR (Tesseract, EasyOCR), super-resolution (EDSR, FSRCNN), OpenCV.

Backend & APIs: Python, FastAPI, Django, DRF, Flask, Node.js, REST, OAuth2, JWT, WebSockets, microservices, Celery, RabbitMQ, Redis Queue.

Databases & Data Eng: PostgreSQL, MySQL, MongoDB, Redis, Elasticsearch, Cassandra, vector DBs, Apache Kafka ETL, data modeling, query optimization.

Cloud & DevOps: Docker, Kubernetes, AWS (EC2, Lambda, S3, Bedrock, RDS, CloudWatch), GCP, Paperspace, NGINX, CI/CD, SonarQube.

Automation & GenAI: N8N, Make.com, Zapier, web scraping (BeautifulSoup, Scrapy, Selenium); Stable Diffusion (SDXL, SSD-1B, OpenJourney), image captioning (BLIP, GIT), LoRA.