Moiz Ahmed

AI/ML Engineer (Gen AI, LLM, NLP)

(512) 595-3050 • moizahmed.swe@gmail.com • Iselin, NJ 08830

Professional Summary

Senior AI/ML Engineer with 10+ years of hands-on experience in Machine Learning, NLP, and Generative AI including designing ML systems, building scalable Python APIs, and deploying models and agents with robust MLOps pipelines. Combines deep AI infrastructure knowledge with full-stack capabilities to support cross-functional AI teams. Skilled in backend systems using FastAPI and Flask, and frontend dashboards using React, Tailwind, and Next.js for real-time model feedback. Proven success in leading model lifecycle automation, edge deployment, and cloud-native GenAI infrastructure. Strong collaborator with cross-functional teams, product stakeholders, and engineering leaders. Trusted technical advisor with success in delivering production-ready AI solutions that accelerate business innovation and reduce time-to-market.

Core Skills

Languages & Frameworks: Python, Java, JavaScript, TypeScript, SQL, React, Next.js, Tailwind
Machine Learning & Deep Learning: TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, LightGBM, CatBoost, H2O.ai, CNN, RNN, LSTM, Transformers, GANs, OpenAI gym, CrewAI, Autogen, Foundation Models(DALLE, CLIP, Stable Diffusion)
Natural Language Processing & LLMs: GPT-4, GPT-3.5, BERT, RoBERTa, T5, FLAN-T5, LLaMA, MPT, Falcon, Sentence Transformers, Hugging Face Transformers, LangChain, Langgraph, LlamaIndex, SpaCy, NLTK, Word2Vec, GloVe, FastText
Generative AI & RAG Pipelines: LangChain, Agentic Workflows, Ollama, Prompt Engineering, Function Calling, Tool Use, Multi-Agent Systems, Chain-of-Thought Reasoning, Knowledge Graphs, Fine-tuning (QLoRA, LoRA, PEFT), Retrieval-Augmented Generation (RAG)
Vector Databases & Search: Pinecone, Weaviate, Milvus, FAISS, ElasticSearch, Azure Cognitive Search, Google Matching Engine
Data Engineering & Big Data: Apache Spark, Databricks, Hadoop, Kafka, Flink, Dask, Hive, HDFS
Databases & Storage: PostgreSQL, MySQL, Redis, Dynamo DB, MongoDB
Model Evaluation & Explainability: SHAP, LIME, Captum, What-If Tool, Fairness Indicators, CrossValidation, A/B Testing
MLOps & Cloud: Docker, Kubernetes, MLflow, Airflow, Kubeflow, GitHub Actions, AWS (SageMaker, Lambda, Bedrock, Comprehend, Transcribe, Glue, S3, Step, SNS, SQA, etc.), GCP, Runpod, Terraform, Saalad Cloud
Data Processing & Feature Engineering: Pandas, NumPy, OpenCV, FeatureTools, Synthetic Data Generation
Visualization & Dashboards: Matplotlib, Seaborn, Plotly, Dash, Streamlit, Altair, Bokeh, ggplot
AutoML & Tuning: Google AutoML, Azure AutoML, AutoKeras, H2O AutoML, TPOT, Optuna, Ray Tune, GridSearchCV
Dev Tools & IDEs: Jupyter, VS Code, PyCharm, RStudio, Colab, Git, GitHub, Bitbucket
Backend Tools: FastAPI, Flask, Django, REST APIs, WebSockets, Middlewares, Auth & Security encryption
Expertise: LLM Infrastructure, Backend API Development, Hybrid Cloud-Edge MLOps, DRL, CV/NLP Systems, GenAI/Agentic AI product

Professional Experience

Senior Full Stack Machine Learning Engineer (LLM Infra + Platform)

Trace Machina | Sep 2023 – Present

- Architected and led the development of ECOps, a comprehensive AI training and deployment platform that supports data analysis, distributed model training across heterogeneous edge devices, cross-site model sharing, deployment, performance monitoring, and drift detection.
- Achieved a **10% reduction in model training and deployment time** and contributed to **8% improvement in energy efficiency** across operational sites through automated retraining and optimized model rollout.
- Designed and deployed scalable FastAPI microservices to serve LLM inference, including JWT-secured endpoints, async I/O, and streaming completions for long context responses.
- Architected and maintained a modular RAG pipeline using LangChain, Pinecone, and custom retriever logic with metadata filters and hybrid search across vector + keyword indices.
- Developed a monitoring dashboard using React + Tailwind + Next.js to visualize model performance, latency distribution, version tracking, and real-time token streaming.
- Tuned vector database (Pinecone, FAISS) configurations using ANN and quantization techniques, achieving 25–35% latency reduction and improved semantic recall.
- Integrated observability with Grafana and Prometheus to track retrieval latency, embedding drift, and vector store query health.
- Led MLOps workflows in Databricks and MLflow for the continuous training, evaluation, and production rollout of domain-specific models.
- Employed prompt injection guards, fallback routing, and response throttling to ensure security and resilience of GenAI endpoints under load.
- Automated test suites using Pytest and integrated model evaluation metrics (BLEU, cosine similarity, groundedness checks) into CI pipelines.

Senior Machine Learning Engineer

Amazon— Remote | Aug 2021 – Sep 2023

- Redesign AI methdology from machine learning to Deep Reinforcement Learning (DRL) using **PPO** that results in additional **10% savings in energy** for buildings.
- Developed a **Gen AI-based chatbot** feature, fine-tune the **Qwen/QwQ-32B** model for voice and chat assistance related to building's HVAC behavior and fault diagnositics this results in boosting sales by **13%.**
- Designed a hybrid MLOps solution using AWS services, enabling initial training on SageMaker and subsequent training on Jetson Nano edge devices resulting in reduce **release cycle time by 30%** and **site visits by 40%.**
- Implemented a feedback system using AWS SageMaker for continuous model performance monitoring and retraining, ensuring sustained model accuracy and this results in additional **energy saving of 4.5%** due to accurate predictions.
- Designed and managed scalable FastAPI-based ML model APIs with JWT authentication and integrated Redis caching, ensuring secure and low-latency inference pipelines.
- Architected MLOps workflows using MLflow, SageMaker to support model registry, drift detection, automated retraining, and continuous deployment on AWS ECS.
- Led the development of a React monitoring dashboard to visualize model versions, metrics, drift thresholds, and real-time alerts.
- Optimized the deployment lifecycle by 30% through templated Helm stacks, GitHub Actions CI/CD pipelines, and Dockerized GenAI microservices.
- Integrated RAG-based LLM endpoints using LangChain and Pinecone for retrieval-augmented medical diagnostics, reducing the average patient triage time by 20%.

Senior Software Engineer

Zammo.ai — San Jose, CA | Jul 2017 – Aug 2019

- Implemented MLOps pipeline for soil detection model, including a monitoring loop, staging and production account, saved 9 hours/month by minimizing downtime and automating retraining.
- Led the development of a motor bearing failure prediction system, achieving a 3.67% reduction in downtime through anomaly detection techniques.
- Streamlined CI/CD and ETL processes, enhancing data management and feature engineering, yielding a monthly savings of 10 hours.
- Engineered a solar GHI forecasting feature, resulting in daily savings of PKR 0.5 million via next-day operations.
- Accomplished cost savings of 156,429 PKR/MW/Month using a LSTM to predict cleaning routines and detect soiled panels.
- Reduced training time by 52% with detailed technical documentation and product demos for new customers and sales/marketing teams.

Software Engineer

Turn Hatch— Remote | Jan 2016 – Jun 2017

- Conducted competitive pricing analysis and data-backed bidding strategies for 5 solar energy projects, successfully securing contracts under market rates with a **maintained profit margin of 2.8%.**
- Identified inefficiencies in energy consumption patterns by leading comprehensive energy audits, **driving a 10% cost reduction** through data-driven insights and actionable recommendations.
- Delivered high-impact energy analyses that supported the optimal design of large-scale utility solar PV plants, contributing to an **annual performance ratio of 83%** through advanced data modeling and simulation tools.

Certifications

- AWS Certified Machine Learning Specialty — Udemy
- Generative AI with Vertex AI — Google
- TensorFlow Developer Certificate — ZTM
- Full-Stack React & FastAPI Bootcamp — Udemy
- YOLOv8 & Object Detection — Udemy
- MLOps with Kubernetes & MLflow — Udemy

Education

Bachelor of Science in Computer Science

Rutgers University–New Brunswick | 2012 – 2016 | New Brunswick, NJ