Moiz Ahmed

Iselin, NJ 08830 | moizwork15@gmail.com | moiz-wine.vercel.app | (510) 985-4236

Professional Summary

Results-driven **ML/MLOps Engineer** with 10+ years of experience building, deploying, and scaling ML/AI systems in multi-cloud environments (AWS, Azure, GCP). Expert in **MLOps pipelines, CI/CD, observability, infrastructure-as-code, container orchestration, and LLMOps**. Proven track record of reducing costs, improving system uptime, and accelerating ML lifecycle automation. Adept at integrating AI models into production workflows with measurable business outcomes.

Experience

Senior ML/MLOps Engineer Feb 2023 – Present

Oncotelic *Agoura Hills, CA*

- Architected and deployed a Generative AI-powered Clinical Exam Preparation and Simulation System using Azure OpenAI (GPT models) and AWS Lambda@Edge APIs, reducing editorial turnaround time by 40%.
- Delivered end-to-end MLOps consulting solutions across AWS SageMaker, Azure ML, and GCP Vertex AI, standardizing model lifecycle management (build, test, deploy) for text, vision, and multi-modal generative models.
- Ensured the safety and reliability of AI-driven systems in high-risk environments.
- Conducted security audits on LLM pipelines to identify and mitigate adversarial threats.
- Collaborated with engineering teams to align AI solutions with robust security protocols.
- Developed and deployed custom ML/LLM pipelines using Python and Docker, integrating model APIs into content workflows and analytics systems.
- Developed internal automation tools to accelerate operational workflows and task execution.
- Optimized prompt engineering, model versioning, and latency tuning strategies to improve user-facing applications powered by LLMs and foundation models.
- Defined and tracked SLIs/SLOs, reducing downtime and improving MTTR through structured incident response and automated rollback strategies.
- Designed post-mortem templates and facilitated blameless incident reviews; introduced DevOps maturity models and technical scorecards tailored to client transformation journeys.

Senior AI Software Engineer Oct 2020 – Apr 2023

InfoStack *Brooklyn, NY*

- Led development of Crane-GPT, an intelligent assistant to reduce crane operator workload and errors. Supported research and cross-functional teams to deliver customized ML-driven solutions.
- Helped balance service quality, cost, and delivery speed through intelligent system design.
- Collaborated with data science teams to productionize BERT-based NLU services using Ray Serve and Python, reducing inference latency by 30% through optimized batching and model caching strategies.
- Built and integrated Retrieval-Augmented Generation (RAG) pipelines for enterprise search use cases, leveraging Python, vector stores, and large language models for contextual grounding and real-time retrieval.
- Delivered a GCP-based containerization project for Richemont's digital stack, including secure CI/CD pipelines and container security policies.
- Piloted Vertex AI pipelines to automate retraining of product recommendation models, resulting in a 25% increase in relevancy during A/B testing phases.
- Worked with client architects to improve PostgreSQL backup and restoration strategies; built dashboards to monitor uptime, latency, and reliability via Grafana and custom Python scripts.
- Integrated ML models into backend APIs and streaming pipelines using Python and gRPC, enabling real-time recommendations and smart routing logic.

Senior Full Stack Engineer Oct 2017 – Jul 2020

SlashNext *Pleasanton, CA*

- Designed and developed full-stack web applications using Python on the backend and React on the frontend, integrating data via GraphQL and REST APIs.
- Built and optimized PostgreSQL-backed services, focusing on efficient query design, indexing strategies, and schema versioning.
- Migrated enterprise workloads from GCP to Azure through multi phase delivery.

– Deployed scalable ECS-based infrastructure with automated CI/CD pipelines; contributed to disaster recovery processes and system validation scripts.

AI Software Engineer Aug 2015 – Sep 2017
Algo.ai *El Segundo, CA*

– Improved system performance and uptime by automating deployment and monitoring.
– Refactored and optimized data pipelines to enhance business continuity and reduce failures.
– Contributed to operational resilience through data engineering and infrastructure upgrades.

Education

Southern Arkansas University Aug 2014 – May 2015
Master's in Computer Science *Magnolia, AR*
Contributed to research on machine learning techniques for healthcare under Dr. Hong Cheng.

University of Texas at San Antonio (UTSA) Aug 2010 – May 2014
Bachelor's in Computer Engineering *San Antonio, TX*

Technical Skills

Languages: Python, C++, JavaScript, TypeScript, MATLAB, SQL
MLOps & Cloud: MLflow, Airflow, Kubeflow, AWS (SageMaker, ECS, EKS), Azure ML, GCP Vertex AI
Generative AI/LLMOps: OpenAI/Azure OpenAI, Hugging Face, LangChain, VectorDB, RAG Pipelines, Ray Serve, Diffusers
DevOps/Infra: Docker, Kubernetes, Terraform, CI/CD, Grafana, Prometheus
Databases: PostgreSQL, MySQL, MongoDB

Selected Projects

AI-Assisted Clinical Exam Preparation and Simulation System
Tech Stack: Python, Meditron LLM, LoRA, QLoRA, PEFT, Transformers

- Built a domain-specific model to support USMLE clinical reasoning tasks, integrating it into an automated MLOps pipeline for data preprocessing, fine-tuning, deployment, and monitoring.
- Reduced annotation workload by **40%** while maintaining **92%+ reasoning accuracy** through reproducible workflows, experiment tracking, and containerized deployment across environments.

CraneGPT: Industrial AI Chatbot
Tech Stack: LLaMA-2, LangChain, VectorDB, ReactJS

- Developed a multilingual industrial chatbot for crane operators with semantic search and context-aware responses, ensuring continuous integration and scalable deployment via containerized microservices.
- Implemented observability and feedback loops to monitor response quality, achieving **75% faster task completion** and **85% user satisfaction** while enabling continuous retraining and version control.

Intelligent Meeting Assistant
Tech Stack: STFT, PyTorch, pyannote-audio, scikit-learn, NoiseReduce

- Engineered a speech diarization and clustering pipeline with automated data preprocessing, noise reduction, and inference pipelines integrated into an MLOps workflow for reproducibility and scale.
- Designed containerized deployment and monitoring dashboards to achieve **92%+ segmentation accuracy**, with retraining pipelines ensuring robust performance across diverse meeting audio datasets.

Dialogue Summarization using LLMs
Tech Stack: Python, Meditron LLM, LoRA, QLoRA, PEFT, Transformers

- Fine-tuned open-source LLMs on dialogue corpora (SAMSum, MultiWOZ) using parameter-efficient training techniques integrated into automated ML pipelines for scalability and reproducibility.
- Applied LoRA/QLoRA in a CI/CD-enabled environment to accelerate experimentation and deployment, improving summary coherence by **19%** while ensuring seamless integration into production APIs.

NeuroGenesis: Synthetic Brain Tumor Scan Generation
Tech Stack: PyTorch, DCGAN, NumPy, Matplotlib

- Generated synthetic MRI images with DCGANs to augment training datasets, integrating generation workflows into an automated MLOps pipeline for dataset versioning and reproducible augmentation.
- Optimized adversarial training with monitoring and evaluation pipelines, improving diversity and realism to **90%+ radiologist-validated quality** while reducing data scarcity issues in ML pipelines.

Breast Cancer Disease Detection System
Tech Stack: Python, ANN, Keras, Pandas, Scikit-learn, Seaborn

- Developed and deployed an ANN-based cancer detection system using automated pipelines for preprocessing, training, validation, and deployment in containerized environments.
- Improved accuracy from **95% → 99%** through data augmentation and k-fold validation, with MLOps practices enabling continuous model evaluation, monitoring, and retraining.

Publications (Submitted)

- **Dual Model EEG Emotion Estimation with Saliency-Based Feature Fusion**. *Multi-agent and Grid Systems (under review)*
  - Developed a hybrid deep learning model combining ResNet50 (PSD features) and Transformer (DE features).
  - Applied saliency-based late fusion to enhance emotional state recognition, achieving **92%** (SEED-IV) and **91%** (DEAP).
  - Integrated robust EEG preprocessing workflows (band-pass filtering, downsampling) aligned with reproducible ML pipelines.
- **Multi-Branch Deep Learning Framework for Biometric Identification and Cognitive State Inference**. *eNeuro (under review)*
  - Proposed a transformer-based multi-branch model leveraging ERP, time-frequency, and spatial EEG features.
  - Focused on olfactory response signals as early biomarkers for Alzheimer's and MCI.
  - Achieved **87% classification accuracy** with a macro F1-score of **0.88**.
  - Demonstrated clinical viability of non-invasive EEG-based cognitive assessment within reproducible research pipelines.
- **Multimodal EEG-Based Classification of Alzheimer's and MCI**. *Brain-Apparatus Communication: A Journal of Bacomics (under review)*
  - Designed a multimodal transformer integrating ERP, time-frequency, and spatial EEG patterns.
  - Employed olfactory response features for early-stage Alzheimer's and MCI detection.
  - Reached **87% accuracy** with macro F1-score of **0.88**.
  - Highlighted clinical potential of automated EEG pipelines in cognitive health monitoring.
- **Physics-Informed Neural Network Framework for Physically Consistent Solar Irradiance Forecasting**. *Engineering Applications of AI (under review)*
  - Integrated solar physics PDE constraints and monotonicity priors into a PINN-based forecasting framework.
  - Forecasted Global Horizontal Irradiance (GHI) with an $\mathbf{R^2}$ **score of 0.989**.
  - Combined meteorological features with domain-specific physics for interpretable predictions.
  - Improved robustness and physical consistency compared to conventional deep learning approaches.