

Abdul Fahed Mohammad

(312) 725-8993 | PLAINFIELD, IL 60544 | abdulmohammad429@hotmail.com | [LinkedIn](#)

PROFESSIONAL SUMMARY

Senior **ML/MLOps Engineer** **Data** Scientist with over 10 years of experience designing and deploying production-grade AI/ML systems across industries. Proven expertise in **NLP**, **Computer Vision**, and **Generative AI applications**, including fine-tuning, quantization, and Retrieval-Augmented Generation (RAG) pipelines. Skilled in containerized deployment (Docker, Kubernetes), CI/CD, MLOps workflows, and real-time inference on AWS, GCP, and Azure. Strong leadership experience—currently leading a team of four engineers and delivering end-to-end data science and ML solutions from research to production.

EXPERIENCE

Senior Machine Learning Engineer

Sep 2023 – Present

Hyatt Hotels Corporation

Chicago, IL

- Engineered a synthetic data-generation pipeline for training and fine-tuning conversational models with dynamic, customer-driven inputs—resulting in a more natural, human-like chatbot experience. Deployed and quantized models on AWS SageMaker for real-time inference, reducing response latency by ~30%.
- Enhanced a multimodal chatbot using specialized agents for market trends retrieval, text-to-SQL conversion, graph visualizations, and contract generation. Implemented LangChain-based RAG, model fine-tuning (BERT, LLaMA 3.1), Whisper for ASR, and prompt engineering for a real estate use case on GCP and Azure.
- Built an interactive clothing-visualization pipeline for apparel customization using VLM feedback loops, ControlNet (segmentation via depth/edges/normals), Microsoft Florence2, Meta SAM, Stable Diffusion, LoRAs, and ComfyUI. Packaged services with Docker and deployed via Azure AI Studio.
- Designed and implemented an end-to-end ETL workflow ingesting CRM data (HubSpot, Salesforce) via Kafka, processed with DBT transformations, and persisted to a centralized data warehouse. Enabled downstream statistical analysis and real-time insights served through a chatbot interface.

Senior ML/MLOps Engineer

May 2022 – Aug 2023

Silicon Valley Bank

Remote

- 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.

MLOps/Machine Learning Engineer

Oct 2020 – Apr 2022

Western Union

Denver, CO

- Built a resume-screening application using spaCy NER and Sentence-Transformers for semantic matching, deployed as a Flask app. Improved deployment efficiency and throughput by ~70%.
- Implemented a Retrieval-Augmented Generation assistant with LangChain and Gradio for internal documentation and client interactions.
- Integrated facial detection and expression analysis using OpenCV for UX research and product features.
- Developed a phishing-detection pipeline extracting 35+ features from email headers and URLs; models deployed in Docker on AWS EC2 and achieved a ~25% increase in detection accuracy.

- Designed sentiment-analysis systems for customer reviews leveraging LLMs, increasing classification effectiveness by ~45%.
- Optimized pipelines and feature engineering workflows on AWS S3, Redshift, and SageMaker, reducing model training time by 35%.

Data Scientist

Jun 2013 – Dec 2018

Genpact

Danville, IL

- Delivered end-to-end machine learning solutions covering data generation, model training, fine-tuning, evaluation, experiment tracking, deployment, and iteration.
- Led projects extracting financial data from SEC filings and documents, including parsing of tables, headers, structures, and named entities.
- Built scalable ETL workflows with Airflow and Spark to process terabytes of structured and unstructured data from APIs, logs, and SQL/NoSQL databases.
- Designed NLP pipelines with spaCy and Hugging Face Transformers for entity extraction and classification, cutting manual review time by 40%.
- Managed the complete model lifecycle, including feature engineering, hyperparameter tuning, optimization, retraining, and deployment.
- Created interactive dashboards in Tableau and Plotly Dash to track KPIs and insights, reducing reporting cycles from weekly to daily.
- Built a proof-of-concept for a fully managed machine learning pipeline on Google Cloud Platform.
- Designed and executed A/B testing frameworks to measure the business impact of data science initiatives.
- Evaluated Google Cloud Vision API for image tagging and Natural Language API for sentiment analysis on reviews.
- Employed Splunk for log monitoring, dashboard creation, and real-time alerting.
- Acted as a key contributor in cross-functional teams with Data Scientists, Engineers, and Product Managers.
- Achieved a 20× speedup in document processing by integrating Apache Spark for distributed data handling.

EDUCATION

Valparaiso University

May 2019 – May 2020

Master's in Computer Science

Valparaiso, IN

University of Maryland - College Park (UMD)

Aug 2011 – May 2015

Bachelor's in Computer Science

College Park, MD

CERTIFICATIONS

- **AWS Certified Machine Learning - Specialty.** *Jul 2023 – Jul 2026*
- **AWS Certified Solutions Architect - Associate.** *Jul 2023 – Jul 2026*

TECHNICAL SKILLS

Languages & Frameworks: Python, Fast API, Flask, Django, C++, JavaScript, TypeScript, React, Node, Flutter, Unreal Engine

MLOps & Data: PyTorch, TensorFlow, Keras, Scikit-learn, Hugging Face, LLM fine-tuning, ONNX, DeepSpeed, Quantization, Pruning, Apache Spark, Splunk, Data loader, Kafka

MLOps & Cloud Docker, Kubernetes, AWS (SageMaker, Glue, EC2, Lambda, S3, CloudWatch, Bedrock), GCP (Vertex AI, BigQuery), Azure (Azure ML, Azure AI Studio, Azure AI Search, Azure OpenAI Service), MLflow, Airflow, DBT, Databricks, Terraform, CI/CD, Grafana, Prometheus, GitHub Actions, Jenkins, Circle CI, Ray Serve

Generative AI/LLMOps: OpenAI/Azure OpenAI, Hugging Face, LangChain, VectorDB, RAG Pipelines, Ray Serve, Diffusers

NLP & CV: spaCy, Sentence-Transformers, LangChain, LangGraph, LlamaIndex, Whisper, BLIP, CLIP, YOLO, SAM, ControlNet, Stable Diffusion, ComfyUI, Streamlit, Gradio, FFmpeg, Blender, OpenCV, Matplotlib, CNN, RNN, LSTM, GAN, ViT, U-Net, LoRA, QLoRA, PEFT, RLHF, DPO, NumPy, Pandas, Matplotlib, Seaborn, SQLAlchemy, XGBoost, Stable Diffusion, Multi-Agent Systems, Fine-tuning (QLoRA, LoRA, PEFT),

Databases: SQL, PostgreSQL, MongoDB, MySQL, Pinecone, VectorDBs, Redis, Firebase, SQLite, Elastic Search

Intelligent Data Analysis Platform

Tech: LangGraph, Streamlit, Pinecone, SQLite, Docker, React

- Delivered an end-to-end data platform with CSV ingestion, schema inference, Text-to-SQL LLM agent, and RAG-based automated insight generation.
- Implemented natural-language-driven charting and a chat orchestrator for conversational analytics.

AI Video Generator

Tech: FFmpeg, Flux, ComfyUI, MoviePy, Docker

- Built a microservice architecture to generate product advertisement videos from images, scripts, and meta-data.
- Employed BLIP + LLaMA for automated captioning and storyboarding.

FadeAway (Background Remover)

Tech: PyTorch, BiRefNet, Flask, Docker

- Implemented BiRefNet-based mask generation with Canny edge preprocessing to accelerate inference and improve blending; reduced inference time by **50%** with CPU/GPU optimized builds.

PhishGuard (Phishing Detection)

Tech: Flask, Docker, scikit-learn, AWS

- Developed browser extension + backend service to analyze email headers and URLs; deployed ML models on AWS and improved detection accuracy by **25%**.

Meet MSO (Conversational Game)

Tech: Whisper, PyTorch, Piper, TinyLlama, Unreal Engine

- Built a conversational game with emotion-aware responses, on-device STT (Whisper), TTS (Piper / XTTS-v2), and low-latency multithreaded interaction.

Resume Screening Platform

Tech: spaCy, Sentence-Transformers, Flask, Docker, React

- Designed resume parsing with weighted semantic scoring and a user-friendly interface for HR teams.

3D Mesh Conversion Tool

Tech: Blender Python API, Gradio, Docker

- Automates conversions between USD/FBX/OBJ/STL/GLTF/PLY formats with texture handling and packaging for distribution.