Musthaq Ahmed Gaffoor

Ph: +971 58 226 9953
Email: musthaq258@gmail.com
Dubai, United Arab Emirates

Professional Summary:

Machine Learning Engineer with 5+ years of experience in **MLOps, LLMOps, and Generative AI system design**, specializing in **cloud-native, scalable AI infrastructure**. Proven expertise in building and productionizing **LLM-based systems** (OpenAI, AWS Bedrock, Vertex AI) and architecting **end-to-end ML pipelines** for training, evaluation, and deployment. Skilled in **Kubernetes, CI/CD, and DevOps** workflows to deliver high-reliability AI services. Adept at bridging research and production, driving observability, governance, and responsible AI practices for enterprise-scale models.

Core Qualifications:

- **Languages & Frameworks:** Python, FastAPI, PyTorch, TensorFlow, LangChain, Hugging Face Transformers, vLLM
- **ML/LLMOps Tools:** MLflow, Vertex AI, Kubeflow, Ray, KServe, BentoML, Weights & Biases
- **Cloud Platforms:** GCP (Cloud Run, GKE, Vertex AI, BigQuery, Cloud Functions), AWS (Bedrock, Lambda, SageMaker)
- **Automation & Infrastructure:** Kubernetes, Terraform, Ansible, Docker, Airflow, Argo CD, CI/CD Pipelines
- **GenAI & LLMs:** Fine-Tuning, RAG Pipelines, Vector Databases, Model Versioning, Prompt Monitoring, LLM Observability
- **Data & Monitoring:** Feature Stores, Drift Detection, Cost & Latency Tracking, Audit Logging
- **Security & Governance:** IAM, Secrets Management, Data Encryption, Model Cards, Ethical AI Compliance
- **Soft Skills:** Analytical Thinking, Technical Documentation, Cross-Functional Collaboration, Mentorship

Professional Experience:

Teaching Assistant *(Sept 2022 - Oct 2025) - Dublin City University, Dublin, Ireland*

- Architected **end-to-end ML and LLM pipelines** for scalable GenAI solutions using **GCP Cloud Run, Vertex AI, and Dockerized services**.
- Designed and deployed **LLMOps workflows** including model versioning, CI/CD automation, inference optimization, and data observability.
- Fine-tuned and productionized **GPT-4 and Hugging Face models**, improving latency by 28% through distributed inference using **Ray Serve**.

- Built **Retrieval-Augmented Generation (RAG) pipelines** integrating **LangChain + Pinecone vector stores** for contextual knowledge retrieval.
- Integrated **MLflow + KServe** for model registry and deployment traceability within a secure GCP environment.
- Collaborated with data engineers to develop **feature stores and automated data ingestion pipelines**.
- Implemented **AI governance frameworks**, including bias audits, hallucination detection, and model drift monitoring.
- Authored detailed **runbooks, design docs, and architecture diagrams** for reproducible AI workflows.

Associate System Analyst *(May 2021 - Aug 2022) - Practically, Hyderabad, India*

- Built **containerized ML microservices** using Docker, Kubernetes, and Python APIs for analytics automation.
- Developed **Airflow DAGs** to orchestrate training and model retraining workflows for AI-driven personalization systems.
- Designed **CI/CD pipelines** integrating GitHub Actions and MLflow for continuous integration and deployment of models.
- Collaborated with DevOps to optimize compute allocation and reduce model inference cost by 15%.

Data Science Intern *(Jan 2021 - May 2021) - Practically, Hyderabad, India*

- Assisted in **model evaluation and fine-tuning** of classification and NLP models.
- Implemented **data preprocessing pipelines** using Pandas and TensorFlow Data API.
- Supported deployment of early AI prototypes in a secure containerized setup.

Education:

- **Masters in Computing** - Artificial Intelligence, Dublin City University, Dublin, Ireland
- **Masters in Computing** - Computer Applications, APJ Abdul Kalam Technological University, India
- **Bachelors in Computing** - Computer Applications, Bharathiar University, India

Certifications:

- **RedHat Certified System Administrator** (170-123-632)
- **Data Science & Machine Learning**: Making Data-Driven Decisions (MIT Schwarzman College of Computing)
- **NASSCOM Python** (Futureskills Prime)