## Khalid Kassem

MLOps Engineer — Remote
Email: khalid.kassem21@example.com | Phone: +971545800213

## Professional Summary

Experienced MLOps Engineer with a strong track record of building production-grade machine learning systems, from data engineering and model training to deployment and monitoring. Skilled in both research prototyping and scalable MLOps practices. Proven ability to lead cross-functional teams and deliver business impact.

## Technical Skills

AWS (S3, EC2, SageMaker), Time Series, GCP (BigQuery), Azure ML, CI/CD (Jenkins/GitHub Actions), Hugging Face Transformers, CUDA, PyTorch, Keras, Reinforcement Learning, OpenCV, SQL, Docker, Linux

## Professional Experience

**MLOps Engineer — AIWorks (2 yrs)**
• Designed and implemented an anomaly detection system for fraud detection using autoencoders and XGBoost for post-filtering.
• Optimized model inference using TensorRT and mixed precision; achieved 2.5x throughput improvement on GPU.

**NLP Engineer — EdgeAI Solutions (1 yrs)**
• Implemented knowledge distillation to create lightweight transformer models for edge deployment with 3x speedup.
• Built recommendation system using collaborative filtering + content-based features; increased CTR in A/B test by 12%.
• Implemented continuous training pipeline using DVC and GitHub Actions to automate model retraining and versioning.
• Optimized model inference using TensorRT and mixed precision; achieved 2.5x throughput improvement on GPU.

**MLOps Engineer — DataForge (1 yrs)**
• Fine-tuned multilingual speech recognition model (wav2vec2) for domain-specific calls with 5% error reduction.
• Optimized model inference using TensorRT and mixed precision; achieved 2.5x throughput improvement on GPU.
• Designed and implemented an anomaly detection system for fraud detection using autoencoders and XGBoost for post-filtering.
• Implemented continuous training pipeline using DVC and GitHub Actions to automate model retraining and versioning.

**AI Engineer — EdgeAI Solutions (1 yrs)**
• Optimized model inference using TensorRT and mixed precision; achieved 2.5x throughput improvement on GPU.
• Led an end-to-end image segmentation project for medical imagery; built data pipeline, trained U-Net variants, and reduced labeling time by 40%.
• Built recommendation system using collaborative filtering + content-based features; increased CTR in A/B test by 12%.

## Selected Projects

• Deployed scalable inference service on AWS with autoscaling, containerization (Docker), and serverless endpoints; reduced latency to <120ms.
• Fine-tuned multilingual speech recognition model (wav2vec2) for domain-specific calls with 5% error reduction.
• Built recommendation system using collaborative filtering + content-based features; increased CTR in A/B test by 12%.

## Education

BSc in Computer Engineering — University of Science, 2016

## Certifications

• TensorFlow Developer Certificate
• Certified Kubernetes Application Developer (CKAD)

## Languages

Arabic (Native), English (Fluent)

*Generated sample resume — 2025-09-30*