# Feras Mahmood

## GenAI and ML Developer

Oakville, ON | (365)7395522 | technocratz979@gmail.com
linkedin.com/in/feras-mahmood | github.com/Ferasman979 | ferasmahmood.com

## Experience

**Sheridan College, Centre of Applied AI** Oct 2025 – PRESENT
**AI/ML Developer** *Oakville, Ontario*

- **Production RAG Systems & Model Optimization:** Architected and deployed retrieval-augmented generation (RAG) agents serving 15,000+ users, achieving 87% answer accuracy (up from 71% baseline) through chunk-size optimization (512 → 256 tokens) and hybrid search implementation combining dense and sparse retrieval.
- **MLOps & A/B Testing Framework:** Automated the ML lifecycle using Databricks and MLflow, and executed an A/B testing framework (Llama-2 variants). Implemented statistical significance testing ($p < 0.05$) and automated rollbacks while tracking latency and hallucination rates, resulting in a 23% engagement uplift.
- **High-Availability Production ML Services:** Deployed real-time ML inference services with Scala and Spark serving 20,000+ concurrent users, maintaining 99.9% uptime SLA through horizontal pod autoscaling (HPA), circuit breaker patterns, and graceful degradation strategies. Reduced p99 latency from 2.1s to 450ms via model quantization (FP32 → INT8) and request batching.
- **Distributed Model Training & GPU Optimization:** Implemented distributed training with PyTorch FSDP on 8x T4 GPUs, reducing BERT training time by 77%. Optimized CUDA kernels to reduce memory by 40% via mixed-precision training.
- **Scalable ML Microservices Architecture:** Engineered cloud-native microservices on GKE with gRPC (<50ms latency), implementing Model Context Protocol (MCP) servers and Istio-based canary deployments for safe rollouts.
- **ML System Observability & Monitoring:** Deployed Prometheus, Grafana, and Tempo for deep visibility into LLM inference, enabling distributed tracing and automated retraining triggers for model drift (PSI > 0.2).

**Paradigm Electronics Inc.** May 2024 – Aug 2025
**Data Analyst / Developer** *Mississauga, ON*

- **Scalable Data Pipeline Engineering:** Optimized Airflow ETL pipelines (50+ DAGs, 2M+ transactions), achieving a 40% runtime reduction (20 → 12 min) through query optimization and incremental processing, improving data freshness for downstream analytics.
- **ML-Powered Anomaly Detection:** Built an automated anomaly detection system using Isolation Forest, identifying $15k in misallocated costs and triggering real-time Slack alerts for critical deviations.
- **Predictive Analytics & Revenue Optimization:** Conducted statistical analysis and built predictive models (XGBoost, linear regression) on product quality metrics correlating manufacturing defect rates with downstream revenue impact, delivering actionable insights through interactive BI dashboards that contributed to documented 20% revenue growth ($200k) in Q2 2024 through quality-driven product mix optimization.
- **Cloud Infrastructure & Automation:** Deployed Node.js/EJS production apps on AWS EC2 with auto-scaling groups ensuring 99.5% availability, automated critical background jobs using Bash scripts with cron scheduling, and implemented Infrastructure as Code (IaC) practices reducing deployment time from 2 hours to 8 minutes.

## Technical Skills

**Languages**: Java, Python, C++ (CUDA), JavaScript, TypeScript, Scala, SQL, Full-Stack Development
**AI/ML & GenAI**: PyTorch, Distributed Training (FSDP), YOLOv8, Transformers, GenAI Architecture
**Cloud & DevOps**: Azure (AKS, DevOps), GCP (GKE, Cloud Run), AWS (Glue, CloudFormation), Terraform
**Data & Observability**: Scalable ETL/ELT, ML Data Streaming, Datadog, Grafana, Prometheus, Tempo

## Education

**Sheridan College** Sep 2021 – Apr 2026
**Bachelor's Degree in CS (Data Analytics)** *Oakville, ON*

- Achievements: Secured First Place in Capstone Showcase, recognized for innovation, technical excellence, and real-world impact in AI-powered sports analytics