

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 AI/ML Developer

Oct 2025 – PRESENT
Oakville, Ontario

- **Production RAG Systems & LLM Optimization:** Architected and deployed retrieval-augmented generation (RAG) agents for serving **15k+** users, improving Mean Cosine Similarity (**87%**) and HHEM faithfulness scores through hybrid search semantic retrieval and token chunk-size optimization (**512 to 256** tokens).
- **LLM Evaluation & Inference Optimization:** Executed A/B testing on Llama-2 variants and reduced hallucination rates. Optimized performance via quantization (FP32 to INT8) and request batching, reducing p99 latency from **2.1s to 450ms**.
- **High-Availability MLOps & Scalable Serving:** Automated ML lifecycles using Databricks and MLflow with automated rollbacks. Deployed scalable ML inference services (Scala/Spark) for 10k+ users, achieving 99.9% uptime through Kubernetes autoscaling and fault-tolerant service design.
- **Distributed Model Training & GPU Optimization:** Built a Video-Text Transformer with a custom CUDA kernel. Implemented PyTorch FSDP on dual T4 GPUs, reducing memory usage by **36% (14GB to 9GB)**.
- **Scalable MCP Server Architecture:** Engineered cloud-native microservices on GKE and EKS with gRPC, implementing Model Context Protocol (MCP) servers and Istio-based canary deployments for safe rollouts.
- **ML System Observability & Monitoring:** Deployed Prometheus, Grafana, and Tempo on Azure for deep visibility into LLM inference, enabling distributed tracing and automated retraining triggers for model drift (**PSI greater than 0.2**).

Paradigm Electronics Inc. Data Analyst / Developer

May 2024 – Aug 2025
Mississauga, ON

- **Scalable Data Pipeline Engineering:** Optimized Airflow ETL pipelines, achieving a **40% runtime reduction (20 to 12 min)** through query optimization and incremental processing, improving data freshness for downstream analytics.
- **ML-Powered Anomaly Detection:** Built a serverless cost anomaly detection model (Isolation Forest) using AWS SageMaker identifying **\$15k** in misallocated costs and triggering real-time Slack alerts for critical deviations.
- **Predictive Analytics:** Engineered an XGBoost Regressor model to forecast product demand based on 3 years of seasonal data, achieving **89% accuracy**. Deployed interactive BI dashboards providing executive visibility into **\$1M+ monthly sales** across three subsidiaries.
- **Cloud Infrastructure & Automation:** Deployed Node.js/EJS production apps on AWS EC2 with auto-scaling groups ensuring **99.5% availability**, implemented Infrastructure as Code (IaC) practices reducing deployment time in CI/CD pipelines from **15 minutes to 8 minutes**.

TECHNICAL SKILLS

Languages: Java, Python, C++ (CUDA), JavaScript, TypeScript, Scala, SQL, Full-Stack Development
AI/ML & GenAI: AI Context Engineering, GenAI Architecture, Product Development Design
Cloud & DevOps: Azure (AKS, DevOps), GCP (GKE, Cloud Run), AWS (EKS, SageMaker), Terraform
Data & Observability: Scalable ETL/ELT, ML Data Streaming, Datadog, Tempo, Grafana, Prometheus

EDUCATION

Sheridan College Bachelor's Degree in CS (Data Analytics)

Sep 2021 – Apr 2026
Oakville, ON

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