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 & LLM Optimization:** Architected and deployed retrieval-augmented generation (RAG) agents serving **15k** users, improving Mean Cosine Similarity (**87%**) and HHEM faithfulness scores through hybrid search and token chunk-size optimization (**512** to **256** tokens). Successfully mitigated hallucinations by balancing dense semantic retrieval with sparse keyword matching.
- **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** and driving a **23% engagement uplift**.
- **High-Availability MLOps & Scalable Serving:** Automated ML lifecycles using Databricks and MLflow with automated rollbacks. Deployed inference services (Scala/Spark) serving **10,000+ users**, maintaining **99.9% uptime** via HPA, circuit breakers, and graceful degradation.
- **Distributed Model Training & GPU Optimization:** Built a compact Video-Text Transformer with a custom CUDA C++ kernel to accelerate embedding fusion. Implemented PyTorch FSDP on dual T4 GPUs for scalable multimodal training, reducing per-GPU memory usage from **14GB to 9GB** (**∼36% reduction**).
- **Scalable MCP Server Architecture:** Engineered cloud-native microservices on GKE with gRPC (less than **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 greater than 0.2**).

Paradigm Electronics Inc. May 2024 – Aug 2025
Data Analyst / Developer *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: 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