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 for serving **20k+** 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:** Designed and deployed LangGraph agents integrating LlamaIndex to process PDF streams from a MongoDB buffer, and utilized Puppeteer for web research, reducing hallucination rates and improving latency from 2.1s to 450ms.
- **High-Availability MLOps & Scalable Serving:** Assisted in orchestrating training and inference pipelines on Amazon EKS using Argo Workflows; deployed an Isolation Forest model via KServe, integrated SNS alerting, and achieved **99.9%** uptime through KEDA autoscaling.
- **ML Pipeline & Deployment:** Contributed to optimizing the ML training pipeline, reducing ingestion-to-train latency from **15 min** to **7 min** through Parquet-based preprocessing improvements, and helped migrate training jobs to Spot instances with checkpointing, cutting compute costs by **70%**.
- **Scalable MCP Server Architecture:** Designed and Deployed scalable MCP servers on Azure Container Apps with embedded middleware, eliminating API Gateway dependency and enabling independent tool scaling with high-volume trace streaming via gRPC (OTLP) to Tempo.
- **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. — May 2024 – Aug 2025
Data Analyst / Developer — *Mississauga, ON*

- **Data Pipeline Engineering:** Optimized Airflow ETL DAGs, achieving a runtime reduction from **12 minutes** to **8 minutes** through parallel processing for automated daily, monthly, and yearly reporting for Global Sales Operations.
- **Business Intelligence & Financial Impact:** Developed and deployed Key Performance Indicator (KPI) dashboards for Production and Engineering teams, highlighting products with the highest quality risks to drive operational efficiency and reduce defect-related costs.
- **Predictive Demand Modeling:** Streamlined distribution planning by forecasting 6-month product demand across all SKUs using 3 years of historical data, powered by **PySpark** distributed regression.
- **Cloud Infrastructure & Automation:** Deployed Node.js/EJS apps on AWS ECS achieving **99.5% uptime**; optimized GitHub Actions CI/CD pipelines with IaC, reducing deployment time from **15 to 8 minutes**.

Technical Skills

Languages: Java, Python, C++ (CUDA), JavaScript, TypeScript, Scala, SQL, PostgreSQL
Tools & Frameworks: LangChain, LlamaIndex, HuggingFace, PyTorch, FastAPI, Node.js, EJS
Cloud & DevOps: Azure (AKS, DevOps), GCP (GKE, Cloud Run), AWS (EKS, SageMaker), Terraform
Data & Observability: Kinesis Firehose, MLFlow, Datadog, Tempo, Grafana, Prometheus

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