Feras Mahmood

AI/ML Engineer

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

Experience

Sheridan College Oct 2025 – PRESENT
AI Software Developer *Oakville, Ontario*

- **Production RAG Pipeline:** Engineered a hallucination-resistant retrieval platform on **Azure AI** with LangChain/LangSmith orchestration and tracing, combining semantic chunking, BM25 and RRF hybrid search via **Elasticsearch**, cross-encoder reranking, and parent-context injection to deliver grounded responses at scale.
- **Agentic AI Platform:** Developed a multi-tenant RAG chatbot (**10k+ users**) on Qdrant and Vertex AI with role-based access control, agentic tool-calling workflows (LangGraph, CrewAI), and input/output guardrails; implemented human-in-the-loop feedback loops that persisted user corrections to improve response accuracy over time.
- **LLM Evaluation & Guardrails:** Built production LLM safety layers with NeMo Guardrails (topical/safety rails), Llama Guard (content classification), and PII-scrubbing parsers; validated retrieval quality via RAGAS (faithfulness, context precision) to enforce model-risk governance and regulatory compliance at scale.
- **Semantic Matching Platform:** Built a production NLP matching service serving **15k+ users** using SentenceTransformers embeddings, pgvector similarity search, and Redis caching with top-5 latency **under 200ms**; deployed on Kubernetes with GitOps (Argo CD), Helm IaC, and Prometheus/Grafana dashboards to track latency and uptime SLAs.
- **AI Inference Platform:** Collaborated with cross-functional developers to engineer a SageMaker-hosted inference platform serving image and audio processing models; implemented SQS-based queuing to batch-process user records and configured CloudWatch dashboards to monitor model latency, error rates, and throughput across clinical workflows.

Paradigm Electronics Inc. May 2024 – Aug 2025
Data Analyst – **Co-op** *Mississauga, ON*

- **Data Pipeline Engineering:** Optimized Airflow ETL/ELT ingestion DAGs, achieving a runtime reduction from **9 minutes** to **3 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 using PowerBI for Production and Engineering teams, highlighting products with the highest quality risks to drive operational efficiency and reduce defect-related costs.
- **Analytics Data Modelling:** Modelled production-grade dbt pipelines on BigQuery, enforcing schema tests and data quality checks that reduced data incidents by **35%** and improved stakeholder confidence in KPI reporting for Sales and Engineering teams.
- **Cloud Infrastructure & Automation:** Deployed Node.js/EJS apps on AWS ECS with an Nginx reverse proxy; implemented parallel Docker builds in GitHub Actions CI/CD, significantly reducing pipeline execution time.

Technical Skills

Languages: Python, SQL, Java, C++, JavaScript, Scala
ML & AI: PyTorch, TensorFlow, HuggingFace Transformers, LangChain, LlamaIndex, CrewAI, DeepEval, FastAPI
Cloud & Platform: GCP (Vertex AI, BigQuery, Cloud Composer, Pub/Sub), AWS (SageMaker, ECS, EKS), Kubernetes, Helm, Terraform, Docker, GitHub Actions, Argo CD
Data & Infra: PostgreSQL, pgvector, Redis, Spark, Airflow, Databricks, Prometheus, Grafana

Education

Sheridan College
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
- Technical Implementation: Architected a computer vision pipeline using PyTorch and OpenCV to analyze player biomechanics, deployed via CloudRun