Ahmed Radwan

+1-437-440-1525 | ahmedyradwan02@gmail.com | Portfolio | LinkedIn | GitHub

Technical Skills

Languages: Python, SQL, Java
Python Packages: Hugging Face Transformers, LangChain, OpenAI API, vLLM, Accelerate, Whisper, PyTorch, TensorFlow, JAX, Scikit-learn, XGBoost, PySpark, Pandas, NumPy, Matplotlib
Concepts: RAG, Embeddings, Vector Search, Prompt Engineering, Instruction Tuning, Fine-Tuning (LoRA), Experiment Tracking, Model Evaluation, Distributed Training, Quantization, Model Compression, Self-Supervised Learning, Few-Shot Learning
Tools: Git, Singularity, Docker, Linux

Education

York University — Toronto, ON
M.Sc. in Computer Science; GPA: 3.96/4.0 — *Sep 2024 – Present*

King Abdulaziz University — Jeddah, Saudi Arabia
B.Sc. in Computer Science; GPA: 4.98/5.0 — *Sep 2019 – Jun 2024*

Experience

Applied Machine Learning Researcher — Sep 2025 – Present
Vector Institute — *Toronto, ON*

- **Reduced GPU costs by 50%** by engineering distributed inference with Accelerate + vLLM, enabling A40 deployment via optimized sharding
- **Improved annotation quality by 40%** by building a custom Human-in-the-Loop validation tool for a **4,958-question** audio-vision benchmark (60+ hours multimodal video)
- **Developed a Multi-Agent RAG system** replicating SOTA multimodal reasoning architectures; implemented retrieval + orchestration for benchmark-matched performance
- **Authored interpretability / explainability survey** across agentic system layers; defined fairness evaluation guidance aligned with EU AI Act and NIST RMF-style risk framing

Graduate Research Assistant — Aug 2024 – Present
York University — *Toronto, ON*

- **Achieved 92% accuracy** for activity recogntion from Wi-Fi signals using self-supervised contrastive learning, eliminating the need for labeled data.
- **Improved inference 17x with 20% smaller models** by compressing time-series representations, maintaining 90%+ fidelity for edge deployment.
- **Reduced 90% of labeled data needs** via few-shot meta-learning, adapting during deployment with fewer than 100 examples.
- **Enhanced robustness by 25%** via masking and noise injection, and built diffusion-based RF reconstruction for missing regions.

Research Assistant — Feb 2024 – Oct 2024
King Abdullah University of Science and Technology (KAUST) — *Thuwal, Saudi Arabia*

- **Reduced memory by 75%** while maintaining 90%+ accuracy using TinyML NLP quantization + compression for edge devices
- **Achieved 30% efficiency gain** over federated learning by implementing **Split Learning** for privacy-preserving NLP; measured efficiency and CO_2-aware tradeoffs

Research Engineer — Sep 2023 – Jun 2025
Asas.Ai — *Remote*

- **Improved coherence by 45%** for story generation by building an LLM-powered workflow (OpenAI API) with multimodal processing and evaluation-driven iteration
- **Boosted Arabic creative-writing performance by 35%** via instruction tuning on curated Arabic datasets and systematic prompt/data ablations

Selected Projects

TinyEco2AI-NLP | ***TensorFlow/Keras, Quantization, Federated & Split Learning, Eco2AI***

- **Published (EuCNC 2025):** Implemented centralized, federated, and split learning pipelines for sentiment classification with quantization + wireless-noise simulation; tracked energy/CO_2 and privacy via reconstruction error

Fairness-Aware Medical Imaging Bias Mitigation | ***PyTorch, Distillation, Ensembles***

- Improved demographic fairness across sensitive subgroups using adversarial fine-tuning, ensembles, and knowledge distillation on HAM10000 and CIFAR-10 while maintaining accuracy

Awards

1st Place – 2024 Student Games, International Society of Automation (Aramco-sponsored)
Top 1% – KAUST AI Academy (Selected from 10,000+ applicants)