Mahmoud El-Sherbiny

Cairo, Egypt | +20 1004006770 | mahmoud.elsherbiny.work@gmail.com
LinkedIn | GitHub

Summary

Machine Learning Engineer who architects and productionizes LLM, RAG, and RL systems from research to deployment. Pioneered Arabic NLP pipelines including ASR and OCR that slashed transcription errors by 25% and digitized prescriptions at 92%+ accuracy. Hardens models via quantization, pruning, and LoRA to cut GPU memory by 75% and enable low-resource inference. B.S. Computer Science, GPA 3.72/4.0.

Skills

ML / AI: PyTorch, TensorFlow, Scikit-learn, HuggingFace, LangChain, FAISS, LLMs, RAG, RL, NLP, generative AI, fine-tuning, quantization, pruning, ASR, OCR

Tools: Python, SQL, C++, Java, FastAPI, Flask, React, TypeScript, PostgreSQL, REST APIs, Docker, Git, Linux, Pandas, NumPy

Languages: Arabic (Native), English (Fluent)

Experience

MedDigi-Verse (Google Play | App Store) **Feb 2024 – Present**
Co-Founder and AI Engineer *Alexandria, Egypt*
- Shipped an AI-powered pharmacy assistant to Google Play and App Store featuring prescription OCR with 92%+ accuracy and semantic product search over 25K+ items.
- Orchestrated cross-functional collaboration across a 20-member startup team spanning machine learning and full-stack development.

Edafa Telecom Solutions **Jul 2025 – Aug 2025**
Software Engineering Intern *Cairo, Egypt*
- Engineered a real-time health monitoring dashboard using Flutter, Spring Boot, and PostgreSQL with RESTful APIs.
- Devised a multi-role market simulation system supporting five user roles with role-based access control and scalable backend architecture.

PSCwork **Feb 2025 – Jul 2025**
Client Relationship Manager *Remote*
- Steered onboarding and support operations for 200+ clients while driving workflow improvements across teams.

Projects

HackOps - AI Cybersecurity Platform (GitHub) **2025 – 2026**
- Pioneered an RL-based attack framework combining Recurrent PPO and Qwen2.5-7B LLM using LoRA and 4-bit quantization, improving model performance by 10% over baseline under a 500K-step training regime.
- Productionized a full-stack platform using Flask, React, MariaDB, and Docker with scoring, logging, and leaderboard modules.

JobFit - AI Career Copilot (GitHub) **2026**
- Built a job-matching pipeline using Apify, multi-factor scoring, and Gemini, achieving 100% Precision@5 and 0.942 NDCG@5.
- Developed evidence-grounded resume tailoring with claim verification, achieving 100% fact preservation and 0% unsupported claims across 225 evaluations.

Audio Retrieval - Arabic Intelligence Pipeline **2025**
- Optimized Whisper-Small (244M) using QLoRA with adaptive preprocessing (denoising, voice activity detection, chunking), reducing transcription word error rate by approximately 25%.
- Architected a retrieval-augmented generation (RAG) semantic search pipeline using FAISS and BGE-M3 embeddings with 4-bit quantization, reducing GPU memory utilization by approximately 75%.

Islam Web - Arabic QA System **2025**
- Forged an Arabic question answering (QA) system over 432K+ fatwas using hybrid RAG with FAISS, BM25, and reranking pipelines.
- Hardened Qwen2.5-1.5B for production via fine-tuning on 10K QA samples with pruning and quantization, reducing model memory footprint from 6GB to 1.2GB.

Education

Egypt-Japan University of Science and Technology **Jun 2026**
Bachelor of Science in Computer Science, GPA 3.72/4.0 *Alexandria, Egypt*
- Relevant Coursework: Deep Learning, NLP, Reinforcement Learning, Computer Vision, Distributed Systems, Algorithms

Leadership

IEEE Computer Society - EJUST **2024 – 2025**
Head of AI and ML Community *Alexandria, Egypt*
- Spearheaded a 10-member team mentoring 100+ students in AI/ML and staging 6+ technical workshops.

Certifications

Machine Learning Specialization (DeepLearning.AI & Harvard University), DEPI Generative AI Track