

Hossam Elsherbiny

AI Engineer

Cairo, Egypt | hossam.elsherbiny95@gmail.com | +20 114 721 1427 | hossam-elsherbiny.me
github.com/H-Elsherbiny | linkedin.com/in/h-elsherbiny

Professional Summary

AI Engineer specializing in production-ready AI systems across LLMs, Agentic AI, machine learning, and time-series forecasting. Proven ability to design scalable ML pipelines, optimize model performance, and build applications that deliver measurable business impact. Experienced in translating cutting-edge AI research into reliable, production-ready solutions.

Experience

AI Engineer, Freelance Oct 2024 – Present

- Fine-tuned the FLUX.1-dev text-to-image model using LoRA on NVIDIA A100 GPU, achieving high-fidelity custom character generation.
- Built AI pipelines for Egyptian Arabic speech recognition, NER extraction, and video-based gesture detection using Whisper, MARBERT, GPT-4o-mini, MediaPipe, and OpenCV.

AI Engineer, Acrow – Cairo Nov 2024 – Oct 2025

- Designed time-series forecasting pipelines for 3,000+ SKUs using SARIMA, Prophet, XGBoost, CatBoost, LightGBM, and neural networks, reducing forecast error from 80% to 55%.
- Engineered a GUI-based forecasting application that automated the end-to-end pipeline, reducing execution time from 3 days to 1.5 hours and increasing profit by 10.5%.

Education

9-Month Professional Program – AI & ML Track, Information Technology Institute (ITI) Oct 2025 – Present

B.Sc. in Computer Engineering – Faculty of Engineering, Zagazig University Oct 2020 – Jul 2025

Projects

CareFlow Clinical Assistant [GitHub](#) | [Portfolio](#)

- Architected a 10-microservice multimodal AI clinical copilot integrating Arabic ASR, MedGemma VLM, MedSAM 2.0, and BGE-M3 RAG, reducing physician documentation time by up to 40% through automated clinical workflows.
- Implemented a medical vision pipeline combining radiology interpretation, ROI segmentation, and layout-aware OCR, achieving 96% precision and 98% recall for automated PII removal with generating evidence-grounded clinical summaries.

EnergiQ Platform [GitHub](#) | [Portfolio](#)

- Deployed an end-to-end OpenAI-powered AI agent to serve 7,000+ users, integrating RAG using internal documentation (Recall@3 = 0.95) and function calling for personalized backend data retrieval (Intent F1 = 0.94).
- Integrated voice interactions using OpenAI Whisper and TTS APIs, enabling real-time conversational support and reducing inference latency to 78 ms/token.

DeepFM Architecture Implementation [GitHub](#) | [Portfolio](#)

- Reproduced the *DeepFM Paper* from scratch for Click-Through Rate (CTR) prediction on the Criteo dataset (~45M samples, 39 features), reproducing the published benchmark results (AUC $\approx$ 0.801, LogLoss $\approx$ 0.456).
- Conducted 6 ablation studies and optimized the model for high-dimensional, imbalanced data, improving performance to AUC $\approx$ 0.802 and F1-score $\approx$ 0.56.

Technical Skills

Programming Languages: Python, SQL, C++

ML Frameworks: PyTorch, TensorFlow, Scikit-learn

LLMs & Agentic AI: Transformers, Hugging Face, LangChain, LangGraph, LangSmith, RAG

Computer Vision (CV): OpenCV, YOLO, Vision Transformers (ViT)

MLOps & Cloud: Docker, Kubernetes, MLflow, Git, AWS

Big Data & Databases: PySpark, Kafka, PostgreSQL, MongoDB, ChromaDB

Backend: FastAPI, REST APIs

Additional Information

Military Status: **Exempted** (Certificate available upon request).