

Ahmed Eladl

AI Engineer

+201090717037 ahmedkhaled553311@gmail.com linkedin.com/ahmed-eladl github.com/ahmed-eladl Google Scholar huggingface.co/AhmedEladl ahmed-eladl.link

Summary

AI Engineer with 2+ years of experience building and deploying Generative AI systems – LLMs, RAG pipelines, and multi-agent frameworks – across voice AI, document intelligence, and research applications. Skilled in fine-tuning and optimizing models for production performance, deploying and maintaining them on cloud-native infrastructure. Brings a strong research foundation, with multiple peer-reviewed publications

Skills

Programming: Python, Java, SQL, C, C++, JavaScript, HTML, CSS, R, Verilog

AI/ML Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace, LangChain, LangGraph, CrewAI, OpenCV, NLTK, vLLM, Pandas, NumPy

API, Deployment & Cloud: FastAPI, Flask, REST APIs, Kubernetes, KServe, AWS, Google Cloud Platform (GCP), Microsoft Azure, Git, GitHub Actions (CI/CD), Ngrok

Concepts: Generative AI, LLMs, RAG, Deep Learning, Model Optimization, Model Deployment & Serving, Model Evaluation, Unit Testing (pytest), Fine-Tuning Techniques, Prompting Techniques, Multi-Agent Systems, Agentic AI, Machine Learning, Reinforcement Learning, Computer Vision, GANs, Big Data, Agile, Scrum

Experience

Ebra.ai

Dec 2025 – Aug 2026

R&D AI Engineer (Contract)

Remote, Saudi Arabia

- Built and deployed an end-to-end AI voice call system (VAD, STT, LLM, TTS), optimizing the pipeline to achieve ~400 - 800ms inference latency for real-time conversations.
- Collected and curated Saudi dialect audio data, building a full pipeline to clean, segment, and prepare it for fine-tuning multiple open-source TTS and ASR models for high-fidelity, low-latency Saudi Arabic speech synthesis and transcription; deployed and maintained a large language model in production using GCP for low-latency inference.
- Deployed and maintained an LLM in production on GCP, enabling low-latency inference.

American University in Cairo

Dec 2025 – Feb 2026

Research Assistant (Contract)

Hybrid, Cairo

- Processed and analyzed SEC filings using NLP pipelines for large-scale text extraction and preprocessing.
- Built labeled dataset of AI-related disclosures and engineered features for analysis.
- Applied text mining techniques (TF-IDF, keyword extraction, co-occurrence analysis) to identify patterns.

Locii Solutions

Jul 2025 – Dec 2025

AI Engineer (Contract)

Remote, Cyprus

- Built an AI system to extract text from various file types, normalize/format the output, and store it in a database.
- Developed a multimodal Retrieval-Augmented Generation (RAG) pipeline for PDFs and images using OCR to enable semantic retrieval and context-aware generation.
- Designed and deployed text-to-SQL agents to translate natural language queries into executable database queries and visualize results and figures.

Lothgha App

Feb 2025 - Apr 2025

AI Developer (Contract)

Remote, Saudi Arabia

- Built an Arabic ASR system with error detection for mispronounced words and stuttering using TTS/STT
- Developed automated monthly analytics reports to monitor performance.
- Developed an AI model for detecting stuttering in various words and submitted a research paper to Interspeech.

Research and Innovation Hub - AIU

Aug 2023 - Jan 2025

AI Research Assistant (Intern → Assistant) (Part-time)

New Alamein City, Egypt

- Co-authored ~5 published research papers applying AI to medical/physiological data, plus 2 computer-science papers (including graduation project, as primary author), presented at conferences.
- Applied Wavelet Transforms to blood-pressure time series to generate scalograms, then trained a CNN to classify healthy vs. sick subjects, achieving 94% test accuracy.
- Used saliency maps to identify the frequency peaks most relevant to CNN predictions and calculated spectral densities at those peaks.

Education

Alamein International University – AIU

BSc. in Computer Science - Artificial Intelligence Science Program

Oct 2020 - Jun 2024

New Alamein City, Egypt

- **Relevant Coursework:** Data Structures and Algorithms, Prob & Stat, Linear Algebra, Deep Learning, Natural Language Processing, Advanced Machine Learning, Intelligent Recommender System
- **Full Academic Scholarship, President El-Sisi** | Fall2020 - Fall2021

Selected Projects

ASRX – Universal ASR Audio Alignment Engine github.com/ahmed-eladl/asrx | Aug 2026 - Sep 2026

- Built and published an open-source Python engine that aligns transcripts from any ASR model or LLM with audio to generate millisecond-level word timestamps, confidence scores, VAD segments, and speaker labels.
- Architected a modular pipeline supporting pluggable backends (Wav2Vec2, Meta MMS, NVIDIA NeMo, Silero VAD, Pyannote) with standardized CLI/API JSON outputs for streaming and batch speech pipelines.

Magpie TTS – Saudi & Emirati Arabic Fine-tuning [HF: Saudi] | [HF: Emirati] | Jul 2026 - Aug 2026

- Fine-tuned NVIDIA's Magpie TTS Multilingual 357M on conversational speech, adapting the architecture to ar-SA and ar-AE dialectal phonetics and non-standard orthography.
- Engineered an end-to-end speech data pipeline (22.05 kHz resampling, noise cleaning, Silero-VAD utterance segmentation, and speaker diarization); published fine-tuned checkpoints and inference demos on Hugging Face.

AI Medical Assistant for Diagnosis & Treatment [Freelance](#) | [Saudi Company](#) | Feb 2025 - Mar 2025

- Used Llama 3's function-calling to parse patient symptoms and vitals into ICD-10 codes and suggest clinical actions; recommendations were validated by physicians.
- Integrated the Serper API to cross-check ICD-10 codes against SFDA (Saudi Food & Drug Authority) resources and recommend approved medications in real time.

AIQ – RAG-based MCQ Generation for Dynamic MOOC Assessments (Graduation Project) | AIU, Sep 2023- Jun 2024

- Designed and deployed a parameter-driven web interface (FastAPI + ngrok) using Llama 3-70B few-shot prompting to automatically generate MCQs from MOOC transcripts.
- Implemented RAG components (retriever + LLM reader) backed by a vector database for efficient retrieval; used Llama 3-70B for MCQ ranking and Whisper for robust video→text extraction.
- Evaluated prompting strategies (Zero-Shot, Few-Shot, RAG) with RQUGE metrics and LLM-as-a-judge validation, achieving a top RAG score of 4.46/5 and 83.3% accuracy.

Improving Mistral 7B model with Attention Sinks and Parameter-Efficient Fine-Tuning | Nov2023-Jan2024

- Fine-tuned Mistral Instruct 7B (transformer with attention sinks) on 1M tokens from the Dolly 2.0 dataset using parameter-efficient fine-tuning (QLoRA); evaluated outputs for quality and diversity using BLEU/ROUGE against Dolly 2.0 as reference.
- Developed an application that utilizes the fine-tuned model to dynamically generate task-specific instructions based on user inputs.

Publications

Published under name variants: Ahmed El-Adl / Ahmed K. El-Adl / Ahmed Khaled Eladl. Full list on Google Scholar (see header).

- **Ahmed El-Adl** et al. *Leveraging a Deep Neural Network for Identification of At-Risk Individuals Chronically Exposed to Organophosphate Pesticides Using LC-MS Chromatograms*. Canadian Journal of Physiology and Pharmacology, 2026.
- **Ahmed El-Adl** et al. *Leveraging Convolutional Neural Networks for Classification of LC-MS/MS Proteomic Profiles and Identification of Impending Carcinogenesis*. American Society of Pharmacology and Experimental Therapeutics, 2025.
- **Ahmed Khaled Eladl**, Mohamed Soliman, Islam Elkabani, Laila Shoukry. *AIQ – RAG-based Generation of Questions for Dynamic MOOC Assessments*. Int'l Conference on Universities & AI Technology, 2024, Amman, Jordan.
- Moamen Ayman, **Ahmed Khaled Eladl**, Laila Shoukry. *MyClassBot: A RAG-based Audio Chatbot Guiding Students and Faculty through Their Schedules*. Int'l Conference on Universities & AI Technology, 2024, Amman, Jordan.
- *Plus 4 additional peer-reviewed publications applying deep learning to physiological/biomedical signal analysis (2023–2024) – full list on Google Scholar.*

Courses

- **Generative AI with Large Language Models** – DeepLearning.AI, 16hrs | Apr–May 2024
- **Advanced Computer Vision with TensorFlow** – DeepLearning.AI, 20hrs | Mar–Apr 2024
- **Introduction to Machine Learning in Production** – DeepLearning.AI, 15hrs | Feb–Mar 2024
- **Natural Language Processing Specialization** – DeepLearning.AI, 120hrs | Jun–Sep 2023
- **Neural Networks and Deep Learning** – DeepLearning.AI, 20hrs | Dec 2022–Jan 2023
- **Machine Learning for All** – University of London, 20hrs | May–Jun 2022