Abdullah Magdy Elbarrany

+20-1023692128
abdullah.elbarrany@gmail.com
GitHub Profile
LinkedIn Profile

Professional Summary

Machine Learning Engineer with 3+ years of experience designing and deploying production-grade AI systems across healthcare and real-time applications. Architected and shipped an LLM-powered clinical support assistant serving approximately 1,000 sessions per week (12,000+ conversations per month, 60,000 LLM calls monthly), automating 30% of coaching workload while maintaining high clinical accuracy and safety standards. Strong expertise in AI agent systems, Retrieval-Augmented Generation (RAG), groundedness evaluation, real-time voice agents (STT/TTS), async backend architecture (FastAPI), and scalable deployment on Google Cloud Platform (Cloud Run). Focused on reliability, safety guardrails, and translating research into production-ready systems.

Experience

•Machine Learning Engineer *Mar 2024 – Present*

mDoc Remote

- Architected and deployed a production LLM-powered health assistant serving approximately **1,000 sessions per week** (12,000+ conversations per month, 60,000 LLM calls monthly), automating **30% of coaching workload**.
- Designed end-to-end AI system architecture using **FastAPI microservices**, async processing, **Pub/Sub messaging**, MongoDB, and deployment on **Google Cloud Run**.
- Built AI agent workflows using **LangChain**, implementing tool-calling, structured outputs, conversation memory, and vector-based **RAG pipelines**.
- Implemented caching strategies for dynamic prompting rules to reduce inference latency and token consumption.
- Integrated **real-time voice capabilities** using STT/TTS pipelines and GPT real-time APIs, including audio preprocessing and streaming responses.
- Engineered a production risk-detection module achieving **94% clinical accuracy**, with monitored false-positive and false-negative rates and automated escalation workflows.
- Implemented **groundedness scoring** and clinical accuracy evaluation pipelines to measure response quality and reduce hallucination risk.
- Designed safety guardrails including prompt injection detection, non-health query classification, fallback routing, and human-in-the-loop escalation.
- Built monitoring and evaluation dashboards tracking clinical accuracy, conversational quality metrics, token usage, and system reliability.
- Evaluated multiple LLM providers prior to production selection; documented real-world performance trade-offs in published technical writing.

•Machine Learning Engineer *May 2023 – Dec 2024*

Zedzag On-site

- Led a team of four engineers delivering AI-enabled features for a production smart home management platform.
- Designed system architecture for real-time computer vision pipelines integrating high-throughput inference into mobile applications.
- Owned full ML lifecycle including data preparation, model training, optimization, deployment, and API integration.
- Integrated backend systems using Firebase and cloud services to enable scalable automation and device orchestration.
- Containerized AI services and supported structured deployment workflows.
- Partnered with product and engineering stakeholders to prioritize features and deliver production-ready AI capabilities on schedule.

Technical Skills

AI & LLM Systems: Large Language Models (LLMs), AI Agents, LangChain, Prompt Engineering, Tool Calling, Retrieval-Augmented Generation (RAG), Embeddings, Vector Search, Groundedness Scoring, Guardrails, Human-in-the-Loop Systems

Voice & Real-Time AI: Speech-to-Text (STT), Text-to-Speech (TTS), GPT Real-Time APIs, Audio Preprocessing, Streaming Responses

Backend & Architecture: Python, FastAPI, Async Processing, REST APIs, Microservices, Pub/Sub Messaging, MongoDB, Containerization (Docker)

Cloud & DevOps: Google Cloud Platform (Cloud Run), AWS (Foundational), Monitoring, Logging

ML Frameworks: PyTorch, TensorFlow, Scikit-learn, YOLO, OpenCV

Selected Projects

•AI-Powered Museum Guide (Statue Vision)

GitHub

- Built an object recognition system integrating computer vision with backend APIs for real-time inference.
- Designed a Flask-based service layer to serve AI predictions and contextual information dynamically.
- Integrated mobile and backend systems to enable seamless AI-assisted user experiences.
- Tech: Python, OpenCV, Flask, Firebase, Flutter.

•Abnormal Behavior Analysis for Poultry Farms

- Developed a real-time AI monitoring system achieving **99.7% precision** using YOLOv8 and sequence models (BI-LSTM, GRU).
- Designed modular ML pipelines for scalable training, evaluation, and deployment.
- Demonstrated high-accuracy anomaly detection in high-volume video streams.

•BuckTracker – AI Currency Tracking System

GitHub

- Developed OCR and CNN-based system achieving **92% accuracy** in multi-currency serial number recognition.
- Integrated computer vision pipelines into structured tracking workflows for production-style environments.

•HomeIoT – Smart Home Automation Platform

- Led development of a Flutter-based IoT platform integrating authentication, cloud storage, and real-time device control.
- Implemented structured backend communication using Firebase and MQTT-based messaging.
- Enabled automation workflows through API-based device interaction.

Professional Activities & Service

•Presenter – Global Digital Health Forum (Google Nairobi)

- Delivered a technical presentation titled **"Insights and Learnings from Implementation of Digital Health and AI in Global Health"** at a Google-hosted GDHF session in Nairobi, Kenya.
- Presented real-world AI deployment insights, system architecture considerations, and performance trade-offs to a technical and global health audience.

•Peer Reviewer – IEEE & Springer Journals Web of Science | ORCID

Publications & Technical Writing

•From Benchmarks to the Real World: Why Healthcare AI Needs Real-World Data mDoc Blog

•The Use of Pose Estimation for Abnormal Behavior Analysis in Poultry Farms DOI

•BuckTracker: System For Multi Banknotes Tracking DOI

•Abnormal Behavior Analysis for Surveillance in Poultry Farms using Deep Learning IEEE Xplore

•Chicken Behavior Analysis for Surveillance in Poultry Farms Paper (PDF)

Achievements

•First Place – 6th UGRF Special Edition Nile University, Egypt *Aug 2024*

•Finalist – Innov8 Hackathon Bahrain *2023*

•Best Paper Award – IMSA International Conference IMSA *2023*

•Finalist – Benha Hackathon Benha University, Egypt *2022*

Education

•Bachelor degree of Computer Science *2019–2023*

University of Greenwich CGPA: 3.6