

OMAR ELGENDY

Machine Learning Engineer | Data Scientist

+971-558980116 | omaratef3221@gmail.com | linkedin.com/in/omaratef3221 | github.com/omaratef3221 | omarelgendy.dev | Dubai, UAE

SUMMARY

Machine Learning Engineer and AI Researcher with 6+ years taking generative-AI systems from prototype to production: agentic AI, RAG, and LLM fine-tuning, deployed on end-to-end MLOps pipelines across AWS and GCP, on a strong classical-ML and time-series foundation. Published researcher with 7 peer-reviewed papers and 560+ citations, including a top-0.1%-cited Computer Science paper on healthcare AI.

EXPERIENCE

Almosafer

Machine Learning Specialist

Dubai, United Arab Emirates

09/2025 – Present

- Re-architected a single-call LLM extraction system for airline booking emails into a multi-step agentic workflow (LangGraph, LangChain) with structured output validation, self-correction loops, and fallback handling, significantly improving extraction accuracy and reliability.
- Built an agentic voice AI system (Retell AI) with LLM-driven dialogue management and tool calling to assist customer service, optimizing response quality and call-handling latency for real-time customer interactions.
- Designed a post-call GenAI pipeline using LLM-based summarization, classification, and entity extraction to convert raw call transcripts into structured records automatically.
- Deployed a time-series forecasting system for flight add-on revenue, enabling revenue optimization decisions that increased add-on revenue by 55%.
- Developed an XGBoost financial-forecasting model for budgeting that cut prediction error (MAPE) by 80% versus the prior baseline.
- Architected end-to-end MLOps pipelines on AWS SageMaker and MLflow, automating model training, versioning, and deployment across these systems.

University of Sharjah

Research Assistant (Part-time)

Sharjah, United Arab Emirates

06/2024 – Present

- Led a 22-condition LLM adaptation study comparing LoRA, QLoRA, and full fine-tuning across 4 open-weight models (Llama-3.1, Qwen2.5, Jais-2, Fanar) for Arabic medical QA, building a two-stage domain and task fine-tuning pipeline on multi-GPU AWS (4× A10G) with multi-seed reruns, bootstrap CIs, and McNemar significance testing; first-author paper under review at *Cluster Computing* (Springer, Q1).
- Designing controlled architecture-modification experiments on decoder LLMs (590M to 13B scaling ladder), replacing transformer FFN layers with learnable RBF/KAN variants, and built the distributed training stack (DeepSpeed ZeRO-3, 8-bit optimizers, gradient checkpointing, SLURM orchestration) enabling full fine-tuning of up to 9B-parameter models on 4× A10G GPUs.
- Built “Welly”, the bilingual Arabic/English RAG chatbot for My FarmWell, a University of Sharjah × IBM app (IBM Sustainability Accelerator), answering farmer questions on irrigation, water conservation, and local regulations.
- Engineered and launched the retrieval-augmented generation pipeline over region-specific water, climate, and groundwater data, grounding LLM responses in IBM Environmental Intelligence and IBM Cloud sources for accurate, farm-level guidance.
- Co-authored 2 papers under review on ML for groundwater-level forecasting (time-series, AutoML), 2026.

Virtue

Machine Learning Engineer

Dubai, United Arab Emirates

01/2023 – 03/2025

- Developed and integrated production-grade LLM applications for chatbots, document Q&A, and intelligent assistants using Transformers, LangChain, and RAG, achieving over 85% answer accuracy on internal mental-health domain evaluations.
- Deployed serverless AI APIs using Docker and AWS Lambda, reducing hosting costs by 25% and maintaining sub-3-second response latency under 1000 RPS.
- Built scalable MLOps pipelines on AWS SageMaker, MLflow, Airflow, and TorchServe, cutting deployment time by 30% and automating versioning, CI/CD, and monitoring.
- Fine-tuned transformer models with QLoRA and LoRA across tourism, agriculture, and healthcare, improving task accuracy by up to 18%.
- Optimized LLM summarization and multi-turn dialogue through systematic prompt engineering.

Virtue

Junior Machine Learning Engineer

Dubai, United Arab Emirates

02/2021 – 01/2023

- Built and deployed machine learning models for user behavior prediction using mood analysis, psychological profiling, and sentiment analysis, improving engagement insights by 30%.
- Developed and trained classic machine learning models (random forests, XGBoost, SVMs) for structured data tasks, improving prediction accuracy by up to 22% over baseline.

- Led end-to-end ML workflows — data preprocessing, EDA, model training, tuning, and evaluation — using Scikit-learn, TensorFlow, and PyTorch, increasing model precision by 20%.
- Implemented and orchestrated experiment tracking and model versioning for better reproducibility and lifecycle management.

University of Sharjah

Sharjah, United Arab Emirates

Research Assistant

08/2020 – 02/2021

- Participated in and authored 7 peer-reviewed AI research papers including 2 in Q1-ranked journals with 560+ total citations and an h-index of 4, focusing on deep learning, NLP, and clinical applications.
- Developed and contributed BERT-based models and Arabic datasets for Arabic fake news detection, achieving 90%+ classification accuracy and published in high-impact NLP journals.
- Planned and evaluated deep learning architectures for Alzheimer's detection using MRI and for heart failure prediction using structured medical datasets.

PROJECTS

Podcast Summarizer Agent | *LangGraph, LangChain, LLM Agents, RAG, MongoDB, Docker* | [GitHub](#)

- Built an end-to-end agentic AI application that autonomously discovers, transcribes, and summarizes podcast episodes, orchestrated as a LangGraph state machine with an LLM reasoning node, a tool node, and conditional routing for retries and error recovery.
- Integrated third-party APIs as LLM-callable tools with dynamic tool selection and structured output handling (SerpAPI search, YouTube Transcript API for automatic speech-to-text captions, MongoDB for persistent storage and metadata management).
- Containerized the service with Docker and deployed it as a REST API with health-check endpoints, a scheduled cron worker, and a Streamlit web interface for non-technical users.

Learn PyTorch in 5 Projects | *PyTorch, CNNs, Transfer Learning, BERT* | Published on freeCodeCamp | [GitHub](#)

- Authored and presented a full-length PyTorch course **published on the freeCodeCamp.org channel**, teaching applied deep learning through five hands-on, end-to-end projects.
- Covered tabular classification with feed-forward networks, image classification with custom CNNs, transfer learning with pretrained vision models, audio classification, and text classification by fine-tuning BERT.
- Maintained the companion open-source repository of reproducible Jupyter notebooks covering data loading, training loops, and evaluation, used as a public learning resource by the ML community.

Text-to-SQL LLM (Qwen2 Fine-Tune) | *Transformers, PEFT, Fine-Tuning, Hugging Face* | [GitHub](#)

- Fine-tuned Qwen2-Instruct on 10K natural-language-to-SQL pairs (DuckDB text2SQL corpus) to translate plain-language questions into executable SQL, enabling natural language interfaces over relational databases.
- Built a reproducible, configurable CLI training pipeline (model ID, dataset, epochs, batch size, learning rate) with checkpointing, step-level logging, and automated run-completion notifications.
- Evaluated with accuracy, precision, recall, and F1, and published the trained model to the Hugging Face Model Hub with a model card, usage examples, and documented limitations for downstream reuse.

Arabic Medical QA: LoRA vs. Full Fine-Tuning | *PyTorch, PEFT, SLURM, AWS Multi-GPU* | [GitHub](#)

- Open-sourced the research codebase behind the *Cluster Computing* (Springer) submission: a configurable two-stage pipeline (domain-adaptive continual pre-training, then supervised fine-tuning) spanning a 22-condition grid across 4 open-weight models (Llama-3.1, Qwen2.5, Jais-2, Fanar) with LoRA, QLoRA, full fine-tuning, and zero-shot baselines.
- Implemented a deterministic log-probability evaluation protocol that removes decoding confounds, plus an automated data-cleaning pipeline reporting accuracy, macro-F1, and per-medical-specialty metrics.
- Engineered memory-efficient distributed training (8-bit AdamW, gradient checkpointing, bfloat16) with SLURM job arrays on AWS A10G nodes, and automated experiment tracking, model versioning, and checkpoint publishing through Weights & Biases and the Hugging Face Hub.

SKILLS

ML & AI: Python, TensorFlow, PyTorch, XGBoost, Scikit-learn, Keras, Neural Networks, C/C++

LLMs & NLP: Transformers, LangChain, LangGraph, RAG, ReAct, NLTK, Fine-Tuning, LoRA, QLoRA, RLHF

MLOps & Cloud: AWS, AWS SageMaker, MLflow, Airflow, Docker, CI/CD, GCP, Azure, TorchServe, Databricks, Kubernetes

Monitoring & Infra: Prometheus, Grafana, Bash, Linux, Git, GitHub, Terraform

Data & Backend: SQL, PostgreSQL, MongoDB, FastAPI, Flask, NumPy, Pandas

Soft Skills: Decision-Making, Teamwork, Creativity, Adaptability, Continuous Learning, Leadership, Problem-Solving, Emotional Intelligence, Ownership, Agile, Attention to Detail

EDUCATION

University of Sharjah

Sharjah, United Arab Emirates

M.Sc. in Computer Engineering, minor Artificial Intelligence

09/2020 – 12/2024

- Thesis: Optimizing Large Language Models for Arabic Text Summarization utilizing RBF Networks

University of Sharjah

Sharjah, United Arab Emirates

B.Sc. in Computer Engineering

09/2016 – 08/2020

- Senior Project: EEG-based Computer Command Input for paralyzed people

LANGUAGES

Arabic: Native **English:** Fluent **German:** Beginner

CERTIFICATIONS

AWS Certified Machine Learning – Specialty | Amazon Web Services (AWS)

AWS Certified Cloud Practitioner | Amazon Web Services (AWS)

PUBLICATIONS

Breast cancer detection using AI techniques: A systematic literature review — *Q1, Artificial Intelligence in Medicine (ScienceDirect)*, 2022. A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, **O. Elgendy**. **428 citations**. doi.org/10.1016/j.artmed.2022.102276

Arabic fake news detection based on deep contextualized embedding models — *Q1, Neural Computing & Applications (Springer)*, 2022. A. B. Nassif, A. Elnagar, **O. Elgendy**, Y. Afadar. **122 citations**. doi.org/10.1007/s00521-022-07206-4

Enhancing RLHF in LLMs: Comparing BERT, XGBoost, and Deep Models for Bias Detection — *IEEE DeSE*, 2025. W. Torfa, **O. Elgendy**, A. B. Nassif. doi.org/10.1109/DeSE68208.2025.11367955

Death/Recovery Prediction for COVID-19 Patients using Machine Learning — *International Journal of Systems Applications, Engineering & Development*, 2020. **O. Atef (Elgendy)**, A. B. Nassif, M. A. Talib, Q. Nasir. **7 citations**.

Alzheimer Detection using Different Deep Learning Methods with MRI Images — *IEEE ASET*, 2023. **O. Elgendy**, A. B. Nassif. ieeexplore.ieee.org/document/10180640

Text Toxicity Level Detection using Deep Contextualized Embedding Models — *OkIP AHIT*, 2024. **O. Elgendy**, A. B. Nassif, B. Soudan. doi.org/10.55432/978-1-6692-0007-9_3

Heart Failure Prediction using Machine Learning with Meta-heuristic Feature Selection — *OkIP AHIT*, 2024. **O. Elgendy**, A. B. Nassif, B. Soudan. doi.org/10.55432/978-1-6692-0007-9_4

Full citation metrics: bit.ly/omarelgendy-research