Hossam Gamal Mostafa

AI Engineer

New Cairo, Egypt | +20 112 524 8097 | e-hossam96@outlook.com
hossammostafa.ai | github.com/e-hossam96 | linkedin.com/in/e-hossam96

Profile Summary

AI Engineer and NLP Data Scientist with hands-on experience in system design and cloud deployment. Architects scalable, reliable AI applications, leads technical discussions with clients, and delivers production-ready LLM solutions that balance performance, cost, and maintainability.

Experience

Xceed Contact Center Jun 2025 – Jul 2026
Senior AI Developer *Full-time • Hybrid • Cairo, Egypt*

- Contributed to the Center of Excellence's AI strategy, driving the company's digital transformation.
- Led end-to-end AI projects with agile methods, including **RAG applications** and **database AI agents**, plus services for machine translation, sentiment analysis, and voice transcription.
- Designed and co-managed cloud infrastructure for AI applications, optimizing for scalability and cost.
- Directed AI development processes to meet industry standards and ISO certification requirements.

Udacity Sep 2022 – Sep 2025
Session Lead *Contract • Remote*

- Delivered weekly Connect sessions for the DECI program covering Python, web development, data analysis, AI, cybersecurity, digital arts, electric circuits, embedded systems, and computer architecture.
- Mentored students and tracked their progress during sessions and office hours.
- Achieved a **4.6** average mentor rating and an **82%** average graduation rate across the last two cohorts.

ASAP Systems Jan 2025 – May 2025
AI Developer *Full-time • On-site*

- Built an **AI forecasting solution** predicting stock quantities and costs for BarCloud customers, pairing a zero-shot forecaster fine-tuned on in-house data (new customers) with Facebook Prophet (existing customers).
- Exposed forecasts through the forecasting dashboard and the BarCloud chat interface via LLM tool calls.
- Designed a **text-to-SQL solution** that turns users' questions into correct, safe MS SQL queries and returns summarized results.
- Enhanced **RAG solutions** for BarCloud feature questions, adding LLM serving (LlamaCpp, vLLM, Ollama) and intent classification that routes voice commands to the right pages and sections.

iSemantics Jul 2022 – Sep 2024
NLP Data Scientist *Full-time • Hybrid • Cairo, Egypt*

- Researched and implemented NLP solutions in information retrieval (multilingual NER, relation extraction) and social sciences (sentiment and aspect-based sentiment analysis, propaganda detection, offensive language identification), improving F1 and precision by **20%+** over strong baselines.
- Led discussions with technical and non-technical clients on integrating AI into their businesses, winning over **3 clients**.
- Developed training plans for incoming and current junior AI developers.
- Conducted **6 intern interviews**, assessing technical fit and alignment with company values.
- Represented the company at Techne Summit Cairo 2024.

British University in Egypt (BUE) Oct 2021 – Sep 2022
Teaching Assistant of Mathematics *Full-time • On-site • Cairo, Egypt*

- Delivered tutorials in Probability, Statistics, Numerical Methods, Differential Equations, and Calculus.
- Supported professors with course logistics: attendance, exam marking, and course archiving on the university's cloud system.
- Achieved a **4.63/5** average student satisfaction score.

Projects

Product Reviews Analysis – Aspect-Based Sentiment Application reviews.hossammostafa.ai • Nov 2026
Personal project • FastAPI, React, TypeScript, llama.cpp, SQLite, Docker, AWS EC2

- Built a **full-stack application** that scores product reviews from 1 to 5 on each aspect customers mention, with charts and CSV export.
- Served a **2B-parameter open-weight LLM** with llama.cpp using JSON-schema-constrained output, so answers are always valid and limited to the aspects the user confirmed.
- Cut scoring time per review by **40%** (4.0 s to 2.4 s) by profiling LLM generation, trimming the output format, and moving to a 4-bit model; runs at about **1.2 s per review** on a CPU-only AWS server.
- Designed the backend for long-running work: background jobs with live progress, SQLite sessions with migrations, rate limiting, and clean failure handling when the model server goes down.
- Deployed on AWS EC2 with Docker Compose, nginx, and Caddy (automatic HTTPS) on a custom domain, with a status page showing when the demo is online.
- Wrote **200+ automated tests** (pytest, Vitest) covering API flows, failures, and UI states.

Skills

Programming Languages: Python, TypeScript, C, SQL, HTML/CSS

LLMs & Agents: RAG, AI agents, prompt engineering, fine-tuning, LLM evaluation, OpenAI, LangChain, LangGraph, LlamaIndex, CrewAI

LLM Serving & Inference: vLLM, llama.cpp, Ollama, ONNX, quantization, constrained decoding (Outlines, XGrammar, JSON schema)

ML & Deep Learning: PyTorch, PyTorch Lightning, Transformers, Sentence-Transformers, LoRA, scikit-learn, GluonTS

Backend & Frontend: FastAPI, REST APIs, React

MLOps & Data: Weights & Biases, DVC, Pandas, Matplotlib, Plotly

Databases: PostgreSQL, MS SQL, SQLite, MongoDB, Neo4j

Vector Stores: Qdrant, ChromaDB, Weaviate, Pinecone, FAISS

Cloud & DevOps: AWS (EC2), Azure, GCP, Docker, Docker Compose, nginx, Caddy, Git, GitHub Actions

Testing & Tools: pytest, Vitest, Jira, Postman

Professional: Leadership, mentorship, project management, agile workflows, problem solving, teamwork

Spoken Languages: Arabic (native), English (fluent)

Education

Mansoura University Oct 2026 – Present
M.Sc. in Computer and Systems Engineering *Mansoura, Egypt*

Alexandria University Sep 2015 – Jul 2019
B.Sc. in Nuclear and Radiation Engineering *Alexandria, Egypt*

- **Grade:** 80% (Very Good), ranked **3rd** in class.
- **Coursework:** Advanced Calculus, Linear Algebra, Differential and Partial Differential Equations, Numerical Methods, Probability, Statistics, Programming (Fortran).

Sohag University Sep 2014 – Jul 2015
Preparatory Engineering Year *Sohag, Egypt*

- **Grade:** 75% (Very Good). **Coursework:** Calculus, Programming (C++).

Professional Development

McKinsey & Company Nov 2023 – Apr 2024
McKinsey Forward Program *Online* • *Certificate*

- Program building transferable business, leadership, and problem-solving skills: adaptability, digital transformation, communication, critical thinking, and agile mindset.
- Solved the “Hustleton City Traffic Problem” case using SMART problem statements, structured thinking, and issue prioritization.