Zeyad Diaa

+201111676128 | Obour, Cairo, Egypt | zeyaddiaa2@gmail.com
linkedin.com/in/zeyad-diaa | github.com/Zeyad-Diaa12

Summary

AI Engineer with 2+ years designing, building, and deploying production machine learning and Large Language Model (LLM) systems, Agentic AI, Retrieval-Augmented Generation (RAG), and enterprise workflow automation. Currently consulting on enterprise AI agents that automate HR and ERP operations. Backed by production backend engineering, owns systems end-to-end from data pipeline and model to the deployed service, and ramps quickly into unfamiliar domains.

Experience

Raya International Services 01/2026 – Present
Consultant AI Engineer *Cairo, Egypt*

- Built and deployed an end-to-end AI-powered Applicant Tracking System (ATS) with multilingual AI interviewers, automated candidate evaluation pipelines, and hiring recommendations, reducing screening time by 80%
- Developed Agentic AI and Large Language Model (LLM) solutions for Oracle Fusion ERP workflows, reducing Oracle Fast Formula development time by 70%
- Designed enterprise AI agents automating HR and ERP operations using NLP and prompt engineering, reducing manual effort by 60%
- Delivered Prompt Engineering training workshops to support enterprise AI adoption

Blackstone eIT 09/2025 – 01/2026
AI Engineer *Cairo, Egypt*

- Developed an Agentic Retrieval-Augmented Generation (RAG) pipeline using LLMs and NLP to extract structured data from architectural building layouts, improving extraction accuracy by 20%
- Fine-tuned and benchmarked the Qwen-VL 2.5 vision-language model for Arabic OCR, reducing Word Error Rate (WER) from 42.32% to 15.94% and Character Error Rate (CER) from 9.73% to 4.25%
- Built AI workflow automation in n8n, integrating APIs, retrieval systems, and recommendation pipelines for sales management applications

A-Technologies 05/2024 – 08/2025
Junior .NET Developer *Cairo, Egypt*

- Developed scalable .NET Core microservices and REST APIs for an e-commerce platform, maintaining 99.9% uptime
- Integrated MeiliSearch full-text search, reducing search response time by 45%
- Applied Agile methodologies, Git workflows, and software engineering best practices to build scalable backend systems

Technical Skills

Programming Languages: Python, C#, C/C++, SQL, JavaScript, TypeScript

AI & Machine Learning: Machine Learning, Deep Learning, NLP, Computer Vision, OCR, Model Fine-Tuning, LoRA/PEFT, Model Evaluation & Benchmarking, PyTorch, TensorFlow, Hugging Face Transformers, Scikit-Learn, PaddleOCR, Detectron2, OpenCV, Whisper

LLM & Agentic Systems: Large Language Models (LLMs), Agentic AI, Multi-Agent Systems, Retrieval-Augmented Generation (RAG), Vision-Language Models (VLMs), Prompt Engineering, Embeddings, Semantic Search, LangChain, CrewAI, Sentence Transformers, ChromaDB, Qdrant, Azure OpenAI, OpenRouter

Data Science & Analytics: Exploratory Data Analysis, Data Visualization, Clustering, Unsupervised Learning, Dimensionality Reduction (PCA), Genetic Algorithms, Multi-Objective Optimization, Pandas, NumPy, Matplotlib, Seaborn, Plotly

Backend & APIs: ASP.NET Core, FastAPI, REST APIs, Microservices, RabbitMQ, Kafka, Streamlit, Gradio

Databases & Data Engineering: PostgreSQL, MySQL, SQL Server, MongoDB, Oracle, Redis, Elasticsearch, MeiliSearch, Vector Databases, Apache Spark

Tools & Core Concepts: Git, GitHub, Linux, Jupyter, n8n, Software Engineering, System Design, Agile Development, Database Design, API Development, Data Security, Version Control, SDLC

Projects

PolySumm: Multimodal RAG for Scientific Paper Summarization GitHub Repository | *Python, PyTorch, Hugging Face Transformers, LoRA/PEFT, LangChain, FastAPI, ChromaDB, PaddleOCR, Detectron2, Docker*

- Built a multimodal Retrieval-Augmented Generation (RAG) system that summarizes scientific papers by fusing text, figures, and tables, deployed as Dockerized FastAPI microservices behind a React frontend
- Engineered the document AI ingestion pipeline: Detectron2 and LayoutParser layout detection with PaddleOCR (PP-OCRv3) text and table recognition converting PDFs to structured JSON, then semantic chunking and Sentence-Transformer embeddings indexed in a ChromaDB vector database
- Fine-tuned and served modality-specific models: PEGASUS-X for abstractive text summarization (Hugging Face Trainer, ROUGE evaluation) and PaliGemma-3B with LoRA/PEFT adapters for figure captioning, plus a Qwen vision-language model for table-to-text generation

Agentic Analysis: Multi-Agent AI Data Analyst GitHub Repository | *Python, CrewAI, LLMs, Agentic AI, Prompt Engineering, Pandas, Matplotlib, Seaborn, Streamlit*

- Built a multi-agent AI data analysis system with CrewAI in which three role-specialized Large Language Model (LLM) agents run as a sequential crew with task-chained context, turning an uploaded CSV or Excel dataset into an executive-ready PDF report
- Developed five custom agent tools (data loader, pandas-backed data query, chart discovery, chart generation, PDF report writer) enabling automated exploratory data analysis and programmatic chart generation
- Implemented a conversational chat agent that answers natural-language questions over the dataset by autonomously selecting the appropriate tool per query, and deployed the app to Streamlit Cloud with model-agnostic LLM routing via OpenRouter

Video RAG Assistant: Question Answering over Video GitHub Repository | *Python, PyTorch, OpenAI Whisper, Sentence Transformers, Qdrant, OpenCV, FFmpeg, LLMs, RAG*

- Built an end-to-end multimodal RAG assistant for question answering over video, orchestrating a five-stage pipeline: FFmpeg audio extraction, Whisper speech-to-text transcription with timestamps, OpenCV keyframe extraction, embedding generation, and vector indexing
- Implemented semantic search over transcript segments using 384-dimensional all-MiniLM-L6-v2 embeddings with cosine similarity and top-k retrieval in a Qdrant vector database, grounding LLM-generated answers in timestamped source segments and keyframe thumbnails

Large Route Optimization using Genetic Algorithms GitHub Repository | *Python, Genetic Algorithms, Multi-Objective Optimization, NetworkX, Folium, Plotly, Pandas, Streamlit*

- Designed a multi-objective genetic algorithm for multi-truck delivery route optimization (vehicle routing), minimizing cost, travel distance, and deadline violations under truck weight, size, and capacity constraints
- Implemented the full evolutionary stack from scratch: permutation chromosome encoding, order and edge crossover, five mutation operators, tournament and exponential rank-based selection, and elitism survivor selection
- Built a Streamlit dashboard with Folium geospatial route maps, NetworkX graph modeling of the city network, and Plotly fitness-per-generation tracking for hyperparameter experimentation

Education

- **Helwan University** Cairo, Egypt
  Bachelor of Science in Computer Science and Artificial Intelligence *GPA: 3.7/4.0*