Adham Ehab

UAE, Dubai (Residency Visa) | +971 565689691 | createdbyadham@gmail.com | linkedin.com/in/adhamehab | github.com/createdbyadham | Personal portfolio

Summary

UAE Resident with valid visa, available to join immediately. AI Engineer with experience building production-grade **Generative AI & ML systems**, backend services, and data pipelines. Skilled in Python, FastAPI, and end-to-end data engineering, with specialized expertise in LLMs, RAG/CAG/MCP, and multi-agent orchestration. Delivered scalable AI solutions and real-world projects, including Compass, a multi-agent productivity system.

Technical Skills

AI/ML Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace, LangChain, vLLM, ONNX Runtime
Backend & Infrastructure: FastAPI, Node.js, Docker, AWS (S3 & Lambda), Redis, REST, GraphQL, WebSockets
Development & Tools: PostgreSQL, MongoDB, SQLite, ChromaDB, pgvector, Git, CI/CD, Pytest

Experience

AI & Software Engineer (Contract) - Al Qemah — Dec 2025 - Present

Sharjah, UAE

- Built a field service management system using **FastAPI** and **PostgreSQL**, replacing paper forms with a fully digital workflow, and wrote automated unit tests using Pytest.
- Made a structured data ingestion system to capture the following data: (geolocation, timestamps, standardized service logs), creating a clean dataset ready for future **predictive models**.
- Developed a React Native (Expo) app for field technicians using TanStack Query for offline-first data synchronization and digital signature capture, integrated with a React/Vite admin dashboard.
- Automated regulatory compliance reporting by building a **dynamic document generation service**, reducing manual data entry by **20+ hours/week** through auto-populated templates and image optimization.

AI Engineer Intern - Beetleware — Aug 2025 - Nov 2025

Remote

- Selected as **1 of 20 finalists** from **1000+ applicants** for Beetleware's AI internship.
- Built a **distributed BERT training pipeline** for Yelp Review Full dataset using **Ray across 3 GPU workers**. Solved worker connectivity using Zerotier Network and migrated to Linux for NCCL support.
- Won **1st place** on public leaderboard and **2nd place** on private leaderboard in a Kaggle competition. Engineered a **Stacking Ensemble pipeline** using Stratified K-Fold OOF predictions from 8+ base models (XGBoost, LightGBM, CatBoost). Used **Optuna** for hyperparameter optimization.
- Built two end-to-end AI agents: a **Text-to-SQL agent** for database querying and a **medical chatbot** using Retrieval-Augmented Generation (RAG) and prompt engineering.

AI Engineer - Independent Consultant — Jan 2025 - Nov 2025

Remote / Various Clients

- **Predictive Analytics System (MC - Egypt):** Engineered a student tracking system with an adaptive ML model to flag at-risk students, reducing dropout risks by **20%** through continuous retraining cycles.
- **Enterprise RAG Pipeline (WhiteWood - UAE):** Designed a knowledge-base augmented RAG system combining vector retrieval with structured data relationships for context-aware business querying.
- **Open Source Fine-Tuning:** Fine-tuned LLaMA and Mistral models using **QLoRA** for domain-specific tasks, and deployed high-throughput inference pipelines using **vLLM** and **ONNX Runtime**.

AI & Robotics Intern - GateIn Technology — June 2024 - July 2024

New Mansoura, Egypt

- Developed an AI-powered classroom assistant capable of automating attendance using **computer vision** and face recognition with **95% accuracy**.
- Integrated NLP models to automatically generate quizzes and answer student queries in real-time using **OpenAI APIs**.
- Reduced manual attendance processing time by **90%** through intelligent content generation and automation.

Projects

Compass - AI Task Assistant | *Go, Python, RAG, MCP, React* | Demo | Github

- Built a **multi-agent system** where specialized agents (Planner, Task, etc.) collaborate to decompose complex requests via a custom **Model Context Protocol (MCP)** server for dynamic tool invocation.
- Implemented **Semantic Caching** using **Redis** to intercept and serve semantically similar queries, cutting down response latency to **<100ms** and reducing API costs by **30%** by bypassing redundant inference.
- Optimized token usage by indexing 40+ tool definitions inside **pgvector**, dynamically retrieving only relevant functions per turn through a RAG pipeline to **cut input tokens by 50%** and **latency by 40%**.
- Developed an execution layer enabling agents to trigger **simultaneous actions** (e.g., creating a Todo while updating a Habit) without blocking the central orchestrator.

Trace - Intelligent Receipt Analysis Platform | *FastAPI, React, PaddleOCR, ChromaDB, Docker* | Demo | Github

- Engineered a cost-effective ingestion pipeline using **PaddleOCR** and a **Quantized SLM (Phi-3.5 4-bit)** to structure raw OCR text into type-safe JSON, reducing ingestion costs by **90%**.
- Architected a **Hybrid RAG pipeline** combining Dense Vector Search (ChromaDB) and Sparse Keyword Search (BM25), to fix retrieval issues where semantic search failed on exact keyword matches.
- Achieved **91%+ Context Precision** (benchmarked via **RAGAS**) by implementing a **Cross-Encoder Reranking** layer and **Parent Document Retrieval** Item-level chunking to optimize context window relevance.
- Deployed a fully containerized stack using **Docker Compose** with persistent volumes for vector stores and model weights, ensuring 100% data privacy and offline capability.

CAPTCHA-Solver (Computer Vision) | *Python, YOLO11, OpenCV, Synthetic Data* | Demo

- Engineered a two-stage vision system to solve image-based CAPTCHAs using **YOLO11**, Stage 1 identifies target icons/order, and Stage 2 locates and clicks targets in the required order. achieving **near 100% accuracy**.
- Built a data generation engine that **synthesized 10k+** labeled training samples (360° rotation, color variations, scaling, and actual backgrounds from the original CAPTCHA).

LiteDB - SQL editor/viewer | *Tauri, Rust, React, ONNX, PostgreSQL, SQLite* | Github | Landing Page

- Built a high-performance (idle RAM usage ≈ 5MB) light-weight cross-platform application using **Tauri (Rust)**, providing interactive schema visualization and seamless operations for PostgreSQL and SQLite databases.
- Designed a schema-aware **Text-to-SQL AI Agent** that translates natural language into complex queries, via the use of local models (Ollama) or cloud-based models (OpenAI, Azure).
- Engineered an on-device **Vector Search** engine to help developers validate RAG pipelines locally. Utilizing **pgvector** and **ONNX** (Transformers.js), with local embedding models (BGE-Large, all-MiniLM) to execute privacy-first semantic matching using Cosine, L2, and Inner Product metrics.
- **Built an automated CI/CD pipeline** via GitHub Actions to cross-compile Rust binaries for Windows, macOS, and Linux, with an over-the-air (OTA) update system.

Education

New Mansoura University 2021 - 2025
Bachelor of Science in Computer Science *New Mansoura, Egypt*

- Graduation Project: *COMPASS – AI-Powered Productivity System* (Grade: A+)

Certifications

IELTS Academic (Band 7.0) 2025
Listening: 7.5, Reading: 7.5, Writing: 7.0, Speaking: 6.5

Deep Learning Nanodegree 2023
Udacity - Neural Networks, CNNs, RNNs, GANs, Deployment Pipelines