YOUSEF FERWANA

AI Engineer | LLM Fine-Tuning, RAG Systems, AI Agents, MLOps

Kuwait City, Kuwait | +965 5134 3165 | ferwanayosef@gmail.com

linkedin.com/in/yousef-ferwana | yousefportfolio.online | github.com/yh12-10

PROFESSIONAL SUMMARY

AI Engineer with 1+ year of full-time experience building production Large Language Model (LLM) systems: Retrieval-Augmented Generation (RAG), LLM fine-tuning with Parameter-Efficient Fine-Tuning (PEFT) and QLoRA, self-hosted LLM inference with vLLM, and AI agents. Fine-tuned 6 domain LLMs on proprietary datasets, built a production Arabic RAG platform used by 400+ teachers as Chief Technology Officer (CTO), and shipped a 16-agent multi-agent system with code-level hallucination guardrails and LLM evaluation. B.Sc. in Data Science and Artificial Intelligence, GPA 3.95/4.00.

EXPERIENCE

AI & Data Engineer **November 2025 - Present**

Procapita Group, Kuwait City, Kuwait

- Fine-tuned 6 domain-specialist Large Language Models on Qwen models with QLoRA and PEFT using PyTorch and Hugging Face Transformers, trained on 5,000+ enterprise documents for competency generation and role-similarity reasoning.
- Engineered 10+ proprietary fine-tuning datasets, including a flagship 65,000-row JSONL job-title-to-competency dataset, with a Python data pipeline that cleans, deduplicates and validates 10,000+ raw source files.
- Architected a self-hosted LLM inference fleet with vLLM on multiple Ubuntu Linux GPU servers over a private Tailscale network, delivering sub-second latency through OpenAI-compatible APIs and removing 100% of cloud API spend.
- Automated data preparation with a continuous ingestion pipeline that structures incoming records for each retraining cycle, replacing all manual preprocessing and keeping data private on company infrastructure.

PROJECTS

Nabgh - Arabic AI EdTech Platform (Production RAG System) **August 2026 - Present**

Role: Chief Technology Officer (CTO) and Technical Owner | Python, BGE embeddings, ChromaDB, GPT models, Node.js, MongoDB, AWS

- Lead the technology for a live AI teaching platform (iOS, Android, admin dashboard) used by 400+ teachers in Kuwait, owning system architecture, AI features, releases and AWS (EC2, S3) infrastructure.
- Developed a production Arabic Retrieval-Augmented Generation (RAG) pipeline over 249 textbooks (about 28,500 pages, 55,000+ chunks) with hybrid search (BGE embeddings, BM25, Reciprocal Rank Fusion), cross-encoder reranking and a ChromaDB vector database, reaching 100% recall@5 and 0.95-0.99 MRR on a 56-question golden set.
- Reduced hallucinations by enforcing grounding in code with a pre-generation refusal gate, citation and quote verification and page-level grounding; 4,340 of 4,358 generated quiz items (99.6%) proven printable in production.
- Designed an LLM evaluation chain (a fast GPT model drafting, a stronger GPT model as an LLM-as-a-judge blind solver, rule-based checks) at $0.014 per worksheet, raising the independent audit pass rate from 11/44 to 44/44.
- Shipped 150+ RAG releases in about 7 weeks with a multi-agent AI coding workflow (Claude Code) and 1,894 automated tests, and cut OCR cost to $0.0007-$0.0025 per page.

Lead and Competitor Intelligence Platform (LLM Data Extraction) **October 2026 - November 2026**

Role: AI Engineer | n8n, Claude (OpenRouter), JavaScript, Python, Google Apps Script, Google Sheets

- Delivered 2 n8n workflow automations (95 nodes) for a Gulf B2B AI-software company that discover, enrich and score B2B leads and extract competitor facts with an LLM into structured JSON and a shared Google Sheet.
- Prevented hallucinated facts by requiring a verbatim source quote for every LLM-extracted claim and verifying it in code, with QA retries and a deterministic fallback; verified on 166 live businesses with 0 lost IDs, 0 duplicates and 519 automated checks.

Yousef HQ - Local Multi-Agent AI System **October 2026 - Present**

Role: Builder | Claude Code, Hermes Agent, Qwen models, LM Studio, Model Context Protocol (MCP), Python

- Created a multi-agent system of 16 AI agents with tool calling and least-privilege access that scout and verify Gulf B2B leads and produce bilingual Arabic/English content at $0 inference cost on a local GPU.

- Implemented a hallucination guardrail where Python computes every fact and the local LLM only phrases it; a claims validator raised the pass rate from 1/6 to 6/6 and prompt size dropped 82% (17,349 to 3,036 tokens).
- Added human-in-the-loop approvals with fail-closed hooks and a PIN-gated dashboard, covered by 1,825 automated tests, including mutation tests that catch 16/16 safeguard breaks.

Athar Fadwa - Bilingual Islamic App with LLM Assistant **December 2025 - June 2026**

Role: Sole Developer | React, TypeScript, tRPC, Supabase PostgreSQL, Gemini models | atharfadwa.xyz

- Launched a live Arabic/English progressive web app serving 6,236 Quran verses, 29,357 hadiths and 14,109 tafsir entries, with an LLM assistant, push notifications and spaced-repetition memorization.
- Introduced an end-to-end test harness against the live database that caught 5 broken features, and diacritic-insensitive Arabic search that raised matches for a core query from 0 to 1,748.

AuthentiGen - Open-Source Computer Vision Pipeline **April 2026 - June 2026**

Role: Sole Developer | TypeScript, Node.js, sharp, React, MySQL, Docker, GitHub Actions | github.com/yh12-10/authentigen

- Replaced a generative-image prototype with a deterministic 12-stage computer vision pipeline for VFX and data augmentation, cutting per-image API cost to $0 with 12/12 byte-identical reruns at 2.2 images per second.
- Released v3.0.0 under the MIT license with Docker deployment, CI/CD on GitHub Actions and 69 automated tests.

Medicaa - Medical RAG Chatbot (Graduation Project) **July 2025**

Role: AI Engineer | LLaMA models, LangChain, Pinecone, Sentence Transformers, Flask

- Prototyped a RAG chatbot over the Gale Encyclopedia of Medicine with semantic search and local LLaMA-model inference, reaching 84% accuracy vs. 72% for a smaller model in an evaluation with medical students.

EDUCATION

B.Sc. in Data Science and Artificial Intelligence **July 2025**

Al-Ahliyya Amman University, Amman, Jordan | GPA: 3.95/4.00 (Excellent)

TECHNICAL SKILLS

Programming: Python, SQL, JavaScript, TypeScript, C#, HTML5, CSS3

AI/ML: Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), LLM Fine-Tuning, LoRA, QLoRA, PEFT, LLM Evaluation, LLM-as-a-Judge, Prompt Engineering, AI Agents, Multi-Agent Systems, Embeddings, Semantic Search, Reranking, Natural Language Processing (NLP), Arabic NLP, OCR, Computer Vision, Deep Learning

Frameworks & Libraries: PyTorch, Hugging Face Transformers, vLLM, LangChain, Sentence Transformers, scikit-learn, Pandas, NumPy, Flask, Node.js, React, n8n, Claude Code, Model Context Protocol (MCP)

Data & Infrastructure: Vector Databases (ChromaDB, Pinecone), PostgreSQL, MongoDB, SQL Server, Docker, Git, GitHub Actions, CI/CD, MLflow, MLOps, Linux (Ubuntu), Tailscale

Cloud & Visualization: AWS (EC2, S3, AI Services), Microsoft Azure, Power BI

CERTIFICATIONS & LANGUAGES

Certifications: IBM AI Engineering Professional Certificate; IBM Generative AI Engineering; IBM Fine-Tuning LLMs; IBM RAG and Agentic AI; Microsoft Certified: Azure Administrator Associate (AZ-104); Machine Learning in Python (Anaconda); Become a Data Analyst (LinkedIn Learning)

Languages: Arabic (Native), English (Fluent)