Ahmad Faour

Senior AI Engineer

Riyadh, Saudi Arabia | +966-596824729 | ahmad.saleh.faour@gmail.com | linkedin.com/in/ahmad-faour | github.com/ahmadfaour9

Professional Summary

Senior AI Engineer with **5+ years of combined experience** across AI/ML engineering and software development in freelance, internship, part-time, and full-time roles. Proven track record building **LLM-powered systems, agentic workflows, and production-grade AI services** end-to-end, from data preparation and fine-tuning (**PEFT/LoRA, RLHF**) to model optimization, quantization, **ONNX** export, and cloud deployment across **GCP, AWS, and IIS-hosted environments**. Strong at bridging **Python ML stacks** with **ASP.NET Core back ends** and **React front ends** to deliver secure APIs, real-time experiences, and scalable deployments. Hands-on with **RAG pipelines, tool-calling agents, orchestration frameworks, vector memory, n8n, and Elsa Workflows**, with a focus on reliable, business-impacting AI systems.

Education

University of Hull — Hull, United Kingdom
M.Sc. in Artificial Intelligence — *Sep. 2025 – Present*
– Currently pursuing Master's studies in Artificial Intelligence, with focus on computer vision, deep learning, and LLM applications.

University of Kalamoon — Deir Atiyah, Rif Dimashq
B.Sc. in Information Technology Engineering, GPA: 3.08 (83/100) — *Aug. 2018 – Feb. 2024*
– Graduated with distinction; achieved the **2nd highest rank** in the class.
– Relevant coursework: Artificial Intelligence, Machine Learning, Neural Networks, Data Mining, Image and Pattern Recognition, Deep Learning, Natural Language Processing, Data Structures, Algorithms, Database Systems, and Cybersecurity.

Yousef Seman Industrial High School — Yabroud, Rif Dimashq
Diploma in Computer Technology, GPA: 92/100 — *Aug. 2016 – May 2018*
– Top academic performer across grades 10–12; ranked **5th in the governorate** in final 12th-grade exams.
– Key coursework: Computer Programming, Electronics, Digital Logic, Network Fundamentals, Technical Drawing, Applied Mathematics, and Physics.

Professional Experience

AI Engineer (Full-time, On-site) — Dec. 2024 – Present
NVSSoft — *Riyadh, Saudi Arabia*
– Designed and implemented an **ASP.NET Core** application with a secure **REST API** layer using Entity Framework/DbContext, alongside **Python**-backed microservices for model inference.
– Configured **IIS** and production environments for availability and throughput; partnered with DevOps to enable **CI/CD** through automated builds, tests, and multi-stage deployments.
– Built a **React** front end integrated with the .NET back end for real-time AI interactions, reducing response latency and improving user experience.
– Orchestrated **AI agent** workflows with **n8n** and **Elsa Workflows**, including webhooks, scheduled jobs, tool chaining, retries, and human-in-the-loop approval flows.
– Applied **SOLID** principles and clean architecture for maintainability, and authored internal documentation and onboarding runbooks.

AI Training & Model Evaluation Specialist (Part-time, Remote) — Dec. 2025 – Present
micro1 — *Remote*
– Evaluated LLM outputs using structured rubrics covering helpfulness, correctness, safety, and style, and documented actionable feedback to improve response quality.
– Supported **RLHF/RLAIF**-style training workflows by refining prompts, identifying failure modes, and proposing guideline updates to improve consistency.
– Performed quality checks and calibration to reduce evaluator variance while maintaining strict adherence to confidentiality and data-handling requirements.

AI Engineer (Part-time, Remote) — Dec. 2024 – Dec. 2025
Reality AI Lab — *Remote*
– Contributed to **Marvel Model**, a next-generation generative **image** model, by building Python training and fine-tuning pipelines and improving stability and scalability.

– Developed tooling for **data augmentation**, feature extraction, and **custom evaluation metrics** for image-based models; presented progress during sprint reviews.

Data Scientist Intern (Part-time, Remote) Sep. 2024 – Feb. 2025
Darrebni *Homs, Syria*
– Built and deployed predictive models for customer behavior and market trends using advanced **SQL** and machine learning; performed univariate, bivariate, and multivariate analysis for actionable insights.
– Applied **feature engineering**, preprocessing, normalization, dimensionality reduction, and **anomaly detection** to improve decision quality and operations.
– Created interactive **Tableau** dashboards and collaborated with engineering and product teams to integrate insights into workflows.

AI Engineer (Part-time, Remote) Sep. 2022 – Dec. 2024
Freelancer.com *Remote*
– Delivered client NLP solutions for classification, summarization, and sentiment analysis using **Transformers**; built an **Arabic handwriting recognition** system achieving **~85% accuracy**.
– Applied **quantization** and exported models to **ONNX**, reducing model size by **59%** and accelerating inference in production deployments.
– Constructed agentic NLP workflows with **LangChain**; fine-tuned and adapted models using **PEFT** and **RLHF** to improve response quality.
– Optimized Node.js/**Express** back ends, improving throughput and reliability by **~20%**.

Machine Learning Intern (Part-time, Remote) Apr. 2024 – Jun. 2024
CampWallah *Sholay, India*
– Integrated ML models into production, **increasing accuracy by ~25%**, and built optimized data pipelines for real-time inference and secure API delivery.
– Contributed to API design for seamless model deployment and monitoring.

Technical Skills

Programming: Python, C#, JavaScript/TypeScript, SQL, MATLAB

AI/ML Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, Pandas, NumPy, Hugging Face Transformers, ONNX

LLM & RAG Systems: Embeddings (Sentence Transformers / OpenAI embeddings), chunking strategies, retrieval, reranking (cross-encoder), RAG evaluation (RAGAS / human evaluation rubrics), structured outputs, prompt evaluation, prompt versioning, guardrails

Vector Databases: ChromaDB, FAISS, Pinecone, Weaviate

LLM Platforms: Google Vertex AI (Gemini), OpenAI, Anthropic

Agents & Orchestration: LangChain, n8n, Elsa Workflows, tool-calling, planner/executor patterns, long-context retrieval/memory, human-in-the-loop

Backend & APIs: FastAPI, Flask, ASP.NET Core, Entity Framework, REST APIs, WebSockets, Uvicorn, Gunicorn

Techniques: Fine-Tuning (PEFT/LoRA), RLHF, quantization, model optimization, ONNX conversion, multi-task learning, domain adaptation

MLOps/DevOps: Docker, Git, CI/CD, IIS, AWS, GCP, PythonAnywhere, Heroku, MLflow, Weights & Biases, Airflow, Prefect

Observability & Testing: OpenTelemetry, Prometheus, Grafana, logging/metrics/tracing (basic), PyTest

Security: OAuth2, JWT, secrets management (AWS Secrets Manager / GCP Secret Manager)

Data & BI: Tableau, data preprocessing, feature engineering, anomaly detection

Soft Skills: Analytical Thinking, Problem Solving, Team Collaboration, Communication, Adaptability, Time Management

Certifications & Awards

Develop AI-Powered Prototypes in Google AI Studio Feb. 2026
Google *View Credential*
– Designed and deployed AI-powered prototypes using **Google AI Studio** and Gemini models, focusing on rapid experimentation, prompt engineering, and production-ready workflows.

AI Software Engineer Oct. 2025
micro1 *View Credential*
– Certified in **Reward Model Development**, **Autograder/Benchmark Design**, and LLM evaluation workflows supporting high-quality AI systems.

Generative AI with Large Language Models Oct. 2024
Coursera & AWS
– Covered **transformer architectures**, **PEFT/LoRA**, and **RLHF**, with practical cloud deployment of LLM systems.

Introduction to Retrieval Augmented Generation (RAG) Dec. 2024
Duke University & Coursera
– Built RAG pipelines using **Python**, **LLMs**, and vector databases.

Intermediate Machine Learning Feb. 2025
Kaggle
– Applied **XGBoost**, advanced preprocessing, and **cross-validation**.

Feature Engineering Nov. 2024
Kaggle
– Designed **feature engineering** workflows to improve model performance.

AWS EMEA Innovate: Migrate. Modernize. Build. Oct. 2024
AWS
– Covered scalable AI deployments on AWS cloud infrastructure.

Azure DevOps: Intro to CI/CD Mar. 2024
United Latino Students Association
– Integrated ML model deployment into automated CI/CD workflows.

Data Science Internship Mar. 2025
Darrebni
– Worked on **predictive modeling**, **data visualization**, and **SQL pipelines**.

Projects

NexusAI: Intelligent Customer Experience Platform | Live Demo | GitHub Jan. 2026 – Present
– *Next.js 14, React, TypeScript, Node.js/Express, GCP Cloud Run, Vertex AI (Gemini), Firestore, BigQuery, Docker.* Built an enterprise support automation platform for high-volume e-commerce, automating **60–80%** of inquiries through intent classification and **RAG**.
– Delivered a multilingual chatbot UX (English/Arabic) with automatic RTL switching, plus an admin portal for live operations including agent takeover, case handling, order management, and staff administration.
– Implemented secure authentication with **JWT** (HTTP-only cookies) and **RBAC**; enabled real-time monitoring via **Firestore** and analytics via **BigQuery** for latency, error rate, top intents, and traffic trends.

AI Research Intelligence Agent | GitHub Nov. 2025 – Present
– *Python, Asyncio, Streamlit, ChromaDB, GraphRAG, Docling/PyMuPDF, OCR, OpenAI/Anthropic/Gemini/Ollama.* Built a production-grade research intelligence system for discovery, deep analysis, and distribution of AI papers with forensic verification.
– Developed a **Scientific Code Forge** that converts papers section-by-section into modular Python/PyTorch codebases, implementing equations and generating training and evaluation pipelines.
– Implemented a **Repo Mapper** linking paper concepts to GitHub code, and a **Scientific Auditor** that cross-checks claims against extracted table metrics, computes a "Hype Score," and supports multi-agent adversarial debate with moderator consensus reports.

AI Assistant with Voice Communication | Freelancer.com Oct. 2024 – Present
– Architected and led development of a multilingual, dialect-aware AI voice assistant combining **Whisper v3**, a fine-tuned **UBC-NLP/MARBERTv2** model for Arabic dialect classification, and downstream reasoning with GPT-family LLMs.
– Designed a modular inference pipeline capable of **sub-500ms response times** for short utterances, integrating ASR, NLP, and reasoning components through REST APIs.
– Implemented **cosine similarity**-based semantic matching to select contextually optimal responses across dialects.
– Built a full-stack application with a **Flask** API backend and **React** UI using WebSocket-based real-time updates for low-latency streaming.
– Created automated orchestration flows in **n8n** and **Elsa Workflows** to handle event triggers, queue management, retries, and human-in-the-loop review for ambiguous cases.
– Improved dialect recognition F1 score by **12%** over baseline Whisper-based approaches through fine-tuning and dataset curation.

Schema-Aware Query Generator Mar. 2025 – May 2025
– *Python, Streamlit, MS SQL Server, Vanna, ChromaDB.* Built a retrieval-grounded NL-to-SQL assistant with schema-aware guardrails (for example, no `SELECT *`) and safe execution constraints.
– Implemented grounding and validation checks to reduce hallucinated columns and tables and improve reliability for business analytics queries.

Arabic Handwriting Recognition | Flask, TensorFlow, ONNX, PythonAnywhere Jun. 2024 – Jul. 2024
– Developed and deployed an Arabic character recognition system using a **CNN-based TensorFlow model** trained on a custom dataset of 120,000 handwritten samples across 28 character classes.
– Achieved **85% top-1 accuracy** after improving preprocessing, data augmentation (rotation, elastic distortions), and dropout regularization.
– Performed **quantization-aware training** followed by ONNX conversion, reducing model size by **59%** and achieving a **2.3x** inference speedup on CPU-only servers.
– Deployed the system to **PythonAnywhere** with a Flask API endpoint for public testing; handled 5,000+ requests per

day with no downtime.

Mixture-of-Experts Image Captioner | GitHub Sep. 2025 – Oct. 2025

– *Python, Streamlit, ResNet-50, Ollama (Llama3), Docker.* Built a hybrid MoE image-captioning system that routes between fast template captions and richer LLM-generated captions to balance speed, interpretability, and detail.

– Implemented a **vision expert** using ResNet-50 with color and complexity signals, plus a **gating network** using confidence and entropy thresholds to select template vs. LLM generation.

– Shipped a **Streamlit UI** for interactive testing, a CLI for batch runs, and full **Dockerized** deployment with YAML-driven configuration for device, prompts, and logging.

Handwritten Cardiac Prescription Recognition | Multitask Learning, GAN, CRNN, Aug. 2022 – Feb. 2024
CTC Loss

– Designed and implemented a multitask recognition system combining **GANs** for data augmentation, **CRNN** for sequence modeling, and **CTC loss** for variable-length transcription.

– Processed complex prescriptions containing mixed Latin and Arabic scripts, dosage units, and physician signatures.

– Integrated a verification pipeline for detecting hospital seal authenticity and standardizing dosage instructions, reducing interpretation errors by **35%**.

– Achieved **˜92% character-level accuracy** on a 50,000-sample dataset, outperforming a CNN+LSTM baseline by **8%**.

AI-Driven HR Management System | NLP, Machine Learning, Web Development Jan. 2022 – Jun. 2022

– Developed an AI-assisted recruitment tool integrating **TF-IDF** and **NLTK**-based text processing to automatically match candidates to job descriptions.

– Trained and evaluated classical ML models (**KNN, SVM, Random Forest**), with **KNN achieving 98% accuracy** in candidate-job match prediction.

– Designed a secure, role-based web application using **Python/Django** for the backend and **JavaScript** for the interactive UI.

– Enabled recruiters to shortlist candidates in **40% less time** through automated ranking and filtering features.