

MARWAN AL-MASRAT

AI ENGINEER

Phone: +962786834685 • Email: masrat.marwan@gmail.com • LinkedIn: linkedin.com/in/marwan-al-masrat
GitHub: github.com/Marwan-ALMasrat • Portfolio: marwan-almasrat.github.io

PROFESSIONAL SUMMARY

AI Engineer experienced in designing and deploying end-to-end NLP and Retrieval-Augmented Generation (RAG) systems using vector databases and modern LLM frameworks. Skilled in building production-ready AI applications, LLM evaluation, and fine-tuning transformer and YOLO models. Currently expanding expertise in Knowledge Graphs, FastAPI, Docker, and AI observability through the AI.SPIRE program. Seeking an AI Engineer position to build scalable, production-grade AI solutions.

Leadership: [Founded and led Sufra Alqalam \(110 members\), awarded 5,000 JOD by the Jordanian Ministry of Culture.](#)

SKILLS

Languages & Tools: Python, SQL, Git, FastAPI, Docker, Linux, VS Code, Google Colab, Gradio, Streamlit

ML & Deep Learning: PyTorch, TensorFlow, Keras, Scikit-Learn, Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), YOLOv8 (Ultralytics), Transformer Models (Hugging Face), Sentence Embeddings, ColPali

NLP & LLM Systems: LangChain, spaCy, NLTK, Tokenization, Transfer Learning & Fine-Tuning, Prompt Engineering, Retrieval-Augmented Generation (RAG), LLM Evaluation (EM, F1, ROUGE)

Vector Search & Databases: ChromaDB, Weaviate (BM25 + Vector Hybrid Search), Neo4j (Cypher), RDF, PostgreSQL, SQLite, PyMuPDF, Roboflow

AI Systems & LLMops: End-to-End RAG Pipelines, Multimodal Retrieval, Knowledge Graphs (Fuseki, SPARQL), Workflow Automation (n8n), Monitoring & Observability (Prometheus, Grafana), OpenRouter API

Data & Visualization: NumPy, Pandas, Matplotlib, Seaborn, Plotly, Power BI, Exploratory Data Analysis (EDA), Hypothesis Testing, KPI Design

Domains: Retrieval-Augmented Generation (RAG), Multimodal AI, NLP & Transformer Fine-Tuning, LLM Evaluation & Benchmarking, Computer Vision & Object Detection, Knowledge Graphs & Semantic Web, Vector Search & Embedding Systems, Predictive Modeling, Exploratory Data Analysis

Soft Skills: Technical Communication, Cross-Functional Collaboration, Problem-Solving, Adaptability, Leadership, Time Management, Critical Thinking

EXPERIENCE

AI Engineer Trainee • *AI.SPIRE: Applied AI and ML Systems — LevelUp Economy & Istidama Consulting*
Feb 2026 – Present

- Enrolled in a 24-week intensive program covering NLP, machine learning, RAG systems, and Knowledge Graphs in practice.
- Built and deployed end-to-end AI systems integrating NLP pipelines, vector search, and APIs.
- Developed professional skills in technical communication, teamwork, and AI system documentation.

Machine Learning & Deep Learning • Qafza Tech

Dec 2024 – May 2025

- Built a K-Nearest Neighbors classifier to predict mobile user behavior class from usage patterns (app usage time, screen-on time, battery drain, data usage, installed apps, age, gender), achieving 90.7% validation accuracy and a 90.7% F1-score across 5 behavior classes on a 700-row dataset..

Data Scientist Trainee • SHAI for AI

Jul 2024 – Dec 2024

- Built classification and regression models (Scikit-Learn, Random Forests) for prediction challenges.
- Built a customer churn prediction pipeline on a 7,032-customer telecom dataset, engineering a preprocessing workflow (OneHotEncoder for categorical features, stratified train/validation/test splits) and benchmarking Logistic Regression against a tuned Random Forest, reaching 81.3% test accuracy and a 0.86 ROC-AUC with Logistic Regression.
- Conducted in-depth exploratory data analysis (EDA) and statistical inference on large, complex datasets (e.g., Olympic history) to extract actionable insights.

Automation Engineer (n8n) • Legal Firm Project • May 2025 – Sep 2025 — Built n8n automation workflows, including an AI assistant and a legal assistant, for a legal company.

Customer Support Team Leader • iSecure LTD • May 2024 – Jun 2025 — Led the support team, managed escalations, monitored performance metrics, and reported process-improvement insights to management.

Customer Support Specialist • iSecure LTD • Jun 2023 – May 2024 — Monitored security-system alerts, resolved client inquiries, and coordinated with maintenance and emergency services on incident response.

Customer Service Supervisor • Zamzam Mall • 2020 – 2022 — Supervised daily operations, resolved customer and cash-handling issues, and maintained consistent service standards.

PROJECTS

Multimodal Retrieval-Augmented Generation for Scientific Document Retrieval

Python • Weaviate • ColPali (colSmol-500M) • Sentence-Transformers (MiniLM) • Cross-Encoder (ms-marco) • PyMuPDF • Streamlit • Gemini/Groq/OpenRouter APIs

[GitHub Repo](#)

- Designed a dual-path RAG system combining visual page embeddings (ColPali late-interaction) with semantically-chunked text, retrieved via hybrid search (BM25 + vector).
- Added cross-encoder re-ranking, raising Q&A accuracy from 80% to 95% on a 20-question scientific-paper benchmark.
- Replaced fixed-size chunking with semantic chunking, correcting 6 of 20 previously incorrect answers by preserving idea coherence across chunks.

FastAPI Monitoring & Observability Stack

• FastAPI • Docker • Docker Compose • Prometheus • Grafana • Neo4j • Weaviate • spaCy • Hugging Face Transformers • httpx • pytest

[GitHub Repo](#)

- Created a three-endpoint NLP API — Named Entity Recognition (/extract), NL-to-Cypher knowledge graph queries (/kg/query), and Retrieval-Augmented Generation (/rag/answer) on a four-service Docker Compose stack (FastAPI, Neo4j, Weaviate, Python).
- Built a full instrumentation layer: Prometheus counter, histogram, and gauge metrics exposed via /metrics, correlated structured JSON logging with request-ID propagation, and in-flight request tracking via ASGI middleware.

- Stood up a Grafana + Prometheus monitoring stack with a pre-provisioned dashboard visualizing request rate, error rate, p95 latency, and in-flight requests in real time.
- Automated a RAG grounding evaluator (eval_rag_smoke.py) that validates citation grounding against retrieved chunks, backed by a full pytest suite covering middleware behavior, metric correctness, counter increments, and smoke evaluation logic.

NLP Evaluation Lab Production-Grade Evaluation Framework for Transformer Models

Python · Transformers · PyTorch · ROUGE · SQuAD Metrics · Streamlit · Hugging Face Inference API

[GitHub Repo](#)

- Built an evaluation lab for extractive QA and abstractive summarization, implementing EM, Token-F1, and ROUGE-1/2/L from scratch with a full SQuAD-style normalization trace and a faithfulness audit that flags numeric hallucinations ROUGE alone cannot detect.
- Designed a weighted 5-factor decision matrix (data availability, task specificity, compute budget, iteration speed, quality gap) producing an auditable 0–90 fine-tune-vs-pretrained recommendation.
- Configured a 3-tier inference fallback (Hugging Face API → Colab GPU → local CPU), auto-selecting the fastest available backend and falling back silently on failure.
- Deployed live on Streamlit Cloud, turning the fine-tuning decision into a measurable, reproducible process.

Computer Vision-Based Object Detection for Jordanian Football

Python · YOLOv8 (Ultralytics) · Roboflow · Google Colab

[GitHub Repo](#)

- Extracted frames from raw Jordanian Premier League match footage (down to per-second granularity, finer where scenes were complex) and annotated approximately 300 images in Roboflow, applying data augmentation to compensate for incomplete pitch coverage and camera angles inconsistent with standard European broadcast footage.
- Fine-tuned YOLOv8 (Medium and Large variants) on Google Colab across roughly 9 training iterations to identify the best-performing model for detecting players, the ball, referees, and goalkeepers under domain-specific filming conditions absent from typical European football datasets.
- Deployed the trained model (best.pt) for inference on full match videos using the Ultralytics YOLO API, accurately localizing players, ball, referees, and goalkeepers across video frames.

AI RAG Assistant Using LangChain

Python · LangChain · IBM Watsonx AI (LLM + Embeddings) · ChromaDB · Gradio

[GitHub Repo](#)

- Engineered an end-to-end RAG pipeline using LangChain to extract precise, context-aware answers from uploaded PDF documents, combining IBM Watsonx's Mistral-Small-3.1-24B LLM with the slate-125m-english-rtvr-v2 embedding model.
- Built a document-processing flow: PDF parsing (PyPDFLoader), recursive text chunking (1,000-character chunks with 100-character overlap), and vector indexing in ChromaDB for vector search.
- Deployed an interactive Gradio interface allowing users to upload a PDF and ask natural-language questions, with answers generated live from the indexed content.

GenAI Using Reasoning & Prompt Engineering

Python · NLP · Transformers · Generative AI · Prompt Engineering · API Integration

[GitHub Repo](#)

- Built a pipeline analyzing 9,400+ Yelp reviews, performing sentiment classification and extracting actionable business insights using a GenAI reasoning engine.

- Integrated the OpenRouter API for multi-turn conversational analysis in Arabic and automated the workflow from raw feedback to structured recommendations.

Solar Power Generation Prediction System

Python · TensorFlow/Keras · Scikit-Learn · Pandas · Streamlit | Demo: [solarpowerpredictor1.streamlit.app](#)

[GitHub Repo](#)

- Built a feedforward neural network (Keras/TensorFlow, 128-64-1 dense architecture) to predict solar power output from weather data, using EarlyStopping to prevent overfitting.
- Engineered features — including a delta-angle term (angle of incidence vs. zenith) and a temperature-humidity index — and removed redundant, highly-correlated weather variables identified via correlation heatmap analysis.
- Achieved an R^2 of 0.79 and RMSE of 439.7 kW on the held-out test set (R^2 of 0.83 on training, 0.76 on validation), and used permutation importance to validate and interpret feature contributions.
- Deployed the trained model and preprocessing pipeline (serialized via Keras and joblib) as an interactive Streamlit application for real-time solar power forecasting.

EDUCATION

AI.SPIRE: Applied AI and ML Systems

Feb 2026 – Present

Python for AI: Applied NLP & Knowledge Systems · LevelUp Economy & Istidama Consulting

- 24-week intensive program covering NLP, machine learning, RAG systems, and Knowledge Graphs in practice.
- Built and deployed end-to-end AI systems integrating NLP pipelines, vector search, and APIs.

Bachelor of Science in Physics

2018 – 2021

Al-Hussein Bin Talal University · Grade: Very Good

CERTIFICATIONS

- [AWS Certified AI Practitioner Amazon Web Services \(Issued: Feb 2026, Expires: Feb 2029\)](#)
- [Data Science Methodology \(Coursera\)](#)
- [End-to-End RAG with Weaviate \(DataCamp\)](#)
- [Project: Generative AI Applications with RAG and LangChain](#)
- [Fundamentals of AI Agents Using RAG and LangChain](#)
- [AI Agents Using RAG and LangChain](#)
- [Introduction to Deep Learning & Neural Networks with Keras](#)
- [How Transformer LLMs Work](#)
- [NLP – Natural Language Processing with Python \(Udemy\)](#)
- Python for Machine Learning & Data Science Masterclass (Udemy)
- Microsoft Power BI Desktop for Business Intelligence (Udemy)
- [The Complete SQL Bootcamp: Go from Zero to Hero \(Udemy\)](#)