

Abdulmalik Alquwayfli

✉ af.alquwayfli@gmail.com

📍 Saudi Arabia, Riyadh

🌐 Abdulmalik Alquwayfli

🔗 aalquwayfli

🌐 aalquwayfli.com

EDUCATION

BSc, Information Systems, Imam Mohammed Bin Saud University

Sep 2021 – Aug 2025

Reinforcement Learning, XCS234, Stanford University

Jan 2025 – Jul 2025

EXPERIENCE

AI Engineer, NCAI - SDAIA • R&D 📄

Nov 2025 – Present

Publishing research on **multimodal semantic retrieval**, alongside **text-to-SQL agents**

AI Engineer, Mawj AI 📄

Aug 2025 – Oct 2025

- Built a **call categorization system** that turns incoming customer calls into structured dashboard insights, integrating **LLM endpoints** (e.g., ChatGPT) with **transcripts from GCP** and a custom **hierarchical categorization** algorithm to sort calls into nested intent categories like **pricing questions and complaints**.

AI Developer Intern, TAHAKOM • NLP Department 📄

Jul 2025 – Oct 2025

- Architected and deployed continuous **pre-training and post-training** pipelines on **AWS SageMaker Hyperpod**, leveraging Trainium, **GPU clusters**, and Hugging Face Optimum Neuron, covering **SFT, DPO, RLHF** for advanced alignment.
- Engineered a **parameter-efficient fine-tuning** pipeline for the **Qwen-2.5-72B** model using **LoRA** and **DeepSpeed**, ultimately **achieving a 1st place rank on the Open-Arabic-LLM-Leaderboard** 📄 with a score of **75.66**.
- Developed **LangGraph**-based **autonomous agents** with dynamic **tool calling** and real **code execution**, integrating **deep search** and file management. Collaborated with cross-functional teams, including **Data Science, Product, and Video Understanding**, on multiple use cases.

Founding Engineer, Muragib 📄

Feb 2025 – May 2025

- Trained, fine-tuned**, and evaluated 10+ open-source **BERT**-based models (e.g., CAMEL BERT, AraBERT), **achieving over 90% accuracy** with <60 ms latency, optimized for real-time applications.
- Developed and deployed advanced machine learning models for a wide range of **NLP tasks**, including **sentiment analysis**, sarcasm detection, and location distribution analysis.
- Built scalable, **AI-powered analytics** systems that deliver actionable insights from unstructured text, empowering business stakeholders to make fast, data-driven decisions.

Software Engineer VW, Microsoft 📄

Jun 2024 – Jul 2024

- Built an AI-powered **computer vision model** using Custom Vision to **recognize and classify images**.
- Conducted **data preprocessing and feature engineering** for predictive analytics solutions.
- Set up and managed **Azure Data Factory** and **Storage** resources to support data **workflows** and **AI deployment**.

SKILLS

Programming languages: Python, Julia, JavaScript, TypeScript, SQL, Rust

Libraries, tools, and frameworks: TRL, LLaMA-Factory, Unsloth, vLLM, LangGraph, PyTorch, Scikit-learn, NLTK, Pandas, NumPy, Matplotlib, Seaborn, FastAPI, Ollama, Pydantic, wandb, Temporal

Model Architectures: Attention, Transformers, BERT, RoBERTa, DDP, SFT, DPO, PPO, GRPO, RLHF, GPT, RAG, LoRA, Quantization, Multi-Agent Systems, MCP, MoE, CLIP, RNNs, YOLO

DevOps & Environments: Git, Bash, Linux, Ubuntu, Docker, Conda, Jupyter Notebook, AWS SageMaker AI, HyperPod, PyTorch Lightning, LogFire

Databases: PostgreSQL, MongoDB, Qdrant, ChromaDB, Supabase, DBeaver, Alembic, SQLAlchemy

Cloud: Digital Ocean, AWS (SageMaker, Trainium), Alibaba Cloud, TPU, AI Foundry, Microsoft Azure

GPU & HPC: CUDA, cuBLAS, cuSPARSE, cuFFT, Julia (CUDA.jl, GPUArrays.jl), SLURM, HPC

Open Source

CUDA.jl [↗](#), *JuliaGPU*

Fixed silently wrong results and missing methods in the cuBLAS and cuSPARSE wrappers, and added diag/Diagonal for GPU sparse matrices. [#3276](#) [↗](#) [#3277](#) [↗](#) [#3279](#) [↗](#) [#3275](#) [↗](#)

GPUArrays.jl [↗](#), *JuliaGPU*

Traced and fixed index arrays never reaching the device for nested GPU views. [#777](#) [↗](#)

OpenResearch [↗](#), *alphaXiv · merged*

Added the full Arabic locale, 1,146 strings, to the Rust research CLI. [#344](#) [↗](#)

camel_tools [↗](#), *CAMeL-Lab*

Fixes to CAPHI phonetic transliteration and test metadata. [#172](#) [↗](#)

Dagger.jl [↗](#), *JuliaParallel · diagnosis credited by maintainer*

Diagnosed a scheduler regression that placed every task on a single worker; the maintainer shipped the fix. [#663](#) [↗](#)

PROJECTS

nanograd [↗](#), *Autograd engine from scratch*

Builds the computation graph dynamically and backpropagates through it, with loss functions and a small neural-network layer. No dependencies.

notes2vec [↗](#), *Local-first semantic search, in Rust*

Embeds and indexes personal notes for fully offline search. Ships prebuilt binaries for Linux, macOS and Windows.

PUBLICATIONS

LARE: Low-Attention Region Encoding for Text–Image Retrieval [↗](#), *ICML 2026, First Author*

Developed a training-free retrieval method that complements global image embeddings with representations of overlooked low-attention regions, improving fine-grained text–image retrieval in cluttered scenes. Introduced **Dense-Set**, a benchmark for evaluating retrieval of less salient objects in COCO and Flickr30K.

Can Coding Agents Migrate to Post-Quantum Cryptography? [↗](#), *under review*

Built a contract-driven checker that exchanges keys and signatures with OpenSSL, then ran 160 counted agent attempts to migrate a Go file signer from RSA to ML-DSA-44. Eight patches passed local signing and verification while OpenSSL rejected their keys, and checker access changed completion by **0 points**. A single-variable ablation flipped **33 of 40** paired runs from pass to fail by shrinking the context window from 128K to 32K.

Farahidi: Rethinking Arabic Diacritization, *under review*

Showed that the same Arabic diacritization predictions score about **a fifth apart** under two widely used scorers, then compared modeling choices under one official scorer, matched budgets, and a training-test overlap audit. Built **Farahidi**, a 96M-parameter character-level encoder trained from scratch and distilled from an ensemble, reaching **0.95 DER / 2.87 WER** on Fadel-Tashkeela with case endings, the lowest reported on that benchmark, and leading WikiNews without training on Modern Standard Arabic benchmarks.

MINER: Multi-crop Inference-time Enhancement for Rare-Object Retrieval with Frozen Dual Encoders [↗](#), *ACML, First Author*

Developed a training-free inference framework that combines global image embeddings with region-level representations and hubness-corrected rescoring to improve retrieval of small, low-salience objects in cluttered scenes. Evaluated across **CLIP, SigLIP, and SigLIP 2** and introduced **ROCS**, a benchmark for rare-object retrieval.

COURSES & CERTIFICATIONS

AWS AI & ML Scholars [↗](#), *AWS & Udacity*

Jun 2025

LLMOps: Building Real-World Applications With LLMs [↗](#), *Udacity*

Dec 2024

Neural Networks and Deep Learning [↗](#), *DeepLearning.AI*

Sep 2024

Full Stack Deep Learning, *The Full Stack*

Dec 2024

Python for Data Science, AI & Development [↗](#), *IBM*

Jun 2024