

Abdulmalik Alquwayfili

✉ af.alquwayfili@gmail.com 📍 Saudi Arabia, Riyadh 🌐 Abdulmalik Alquwayfili 🌐 aalquwayfili

🌐 aalquwayfili.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

Open Source

CUDA.jl , *JuliaGPU* · 3 pull requests merged

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*