

Sultan Alrashed

+966 55 511 2967 | sultan.m.rashed@gmail.com | [linkedin.com/in/sulrash](https://www.linkedin.com/in/sulrash) | github.com/sulrash

EDUCATION

Columbia University

Master's of Science in Computer Science (In Progress)

Dec. 2027

New York City, USA

University of Manchester

Bachelor of Science (Honours) in Artificial Intelligence - Second Upper-Class

Jul. 2022

Manchester, UK

EXPERIENCE

Research Engineer

King Abdullah University of Science & Technology (KAUST)

Jan. 2025 – September 2026

Thuwal, SA

- Joined the OPTIMAL lab under the supervision of Prof. Francesco Orabona as the lead research engineer facilitating data curation and large language model pretraining, post-training, and evaluation.
- Led work for a **3B LLM pretraining** run on 384 GH200s on 5 trillion tokens: synthetically generated required missing data, designed the staged curriculum and data mixture, selected the framework, and managed execution.
- Built a distributed, asynchronous, thread-safe translation pipeline on SLURM using vLLM with a 7B model, delivering **>10B tokens/day** on V100s, for a total of **400B** tokens of translated text.
- Implemented Deepseek's mHC and Engram papers, among other techniques from recent publications, into a custom fork of Nanochat to train a 1.7B LLM on 800B tokens to validate efficiency gains of recent publications.
- Supervised two Master's students and an intern: (i) mentored development of a domain-specific vision-language model for Arabic OCR deployed in-house to OCR scanned PDFs for our pretraining pipeline, extracting **5B high quality tokens** (ii) guided data extraction from Common Crawl using **datatrove** (iii) Mentored intern on training an **mmbert** classifier to extract conversational data from a pretraining corpus.

Artificial Intelligence Engineer

Saudi Data & Artificial Intelligence Authority (SDAIA)

Jan. 2023 – Jan. 2025

Riyadh, SA

- As one of the founding members of the ALLaM team, I became intimately familiar with the entire LLM pipeline. This includes pretraining, finetuning, deployment, evaluation, and dataset management.
- Added support for general dataset structures (e.g. preference data) into Megatron-LM, allowing us to leverage Megatron for finetuning for an **over 2x training throughput improvement** over our deepspeed based repository. Modified C python bindings code to implement packing strategies.
- Leveraged efficient parallel processing strategies on over a hundred CPU cores in a multinode setup to improve the ALLaM team's previously single threaded pretraining data cleaning pipeline's speed **by 55x**.
- Assisted in implementing vocabulary expansion of pretrained LLMs to improve Arabic tokenization and model performance. Improving ability of pre-existing models to acquire Arabic as a second language.

Research Engineering Fellowship

KAUST & SDAIA Partnership

Apr. 2024 – Oct. 2024

Thuwal, SA

- **Selected from 4000 other employees (0.075%)** to attend KAUST from SDAIA. Coordinating efforts between KAUST and SDAIA to implement an AI-based learning management system for the Ministry of Education. Presented at GAIN 2024 as a headlining product from SDAIA, ran as a pilot in schools.
- Developed and deployed a vector database server to store textbook content using Milvus. The server hosted a long context embedding model alongside a translation model to dynamically translate retrieved content to improve final model generation quality.
- Supported the Kaleem team in SDAIA in implementing a real-time video based AI tutoring system. Developed the backend in Python and Flask, deployed an ensemble of models through vLLM to support the system: three LLMs quantized to FP8, **and a custom engagement tracking model based on two LSTMs**.

We're Scraping the Bottom: Curation as a Path Forward For Training LLMs (KAUST Rising Stars in AI Symposium 2026) Sultan Alrashed, Chadi Helwe, Francesco Orabona. Invited poster synthesizing MixMinMatch and SmolKalam into a single narrative on the redundancy produced by repeated large-scale web scraping, and on curation as the more efficient path forward.

Mix, MinHash, and Match: Cross-Source Agreement for Multilingual Pretraining Datasets (Arxiv 2026) Sultan Alrashed, Francesco Orabona. Shows that cross-corpus redundancy is itself a free, model-free quality signal: documents independently recovered by multiple scraping pipelines are more likely to be high quality. **Outperforms the strongest baselines by 4–11%** on Arabic, Turkish, and Hindi while beating model-based filters, with all datasets (AraMix, TurMix, HinMix) released openly.

SmolKalam: Ensemble Quality-Filtered Translation at Scale for High Quality Arabic Post-Training Data (Arxiv 2025) Sultan Alrashed, Chadi Helwe, Francesco Orabona. Introduces the largest multi-turn, tool calling, reasoning included Arabic post-training dataset, at around 2B tokens.

Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures (Arxiv 2025) Tyler A. Chang et al. Wrote the Arabic Saudi dialect portion of a multilingual version of the PIQA benchmark.

Beyond the Ideal: Analyzing the Inexact Muon Update (AISTATS 2026 | CPAL 2026 Spotlight Track) Egor Shulgin, Sultan Alrashed, Francesco Orabona, Peter Richtárik. Analysis of Muon's inexact orthogonalized update with performance bounds under an additive-error LMO framework.

Cards Against Contamination: TCG-Bench for Difficulty-Scalable Multilingual LLM Reasoning (EACL Findings 2026) Sultan Alrashed, Jianghui Wang, Francesco Orabona. A contamination-resistant, English/Arabic text-game suite with scalable difficulty.

ReviseQA: A Benchmark for Belief Revision in Multi-Turn Logical Reasoning (ICML 2025 World Models Workshop) Chadi Helwe, Sultan Alrashed, Francesco Orabona. Benchmark testing logical consistency under iterative context updates.

SmolTulu: Higher Learning Rate to Batch Size Ratios Can Lead to Better Reasoning in SLMs (Arxiv 2024) Sultan Alrashed. Preprint showing LR:batch size ratios can boost reasoning in small LMs, includes model release. Scored the highest IFEval score of any sub-3B model at release.

Fineweb-Edu-Ar: Machine-translated Corpus to Support Arabic Small Language Models (Arxiv 2024) Sultan Alrashed, Dmitrii Khizbullin, David R. Pugh. Introduces a machine-translated Arabic corpus derived from FineWeb-Edu (202B tokens) for training/evaluating Arabic SLMs, largest at the time.

ALLaM: A Series of Large Language Models for Arabic and English (ICLR 2025) M Saiful Bari et al. SDAIA's ALLaM paper, a series of bilingual pretrained LLMs. Initially I had a hand in most parts of the pipeline, later on focused on finetuning / alignment as our team size grew.

When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards (ACL 2024) Norah Alzahrani et al. Exploration into how sensitive LLM benchmarks and evaluations are to minor perturbations to structure. Many results are contributed to LM-Harness.

PROJECTS & VOLUNTEERING

Modded-Nanogpt World Record 2026 (github.com/kellerjordan/modded-nanogpt) Set a world record in the modded-nanogpt speedrunning leaderboard through improved model initialization.

Next-Token Agent 2024 (github.com/SulRash/ntaGPT) A project focused on pretraining and finetuning tiny LLMs from scratch to solve ASCII games by generating each frame.

Environment Encoder 2024 (github.com/SulRash/envenc) A proposal and an implementation of an idea to train a reinforcement learning agent to play games given a vision-language model's embeddings, instead of the actual environment. Did not work.

Microsoft's Megatron-Deepspeed Contributions 2023 (github.com/Microsoft/Megatron-DeepSpeed) Fixed bugs in Microsoft's Megatron-Deepspeed repository that led to errors when running LLM pretraining jobs.

Reinforcement Learning Roguelike Solver 2022 (github.com/SulRash/Graduation-Project) For my university honours project, I wrote PPO, DQN, and A2C reinforcement learning agents to compare the effects of perfect and imperfect information on my CLI roguelike game EasyRogue in a report.