Arabic LLM Data Intelligence Internship
Analyzing the Quality, Coverage, and Gaps of Arabic Training Data for Large Language Models

Project Description

The performance of Large Language Models (LLMs) is highly dependent on the quality, diversity, and representativeness of their training data. While extensive research has analyzed English training corpora, systematic evaluation of Arabic LLM training data remains limited. This project aims to analyze Arabic training datasets to estimate their quality, topic coverage, biases, and contribution to model performance.

Inspired by recent research on data influence and contribution estimation, the project will apply similar methodologies to Arabic corpora to understand which data sources improve model performance and where significant gaps remain.

Project Type

Research and Development (Arabic NLP + Data-Centric AI)

Objectives

- *Collect and document major Arabic datasets used in LLM training.*
- *Estimate dataset quality using influence and contribution analysis.*
- *Analyze topic distribution and identify domain gaps.*
- *Detect noise, redundancy, and bias in training data.*
- *Provide recommendations for improved Arabic data curation.*

Methodology

- *Data collection from public Arabic corpora and repositories.*
- *Preprocessing: normalization, deduplication, and metadata tagging.*
- *Topic modeling and clustering for domain analysis.*
- *Approximate Shapley-style or leave-one-out influence estimation.*
- *Evaluation using perplexity and downstream task performance.*

Deliverables

- *Annotated inventory of Arabic LLM training datasets.*
- *Quality and influence analysis report.*
- *Topic distribution and gap visualization dashboard.*
- *Final technical report with actionable recommendations.*

Required Skills

- *Python programming.*
- *Experience in NLP and machine learning.*
- *Familiarity with large-scale text processing.*
- *Interest in Arabic language technologies.*

Preferred Academic Level

Open to senior undergraduate students, MSc students, and fresh graduates in Computer Science, Artificial Intelligence, Computational Linguistics, or Data Science.

Learning Opportunities

Interns will gain hands-on experience in data-centric AI, Arabic NLP, LLM evaluation techniques, and research-oriented analysis. This project offers a strong portfolio opportunity at the intersection of AI and Arabic language research.

Expected Team Size

2–4 interns

Mentors:

Hamdy Mubarak (hmubarak@hbku.edu.qa), Principal Software Engineer

Yifan Zhang (yzhang@hbku.edu.qa), Senior Software Engineer

References:

WHAT'S IN MY BIG DATA?

https://arxiv.org/pdf/2310.20707