Project Title

Multi-modal-dialectal-cultural Arabic Understanding

Project Description:

This project develops linguistics-first resources and methods to better represent Arabic dialects, sociolinguistic variation, and cultural context across modalities (text, speech, and document imagery). The goal is to close critical gaps in dialectal coverage, code-switching, culturally grounded meaning, and annotation standards, enabling stronger LLM training/evaluation, education, digital humanities, and Arabic-focused language technologies.

Project Type: Research or Engineering

Duties/Activities:

Students will choose a stream (or mix streams) and work in a team:

- **Arabic Linguistics & Annotation Stream (Core):** curate and annotate dialect phenomena (MSA vs dialect, code-switching, borrowed words, named entities, idioms/proverbs, politeness/register, region markers); write high-quality **annotation guidelines** and resolve annotator disagreements.
- **Dialect Resources Stream:** build lexicons, dialect maps, parallel variants, minimal pairs, and metadata (region, register, domain, speaker/writer profile where appropriate) to support training and evaluation.
- **Evaluation & Benchmark Stream:** design linguistics-grounded benchmarks for dialect identification, meaning equivalence across dialects, cultural reference understanding, and robustness to spelling variation/Arabizi; include grounded evaluation where relevant.

Required Skills:

Any one of the following is sufficient:

- **Arabic linguistics / dialectology / sociolinguistics** (annotation, analysis, guideline writing)
- **NLP/ML basics** (data processing, evaluation, simple modeling; LLM familiarity is a plus)
- **Engineering** (Python, data pipelines, annotation tooling, basic frontend/backend)

Preferred Intern Academic Level:

Strong BSc (3rd/4th year), MSc, or PhD (excellent Arabic/dialect competence and research maturity valued; strong coding is a plus).

Learning Opportunities:

- Build **publishable Arabic linguistic resources** (guidelines, datasets, lexicons, benchmarks) with real research impact.
- Learn best practices for **annotation design**, inter-annotator agreement, dataset QA, and bias/coverage analysis.
- Gain exposure to how linguistic insights translate into better **LLM evaluation and training data**.
- Contribute to a team workflow that turns linguistic research into usable artifacts for real systems.

Expected Team Size:

3–6 interns, team-based with clear roles (linguistics core + evaluation + tooling support), and shared deliverables.

Mentors

Name: Ehsaneddin Asgari
Email: easgari@hbku.edu.qa