Assignment 4 (LLMs)

General Information Retrieval (including Q/A)

Benchmarking among LLMs

Problem 1: General Information Retrieval (including Q/A)

Information Retrieval and Question Answering Using Arabic Language. This problem introduces you to classical keyword-based retrieval, semantic search using embeddings, and Retrieval-Augmented Generation (RAG). You will apply these techniques using a single Arabic book of your choice (in plain text format) as the reference source.

Tasks (consult and use any reasonable LLM to help you in doing all or part of these tasks)**:**

1. **Book Preparation for rearch**
   - Select a public domain or permitted Arabic book in plain text.
   - Split the text into short paragraphs (2–4 sentences).
   - Generate embeddings using a multilingual or Arabic sentence embedding model (Consult any reasonable LLM for the best embedding system available and how to use it).
   - Index the vectors using a tool like FAISS.
2. **Retrieval System**
   - Build an interface that accepts Arabic queries.
   - Implement:
     a. Classical search (TF-IDF or BM25)
     b. Semantic search (embedding-based)
   - Display the top 5 results from both methods side by side for each query.
3. **RAG System**
   - Extend the interface to answer Arabic questions.
   - For each query:
     a. Retrieve relevant text using semantic search
     b. Use an LLM to generate an answer with the retrieved content (RAG)
     c. Also generate an answer without retrieval (LLM only)
   - Display both answers for comparison.

Deliverables:

1. Description of the selected book: title, source, text, chunking method, the book chunked and embedding/indexing process and result.
2. Results from at least **10 Arabic queries** using both classical and semantic search.
3. **Answers for 10 queries** using both RAG and LLM-only methods.
4. A comparison and reflection discussion:
   - Which retrieval method gave more relevant results
   - Whether retrieval improved LLM-generated answers
   - Conclusions on the value of semantic retrieval and RAG for Arabic text.

It is to be noted that if you select a book subject not very common, there will be a chance to get improved answers.

Prob-2: Benchmarking among LLMs

Preparing benchmark data to test and compare different LLMs in terms of their strength in **Natural Language Technologies (NLTs)** requires careful planning to ensure **validity, fairness, reproducibility, and relevance**. Few LLMs serve the Arabic Language, such as ChatGPT (OpenAI), Gemini (Google), Allam (KSA), Llama (Meta), Jais (UAE), Fannar (Qatar), etc. Please select at least 5 of these LLMs and compare their performance. Below are comprehensive **guidelines** grouped into key stages:

1. Define the Scope and Objectives

- **Select one of the Arabic Natural Language Technologies below to be evaluated (mention below examples of these technologies:**
  - Question/Answering (Q/A)
  - Arabic Summarization
  - English to Arabic Translation
  - Automatic Speech Recognition (ASR)
  - Typewritten Optical Character Recognition (OCR)
  - Handwritten Optical Character Recognition (OCR)
  - Manuscript Optical Character Recognition (OCR)
  - Text to Speech (TTS)
  - Automatic Diacritization (تشكيل آلي) and Automatic Erab (الإعراب الآلي)

  Any other NLP technology of your choice … define

  No two groups will take the same.

- **Determine the language included in your evaluation**:
  - Whenever it applies, select
    - ~ 33% of the data in Modern Standard Arabic (MSA) and
    - ~ 33% of Classical Arabic and
    - ~ 33% in Arabic dialects (Egyptian dialect or other Arabic dialects or mixture between them).
- Determine if your task will be **domain-specific** (medical, legal, social media, etc.) **or generic data**.
- Determine sources **(genres)**: news, conversations, tweets, reviews, etc.

2. Source and Create the Dataset

- Check the literature (you may consult any LLM) for the reasonable amount and specifications of the benchmarking data. Usually, we include both **clean and noisy data** to test robustness.
- However, find below are suggested (reduced) amounts of data to ease the task for you; however, I will expect that you will report the minimum amount of data for professional work in your report.
- Collect raw text/speech/images from real-world usage (web, forums, etc.).
- The members of each group will be the **annotators** to create **gold standards**.
- Define **clear guidelines** for annotation by reviewing similar work (you may also consult any LLM).
  - Ensure **Annotation Quality**
  - Use **multiple annotators** per item.
  - Calculate inter-annotator agreement (**IIA**) (e.g., Cohen's Kappa).
  - Resolve conflicts in annotation, if any, by majority voting.
  - Ensure:
    - **Diversity** (gender, dialects, sentence lengths, topics).

- **Representativeness** (reflects real-world usage).

- **Final Review:**
  - **Clean data:** Remove duplicates, remove very noisy data, and review data after annotation for any possible errors in data ordering or mishandling.
  - **For speech:** Align audio with transcripts.

3. Define Evaluation Metrics

Choose appropriate metrics for each task (consult LLM), find below some examples:

- **For ASR, OCR:** Word Error Rate (WER), Character Error Rate (CER)
- **MT:** BLEU, METEOR
- **Question/Answering/Text Summarization:** ROUGE, Human evaluation

4. Document Everything on your answer that should include:

- Dataset creation process
- Annotation guidelines
- Preprocessing steps
- Known biases or limitations

  You may give all the information about the process that you handled to an LLM to help generate a comprehensive report in no less than 8 pages. You may also evaluate your selected LLM in report generation. You should add any graph or table that illustrates your work.

A Checklist Table

(recommended to be used in your answer)

#	Category	Checklist Item	Status (Yes/No)	Notes
1	Scope Definition	Have you clearly defined the NLT task(s) to be evaluated?		
2	Scope Definition	Have you specified the evaluation goals (accuracy, robustness, etc.)?		
3	Language and Domain Selection	Have you selected relevant languages and dialects?		
4	Language and Domain Selection	Have you included multiple genres and domains?		
5	Data Collection	Is the data licensed for research/benchmark use?		
6	Data Collection	Have you included real-world or synthetic data?		
7	Annotation	Is there a clear annotation schema for the task?		
8	Annotation Quality	Is inter-annotator agreement calculated and acceptable?		
9	Data Balance	Does the dataset reflect demographic and topical diversity?		
10	Preprocessing	Is the data cleaned and normalized consistently?		
11	Preprocessing	Is the preprocessing procedure documented?		
12	Evaluation Metrics	Are appropriate metrics defined for each NLT task?		
13	Evaluation Metrics	Do you include baseline results or evaluation scripts?		
14	Documentation	Is the data creation process fully documented?		
15	Documentation	Are known limitations or biases discussed?		

16	Fairness	Does the dataset include diverse and fair representation?		

Amounts of benchmarking data with some guidelines for different technologies

Technology	Minimum Benchmark Data Size	Rationale
Question/Answers	100 Arabic Q/A pairs with gold answers, each with 5 questions from different topics	Test response quality and factuality across domains and vendors. Test % Hallucinations, if any
Arabic Summarizer	30 source texts + 2 human summaries each	Evaluate content coverage, coherence, and compression
English to Arabic Translation	50 sentence pairs (EN-AR) with reference human translations	Reliable scoring with BLEU, METEOR, or human preference
Automatic Speech Recognition (ASR)	30 audio clips (~ 2-4 mins each) from different background and different speakers	Evaluate WER, noise robustness, dialect handling
Typewritten OCR	90 scanned pages, each 5 pages of different sources (Books [40 pages], magazines [40 pages], newspapers [10 pages])	Simple layout OCR; establish baseline across tools
Handwritten OCR	60 scanned pages, each 5 pages of a different writer	Assess stroke variation, writing styles, segmentation
Manuscript OCR	30 manuscript images, each 1 page of different manuscript (high variation, historical)	Hard to OCR needs focused testing on script types
Text to Speech (TTS)	20 sentences in text form with correct audio	Assess naturalness, pronunciation, and expressiveness. Do human listening sessions
Automatic Diacritization or Erab	50 Arabic sentences (of average 5-15 words)	Evaluate grammatical accuracy and linguistic precision