MedAraBench: Large-scale Arabic Medical Question Answering Dataset and Benchmark

Mouath Abu Daoud, Leen Kharouf, Omar El Hajj,
Dana El Samad, Mariam Al Omari,
Khaled Saleh, Jihad Mallat,
Nizar Habash, Farah E. Shamout

Background, Motivation, and Related Work

Section 1

NYU

Why Benchmark LLMs for Arabic Medical Applications?

- LLMs have high potential for applications in healthcare, but benchmarking is essential for safe deployment.
- Most benchmarks are English-centric, leaving 380M+ Arabic speakers underserved.
- **MedAraBench** serves as the first large-scale Arabic benchmark for medical LLM evaluation, featuring 24,883 MCQs across 19 medical specialties and 5 difficulty levels, addressing real-world clinical relevance using real world data and through bias injection.

Benchmark	Language(s)	Type	Size	Expert Annotation	Difficulty Mapping	Specialty Coverage	Arabic	Public
MedQA	English, Chinese	MCQs	60,000	✓	×	✓	×	✓
MedMCQA	English	MCQs	193,000	✓	×	✓	×	✓
MMLU (USMLE)	English	MCQs	1,800	×	×	✓	×	✓
MMLU Translation	14 incl. Arabic	MCQs	15,000	✓	×	✓	✓	✓
AraMed	Arabic	QA	270,000	✓	×	✓	✓	×
MedArabiQ	Arabic	QA and MCQs	700	×	×	✓	✓	✓
MedAraBench (Ours)	Arabic	MCQs	24,000	✓	✓	✓	✓	✓

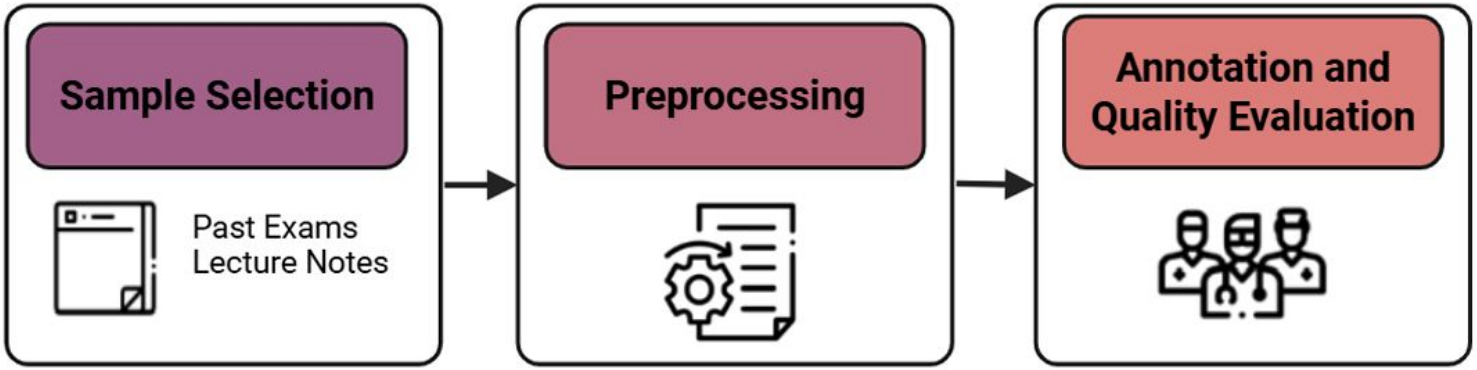
NYU

Data Collection and Cleaning

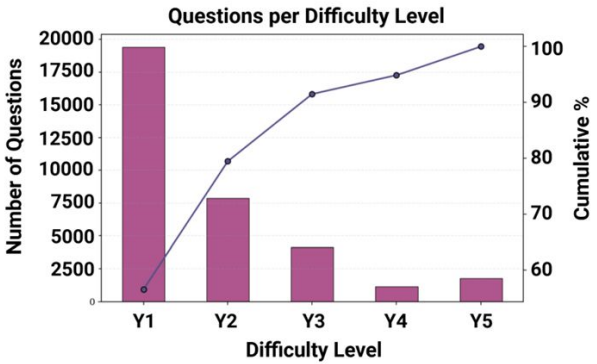
Section 2

NYU

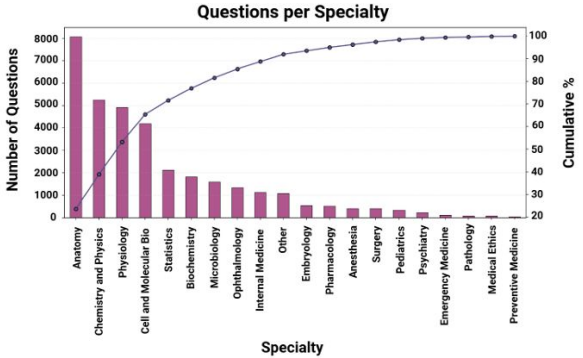
Data Processing Pipeline



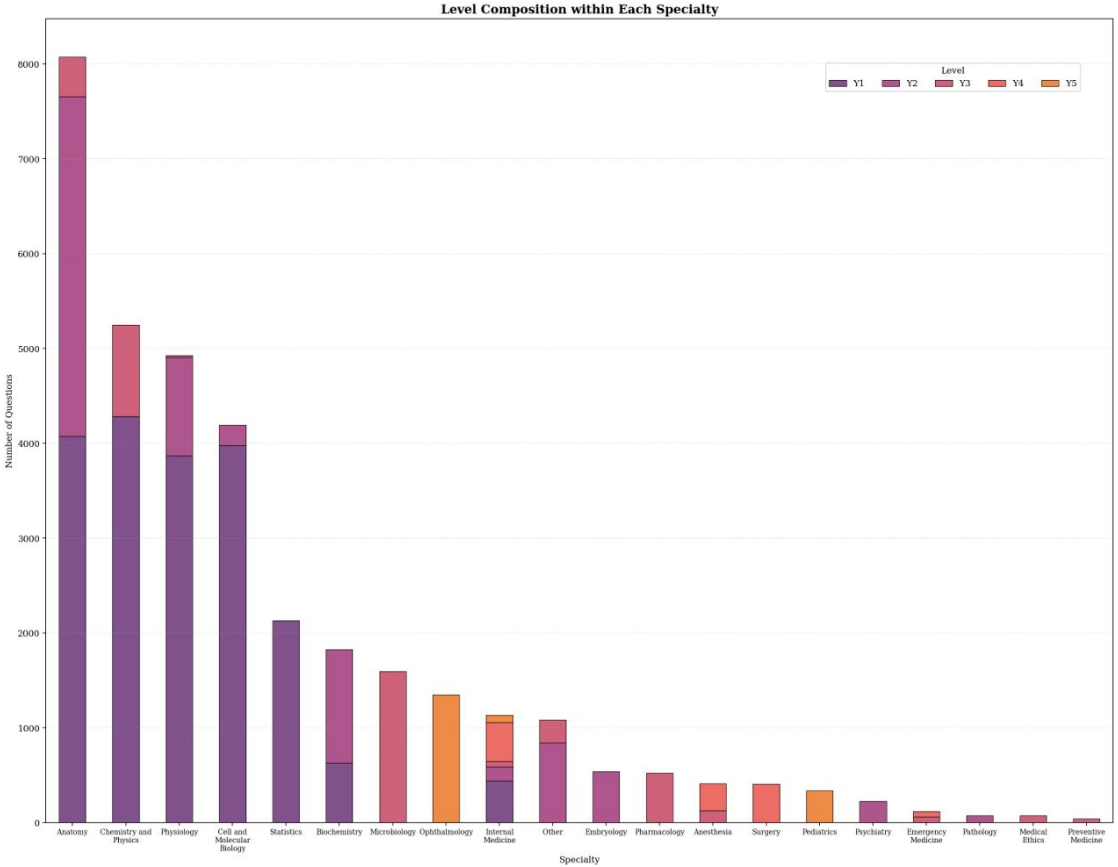
MCQ Coverage



(a)



(b)



MedAraBench Quality Evaluation

Section 3

NYU

Expert Quality Evaluation

Two board-certified clinicians (Anesthesiology and IM) with over 20 years of experience.

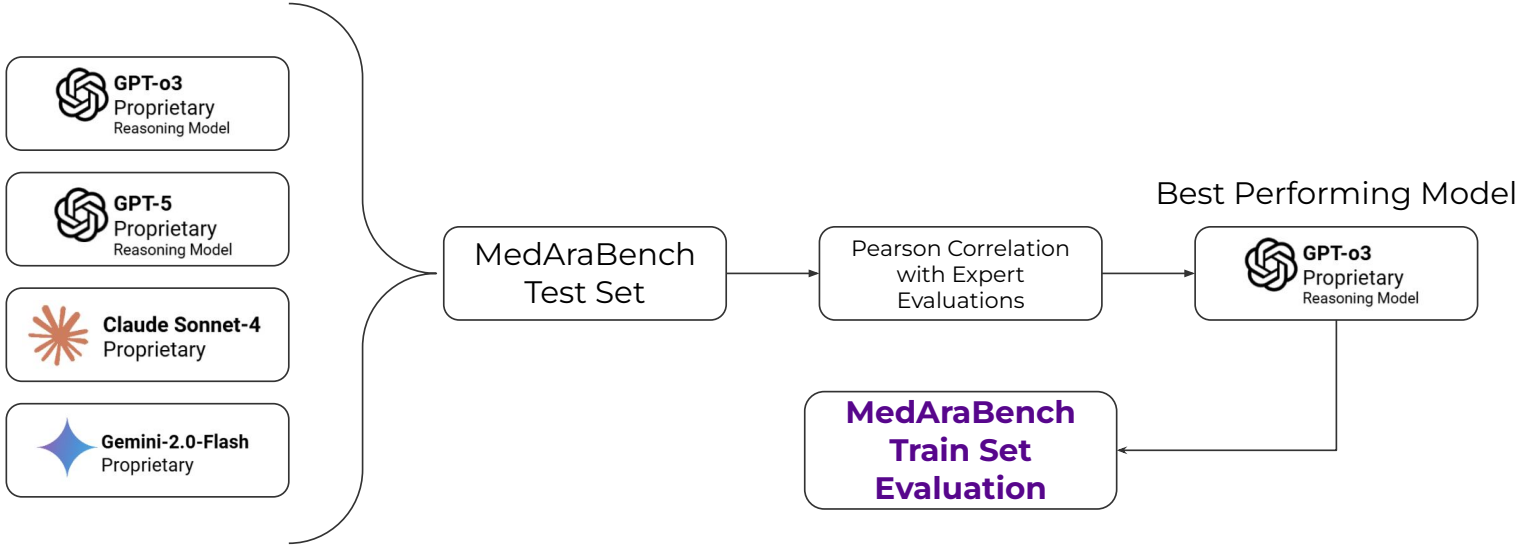
Binary scale assessment of data quality based on:

1. **Medical Accuracy**
2. **Clinical Relevance**
3. **Question Difficulty**
4. **Question Quality**

Sample Size = 378 (based on Cochran's formula)

Metric	Average [standard deviation]	Percent Agreement	Cohen's Kappa
Medical Accuracy	0.722 [0.448]	82.0%	0.555
Clinical Relevance	0.653 [0.476]	65.6%	0.275
Question Difficulty	0.669 [0.471]	65.6%	0.233
Question Quality	0.767 [0.423]	68.3%	0.152

LLM-as-a-Judge Evaluation



Model	Evaluation Metric (average)			
	Medical Accuracy	Clinical Relevance	Question Difficulty	Question Quality
GPT-o3	0.673 [0.469]	0.827 [0.378]	0.588 [0.492]	0.841 [0.366]
Gemini 2.0 Flash	0.717 [0.450]	0.565 [0.496]	0.815 [0.388]	0.774 [0.366]
Claude-4-Sonnet	0.711 [0.453]	0.749 [0.434]	0.576 [0.494]	0.764 [0.425]
GPT-5	0.533 [0.499]	0.610 [0.488]	0.597 [0.490]	0.476 [0.499]

Evaluation results of LLM-as-a-judge applied to the test set

	Expert A				Expert B			
Model	GPT-o3	Claude-4-Sonnet	Gemini-2.0-Flash	GPT-5	GPT-o3	Claude-4-Sonnet	Gemini-2.0-Flash	GPT-5
Medical Accuracy	**0.577**	0.065	0.023	0.043	**0.505**	0.053	0.053	0.068
Clinical Relevance	**0.252**	0.165	0.176	0.071	**0.377**	0.131	0.131	0.104
Question Quality	**0.407**	0.007	0.023	0.044	**0.336**	0.062	0.062	0.033
Question Difficulty	-0.114	-0.070	**0.019**	-0.116	-0.018	**0.039**	**0.039**	-0.049

Pearson correlation coefficients between model and expert ratings.

Dataset	**Medical Accuracy**	**Clinical Relevance**	**Question Quality**	**Question Difficulty**
Training set	0.638 [0.481]	0.821 [0.383]	0.839 [0.367]	0.561 [0.496]
Test set	0.673 [0.469]	0.827 [0.378]	0.588 [0.492]	0.841 [0.366]
P-value	0.0001	0.0362	0.0001	0.0001

GPT-o3 LLM-as-a-judge scores on training and test sets (mean [std]) with t-test p-values.

Benchmarking SOTA LLMS

Section 4

NYU

Models and Prompt

- Proprietary
  - General-purpose
    - GPT-5, GPT-4.1, GPT-o3
    - Claude Sonnet-4
    - Qwen-Plus
    - Gemini-2.0-Flash
- Open-Source
  - General-purpose
    - LLaMa-3.3-70B, LLaMa-3.1-8B
    - DeepSeek V3
  - Medical-focused
    - Apollo Med
    - MedGemma
    - Med42
    - BiMedix-B-
  - Arabic-centric
    - Cohere Command R7B Arabic
    - Fanar-8.7B

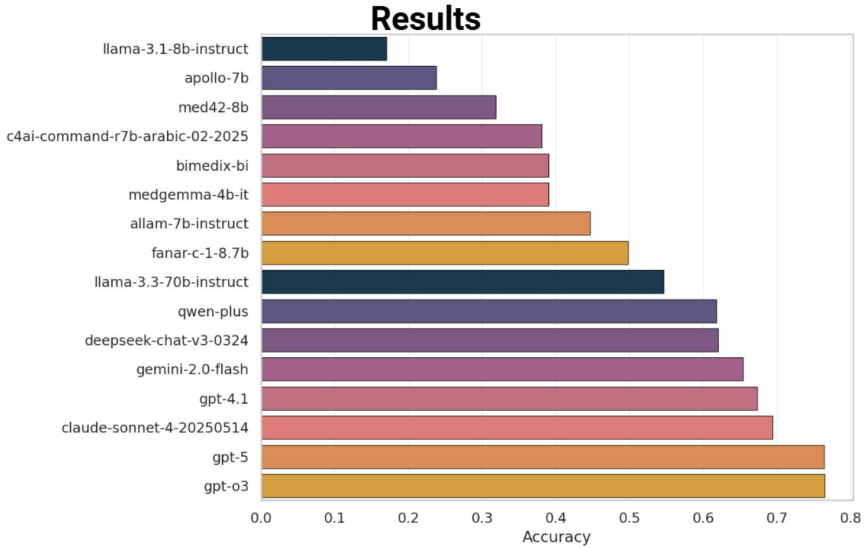
```
"You are an expert medical virtual assistant.
Please provide the correct answer letter (A, B, C, or D)
for the following Arabic medical multiple-choice question.
Question:
{question_text_in_Arabic}

Options:
A:{option_A_in_Arabic}
B:{option_B_in_Arabic}
C:{option_C_in Arabic}
D:{option_D_in Arabic}
Answer:"
```

NYU

Benchmark accuracies, model sizes, and training dataset size for all evaluated LLMs.

Model Type	Model Category	Model	Model Size	Training Dataset Size	Overall Accuracy
Proprietary	General-purpose	claude-sonnet-4-20250514	Unknown	Unknown	0.694
		gemini-2.0-flash	Unknown	Unknown	0.654
		gpt-4.1	Unknown	Unknown	0.673
		gpt-5	Unknown	Unknown	0.764
		gpt-o3	Unknown	Unknown	**0.765**
Open-source	General-purpose	deepseek-chat-v3-0324	8B parameters	1.8 trillion	0.620
		qwen-plus	8B parameters	18 trillion	0.618
		llama-3.3-70b-instruct	70B parameters	15 trillion	0.547
		llama-3.1-8b-instruct	8B parameters	15 trillion	0.170
	Arabic-centric	fanar-c-1-8.7b	8.7B parameters	1 trillion	0.498
		allam-7b-instruct	7B parameters	5.2 trillion	0.447
		c4ai-command-r7b-arabic-02-2025	7B parameters	Unknown	0.381
	Medical	medgemma-4b-it	4B parameters	4 trillion	0.390
		apollo-7b	7B parameters	Unknown	0.238
		med42-8b	8B parameters	15T + 1B	0.318
		bimedix-bi	27B parameters	632 million	0.390



NYU

Few-Shot and QLoRA Fine-tuning

Section 5

NYU

Few-Shot and QLoRA

Few-shot: experiments on LLaMa-3.1-8B-instruct to assess in-context learning capabilities using 3 high-quality sample questions (sample questions were rated highly across all evaluation metrics by expert evaluators)

QLoRA: parameter efficient fine-tuning on LLaMa-3.1-8B-instruct loaded in 4-bit precision and trained on the MedAraBench training set.

Model	Baseline Accuracy	Few-shot Accuracy	QLoRA Accuracy
llama-3.1-8b-instruct	0.170	0.191	0.320

Few-shot and QLoRA fine-tuning performance compared to baseline zero-shot accuracy for Llama-3.1-8B-instruct

Discussion

Section 6

NYU

Key Takeaways

MedAraBench as a new standard

- MedAraBench presents as a more rigorous and representative dataset for Arabic medical reasoning compared to existing baselines, with extensive expert evaluations to ensure data quality.

Arabic Medical Gap

- Proprietary models still outperform medical-focused open-source models, suggesting that current medical LLMs lack sufficient Arabic medical pre-training.

Model Development Utility

- QLoRA results show that the MedAraBench training set successfully improves performance in smaller models (e.g., Llama-3.1-8B), proving its value for both evaluation and fine-tuning.

Thank You!

Read our paper here!

Access our GitHub here!

NYU