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

Anonymous authors
Paper under double-blind review

Abstract

Arabic remains one of the most underrepresented languages in natural language processing research, particularly in medical applications, due to the limited availability of open-source data and benchmarks. The lack of resources hinders efforts to evaluate and advance the multilingual capabilities of Large Language Models (LLMs). In this paper, we introduce MedAraBench, a large-scale dataset consisting of Arabic multiple-choice question-answer pairs across various medical specialties. We constructed the dataset by manually digitizing a large repository of academic materials created by medical professionals in the Arabic-speaking region. We then conducted extensive preprocessing and split the dataset into training and test sets to support future research efforts in the area. To assess the quality of the data, we adopted two frameworks, namely expert human evaluation and LLM-as-a-judge. Our dataset is diverse and of high quality, spanning 19 specialties and five difficulty levels. For benchmarking purposes, we assessed the performance of eight state-of-the-art open-source and proprietary models, such as GPT-5, Gemini 2.0 Flash, and Claude 4-Sonnet. Our findings highlight the need for further domain-specific enhancements. We release the dataset and evaluation scripts to broaden the diversity of medical data benchmarks, expand the scope of evaluation suites for LLMs, and enhance the multilingual capabilities of models for deployment in clinical settings.

1 Introduction

The emergence of Large Language Models (LLMs) has driven transformative progress in Natural Language Processing (NLP) in recent years. They have demonstrated exceptional performance across various renowned benchmarks due to their powerful understanding and reasoning abilities, grounded in the vast amount of knowledge in their training corpora (Brown et al., 2020; Bommasani et al., 2022; Chowdhery et al., 2022). This includes general and domain-specific benchmarks (Wang et al., 2021; Stahlberg, 2020).

However, performance improvements remain variable across underrepresented languages and domains, particularly in high-stakes applications like medicine (Jiang et al., 2025; Yang et al., 2025). For example, Arabic is among the most spoken languages in the world, with over 400 million speakers across the globe. However, it remains underrepresented in the medical domain, mainly due to the unique challenges that it poses with its rich morphology, dialectal variation, and limited expert-annotated resources (Farghaly & Shaalan, 2009). These challenges are compounded by the variability in the language of instruction across medical schools in Arabic-speaking regions. In some countries, such as Syria and Sudan, Arabic is widely used in exams and instructional materials, while in others, English and French remain dominant (Alhamami & Almelhi, 2021). This uneven linguistic landscape underscores the need for robust tools that can reason over Arabic medical text, particularly in zero- or low-resource educational settings.

Several benchmarks have been recently introduced in the medical domain. However, most of them focus almost exclusively on English (Jin et al., 2021; 2019). Recent work began to address this need but remain limited in scope and size (Abu Daoud et al., 2025). Thus,

Table 1: Comparison of MedAraBench with existing medical QA benchmarks, including estimated dataset size.

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	✓	×	✓	✓	×
MedAraBench (Ours)	Arabic	MCQs	24,000	✓	✓	✓	✓	✓

there is a pressing need for large-scale benchmarks to assess and improve LLMs for Arabic-language medical reasoning.

To address those gaps, in this paper, we present MedAraBench, a comprehensive benchmark for evaluating and advancing LLMs on Arabic medical tasks. The dataset consists of curated Multiple-Choice Questions (MCQs) spanning different specialties and difficulty tiers aligned with stages of medical education. We propose a standardized development and evaluation protocol to enable reproducible and clinically meaningful assessment of LLMs. Our key contributions are as follows:

- We introduce MedAraBench, a large-scale Arabic medical benchmark featuring 24,883 MCQs across 19 medical specialties and five difficulty levels. The benchmark includes standardized training and test sets to enable systematic evaluation and advancement of LLMs.
- We perform extensive quality assessment via human expert evaluation, focusing on question clarity, clinical relevance, and medical correctness, as well as automated LLM-as-a-judge analysis.
- We benchmark eight state-of-the-art proprietary and open-source LLMs on the MedAraBench test set in the zero-shot setting to establish baseline performance for future research.

2 Related Work

In recent years, several benchmark datasets have been developed to assess the capabilities of LLMs in the medical domain, driven by the expanding demand for applications that can streamline clinical workflows. Despite this progress, Arabic remains underrepresented in clinical NLP, mainly due to the lack of high-quality data to support building clinical applications in Arabic (Abdelaziz et al., 2025). As such, most existing benchmarks focus on English. For instance, the Massive Multitask Language Understanding (MMLU) benchmark includes question-answer pairs from the US Medical Licensing Exam (USMLE) (Hendrycks et al., 2021). Jin et al. (2020) introduce MedQA, a multilingual benchmark dataset consisting of multiple-choice questions sourced from medical licensing exams in English and Chinese. MedMCQA (Pal et al., 2022) extends these benchmarks to a multilingual evaluation framework but remains limited in Arabic.

Recent work have introduced new resources for medical evaluation in Arabic. Translations of existing datasets, such as of MMLU into 14 languages, including Arabic, by professional human translators (Achiam et al., 2023), provide valuable data but lack necessary nuances for proper integration into clinical practice. AraMed presents an Arabic medical corpus and an annotated Arabic QA dataset sourced from online medical platforms (Alasmari et al., 2024). MedArabiQ presents one of the first Arabic medical MCQ datasets (Abu Daoud et al., 2025), yet lacks specialty coverage, difficulty mapping, and expert evaluation.

Several evaluation frameworks have been proposed to evaluate the performance of clinical AI models. Kanithi et al. (2024) introduce 'MEDIC', a framework for evaluating LLMs from medical reasoning andethics, to in-context learning and clinical safety. Wang et al. (2024) propose testing models on real-world input noise, dialogue interruptions, and reasoning justifications. Despite the growing interest in multilingual evaluation, there remains a critical

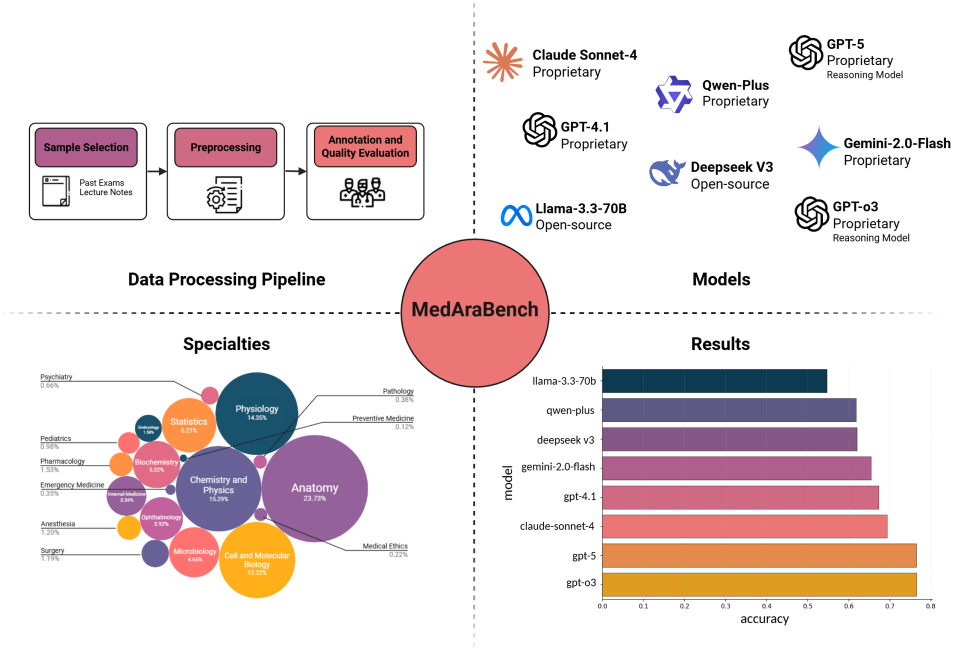


Figure 1: Overview of MedAraBench.

gap in comprehensive, high-quality, and clinically relevant benchmarks for under-served languages. We aim to address this gap by introducing a comprehensive Arabic benchmark. Table 1 provides a structured comparison between MedAraBench and other existing medical benchmarks, covering key dimensions such as language coverage, specialty diversity, expert annotation, and public availability.

3 Methodology

Here, we describe the steps pertaining to data collection, processing, evaluation and benchmarking, to facilitate proper reuse and fair comparison, aligning with best practices for benchmark construction and evaluation. An overview is provided in Figure 1.

3.1 Data Collection and Preprocessing

We compiled a large repository of scanned paper-based exams and lecture notes, hosted on student-led social platforms of regional medical schools. The dataset did not include any personal or real patient data, so anonymization was not necessary, and our data collection complied with privacy and ethical standards. Considering the nature of the documents, we recruited professional typists to digitize the data. We then aggregated the documents to build a single MCQ dataset.

Upon manual inspection by NLP researchers, we observed that several documents exhibited issues such as missing or malformed correct answers, incomplete or duplicated answer choices, non-standard formatting or misaligned fields, and ambiguous answer keys or extraneous non-MCQ content. To ensure dataset quality and model compatibility, we applied strict filtering criteria to remove any questions with such issues. The filtering process was performed manually by five NLP researchers. While data acquisition and preprocessing required extensive effort, it highlights that the dataset is not publicly accessible in structured formats, thus reducing the likelihood of data contamination.

Each question is associated with several annotations: (i) number of answer choices (i.e., ABCD (4 choices), ABCDE (5 choices), and ABCDEF (6 choices)), (ii) difficulty level corresponding to five years of study (Y1 - Y5), and (iii) medical specialty. The questions

fall under 19 medical specialties: Anatomy, Anesthesia, Biochemistry, Cell and Molecular Biology, Chemistry and Physics, Embryology, Emergency Medicine, Internal Medicine, Medical Ethics, Microbiology, Ophthalmology, Pathology, Pediatrics, Pharmacology, Physiology, Preventive Medicine, Psychiatry, Statistics, and Surgery.

We performed a stratified random split of the dataset into training (80%) and test sets (20%). This was to ensure that the medical specialties are represented evenly across the training and test sets (i.e., if the dataset contains 100 Cardiology questions, 80 would be randomly included in the training set, while the remaining 20 would be in the test set). We summarized the dataset in terms of token length distribution, medical specialty distribution, and difficulty level distribution.

3.2 Quality Assessment

To further assess the quality of our dataset, we conducted two analysis: human expert evaluation and using LLM-as-a-judge.

3.2.1 Human Expert Evaluation

We designed our expert evaluation protocol to assess the data according to the following criteria:

1. **Medical Accuracy:** the extent to which the question, options, and correct answer reflect current, evidence-based medical knowledge (Scale: high or low) (Olatunji et al., 2024; Iskander et al., 2024; Rejeleene et al., 2024).
2. **Clinical Relevance:** the practical importance and applicability of the question content to real-world medical practice or education (Scale: high or low) (Iskander et al., 2024; Olatunji et al., 2024).
3. **Question Difficulty:** the complexity required to answer the question correctly (Scale: high or low) (Iskander et al., 2024).
4. **Question Quality:** assessment of the MCQ construction quality (Scale: high or low) following established medical education standards (Al-Rukban, 2006):
    - *Clarity:* question is clear, complete, and unambiguous.
    - *Option Homogeneity:* all distractors are plausible and of similar type.
    - *Single Best Answer* one clearly correct option exists.
    - *No Cueing:* options do not provide clues to other answers.

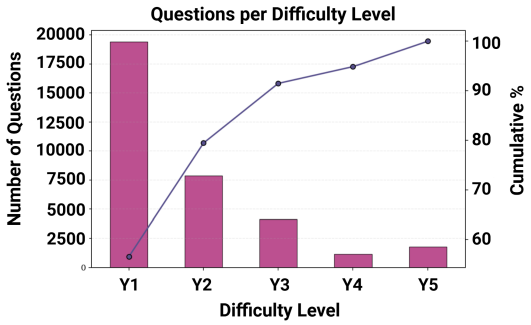
We selected samples for review from the test set and determined the sample size based on Cochran's formula (Cochran, 1977), to create a representative sample size (Hosseini, 2024). We estimated a single proportion at 95% confidence with a ± 5 percentage-point margin of error, using $p = 0.5$ as a conservative assumption when the true quality rate is unknown because it maximizes variance and therefore yields a safe upper bound on sample size. We first calculated an estimated sample size assuming an infinite population using

$$n_0 = \frac{z^2 p(1-p)}{e^2}$$

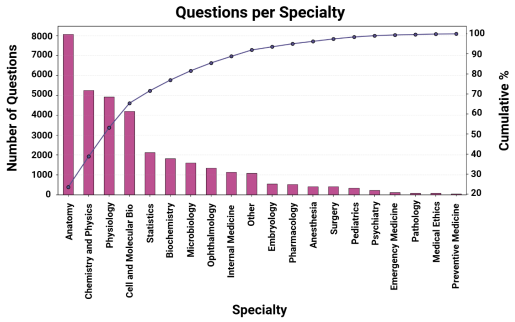
with $z = 1.96$, $p = 0.5$, and $e = 0.05$. However, since the dataset is finite, we then applied the finite population correction using

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}.$$

We recruited two board-certified clinicians specializing in Anesthesiology and Internal Medicine with Arabic clinical fluency and over 20 years of experience each. The reviews were double-blinded to model outputs and data provenance, and were conducted independently with pre-registered instructions on *Qualtrics*.



(a)



(b)

Figure 2: Overview of dataset according to difficulty level and specialties.

3.2.2 LLM-as-a-Judge

Considering that only a subset of the test set was considered for the human expert evaluation, we further introduced an LLM-as-a-judge evaluation protocol as an additional rating of data quality. We prompted three SOTA LLMs (gpt-4-0613, gemini-1.5-pro-latest, and claude-3.5-sonnet-latest) to act as medical education experts and to evaluate the MCQs along the same metrics used in our expert quality evaluation: Medical Accuracy, Clinical Relevance, Question Difficulty, and Question Quality, however on a Likert 1 to 5 scale for the entire test set. This approach gives us the advantage of providing a more nuanced evaluation across a broader set of our data, while also providing insights on the efficacy of LLMs in evaluating the quality of Arabic medical data relative to expert annotators.

3.3 Benchmarking Protocol

To introduce new benchmark results, we evaluated 8 proprietary and open-source models on the test set of the MedAraBench benchmark. We set the models' temperature as 0 to ensure stable outputs as shown in previous work for classification of MCQs (Abu Daoud et al., 2025). We selected one letter response per question.

- **Open-source models:** Llama-3.3-70b-instruct (Grattafiori et al., 2024) and Deepseek-chat-v3-0324 (DeepSeek-AI, 2024).
- **Proprietary models:** Claude-sonnet-4-20250514 (Anthropic, 2025), Gemini-2.0-flash (Google Cloud, 2025), GPT-4.1 (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), GPT-o3 (OpenAI, 2025c), and Qwen-plus (Yang & et al., 2024).

4 Results

In this section, we present a summary of our dataset and the results of our expert quality evaluation, LLM-as-a-judge experiments, and benchmarking experiments.

Table 2: Expert quality assessment results for the representative data subset.

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

Table 3: Evaluation results of LLM-as-a-judge applied to the test set.

Model	Evaluation Metric (average)			
	Medical Accuracy	Clinical Relevance	Question Difficulty	Question Quality
Gemini 1.5 Pro	3.83	2.95	2.15	3.49
GPT-4	3.80	3.16	2.76	3.28
Claude	4.19	3.47	2.53	3.44
Average (Likert 1 to 5)	3.94	3.19	2.48	3.40
Average (Fraction of 5)	0.79	0.64	0.50	0.68

4.1 Dataset Summary

The initial dataset consisted of 34,333 MCQs. The manual filtering process resulted in a reduction of approximately 29% of the initial dataset, yielding 24,883 samples overall. The training set consisted of 19,894 samples, and the test set of 4,989 samples. An overview of the dataset is shown in Figure 2. Additional statistical summaries can be found in Appendix A.

4.2 Expert Quality Assessment

At 95% confidence and ±5% margin, Cochran's formula initially yielded $n_0 = 384$ questions. The final sample size was 378 after adjusting for a finite sample. Hence, our two annotators completed a review of 378 questions as a representative sample of the entire dataset. The results of our data quality assessment and the inter-annotator agreement are summarized in Table 2. Our results show slight to fair levels of agreement across all metrics, with Medical Accuracy having the highest level of agreement with a Cohen's Kappa score of 0.555 and a percentage agreement of 82%.

Additionally, we provide a detailed per-specialty breakdown of the evaluation results in Appendix B. To better assess the results, we investigate individual average and agreement scores for each specialty. Specifically, Figures B1, B2, B3, and B4 show the average accuracy per specialty for each metric, while Tables B1, B2, B3, and B4 show the average accuracy and agreement results per specialty for each metric. All in all, our expert quality evaluations indicate that the data is of high quality with fair levels of agreement across a random sample of our test set.

Table 4: Benchmark accuracies by answer-choice set and overall.

Type	Model	ABCD	ABCDE	ABCDEF	Overall
Proprietary	claude-sonnet-4-20250514	0.702	0.658	1.000	0.694
	gemini-2.0-flash	0.661	0.623	1.000	0.654
	gpt-4.1	0.694	0.588	0.500	0.673
	gpt-5	0.762	**0.774**	1.000	0.764
	gpt-o3	**0.768**	0.754	1.000	**0.765**
	qwen-plus	0.633	0.554	1.000	0.618
Open-source	llama-3.3-70b-instruct	0.562	0.484	0.500	0.547
	deepseek-chat-v3-0324	0.635	0.555	1.000	0.620

4.3 LLM-as-a-Judge Assessment

The results of our LLM-as-a-judge experiments are summarized in Table 3 across all four evaluation metrics. Our results show moderate agreement among different LLMs and comparable results with the expert evaluation scores. To better understand the results in comparison with expert evaluations, we provide a detailed breakdown of the LLM-as-a-judge results in Appendix C. Generally, we see alignment between LLM-as-a-judge evaluation and expert evaluation of the quality of our dataset, with both

4.4 Benchmarking SOTA LLMs

We report all benchmarking results in Table 4. Our results show that reasoning models consistently outperform all others across the benchmark dataset. Specifically, GPT-o3 achieves the highest performance in the ABCD subset with an accuracy score of 0.768, while llama-3.3-70b-instruct performs the poorest with a score of 0.562. Within the ABCDE subset, GPT-5 achieves the highest performance with an accuracy score of 0.774, while llama-3.3-70b-instruct performs the poorest with a score of 0.484. Overall, GPT-o3 and GPT-5 have the highest accuracy scores of 0.765 and 0.764, respectively, while llama-3.3-70b-instruct and qwen-plus have the lowest accuracy scores of 0.547 and 0.618, respectively. Appendix D includes a more detailed breakdown of the accuracy scores per specialty and level for each of the 8 evaluated models.

5 Discussion

Overall, in this study, we present MedAraBench, a new 24k dataset consisting of both training and test sets. We report performance baseline results for SOTA LLMs and highlight critical differences in model capabilities under zero-shot settings. By exposing gaps in Arabic medical understanding, MedAraBench offers useful insights for the development of more inclusive, multilingual, and domain-specialized language models.

The expert quality assessment and LLM-as-a-judge experiments provide valuable insights into the quality of our data and the plausibility of using LLMs to evaluate the quality of medical datasets. Namely, the expert quality evaluation yields average scores ranging from 0.653 - 0.767 across all 4 evaluation metrics, with percent agreements and Cohen's Kappa scores ranging from 0.656 - 0.820 and 0.152 - 0.555, respectively, indicating slight to fair levels of agreement across all metrics. The average evaluation metric scores indicate moderate to high-quality data and fair agreement across annotators, but they indicate the need for the curation of more benchmark datasets of higher quality and clinical relevance to properly assess the readiness of LLMs for clinical deployment. While they are not directly comparable due to varying evaluation scales, we note that the results of LLM-as-a-judge and expert quality evaluation are comparable. These results demonstrate the potential of LLMs to be used for data quality evaluation in the medical domain, pending further alignment with medical standards.

Our benchmark evaluation results coincide with prior research on LLM benchmarking in the medical domain, whereas proprietary models typically outperform open-access models in structured tasks such as multiple-choice QA. This was previously demonstrated by Chen et al. (2025) and Alonso et al. (2024), who demonstrated superior accuracy performance by proprietary models relative to open-source models in medical QA tasks across multiple languages. This was further shown by Abu Daoud et al. (2025), who demonstrated superior performance by proprietary models such as Gemini 1.5 Pro, Claude 3.5 Sonnet, and GPT-4 in Arabic medical MCQ. Our results reinforce those findings, with all proprietary models performing at significantly higher or similar accuracy scores to open-source models. Namely, we observe that the lowest overall zero-shot accuracy score was achieved by llama-3.3-70b-instruct at 0.547, followed by deepseek-chat-v3-0324 and qwen-plus with accuracy scores of 0.620 and 0.618, respectively. On the other hand, we observe that gpt-o3, gpt-5, and claude-sonnet-4-20250514 achieve the highest accuracy scores of 0.765, 0.764, and 0.694, respectively. This is possibly due to the larger training corpora, stronger pretraining on structured datasets, more extensive instruction tuning, and specialized reinforcement learn-

ing pipelines that proprietary models undergo relative to their open-source counterparts. Additionally, our results show significantly higher performance by reasoning models (gpt-5 and gpt-o3) relative to other models, showing the promise and importance of incorporating reasoning and explainability into medical NLP as a whole and Arabic medical NLP specifically.

Building on existing frameworks (Abu Daoud et al., 2025), this study adopts an equivalent evaluation design by utilizing structured Arabic medical MCQs, thus covering a similar task definition and allowing for comparison of different generations of models. This alignment allows us to compare on a high-level legacy LLMs (older models) against contemporary models (newer models). Our results show significant improvement in accuracy scores for all contemporary models relative to legacy models. Namely, claude-3.5-sonnet-20240620 achieved an accuracy score of 0.535 when tested with the MedArabiQ dataset, while claude-sonnet-4-20250514 achieved an overall accuracy score of 0.694 on MedAraBench. Additionally, gpt-4.1, gpt-5, and gpt-o3 achieved accuracy scores of 0.673, 0.764, and 0.765 on MedAraBench, which show significant improvement relative to the 0.535 achieved by gpt-4 in MedArabiQ. While the test sets are different, this indicates a generational improvement that can be attributed to model advances and increased domain-specific training data.

However, despite the significant improvement across generations, the highest performing model achieved an accuracy score of 0.765, which does not match expert-level performance and indicates clear headroom for improvement before being ready for deployment in clinical settings. Furthermore, this allows us to raise important questions about what exactly LLMs are learning. High accuracy does not necessarily indicate deep understanding or clinical reasoning. Instead, models may be leveraging statistical associations and lexical patterns to eliminate implausible answers. For example, frequent exposure to certain disease-treatment pairs during pretraining may allow models to make educated guesses without reasoning through symptom progression or differential diagnosis. This distinction is crucial, particularly in high-stakes applications such as medical education or decision support. Future work should consider evaluating not just answer correctness, but also the rationale behind model choices, possibly through explanation-based tasks or clinician scoring of model justifications.

6 Limitations and Future Work

While our study provides a comprehensive evaluation of LLMs on Arabic medical MCQs and represents a substantial advancement in benchmarking capabilities, several limitations exist. First, the dataset is limited in its capability to evaluate LLMs on classification tasks only due to the nature of the MCQ dataset, preventing the evaluation of LLMs in generative tasks. Additionally, although the data source was not available in a structured digital format and required extensive digitization and cleaning efforts, we cannot certainly rule out contamination. Furthermore, our data assumes fluency in Modern Standard Arabic, which, despite being common in formal settings, may not fully align with linguistic realities. This can affect the generalizability of MedAraBench to learners or practitioners accustomed to dialectal or mixed-language instruction.

Another limitation emerges from our expert quality evaluation experiments. Although our dataset was reviewed by two expert clinicians, we observed occasional inconsistencies across their assessments. We acknowledge that expert disagreement and inherent subjectivity are common in clinical judgment, but recognize the need for broader consensus in future validation efforts. Moreover, our data is text-only in its format, limiting its applicability to accommodating image-based reasoning required in specialties such as radiology and dermatology, and warranting expansions to other data modalities in future work.

In this study, we primarily focused on zero-shot evaluation of model performance, providing an assessment without further adaptation. While this is an important evaluation framework, future work could explore the impact of few-shot and chain-of-thought prompting strategies, as well as fine-tuning as an opportunity to improve model accuracy and performance. Furthermore, future work could warrant the incorporation of dialectal data to enhance model adaptability across diverse Arabic clinical settings.

7 Conclusion

To conclude, we introduced a large-scale Arabic medical benchmark designed to evaluate the zero-shot performance of LLMs on curated MCQs. Covering 19 medical specialties and spanning five difficulty levels, MedAraBench provides a comprehensive and fine-grained lens for assessing Arabic medical reasoning in LLMs. Our benchmark provides an advancement for developing benchmarks in the Arabic language and exposes limitations in the performance of current LLMs in low-resource language tasks and the need for robust multilingual training strategies. Future work should explore fine-tuning strategies and the curation of larger and higher-quality datasets tailored to Arabic medical contexts. We release MedAraBench in hopes of supporting downstream clinical applications, and we hope that it serves as a catalyst for continued research at the intersection of Arabic NLP and medical AI.

Ethics Statement

The authors disclose the use of generative AI tools to assist with LaTeX code cleanup and formatting only, with all content, analyses, and conclusions authored and verified by the researchers involved in this project.

Reproducibility Statement

The MedAraBench benchmark is available at https://anonymous.4open.science/r/medarabench-3BE4/

References

Mariam Essam Abdelaziz, Mohanad A. Deif, Shabbab Ali Algamdi, and Rania Elgohary. A benchmark arabic dataset for arabic question classification using aafaq framework. *Scientific Data*, 2025.

Mouath Abu Daoud, Chaimae Abou Zahir, Leen Kharouf, Walid AlEisawi, Nizar Habash, and Farah Shamout. Medarabiq: Benchmarking large language models on arabic medical tasks. *PMLR*, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mohammed Al-Rukban. Guidelines for the construction of multiple choice questions tests. *Journal of Family Community Medicine*, 2006. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC3410060/`.

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. Aramed: Arabic medical question answering using pretrained transformer language models. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pp. 50–56, 2024.

Munassi Alhamami and Abdullah Almelhi. English or arabic in healthcare education: Perspectives of healthcare alumni, students, and instructors. *Journal of Multidisciplinary Healthcare*, 2021.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938, September 2024. ISSN 0933-3657. doi: 10.1016/j.artmed.2024.102938. URL `http://dx.doi.org/10.1016/j.artmed.2024.102938`.

Anthropic. Claude opus 4 & claude sonnet 4 — system card. https://www.anthropic.com/claude-4-system-card, 2025.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill,

Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions, 2025. URL https://arxiv.org/abs/2402.18060.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

William Gemmell Cochran. Sampling techniques. *john wiley sons*, 1977.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Ali Farghaly and Khaled Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 2009.

Google Cloud. Gemini 2.0 flash | generative ai on vertex ai. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash, May 2025.

A. Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Pedram Hosseini. A benchmark for long-form medical question answering. *arXiv preprint*, 2024.

Shadi Iskander, Nachshon Cohen, and Zohar Karnin. Quality matters: Evaluating synthetic data for tool-using llms. *arXiv preprint arXiv:2409.16341*, 2024. URL https://arxiv.org/abs/2409.16341.

Luyi Jiang, Jiayuan Chen, Lu Lu, Xinwei Peng, Lihao Liu, Junjun He, and Jie Xu. Benchmarking chinese medical llms: A medbench-based analysis of performance gaps and hierarchical optimization strategies, 2025. URL https://arxiv.org/abs/2503.07306.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL https://arxiv.org/abs/2009.13081.

Di Jin, Yansong Pan, Tao Ouyang, and Pascale Fung. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2021. URL https://arxiv.org/abs/2009.13081. Submitted to AAAI 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Praveen K. Kanithi, Clément Christophe, Marco A. F. Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*, 2024.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, and Chinemelu Aka. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*, 2024. URL https://arxiv.org/html/2411.15640.

OpenAI. Gpt-4 technical report. https://openai.com/index/gpt-4-1/, April 2025a.

OpenAI. GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025b.

OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/o3-and-o4-mini-system-card.pdf, 2025c.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Rick Rejeleene, Xiaowei Xu, and Johm Talburt. Towards trustable language models: Investigating information quality of large language models. *arXiv preprint arXiv:2401.13086*, 2024. URL https://arxiv.org/abs/2401.13086.

Felix Stahlberg. Neural machine translation: A review and survey, 2020. URL https://arxiv.org/abs/1912.02047.

Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy Chen. Resilience of large language models for noisy instructions. *EMNLP*, 2024.

Cunxiang Wang, Pai Liu, and Yue Zhang. Can generative pre-trained language models serve as knowledge bases for closed-book qa?, 2021. URL https://arxiv.org/abs/2106.01561.

An Yang and et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. Large language model synergy for ensemble learning in medical question answering: Design and evaluation study. *J Med Internet Res*, 2025.

A Data Analysis

A.1 Token Length Distribution

There are 24,883 questions in our dataset. The questions are moderate in length, with a total average length of 8.0214 characters. The answers vary in format, whereas 24,523 questions have four answer choices (A, B, C, and D), 9801 questions have five answer choices (A, B, C, D, and E), and 9 questions have six answer choices (A, B, C, D, E, and F). The average answer length across the datasets is 16.371 characters. Figure A1 below gives a detailed breakdown of the distributions of text lengths of the entire dataset.

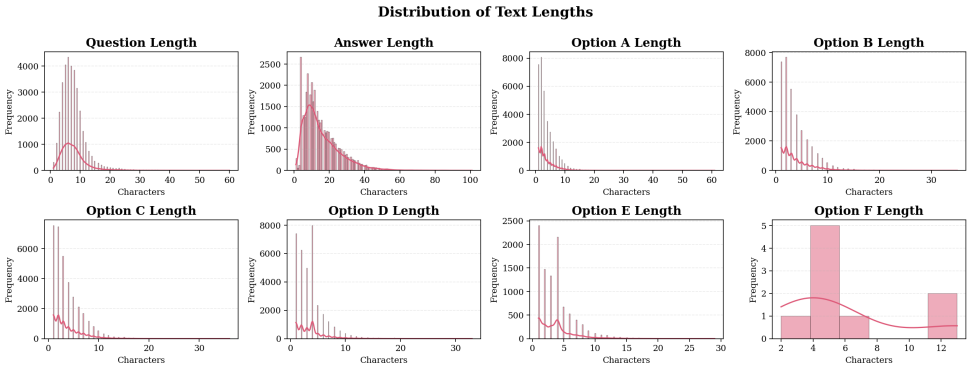


Figure A1: **Distribution of Text Lengths.** (a) Distribution of question length; (b) Distribution of answer length; (c) Distribution of Option A length; (d) Distribution of Option B length; (e) Distribution of Option C length; (f) Distribution of Option D length; (g) Distribution of Option E length; and (h) Distribution of Option F length in MedArabiQ v2 dataset.

A.2 Distribution per Medical Specialty

The largest subsets of the dataset fall under Anatomy (8146 questions - 23.73% of the dataset) and Physiology (4927 - 14.35%), while the smallest subsets fall under Preventive Medicine (41 - 0.12%) and Medical Ethics (76 - 0.22&). A more detailed breakdown of the distribution of questions according to specialty can be shown in Table A1 and Figure 2 (b) above.

A.3 Distribution per Difficulty Level

The largest subset of the dataset fall under Y1 (19414 questions - 56.55% of the dataset) while the smallest subset falls under Y4 (1161 - 3.38%). A more detailed breakdown of the distribution of questions according to level can be shown in Table A2 and Figure 2 (a) above. Additionally, Figure A2 shows the the composition of the five levels across the 19 specialties included in MedAraBench.

Table A1: **Distribution of Questions per Medical Specialty.**

Medical Specialty	Number of Questions	Percentage
Anatomy	8146	23.73%
Anesthesia	413	1.20%
Biochemistry	1826	5.32%
Cell and Molecular Biology	4194	12.22%
Chemistry and Physics	5250	15.29%
Embryology	542	1.58%
Emergency Medicine	121	0.35%
Internal Medicine	1148	3.34%
Medical Ethics	76	0.22%
Microbiology	1597	4.65%
Ophthalmology	1347	3.92%
Pathology	132	0.38%
Pediatrics	337	0.98%
Pharmacology	524	1.53%
Physiology	4927	14.35%
Preventive Medicine	41	0.12%
Psychiatry	225	0.66%
Statistics	2132	6.21%
Surgery	409	1.19%

Table A2: **Distribution of Questions per Difficulty Level.**

Difficulty Level	Number of Questions	Percentage
Y1	19414	56.55%
Y2	7860	22.89%
Y3	4134	12.04%
Y4	1161	3.38%
Y5	1764	5.14%

B Expert Quality Assessment Details

In this appendix, we provide a detailed breakdown of the expert evaluation results introduced in Section 2. While the main text summarizes overall averages and agreement levels across all specialties, here we report per-specialty results to give a more fine-grained view of model performance and annotator consistency.

Figures B1–B4 present the distribution of annotator scores across specialties for each of the four evaluation metrics (Medical Accuracy, Clinical Relevance, Question Difficulty, and Question Quality). The corresponding Tables B1– B4 report the average scores, number of evaluated questions, percentage agreement, and Cohen's Kappa values for each specialty. Together, these results highlight the variability across domains and provide context for interpreting the aggregate quality metrics shown in the main text.

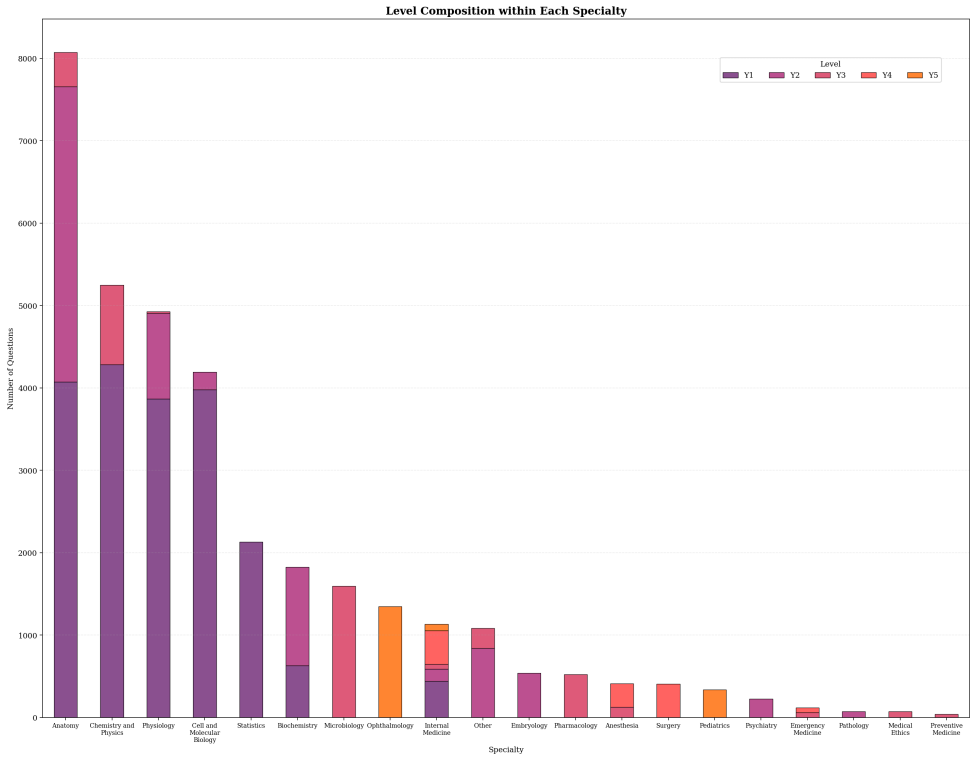


Figure A2: **Level Composition within Each Specialty.** Composition of the 5 different levels (Y1 - Y5) in MedArabiQ v2 dataset.

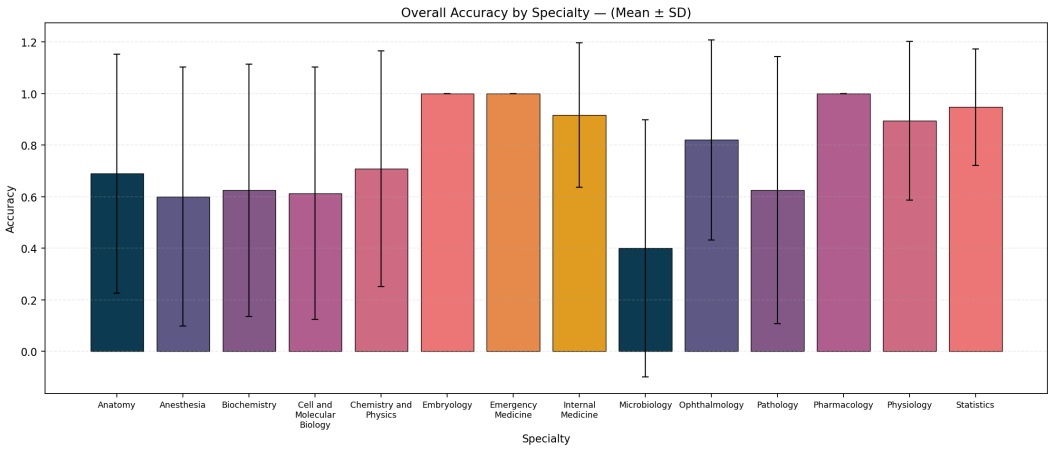


Figure B1: **Overall Annotator Accuracy Scores per Specialty.**

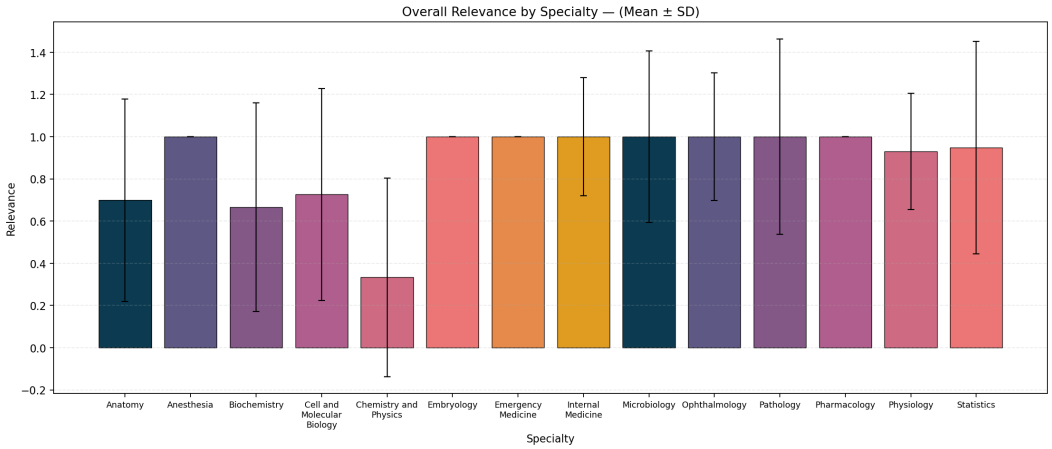


Figure B2: **Overall Annotator Relevance Scores per Specialty.**

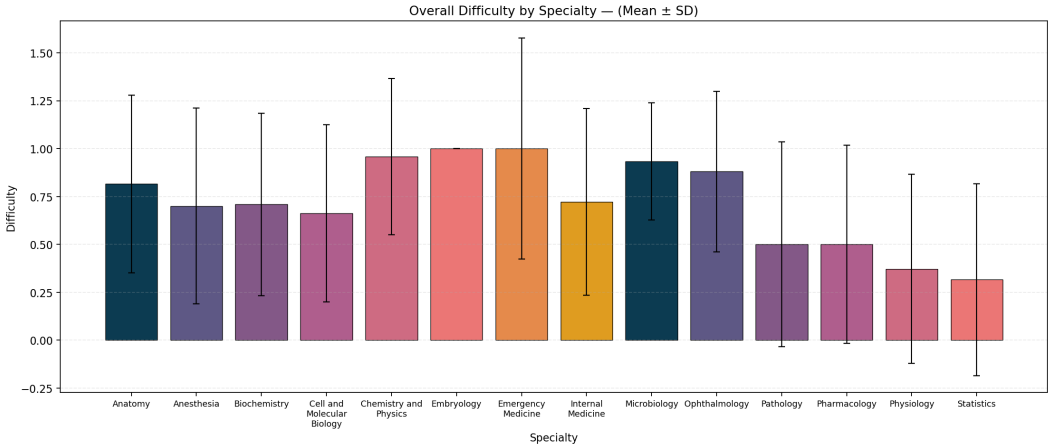


Figure B3: **Overall Annotator Difficulty Scores per Specialty.**

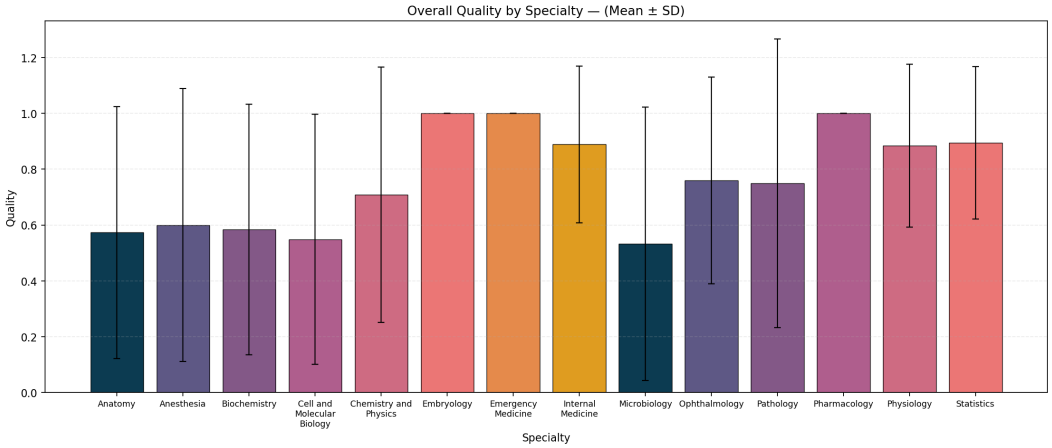


Figure B4: **Overall Annotator Quality Scores per Specialty.**

Table B1: Annotator Accuracy Scores per Specialty.

Specialty	Average [standard deviation]	Number of Questions	Percent Agreement	Cohen's Kappa
Anatomy	0.689 [0.464]	103	0.748	0.431
Anesthesia	0.600 [0.503]	10	0.600	0.167
Biochemistry	0.625 [0.489]	24	0.750	0.471
Cell and Molecular Biology	0.613 [0.489]	62	0.774	0.529
Chemistry and Physics	0.708 [0.457]	48	0.917	0.798
Embryology	1.000 [0.000]	1	1.000	-
Emergency Medicine	1.000 [0.000]	2	1.000	-
Internal Medicine	0.917 [0.280]	18	0.944	0.640
Microbiology	0.400 [0.498]	15	0.733	0.455
Ophthalmology	0.820 [0.388]	25	0.880	0.603
Pathology	0.625 [0.518]	4	0.750	-
Pharmacology	1.000 [0.000]	4	1.000	-
Physiology	0.895 [0.308]	43	0.884	0.380
Statistics	0.947 [0.226]	19	1.000	1.000

Table B2: Annotator Relevance Scores per Specialty.

Specialty	Average [standard deviation]	Number of Questions	Percent Agreement	Cohen's Kappa
Anatomy	0.699 [0.479]	103	0.641	0.225
Anesthesia	1.000 [0.000]	10	1.000	-
Biochemistry	0.667 [0.494]	24	0.708	0.400
Cell and Molecular Biology	0.726 [0.501]	62	0.435	0.095
Chemistry and Physics	0.333 [0.470]	48	0.688	0.286
Embryology	1.000 [0.000]	1	1.000	-
Emergency Medicine	1.000 [0.000]	2	1.000	-
Internal Medicine	1.000 [0.280]	18	0.833	0.000
Microbiology	1.000 [0.407]	15	0.600	0.000
Ophthalmology	1.000 [0.303]	25	0.800	0.000
Pathology	1.000 [0.463]	4	0.500	-
Pharmacology	1.000 [0.000]	4	1.000	-
Physiology	0.930 [0.275]	43	0.884	0.224
Statistics	0.947 [0.504]	19	0.211	0.021

C LLM-as-a-Judge

The models were provided with the full test set of MCQ (stem, options, and correct answer) and instructed to return only valid JSON output. The exact prompt was:

Listing 1: Prompt provided to LLMs

```
You are a medical education expert. Evaluate the following multiple-
    choice
question (MCQ) on a scale of 1 to 5 for each of the following metrics:

1. Medical Accuracy (1=very inaccurate, 5=highly accurate)
2. Clinical Relevance (1=not relevant, 5=highly relevant)
3. Question Difficulty (1=very easy, 5=very difficult)
4. Question Quality (1=poor, 5=excellent)

Important: Return ONLY valid JSON. No explanations, no markdown, no text.
The response must be exactly like this:

{
  "medical_accuracy": <1-5>,
  "clinical_relevance": <1-5>,
  "question_difficulty": <1-5>,
  "question_quality": <1-5>
}

Question stem: {row['Question']}
Options:
{options_text}
Correct answer: {row['Correct Answer']}
```

This setup ensured that LLM outputs were standardized and machine-readable. By aggregating scores across thousands of test questions, we obtained descriptive statistics and

Table B3: Annotator Difficulty Scores per Specialty.

Specialty	Average [standard deviation]	Number of Questions	Percent Agreement	Cohen's Kappa
Anatomy	0.816 [0.464]	103	0.650	0.240
Anesthesia	0.700 [0.510]	10	0.500	0.074
Biochemistry	0.708 [0.476]	24	0.750	0.442
Cell and Molecular Biology	0.661 [0.463]	62	0.710	0.320
Chemistry and Physics	0.958 [0.408]	48	0.625	0.027
Embryology	1.000 [0.000]	1	1.000	-
Emergency Medicine	1.000 [0.577]	2	0.000	-
Internal Medicine	0.722 [0.487]	18	0.611	0.182
Microbiology	0.933 [0.305]	15	0.800	-0.098
Ophthalmology	0.880 [0.418]	25	0.720	0.229
Pathology	0.500 [0.535]	4	1.000	-
Pharmacology	0.500 [0.518]	4	0.750	-
Physiology	0.372 [0.494]	43	0.651	0.281
Statistics	0.316 [0.500]	19	0.368	-0.009

Table B4: Annotator Quality Scores per Specialty.

Specialty	Average [standard deviation]	Number of Questions	Percent Agreement	Cohen's Kappa
Anatomy	0.573 [0.451]	103	0.573	0.044
Anesthesia	0.600 [0.489]	10	0.700	0.348
Biochemistry	0.583 [0.449]	24	0.625	0.143
Cell and Molecular Biology	0.548 [0.448]	62	0.484	-0.120
Chemistry and Physics	0.708 [0.457]	48	0.792	0.496
Embryology	1.000 [0.000]	1	1.000	-
Emergency Medicine	1.000 [0.000]	2	1.000	-
Internal Medicine	0.889 [0.280]	18	0.944	0.640
Microbiology	0.533 [0.490]	15	0.533	0.037
Ophthalmology	0.760 [0.370]	25	0.840	0.432
Pathology	0.750 [0.518]	4	0.750	-
Pharmacology	1.000 [0.000]	4	1.000	-
Physiology	0.884 [0.292]	43	0.860	0.178
Statistics	0.895 [0.273]	19	0.842	-0.075

model-wise distributions that enabled a more fine-grained analysis than binary human ratings alone.

Figure C5 shows box plots of the score distributions for each metric. These plots display the median (horizontal line inside the box), interquartile range (box), and outliers (circles). Gemini and GPT-4 produced broader spreads across the 1–-5 scale, while Claude's ratings were more concentrated toward higher scores, particularly for Medical Accuracy and Clinical Relevance. Question Difficulty was more evenly distributed across models, with medians near 2–-3.

To examine agreement between models, Pearson correlation coefficients were calculated on a per-question basis. Heatmaps for each metric are shown in Figure C6. Correlations ranged from 0.49 to 0.72, indicating moderate consistency between models. For example, Claude generally assigned higher absolute values, but the relative ordering of items remained similar across models.

It is important to interpret these results alongside the clinician ratings, which were binary (high/low). Binary judgments compress intermediate cases into "low," whereas the LLMs frequently assigned middle values (2-–4). This explains much of the apparent discrepancy: questions rated as "low" by clinicians often received intermediate LLM scores. Rather than contradicting clinicians, the LLMs add resolution by distinguishing between poor, moderate, and strong items within the test set.

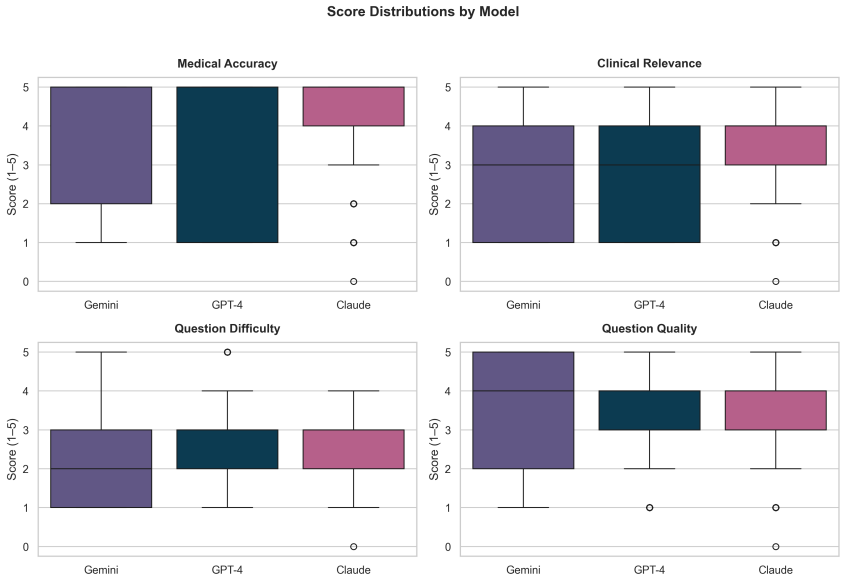


Figure C5: Distribution of LLM ratings (1–5) across four evaluation metrics and three models. Box plots show medians, interquartile ranges, and outliers.

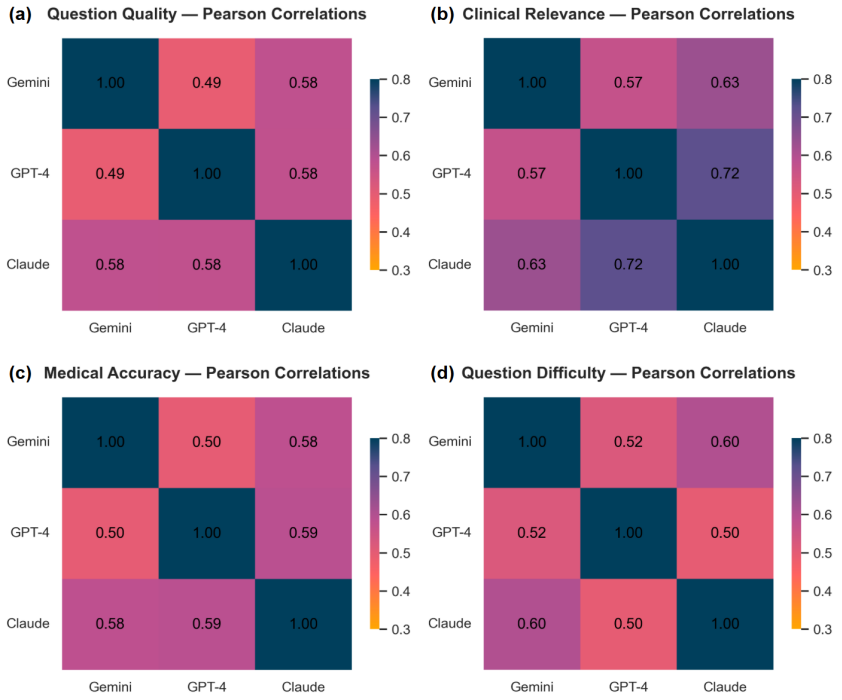


Figure C6: Pearson correlations of Medical Accuracy, Clinical Relevance, Question Difficulty, and Question Quality scores across models.

D Performance per Specialty and Level

To complement the overall evaluation, we analyze model accuracy across different medical specialties and difficulty levels. This breakdown highlights domain-specific strengths and weaknesses, as well as how performance varies across the five curriculum levels (Y1–-Y5).

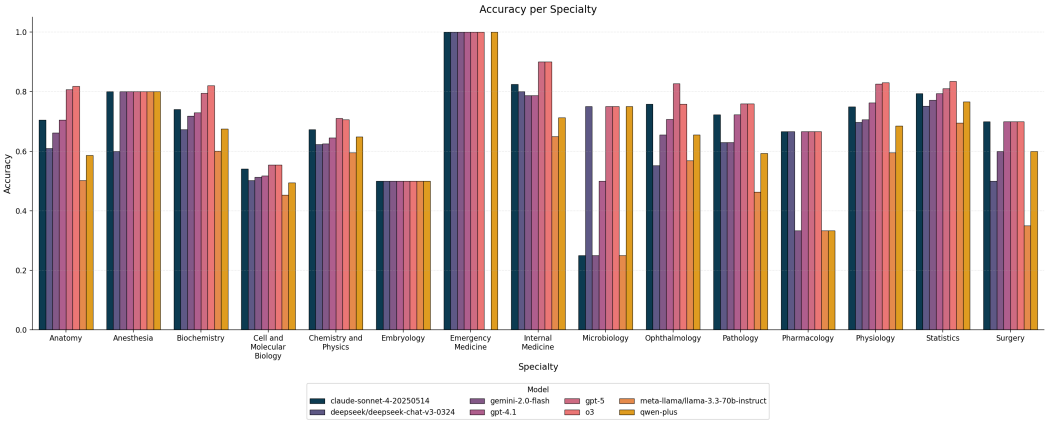


Figure D1: **Accuracy per Specialty across all 8 models.**

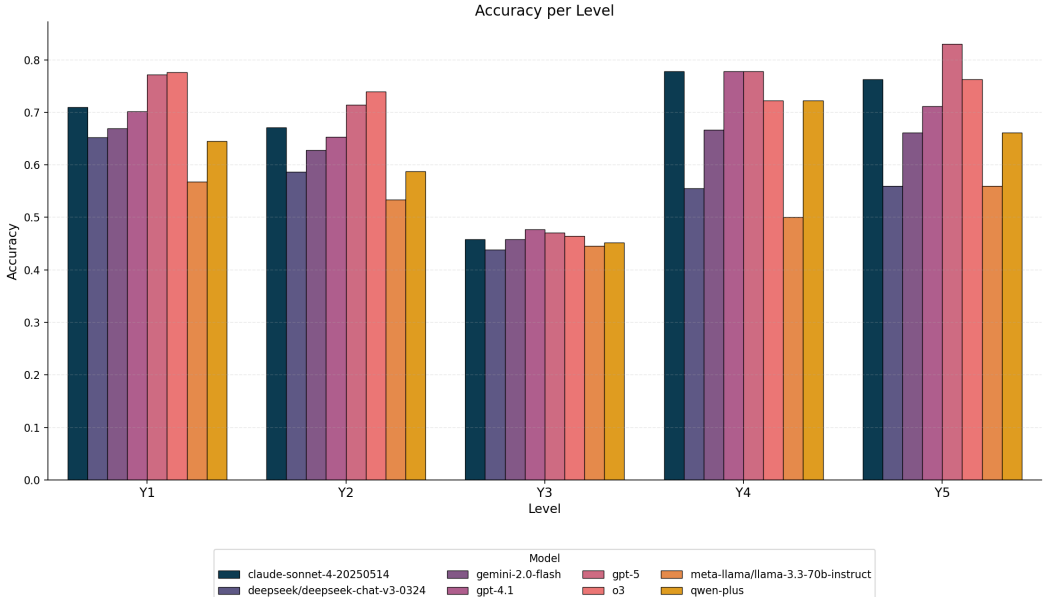


Figure D2: **Accuracy per Level across all 8 models.**

Table D1: Accuracy on MedAraBench (contemporary) vs MedArabiQ (legacy).

MedAraBench		MedArabiQ	
Contemporary Model	Accuracy	Legacy Model	Accuracy
claude-sonnet-4	69.4	claude-sonnet-3.5	53.5
gemini-2.0-flash	65.4	gemini-1.5	57.5
gpt-4.1	67.4		
gpt-5	76.4	gpt-4	53.5
gpt-o3	76.5		
llama-3.3-70b	54.7	llama-3.1-8b	26.2
qwen-plus	61.8	qwen-2.5-7b	38.0
deepseek v3	62.0	deepseek v3	50.5

Table D2: Average Scores of Expert vs LLM evaluation of MedAraBench

Evaluation Metric	**Expert Evaluation (average + std)**	**LLM Evaluation (average + std)**
Medical Accuracy	0.722 [0.448]	0.788 [0.312]
Clinical Relevance	0.653 [0.476]	0.639 [0.281]
Question Difficulty	0.669 [0.471]	0.496 [0.179]
Question Quality	0.767 [0.423]	0.681 [0.245]

E Answer Choice Distribution Balance

After constructing the dataset, we observed small but notable imbalances in answer distributions after constructing our dataset.

This could lead to model bias toward any specific answer position (e.g., always selecting "A"). As such, to address this, we analyzed and adjusted the distribution of correct answer choices across all subsets and splits by making minimal targeted adjustments (e.g., reordering options when possible) to bring the correct answer frequencies closer to uniformity. This refinement was done independently for the training and test sets across each answer format group (ABCD, ABCDE, ABCDEF). This adjustment ensures a balanced representation of correct answers and helps reduce the likelihood that models learn position-based heuristics.

Figures E1–E3 visualize the resulting distributions.

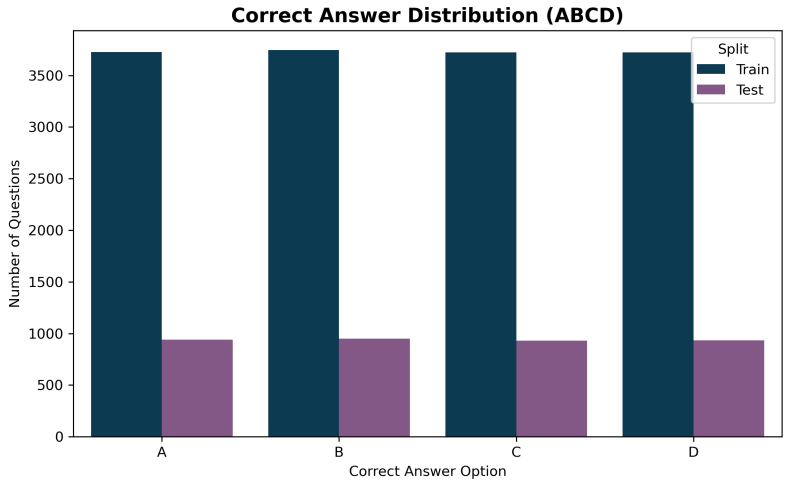


Figure E1: **Answer Choice Distribution for ABCD format (Train/Test)**

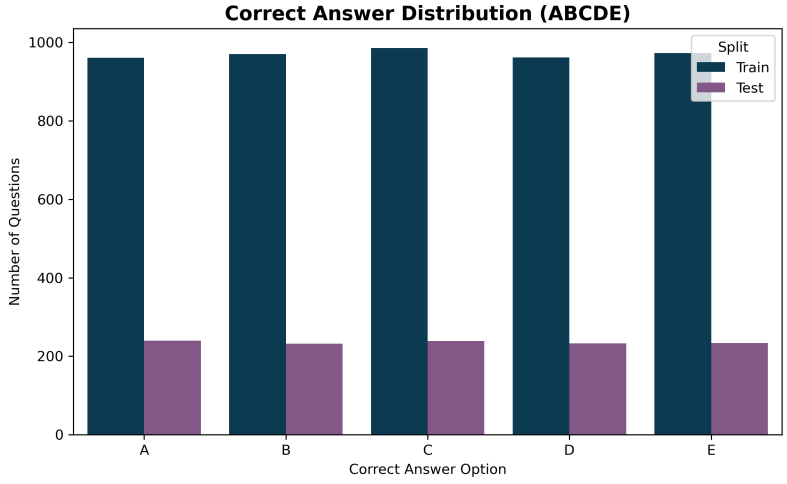


Figure E2: **Answer Choice Distribution for ABCDE format (Train/Test)**

We further validated this balance using chi-square goodness-of-fit tests. Chi-square goodness-of-fit tests confirmed no significant deviation from uniformity in ABCD and ABCDE splits ($p > 0.97$), indicating that the observed answer distributions do not significantly deviate from a uniform distribution. The ABCDEF format was excluded from

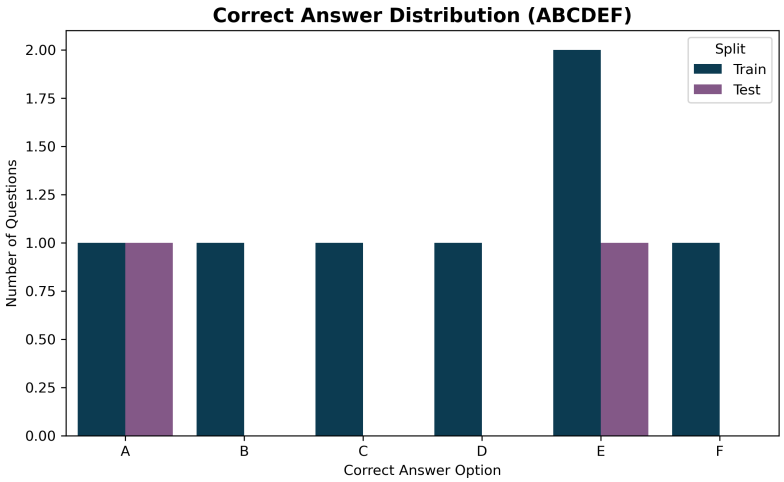


Figure E3: **Answer Choice Distribution for ABCDEF format (Train/Test)**

this analysis in the test set due to insufficient sample size (n = 2). All tests support the conclusion that the distributions do not significantly differ from uniformity.

The resulting χ^2 values and p-values were as follows:

- **ABCD (Train)**: $\chi^2 = 0.099$, $p = 0.992$
- **ABCD (Test)**: $\chi^2 = 0.212$, $p = 0.976$
- **ABCDE (Train)**: $\chi^2 = 0.422$, $p = 0.981$
- **ABCDE (Test)**: $\chi^2 = 0.226$, $p = 0.994$
- **ABCDEF (Train)**: $\chi^2 = 0.714$, $p = 0.982$
- **ABCDEF (Test)**: $\chi^2 = 6.02$, $p = 0.304$

This adjustment ensures a balanced representation of correct answers and helps reduce the likelihood that models learn position-based heuristics.