MedArabiQ: Benchmarking Large Language Models on Arabic Medical Tasks

Mouath Abu Daoud[†] MMA9138@NYU.EDU
Chaimae Abouzahir[†] CA2627@NYU.EDU
Leen Kharouf LK2713@NYU.EDU
Walid Al-Eisawi WA2076@NYU.EDU
Nizar Habash NIZAR.HABASH@NYU.EDU
Farah E. Shamout FARAH.SHAMOUT@NYU.EDU
New York University Abu Dhabi, United Arab Emirates
[†] *Equal contributions*

Abstract

Large Language Models (LLMs) have demonstrated significant promise for various applications in healthcare. However, their effectiveness in the Arabic medical domain remains unexplored due to the lack of high-quality domain-specific datasets and benchmarks. This study introduces MedArabiQ, a new benchmark dataset consisting of seven Arabic medical tasks, covering multiple specialties and including multiple-choice questions, fill-in-the-blank questions, and patient-doctor questions and answers. We first constructed the dataset using past medical exams as well as publicly available datasets. We conducted an extensive evaluation with eight state-of-the-art open-access and proprietary high-resource LLMs, including GPT-4, Deepseek v3, and Gemini 1.5. Our findings highlight the need for the creation of new high-quality benchmarks that span different languages to ensure fair deployment and scalability of LLMs in healthcare. By establishing this benchmark and releasing the dataset, we provide a foundation for future research aimed at evaluating and enhancing the multilingual capabilities of LLMs for the equitable use of generative AI in healthcare.

Data Availability In this article, we present a new benchmark dataset, MedArabiQ, designed to evaluate the performance of LLMs on Arabic medical tasks. We release our data for reproducibility to enable a fair evaluation of language models in the future: https://github.com/nyuad-cai/MedArabiQ

1. Introduction

The recent advent of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP) by demonstrating exceptional performance in various tasks, ranging from language translation to creative writing (Nazi and Peng, 2023). Although LLMs were initially designed for general language understanding, they have since been evaluated for many domain-specific applications, such as education, programming, art, and medicine. They have also been adapted for domain-specific tasks using different fine-tuning strategies and specialized datasets (Kaddour et al., 2023).

The use of LLMs in healthcare has sparked enthusiasm due to the potential to improve diagnostic processes, clinical decision-making, and overall patient care (Nazi and Peng, 2023;

Meng et al., 2024). One particular application of interest is medical education, in which LLMs can generate concise summaries and support interactive learning experiences (Benítez et al., 2023). To this end, several benchmarks have been proposed to assess the capabilities of LLMs in medical reasoning. Despite these advances, challenges remain, including ethical concerns, risks of generating biased or harmful content, and variability in performance in different languages and cultural contexts (Yang et al., 2024; Nazi and Peng, 2023).

Existing benchmarks, such as GLUE and MedQA, cater primarily to English, leaving a significant gap in the evaluation of LLMs for Arabic healthcare applications (Qiu et al., 2024). This is due to multiple factors, including the limited availability of high-quality Arabic datasets for clinical applications, coupled with the unique linguistic complexity of Arabic, especially when considering the many different Arabic dialect regions (e.g., Gulf, Maghreb, Egypt, Levant, among others) in addition to Modern Standard Arabic (MSA) (Salameh et al., 2018). Additionally, although various multilingual models include Arabic in their training data, their performance often falls short in clinical contexts due to insufficient domain-specific resources and a lack of appropriate benchmarks (Nazi and Peng, 2023; Gangavarapu, 2024). Addressing these gaps is critical to unlocking the full potential of LLMs for Arabic-speaking patients and providers, ensuring equitable access to AI-driven advancements in healthcare.

To address these gaps, there is a growing need for frameworks that assess LLM performance in clinical tasks specific to Arabic-speaking populations. By developing benchmarks that reflect real-world clinical interactions, we can ensure more reliable and culturally appropriate LLM deployment in multilingual healthcare systems. In this study, we make several key contributions to address these challenges by introducing MedArabiQ (see Figure 1). First, we developed seven benchmark datasets designed to evaluate LLMs in Arabic healthcare applications while addressing Arabic linguistic complexity along with domain-specific challenges that hinder existing models. We focus on critical healthcare medical tasks, including medical question-answering, clinical dialogue, and ethical decision making. Secondly, we analyze the performance of multilingual and Arabic LLMs, including high-resource proprietary and open-access LLMs, highlighting the impact of linguistic coverage and transparency in training data on healthcare applications. We conduct a comprehensive evaluation to assess model performance, in order to provide a robust foundation for advancing AI-driven solutions for Arabic healthcare tasks.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work provides a strong foundation for advancing LLM evaluation for Arabic medical tasks. While our work focuses on MSA, future efforts could further develop our preprocessing pipeline to explore the use of regional dialects, which are commonly used in clinical communication. MedArabiQ also eliminates the entry barrier for researchers in underrepresented regions by providing annotated data and an evaluation framework. By promoting open data and reproducibility, we believe that this is an important step towards building LLM agents that are more inclusive and globally relevant. Overall, MedArabiQ supports further research in areas like clinical decision support and telehealth, ultimately contributing to the development of AI models that are better suited for broader adoption in global real-world clinical workflows.

2. Related Work

Several benchmarks have been proposed for the evaluation of LLMs for various medical tasks, predominantly in the English language. For example, Gao et al. (2023) introduce Dr. Bench, a diagnostic reasoning benchmark for clinical NLP emphasizing clinical text understanding, medical knowledge reasoning, and diagnosis generation. The benchmark combines English-language data from various sources but primarily consists of in-hospital clinical notes. To assess LLM performance in medical question-answering tasks, a number of benchmarks incorporated samples from medical board exams, such as MedQA. MedQA advances multilingual benchmarking by including questions in traditional and simplified Chinese and English (Jin et al., 2020). The Massive Multitask Language Understanding (MMLU) benchmark uses the US Medical Licensing Exam (USMLE) (Hendrycks et al., 2021) for a subset of tasks. MedMCQA extends these benchmarks to a multilingual evaluation framework (Pal et al., 2022).

Despite the growing interest in Arabic NLP, Arabic medical benchmarks remain limited. Achiam et al. (2023) translated MMLU into 14 languages, including Arabic, with the help of professional human translators. AraSTEM focuses on the task of question-answering and includes a medical subset (Mustapha et al., 2024). AraMed similarly presents an Arabic medical corpus and an annotated Arabic question-answering dataset (Alasmari et al., 2024). Other datasets are also largely skewed towards the task of Arabic medical question-answering and have other limitations, summarized in Table A1. While these resources are helpful, they do not comprehensively cover the spectrum of Arabic medical tasks, highlighting the need for dedicated benchmarking efforts.

Clinically, a number of frameworks have been proposed for evaluating clinical AI models. The 'Governance Model for AI in Healthcare' consists of four main pillars: fairness, transparency, trustworthiness, and accountability (Reddy et al., 2020). Similarly, Dada et al. (2024) propose the Clinical Language Understanding Evaluation framework, which was designed to assess LLMs with real patient data for various modalities, including patient-specific question-answering, hypothesis deduction, and problem and inquiry summarization. Kanithi et al. (2024) introduce 'MEDIC', a framework for evaluating LLMs across five clinically relevant categories: medical reasoning, ethical and bias concerns, data and language understanding, in-context learning, and clinical safety and risk assessment. Although there has been significant progress in the field of benchmarking LLMs in medical tasks, most evaluations have been performed on data in English, and there is no single benchmark to assess LLMs in Arabic on more than one medical task and performance metric. There are over 380 million native Arabic speakers (Eberhard et al., 2024), a number of which are monolingual. With the vast potential of LLMs in healthcare, it is crucial to accommodate Arabic-speaking patients to ensure fair deployment. Further details on related works are present in Appendix A.

3. Methods

In this section, we describe the details of our methodological framework to construct the datasets and evaluate state-of-the-art LLMs. We start by explaining our chosen tasks, models and the experimental setup used to evaluate model performance. In Section 4,

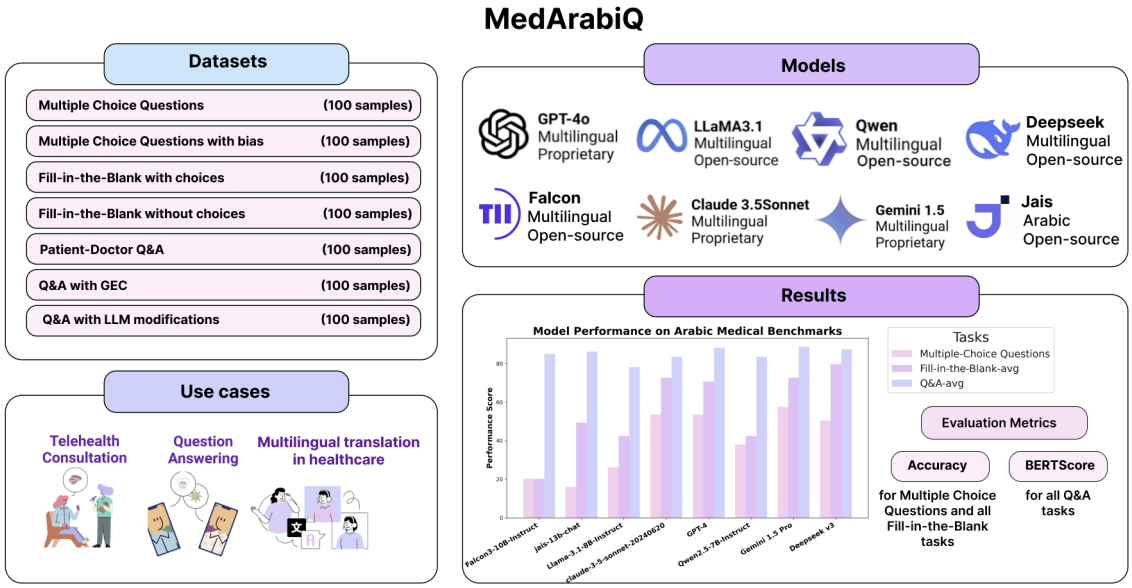

Figure 1: **Overview of MedArabiQ.** We construct seven new benchmark datasets and evaluate eight state-of-the-art LLMs on them.

we further explain our data collection and preprocessing methods. A general overview of MedArabiQ is provided in Figure 1.

3.1. Tasks and Models

In order to develop a reliable framework for evaluating LLMs in Arabic healthcare applications, we focused on telehealth consultation and question-answering as the main use cases. These involve not only medical reasoning ability but also natural patient-doctor dialogue. The LLM should mimic the role of a clinician as closely as possible, which includes possession of medical knowledge and the ability to utilize the knowledge, personalized to the needs and queries of patients. Personalization, however, should not lead to bias or prejudice in the LLM's response, based on the patient's profile.

At the same time, it is important to note that benchmarks used to evaluate LLMs often face the challenge of data contamination. Many modern LLMs are trained on large web crawl datasets sourced from the internet, which can inadvertently include benchmark questions. This overlap inflates performance metrics, making such benchmarks unreliable for fair evaluation (Chunyuan et al., 2024; Dong et al., 2024). Addressing this issue is critical to ensure that benchmark results accurately reflect the true capabilities of LLMs, free from biases introduced by prior exposure to evaluation data.

To mitigate these concerns, we categorize models into two primary groups: those with known training data and those with unknown training data. This initial distinction allows

us to assess contamination risks more reliably for models whose training sources are documented, while acknowledging that for models without such transparency, contamination cannot be easily ruled out.

In addition to categorizing models based on the transparency of their training data, this study further classifies models according to their linguistic coverage. Specifically, we distinguish between models that are broadly multilingual and those that are primarily Arabic-centric multilingual. Although many LLMs include training data from numerous languages, exposure to a language does not inherently confer proficiency (Bender et al., 2021). GPT-4, for example, is known to have been trained on a variety of languages (Achiam et al., 2023). By assessing multilingual and Arabic-centric multilingual models separately, we explore how language coverage and specialization influence performance, particularly in Arabic medical tasks. Our model grouping is shown in Table B1 of Appendix B.

This grouping reflects important distinctions in model transparency and linguistic specialization, but we acknowledge the considerable heterogeneity within each group regarding architecture, scale, training data, and performance. To better account for this, we provide additional analysis in Figure B1 and Table B2 of the appendix, comparing models by training compute (FLOPs) and dataset size. As expected, larger training compute generally correlates with improved performance, particularly for closed models such as GPT-4 and Gemini 1.5 Pro. Training dataset size shows some correlation within open models, though trends are less consistent. These analyses support the rationale for our grouping by offering further context for interpreting performance differences amid diverse model characteristics.

3.2. Experimental Setup

Our framework evaluates LLMs on Arabic medical tasks through seven datasets derived from two sources: (1) private structured medical exams/notes and (2) public patient-doctor dialogues adopted from AraMed (Alasmari et al., 2024), as will be shown in further detail in the following sections.

Evaluation. The evaluation of LLMs in Arabic healthcare applications requires a comprehensive framework that balances technical performance with real-world applicability (Sallam and Mousa, 2024). Our framework assesses medical reasoning, decision-making, and dialogue-based question-answering (Kanithi et al., 2024; Guo et al., 2023). Multiple-choice questions (MCQ) were evaluated on accuracy of the answer, while open-ended questions, such as those in the fill-in-the-blank and patient-doctor Q&A, were assessed using BERTScore to measure semantic alignment (Zhang et al., 2020). We perform zero-shot prompting for all models and tasks.

Rather than applying a uniform temperature across models and tasks, we tuned temperature settings based on empirical testing and guidance from prior work (Du et al., 2025), which recommends task- and model-aware configuration to improve generation quality without introducing evaluation bias. For classification tasks such as MCQ and fill-in-the-blank, we used a low temperature (0.0 for open models, 0.2 for closed models) to promote deterministic, stable outputs. For generative tasks like patient-doctor Q&A, slightly higher temperatures improved coherence and naturalness without degrading factuality. Accordingly, we used 0.4 for open models (e.g., LLaMA 3.1, Qwen 2.5) and 0.2 for closed models (e.g., GPT-4, Gemini 1.5 Pro). We further extended our experiments with a temperature

study across representative models in Appendix B.2. For semantic evaluation, we use the XLM-RoBERTa-Large model in BERTScore computation due to its multilingual training, which includes Arabic, making it more appropriate to use in our setting.

Instruction-tuned Models. We employed instruction-tuned versions of the models described earlier due to their superior ability to interpret and execute task-specific instructions. In contrast, the base versions demonstrated significant limitations in following prompts, even with extensive prompt engineering. This aligns with the literature (Chung et al., 2022; Zhang et al., 2023), which shows that instruction fine-tuning significantly improves model performance and prompt adherence across various tasks and model sizes.

Instruction Prompts. Prompt engineering is crucial for evaluating LLMs. In our prompt engineering experiments, we tested both English and Arabic prompts and found that English prompts were generally more effective. However, for Q&A benchmark tasks, Arabic prompts performed better in open-access models. Based on these findings, we used English prompts for all tasks except for the Q&A datasets. The same prompt was used for each model, though it was customized to each category of tasks as shown in Table C1 of the appendix. This was crucial to ensure the prompt reflected realistic scenarios as closely as possible, tailored to different use cases.

Answer Processing. For MCQ and fill-in-the-blank with choices tasks, models generate both the index of the predicted correct option and the full textual answer. However, some models, particularly open-access ones, exhibit spelling inconsistencies. To ensure accuracy, we evaluated only the first character generated after the phrase "The correct letter is:", comparing it against valid options. If the character does not match a valid choice, then the response is deemed invalid. For open-ended benchmarks, where answers are not constrained to multiple-choice formats, we evaluate the full model-generated response.

3.3. Bias Assessment and Mitigation

To ensure that LLMs can be effectively deployed in healthcare and provide meaningful contributions to the field, it is critical to address their potential to replicate human biases. Completely eliminating bias from LLMs is nearly impossible, as the datasets used to train these models are inherently shaped by human judgment, which is susceptible to bias. Recognizing this challenge, we developed an evaluation framework to systematically assess model resistance to bias, measure susceptibility, and evaluate mitigation strategies. The framework draws on methodologies outlined in recent work (Schmidgall et al., 2024), while adapting key elements to the context of Arabic medical datasets and healthcare applications. These adaptations include culturally relevant bias categories, prompts designed for clinical scenarios, and additional evaluation metrics to align with real-world needs.

We created a structured framework to systematically assess language-model resistance to cognitive biases:

1. **Baseline Testing:** Models are evaluated using the original, unbiased dataset to establish baseline performance metrics.

2. **Bias Testing:** Models are tested with biased variations of prompts. We compute the change in accuracy to assess impact of bias (see Table D1 of the appendix for bias injection examples).

3. **Evaluation of Bias Mitigation:** Mitigation techniques are tested to determine their impact on accuracy. These include:

    - **Bias Education:** Adding warnings to prompts emphasizing evidence-based reasoning (e.g., "Evaluate each patient uniquely, without relying on trends or recent cases").
    - **One-Shot Demonstration:** A single negative example is provided to illustrate incorrect reasoning caused by bias.
    - **Few-Shot Demonstration:** Both negative and positive examples are presented to show correct and incorrect handling of bias.

    See Table E1 in the appendix for an illustration of these mitigation strategies in use.

This framework provides a systematic and reproducible approach to assess and address cognitive biases in LLMs, ensuring their deployment in healthcare contexts is both effective and ethically sound.

4. Dataset Construction

As previously stated, our framework for evaluating LLMs in Arabic healthcare applications integrates clinically validated knowledge and linguistically diverse sources to construct a benchmark that mirrors real-world patient needs and clinical reasoning challenges. We do so by deriving our datasets from two primary sources: past exams and notes from Arabic medical schools, and the AraMed Dataset (Alasmari et al., 2024).

4.1. Data Selection

In our data selection process, we specifically selected data sources that were unlikely to have been included in prior training datasets.

4.1.1. Multiple-Choice Questions

To evaluate the models' medical understanding, we curated a standard dataset with question-answer pairs, covering foundational and advanced medical topics, such as physiology, anatomy, and neurosurgery. From our curated dataset, we selected a random set of 100 multiple-choice questions on an ad-hoc basis by prioritizing: (i) broad coverage of specialties, (ii) varying difficulty levels, and (iii) clarity of question and answer. We then digitized them into CSV files and performed manual verification. The average question length is 24 words.

4.1.2. Multiple-Choice Questions with Bias

Following recent work (Schmidgall et al., 2024), we injected bias in the multiple-choice questions dataset to evaluate how LLMs handle ethical or culturally sensitive scenarios. In particular, we utilized a set of well-defined bias categories (Schmidgall et al., 2024), including (i) confirmation bias, (ii) recency bias, (iii) frequency bias, (iv) cultural bias, (v) false-consensus bias, (vi) status quo bias, and (vii) self-diagnosis bias. By manually

injecting the bias, we ensured relevance to the unique linguistic and clinical challenges of Arabic healthcare contexts. This resulted in a dataset consisting of 100 samples.

4.1.3. Fill-in-the-Blank with Choices

To assess knowledge recall and in-context learning, we manually constructed fill-in-the-blank questions, each accompanied by a set of predefined answer choices. The model was required to select the most appropriate answer from the given options. This approach evaluates the model's ability to recognize correct answers within a constrained set, reducing the reliance on generative capabilities. The resulting dataset consists of 100 samples.

4.1.4. Fill-in-the-Blank without Choices

In this setting, the fill-in-the-blank questions were presented without predefined answer choices, requiring the model to generate responses independently. This evaluation measures the model's ability to recall and generate accurate medical knowledge without additional information. The dataset for this task also comprises 100 samples.

4.1.5. Patient-doctor Q&A

AraMed is an Arabic medical corpus for question-answering that was originally sourced from AlTibbi, an online medical patient-doctor discussion forum (Alasmari et al., 2024). The original dataset consists of 270,000 question-answer pairs, of which 400 were made publicly available. We meticulously selected 100 samples covering the entire range of specialties present originally in AraMed, ensuring quality and avoiding redundancy.

4.1.6. Q&A with Grammatical Error Correction (GEC)

Since the patient-doctor Q&A dataset uses dialectal Arabic, we constructed an additional version with enhanced linguistic quality and consistency. We specifically applied a Grammatical Error Correction (GEC) pipeline tailored for Arabic healthcare texts. Given the morphological complexity and syntactic richness of Arabic, this preprocessing step was essential. First, we used cameltools, an open-source Arabic NLP library, to morphologically disambiguate words and dediacritize text, ensuring uniformity (Obeid et al., 2020). This step was crucial for handling Arabic's inflectional patterns and preparing the dataset for grammatical correction. Then, we employed an existing fine-tuned BERT-based Arabic Grammatical Error Detection (GED) model (Alhafni et al., 2023) to detect agreement errors, incorrect word order, and missing inflections. Each token was tagged with a grammatical label to facilitate structured corrections. Next, we used the GEC model from the same pipleine, built on mBART, to automatically correct detected errors while preserving semantic integrity. The model was fine-tuned on the QALB-2015, QALB-2014, and ZAEBUC corpora to enhance its performance in grammatical corrections (Mohit et al., 2014; Rozovskaya et al., 2015; Habash and Palfreyman, 2022).

4.1.7. Q&A with LLM Modifications

Additionally, to mitigate potential memorization, since some models could have been trained on scraped data from Altibbi, we modified the dataset using an LLM for a more rigorous

assessment. (Dong et al., 2024). In particular, we used GPT-4o to paraphrase our original questions using the prompt, "You are a helpful assistant that paraphrases text while keeping its meaning intact." This approach ensured that the medical concepts remained unchanged while ensuring the models do not rely on memorized content (Zhou et al., 2024).

In summary, we constructed seven new datasets using two primary sources, AraMed and past medical exams, via extensive manual verification to build the MedArabiQ benchmark, all of which are summarized in Table F1 in Appendix F.

4.2. Data Extraction

In the collection of the MCQ dataset in Section 4.1.1 and the fill-in-the-blank questions dataset in Section 4.1.3, we started off by collecting paper-based past exams and lecture notes sourced from a large repository of academic materials hosted on student-led social platforms of regional medical schools. No personally identifiable information or real patient data was included, and thus anonymization was not necessary as our data collection complied with privacy and ethical guidelines. These exams and answers were not readily available in structured digital formats, necessitating a rigorous manual process to ensure clarity and correctness. Given that Arabic medical education is not widely digitized, these exams are not publicly accessible in structured formats. Even if some individual questions exist online, the extensive effort required to compile, format, and structure them into a benchmark dataset significantly reduces the likelihood of contamination. Questions were selected to reflect increasing complexity across different academic years, ensuring that model performance could be assessed at varying levels of medical expertise.

As for the patient-doctor Q&A dataset, our selection process prioritized questions with well-formed queries and meaningful answers, avoiding instances where responses were overly generic (e.g., "Consult a doctor" or "See a specialist"). However, we retained some examples of such cases to reflect real-world user behavior, as patients often seek medical advice even for questions that are better suited for in-person consultations.

Additionally, we maintained proportionality in representation to mirror the dataset's original structure, particularly for categories like reproductive and sexual health. These categories comprised a significant portion of the dataset and are often underrepresented in Arabic-language medical research despite their importance. Given the sensitivity and cultural taboos associated with these topics, we ensured their inclusion to provide a more realistic and balanced evaluation of the model's ability to handle diverse medical inquiries. This approach ensures the dataset reflects the varied nature of medical concerns while maintaining its relevance to real-world healthcare scenarios. We selected 100 questions, allocating an equal proportion of samples to each specialty: cardiology, obstetrics and gynecology, surgery, pediatrics, neurology, oncology, endocrinology, dentistry, otholarhyntology, public health, dermatology, primary care, pulmonology, and psychology.

We also incorporated information about the patient, specifically age and gender, when known, into the query. When this information was unknown or erroneous, it was excluded. Some questions came from an acquaintance of the patient, so no information was known about the patient. The information was usually prepended to the question or inserted after the greeting, in the format of "I am a [man/woman] and I am [x] years old". This

Table 1: **Summary of performance results across all benchmark datasets.** We present the results in terms of accuracy and BERTScore, depending on task, and show best results in bold per row. Due to space constraints, abbreviated model names are used in the table; full names are as follows: Falcon = `Falcon3-10B-Instruct`, Jais = `jais-13b-chat`, LLaMA = `LLaMA-3.1-8B-Instruct`, Claude = `claude-3-5-sonnet-20240620`, GPT-4 = `gpt-4-0613`, Qwen = `Qwen2.5-7B-Instruct`, Gemini = `Gemini 1.5 Pro`, Deepseek = `Deepseek v3`.

Benchmark Datasets	Metrics	Falcon	Jais	LLaMA	Claude	GPT-4	Qwen	Gemini	Deepseek
Multiple-Choice Questions	Accuracy	20.2	16.0	26.2	53.5	53.5	38.0	**57.5**	50.5
Fill-in-the-Blank with choices	Accuracy	20.2	49.4	42.4	72.7	70.7	42.40	72.7	**79.7**
Fill-in-the-Blank without choices	BERTScore	85.1	86.2	78.2	83.6	88.2	83.5	**88.8**	87.4
Patient-Doctor Q&A	BERTScore	84.9	**85.7**	81.2	81.1	84.5	85.2	82.5	82.2
Q&A with LLM modifications	BERTScore	84.8	**85.6**	85.5	83.7	84.5	85.2	82.5	82.3
Q&A with Grammatical Error Correction	BERTScore	84.4	**85.5**	84.9	83.6	84.2	84.9	82.3	82.0

information was useful in adding personalization to the questions, making our benchmark better suited to evaluate LLM performance in real-world medical scenarios.

5. Results

We report all results of the experiments performed on six of our benchmarks in Table 1. Our results show that no single model outperforms all others across all benchmarks. For closed tasks (MCQ and fill-in-the-blank with and without choices), closed models perform best, as expected. Gemini 1.5 Pro achieves the highest accuracy in two out of six benchmark tasks, while Deepseek leads on the fill-in-the-blank with choices task with an accuracy of 79.7. For closed tasks, Gemini and Deepseek perform best with accuracy scores of 57.5 and 79.7, respectively. In open-ended tasks, namely patient-doctor Q&A, Q&A with GEC, and Q&A with LLM modifications, Jais performs best, achieving scores of 85.7, 85.6, and 85.5 respectively. However, models do not consistently excel across both task types, as Jais performs poorly on closed tasks. Among open-access models, Jais is the best-performing, followed by LLaMA, which achieves the highest score on Q&A with LLM Modifications.

Figure 2 (a) compares the performance of open-access and proprietary high-resource models averaged across all benchmark tasks. Among all models, **Gemini 1.5 Pro** achieves the **highest average benchmark score**, followed closely by GPT-4, Deepseek v3 and Claude 3.5, reinforcing the dominance of proprietary high-resource models in NLP tasks. The open-access models, including Llama 3.1, Qwen 2.5, and Falcon, demonstrate competitive but more variable performance, with Qwen 2.5 emerging as the strongest among them. The error bars indicate that while closed models exhibit greater stability across tasks, open-access models show higher performance variability, likely due to their reliance on general-purpose pretraining rather than domain-specific fine-tuning. These findings highlight the

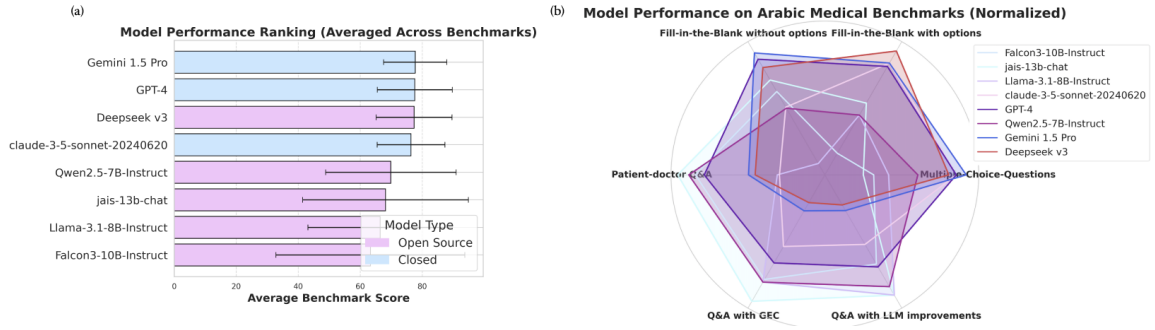

Figure 2: **Overall results of SOTA models on the MedArabiQ benchmark.** (a) Average performance across benchmark datasets. (b) Breakdown of model performance across benchmark tasks.

current dominance of proprietary high-resource models while underscoring the potential of open-access models with further fine-tuning and domain adaptation.

Figure 2 (b) provides a task-specific performance breakdown, revealing significant performance variation across MCQ, fill-in-the-blank, and generative Q&A tasks. Closed models consistently achieve high scores across all benchmarks, particularly excelling in Q&A tasks, including those with GEC and LLM-modified questions. This suggests a strong capability for handling complex medical inquiries, a critical requirement in real-world applications. However, the fill-in-the-blank and multiple-choice tasks exhibit more performance divergence, with open-access models like Qwen 2.5 and Llama 3.1 lagging behind the leading closed based models. Notably, Jais performs better in certain benchmarks but struggles with structured tasks, suggesting limitations in factual consistency and retrieval-based reasoning. The observed discrepancies emphasize the need for fine-tuning on structured medical datasets, particularly for open-access models, to improve accuracy in knowledge-intensive tasks. Figure 3 shows examples of model responses for the closed and open-ended tasks.

Table 2: **Comparison of performance by accuracy score without bias, with bias, and with bias mitigation across the three chosen models.**

Model	Q&A	Q&A with Bias	Bias Education	One-Shot Mitigation	Few-Shot Mitigation
Claude 3.5 Sonnet	53.1	52.0	51.0	**55.1**	52.0
GPT-4	**66.3**	35.7	42.9	46.9	46.9
Gemini 1.5 Pro	55.1	**55.1**	**57.1**	54.1	**59.2**

Table 2 and Figure 4 highlight the impact of bias and different mitigation strategies on model performance. In Figure 4 (a), we compare the accuracy of GPT-4, Gemini 1.5 Pro, and Claude 3.5 Sonnet-20240620 on original questions versus questions injected with bias across various bias categories. The results reveal that, generally, all models experience a decline in accuracy when bias is introduced, with the extent of decrease varying by bias type. Across all bias catgories, Gemini 1.5 Pro demonstrates high resilience to bias, while

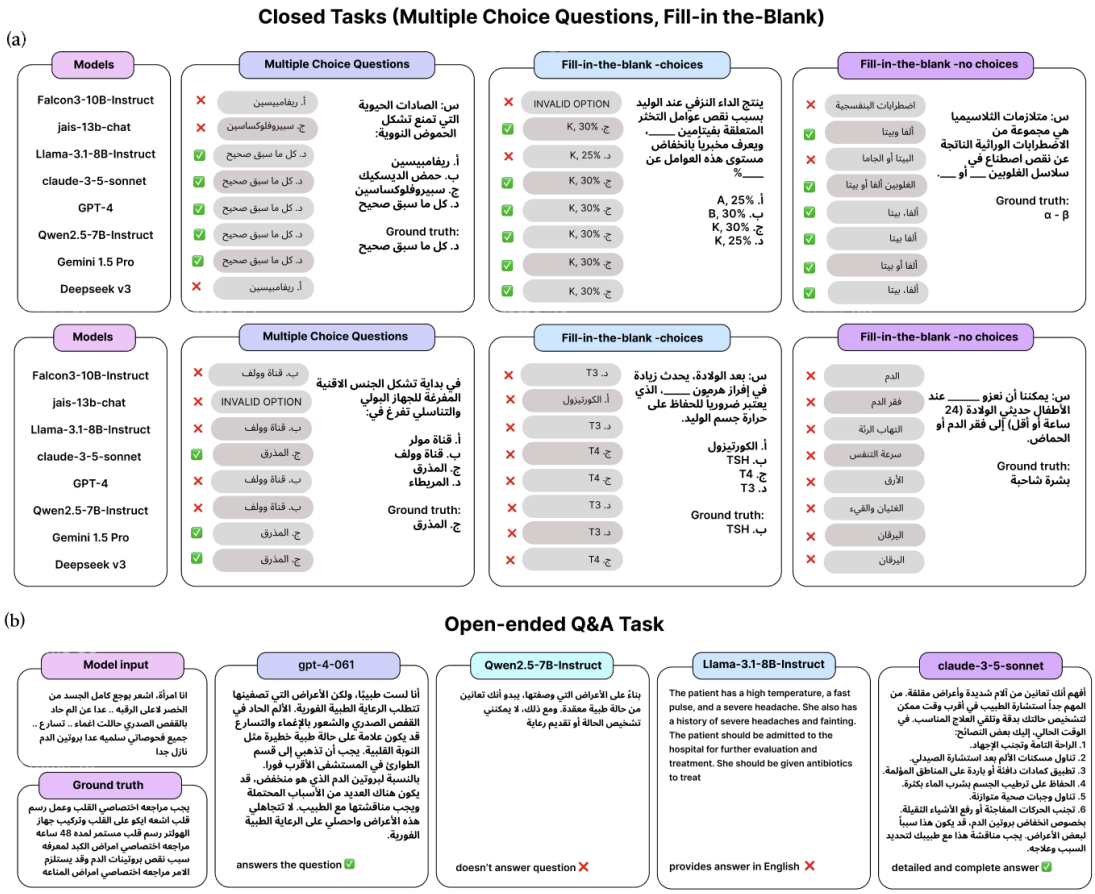

Figure 3: **Performance Samples from Closed and Open-ended Benchmarks.** We report model outputs for samples from (a) MCQ and Fill-in-the-Blank, and (b) patient-doctor Q&A tasks, illustrating differences in accuracy, response validity, and language consistency across proprietary high-resource and open-access models. An English version of this figure is provided in Figure G1 of the appendix.

GPT-4 shows more significant performance drops in False-Consensus Bias and Recency Bias. Additionally, Figure 4 (b) illustrates the effectiveness of different bias mitigation strategies. Notably, all models exhibit improved accuracy with strategies like Few-Shot prompting compared to biased questions without mitigation. Gemini 1.5 Pro outperforms the other models with both Bias Education and Few-Shot Mitigation, reinforcing its robustness in handling biased inputs.

Figure 4 (c) presents a radar plot summarizing model performance across the employed bias mitigation strategies. The plot shows Gemini's consistent performance across all strategies, while GPT-4 and Claude 3.5 Sonnet display larger variability across strategies such as Bias Education and One-Shot prompting. These results further highlight the importance of bias mitigation strategies in enhancing model reliability. Additional information on the per-

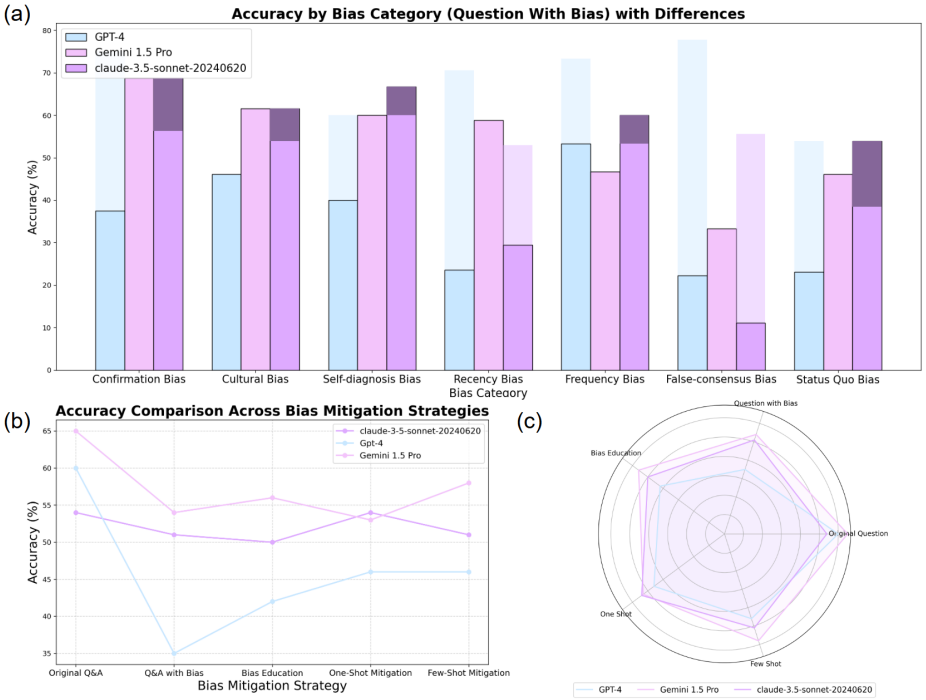

Figure 4: **Overview of Bias Mitigation Strategy Results for top-performing models.** (a) Comparison of accuracy between original questions and questions with bias across different bias categories. The bars represent accuracy percentages, with lighter shades representing a decrease, and darker shades an increase, in accuracy after bias incorporation. (b) Line plot comparing accuracy scores across various mitigation strategies. (c) Radar plot illustrating model performance across different bias mitigation strategies, highlighting relative strengths and weaknesses of each model in handling bias-related challenges.

formance of models by bias category and medical specialty across different bias mitigation strategies can be found in Figures H1 and I1 of the appendix, respectively.

6. Discussion

Our findings are consistent with prior research on medical LLM benchmarking, where proprietary high-resource models outperform open-access models in structured tasks but perform comparably in generative ones. Chen et al. (2025) demonstrated that proprietary high-resource models like GPT-4 and Med-PaLM 2 excel in multiple-choice and factual retrieval tasks, largely due to stronger pretraining on structured datasets and domain-specific knowledge integration. Similarly, Alonso et al. (2024) found that proprietary high-resource models generally achieve higher accuracy in medical question-answering tasks across multiple languages, reinforcing the idea that closed models have better knowledge grounding and factual consistency. Our results support these trends, with Gemini 1.5 Pro and Claude

3.5 Sonnet leading in multiple-choice and fill-in-the-blank tasks, suggesting that API-based models are better suited for clinical decision support and structured question-answering tasks.

Jais underperforms even compared to a random baseline in structured tasks due to severe hallucinations, generating unreliable responses, while Falcon struggles to generalize due to limited exposure to high-quality medical datasets(Penedo et al., 2023b). Figure 3 further illustrates these issues, particularly in the fill-in-the-blank task, where Jais frequently selects invalid options or provides factually inconsistent answers. These results highlight the limitations of open-access models in structured medical applications, where their general-purpose pretraining leads to factual inconsistencies, necessitating domain-specific fine-tuning. Additionally, the relatively lower performance of the Arabic-trained Jais relative to other open-access multilingual models such as Qwen suggests that task specificity may be more critical than language specificity, for Arabic medical tasks, and warrants further exploration in the future.

Figure 4 further illustrates the variability in model performance when exposed to biased content and the effectiveness of different bias mitigation strategies for the top 3 scoring models (GPT-4, Gemini 1.5 Pro, and Claude 3.5 Sonnet-20240620). All models experience an accuracy drop when bias is introduced, with GPT-4 Sonnet showing the most pronounced decline across multiple bias categories. Notably, bias mitigation techniques such as One-Shot and Few-Shot prompting have shown significant improvement in model performance, especially for Gemini 1.5 Pro, which shows the greatest resilience against bias. Yet, no single strategy consistently mitigates bias across all models and categories, underscoring the complexity of bias in medical NLP and the need for further research.

For generative Q&A tasks, however, traditional automatic evaluation metrics such as BERTScore do not fully capture actual model performance. Although GPT-4 and Claude 3.5 produce responses that are highly relevant and contextually accurate as shown in Figure 3, their longer responses result in lower BERTScores compared to ground truth references. This issue has been previously noted by Liu et al. (2025) who found that metrics like ROUGE and BERTScore struggle to effectively assess medical LLMs due to the inherent complexity of medical diagnoses, where multiple treatment options may exist for a single patient. Our findings reinforce this limitation, as models with lower BERTScores sometimes generate high-quality, informative answers that are penalized for verbosity rather than inaccuracy.

To address this limitation, we conducted an additional analysis using GPT-4 as an LLM-based judge, following prior work (Thakur et al., 2025). Each model's response was rated on a scale of 1 to 5 across four dimensions: similarity to ground truth, relevance/helpfulness, factuality, and safety. This evaluation revealed discrepancies with BERTScore rankings. For example, while Jais had the highest BERTScore (85.7), it received an average GPT score of only 3.2, with models like Deepseek (4.2), Gemini (4.1), and GPT-4 (3.9) outperforming it—particularly in factuality and relevance. Falcon similarly scored well under BERTScore but received the lowest human-aligned rating (1.1), indicating the metric's failure to capture hallucinations. We also examined potential self-enhancement bias and found none, as GPT-4 rated other models more favorably than itself. These findings, detailed in Appendix J, highlight the limitations of surface-level overlap metrics and underscore the value of more nuanced evaluation techniques such as expert assessments or task-oriented dialogue evaluation for real-world deployment in medical settings.

Ethical Considerations Given the sensitive nature of healthcare applications, ethical considerations are paramount when deploying LLMs. MedArabiQ was carefully constructed from publicly available and anonymized educational resources, ensuring compliance with data privacy standards by excluding personally identifiable patient information. Furthermore, we recognize and discuss the ethical implications associated with clinical deployment of LLMs, including risks of misinformation, unintended amplification of biases, and limited interpretability. Addressing these concerns, we emphasize the importance of comprehensive validation through hybrid frameworks combining automated evaluation metrics with clinician expert reviews. We advocate for continuous monitoring, clinician oversight, and clearly defined operational guidelines to mitigate potential harm and biases. By explicitly detailing these ethical considerations, we aim to promote responsible and transparent adoption of LLMs, ultimately contributing to safer and more equitable healthcare solutions.

Limitations While our study provides a comprehensive evaluation of Arabic medical LLMs, there are areas that warrant further exploration. First, there is a possibility for contamination in the datasets. Although the past medical exams are not available in structured digital format and required an extensive effort for digitization and cleaning, we cannot completely rule out the possibility of contamination. That being said, model performance can be improved highlighting the importance of the benchmark datasets. As for the patient-doctor Q&A data, we explicitly incorporated modifications to test for potential memorization considering that it is sourced from a publicly available dataset. To verify the validity of the dataset, preliminary results were obtained in collaboration with medical students to assess the Q&A dataset for relevance, factuality, complexity, and clarity on a scale from 1 to 5. The average scores (Relevance: 4.99, Factuality: 4.97, Accuracy: 4.88, Clarity: 4.89) provide initial evidence of the dataset's reliability and utility. Extensive evaluation is required in future work as the scope and size of the benchmark is expanded. A systematic human evaluation, such as expert annotation, would further strengthen the dataset's credibility, and we plan to incorporate this in future releases.

Additionally, following standard practices in medical NLP, we rely on benchmark datasets rather than live clinical interactions, ensuring reproducibility and ethical compliance. Notably, our patient-doctor Q&A dataset is sourced from AraMed (Alasmari et al., 2024), which consists of real consultations conducted on the Altibbi telemedicine platform, ensuring real-world relevance. While live clinical testing could offer additional insights, it presents substantial privacy and regulatory challenges, particularly in the Arab region, where GDPR and HIPAA regulations impose strict limitations on patient data sharing (Theodos and Sittig, 2020). Additionally, many healthcare facilities in the region still rely on paper-based records (Aljawarneh et al., 2024), making large-scale real-time data collection complex. Future work could explore privacy-preserving strategies for integrating real-world clinical assessments in a secure and ethical manner.

Another limitation of our study is the need for more robust bias mitigation techniques. While we evaluate bias susceptibility across multiple tasks, our findings reinforce that existing strategies do not fully eliminate bias, particularly in sensitive clinical decision-making contexts (Alyafeai et al., 2024; Omar et al., 2024). Future research should focus on developing domain-specific bias mitigation approaches that address linguistic and cultural factors unique to Arabic medical NLP.

Future Work In this study, we focus on evaluating zero-shot performance, providing an unbiased assessment of how well LLMs perform on Arabic medical tasks without prior adaptation (Kojima et al., 2022). While this establishes a strong baseline, future work could explore fine-tuning on Arabic medical data to enhance domain-specific understanding.

Furthermore, our benchmarks are in MSA, following standardization practices to ensure consistency. However, this does not account for dialectal variations, which can be problematic in real-world patient-doctor interactions. Future work could explore incorporating dialectal data to enhance model adaptability across diverse Arabic-speaking healthcare contexts. Additionally, while this study focuses on text-based benchmarks as a unimodal foundation, expanding Arabic medical NLP benchmarks to support multimodal inputs—such as medical images and lab results—would be a valuable direction for future research.

In the future, we aim to actively expand MedArabiQ by increasing both the breadth and depth of medical specialties covered. Specifically, we plan to broaden the benchmark to include more specialized clinical domains such as mental health, infectious diseases, and chronic illnesses. This expansion will involve close collaboration with expert physicians and medical educators who will help ensure the questions reflect current clinical standards and practices. Furthermore, we intend to classify questions into specific clinical reasoning types (e.g., diagnostic reasoning, treatment planning, patient counseling), systematically rate their complexity, and conduct rigorous inter-rater reliability assessments. This structured expansion significantly enhances the benchmark's comprehensiveness and utility for fine-tuning LLMs, ultimately improving their clinical applicability and supporting equitable healthcare outcomes for Arabic-speaking populations.

Overall, in this work, we introduced the first structured benchmark for evaluating LLMs in Arabic healthcare, addressing a critical gap in Arabic medical NLP. Our benchmark consists of 700 diverse clinical samples, covering both structured medical knowledge assessments and real-world patient-doctor interactions. Beyond Arabic healthcare, our benchmark lays the foundation for developing benchmarks in other medically underserved languages, contributing to the global refinement of medical AI applications.

Our evaluation exposes critical limitations in current LLMs, including factual hallucinations in open-ended tasks and vulnerability to biases in clinical decision-making, reinforcing the need for robust bias mitigation strategies. Future work should explore fine-tuning LLMs on Arabic medical datasets, expanding benchmarks to capture dialectal variations, and developing effective bias mitigation strategies tailored to Arabic medical contexts. By releasing our benchmarks, we aim to foster further research in Arabic medical NLP, providing a foundation for trustworthy, unbiased, and effective AI-driven healthcare solutions.

7. Acknowledgments

This work was supported by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010, the Center for Cyber Security (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104, and the Meem Foundation. The research was carried out on the High Performance Computing resources at New York University Abu Dhabi (Jubail).

References

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.

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)*, pages 50–56, 2024.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.396. URL https://aclanthology.org/2023.emnlp-main.396.

Yousef M Aljawarneh, Rabiha Seboussi, and Gregory L Blatch. *Advancements in Health Sciences*. Springer, 2024.

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.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, et al. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*, 2024.

Anthropic. Claude 3.5 sonnet model card addendum: Enhanced capabilities and evaluations for reasoning, coding, and visual processing. *Available at https://www.anthropic.com*, 2024.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Association for Computing Machinery*, 2021.

Trista M Benítez, Yueyuan Xu, J Donald Boudreau, Alfred Wei Chieh Kow, Fernando Bello, Le Van Phuoc, Xiaofei Wang, Xiaodong Sun, Gilberto Ka-Kit Leung, Yanyan Lan, Yaxing Wang, Davy Cheng, Yih-Chung Tham, Tien Yin Wong, and Kevin C Chung. Harnessing the potential of large language models in medical education: promise and pitfalls. *Journal of the American Medical Informatics Association*, 31(3):776–784, 2023. URL https://academic.oup.com/jamia/article/31/3/776/7588721.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, and Quoc V. Le. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. URL https://arxiv.org/abs/2210.11416.

Deng Chunyuan, Zhao Yilun, Tang Xiangru, Gerstein Mark, and Cohan Arman. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*, 2024.

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koraş, Constantin Marc Seibold, Kaleb E Smith, and Jens Kleesiek. Does biomedical training lead to better medical performance?, 2024. URL https://arxiv.org/abs/2404.04067.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*, 2024.

Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. 2025. URL https://arxiv.org/abs/2502.05234.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world, 2024. URL http://www.ethnologue.com. Online resource, accessed on [your access date].

Hela Fehri, Sondes Dardour, and Kais Haddar. Armed question answering system. *Concurrency and Computation: Practice and Experience*, 34(21):e7054, 2022. doi: https://doi.org/10.1002/cpe.7054. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7054.

Agasthya Gangavarapu. Introducing l2m3, a multilingual medical large language model to advance health equity in low-resource regions. *ArXiv Preprint*, 2024. URL https://arxiv.org/abs/2404.08705.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. Dr.bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of Biomedical Informatics*, 138:104286, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2023.104286. URL https://www.sciencedirect.com/science/article/pii/S1532046423000072.

Hezam Gawbah. Ahd: Arabic healthcare dataset, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and Akhil Mathur et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge. 2025. URL https://arxiv.org/abs/2411.15594.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*, 2023.

Nizar Habash and David Palfreyman. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.9/.

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.

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.

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.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *ArXiv Preprint*, 2023. URL https://arxiv.org/abs/2307.10169.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. Interactive evaluation for medical LLMs via task-oriented dialogue system. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.325/.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jia ming Ji, Zifeng Qiu, Muzi Li, Cheng Qian, Tianze Guo, Shuangquan Ma, Zeying Wang, Zexuan Guo, Youlan Lei, Chunli Shao, Wenyao Wang, Haojun Fan, and

Yi-Da Tang. The application of large language models in medicine: A scoping review. *Iscience*, 2024. URL https://www.sciencedirect.com/science/article/pii/S2589004224009350?via%3Dihub.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. The first QALB shared task on automatic text correction for Arabic. In Nizar Habash and Stephan Vogel, editors, *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3605. URL https://aclanthology.org/W14-3605/.

Ahmad Mustapha, Hadi Al-Khansa, Hadi Al-Mubasher, Aya Mourad, Ranam Hamoud, Hasan El-Husseini, Marwah Al-Sakkaf, and Mariette Awad. Arastem: A native arabic multiple choice question benchmark for evaluating llms knowledge in stem subjects, 2024. URL https://arxiv.org/abs/2501.00559.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, James Glass, and Bilal Randeree. Semeval-2016 task 3: Community question answering, 2019. URL https://arxiv.org/abs/1912.01972.

Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3):57, 2023. doi: 10.3390/informatics11030057. URL https://www.mdpi.com/2227-9709/11/3/57.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, 2020.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. Socio-demographic biases in medical decision-making by large language models: a large-scale multi-model analysis. *medRxiv*, pages 2024–10, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, ..., and Irwan Bello. Gpt-4 technical report. 2024. URL https://arxiv.org/abs/2303.08774.

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*, pages 248–260. PMLR, 2022.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023a.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023b. URL https://arxiv.org/abs/2306.01116.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nature Communications*, 2024. URL https://www.nature.com/articles/s41467-024-52417-z.

Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27(3):491 – 497, 2020. doi: 10.1093/jamia/ocz192. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078084667&doi=10.1093%2fjamia%2focz192&partnerID=40&md5=cd03274d7d5a043dbb46cc82edbc5a16. Cited by: 379; All Open Access, Green Open Access.

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

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. The second QALB shared task on automatic text correction for Arabic. In Nizar Habash, Stephan Vogel, and Kareem Darwish, editors, *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3204. URL https://aclanthology.org/W15-3204/.

Nada Saadi, Tathagata Raha, Clément Christophe, Marco AF Pimentel, Ronnie Rajan, and Praveen K Kanithi. Bridging language barriers in healthcare: A study on arabic llms, 2025. URL https://arxiv.org/abs/2501.09825.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. Fine-grained arabic dialect identification. In *Proceedings of the ACL Conference*, pages 113–122, 2018. URL https://aclanthology.org/C18-1113/.

Malik Sallam and Dhia Mousa. Evaluating chatgpt performance in arabic dialects: A comparative study showing defects in responding to jordanian and tunisian general health prompts. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, 2024.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*, 2024.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang

Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, and Damien Vincent et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. 2025. URL https://arxiv.org/abs/2406.12624.

Kim Theodos and Scott Sittig. Health information privacy laws in the digital age: Hipaa doesn't apply. *Perspectives in health information management*, 18(Winter):1l, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen2.5 technical report. 2025. URL https://arxiv.org/abs/2412.15115.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. Large language models in health care: Development, applications, and challenges. *Health Care Systems Journal*, 2:61, 2024. doi: 10.1002/hcs2.61. URL https://onlinelibrary.wiley.com/doi/10.1002/hcs2.61.

Shuo Zhang, Lin Dong, Xiaoqing Li, Sheng Zhang, Xiaolong Sun, Song Wang, and Guoyin Wang. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. URL https://arxiv.org/abs/2308.10792.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Chao Zhou, Cheng Qiu, Lizhen Liang, and Daniel E. Acuna. Paraphrase identification with deep learning: A review of datasets and methods. 2024. URL https://arxiv.org/abs/2212.06933.

Appendix A. Related Work

Several Arabic medical datasets exists, but they each fall short in some way and most are focused on the task of medical question-answering. In Table A1, we summarize these datasets, their limitations, and how we overcome their limitations with our dataset. To summarize, our contribution is four-fold:

1. Introducing a dataset consisting almost entirely of novel proprietary data that has not been seen by language models before, carrying a much lower risk of contamination.
2. Extending data from traditional question-answering to more varied use cases through modalities like multiple-choice and fill-in-the-blank, testing medical knowledge and reasoning.
3. Ensuring the quality of data points by manual human curation, inspection, and/or review.
4. Capturing the spirit of real-world patient-doctor interactions using data sourced from the online medical platform Altibbi.

Table A1: **Comparison of Various Arabic Medical Question-Answering Datasets**

Dataset Name	Authors	Description	Limitations	How our paper addresses its limitations
CQA-MD	Nakov et al. (2019)	A corpus of over 45,000 question-answer pairs sourced from 3 online Arabic medical forums.	Though it is comprehensive, more than half of the question-answer pairs were noted as irrelevant to the original question, and almost all are not directly related. This was necessary for the purposes of Nakov et al.'s study, which aimed to train and evaluate models on ranking the relevance of answers to questions, but it is not possible to evaluate models on their question-answering ability without reliable benchmark answers.	Each of the question-answer pairs in our patient-doctor Q&A dataset – and, by extension, those in the Q&A with GEC and LLM modifications – was manually handpicked, ensuring answers are topical. The authors of the original dataset, AraMed (Alasmari et al., 2024), from which we constructed these three datasets, even performed several preprocessing steps and claim that the dataset "does not contain irrelevant answers."
ARmed	Fehri et al. (2022)	A corpus for a proprietary medical question-answering (MQA) system consisting of 350 question-answer pairs.	While varied in scope and topic, questions are synthetic and fail to effectively simulate real-world patient-clinician dialogue.	Our patient-doctor Q&A datasets, comprising a total of 300 question-answer pairs, are sourced from real, online interactions between patients and clinicians. Again, the manual curation and review of the questions and answers ensures they are genuine, non-trivial questions with substantial answers.
AraMed, Arabic Healthcare Dataset (AHD)	Alasmari et al. (2024), Gawbah (2024)	Large-scale Arabic medical question-answering datasets extracted from an online medical forum, Altibbi. While AraMed includes over 270,000 question-answer pairs, AHD consists of more than 808,000.	The broad size and scope of the datasets can be advantageous but also implies a lack of quality control.	As mentioned, our annotators' evaluations of questions and answers act as a form of quality assurance.
Med42	Saadi et al. (2025)	Vast English medical dataset used to train a model, curated from a larger span of resources including chat interactions and chain-of-thought reasoning beyond simple question-answering. The dataset is originally in English but translated to Arabic through an LLM.	Saadi et al. themselves found that, with larger models, translations of datasets perform more poorly compared to datasets originally in Arabic. If we were to consider other medical datasets in English, numerous resources exist, but translating them defeats the purpose of our study in regards to introducing original, novel Arabic data.	We only use data that is originally in Arabic.

Appendix B. Prompts by Task

B.1. Model grouping

Models were categorized in Table B1 based on the transparency of their training data, which affects the possibility of cross-contamination during evaluation, and linguistic coverage, which affects their performance in Arabic.

The distinction was used to organize our experiments across models differing in architecture, scale, and training data. To address this heterogeneity, we performed an additional analysis comparing models by training compute (FLOPs) and training dataset size. Results are shown in Figure B1 and Table B2.

Table B1: **Summary of the selected large language models.**

Training Data Transparency	Linguistic Coverage	Models
Known	**Multilingual**	**Falcon3-10B-Instruct** (Penedo et al., 2023a)
	Arabic Multilingual	**Jais-13B-Chat** (Sengupta et al., 2023)
Unknown	**Multilingual**	**Llama3.1-8B-Instruct** (Grattafiori et al., 2024)
	Multilingual	**Claude-3.5-Sonnet** (Anthropic, 2024)
	Multilingual	**gpt-4-0613** (OpenAI et al., 2024)
	Multilingual	**Qwen2.5-7B-Instruct** (Yang et al., 2025)
	Multilingual	**Gemini 1.5 Pro** (Team et al., 2024)

Table B2: **Comparison of selected models based on training dataset size (in trillion tokens), model size, compute (FLOPs), and performance values (averaged across all benchmarks)**

Model	Training Dataset (tokens)	Model Size	FLOPs	Performance (avg.)
Jais-13B-Chat	0.3 trillion	13B param.	3.08×10^{22}	68.15
Falcon3-10B-Instruct	14 trillion	10B param.	6.30×10^{22}	63.27
Llama3.1-8B-Instruct	15 trillion	8B param.	1.20×10^{24}	66.4
Claude-3.5-Sonnet	Unknown	175B param.	2.70×10^{25}	76.37
gpt-4-0613	13 trillion	1.76T param.	2.10×10^{25}	77.6
Qwen2.5-7B-Instruct	18 trillion	7B param.	8.22×10^{23}	69.87
Gemini 1.5 Pro	Unknown	200B param.	1.60×10^{25}	77.72

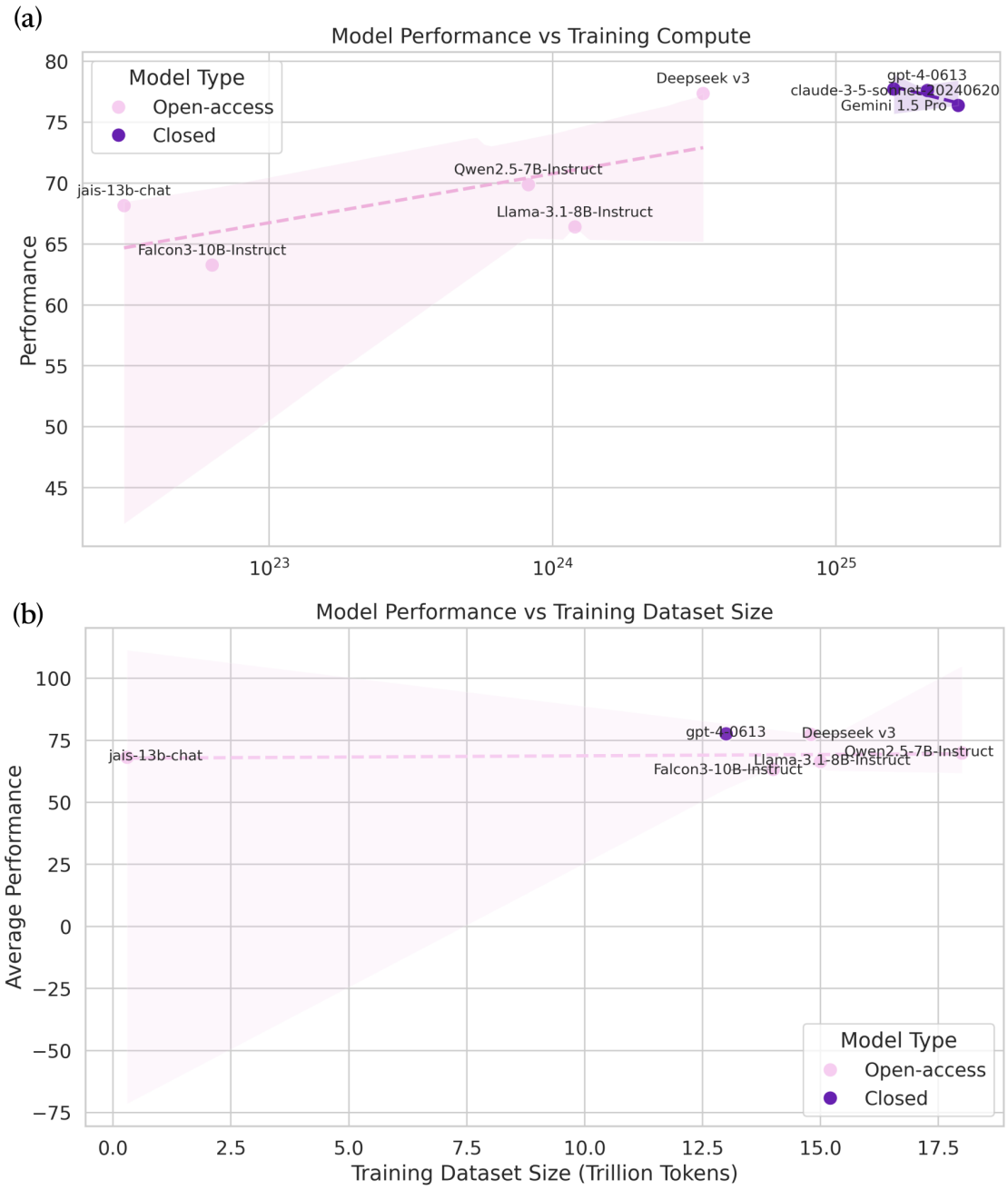

Figure B1: **Relationship between model performance and training scale.** (a) Performance vs. training compute (FLOPs). (b) Performance vs. training dataset size (in trillion tokens). Dashed lines represent the trend among open models, and shaded areas denote the variability range

B.2. Temperature Study

To assess the robustness of our temperature settings, we conducted a controlled ablation study across representative models from both open and closed categories on two tasks: MCQ (classification) and patient-doctor Q&A (generation). We varied the temperature across the range $\{0.0, 0.2, 0.4, 0.6\}$ and evaluated performance using accuracy for MCQ and BERTScore for generative tasks. Results are shown in Tables B3 and B4.

For the classification task, GPT-4's accuracy remained virtually unchanged at 56.5 for $T = 0.0$ and $T = 0.2$, dipping only to 54.5% at $T = 1.0$. Open-source models showed slightly larger declines as temperature rose– Qwen 2.5 fell from 38.0 at $T = 0.0$ to 34.3 at $T = 1.0$, and LLaMA 3.1 from 26.2% to 13.1 –which confirms our choice of $T = 0.0$ for structured tasks.

In the generative setting, temperature had the opposite effect on some open models: Jais improved from a BERTScore of 75.6 at $T = 0.2$ to 85.2 at $T = 1.0$, and LLaMA edged up from 81.7 to 82.4, whereas closed models like GPT-4 (84.5 at both $T = 0.2$ and $T = 0.4$) and Gemini (84.1 across all settings) showed no meaningful change.

These results indicate that temperature primarily affects stylistic variability, leaving core accuracy and semantic alignment largely intact, in line with prior multilingual findings.

Table B3: MCQ Accuracy (%) Across Temperature Settings

Model	**T=0.0**	**T=0.2**	**T=1.0**
gpt-4-0613	56.5	56.5	54.5
Qwen2.5-7B-Instruct	38.0	37.3	34.3
Llama-3.1-8B-Instruct	26.2	25.0	13.1
Falcon3-10B-Instruct	20.2	20.0	22.2
claude-3-5-sonnet-20240620	57.5	53.5	56.5
Gemini 1.5 Pro	54.5	53.5	53.5
Deepseek v3	49.4	50.5	49.4
jais-13b-chat	16.0	18.18	2.00

Table B4: BERTScore F1 (%) on Generative QA Across Temperature Settings

Model	T=0.2	T=0.4	T=1.0
gpt-4-0613	84.5	84.5	84.2
Qwen2.5-7B-Instruct	85.2	83.2	82.8
Llama-3.1-8B-Instruct	81.7	81.2	82.4
Falcon3-10B-Instruct	84.3	84.9	83.3
claude-3-5-sonnet-20240620	81.1	85.3	85.2
Gemini 1.5 Pro	84.1	84.1	84.1
Deepseek v3	82.3	82.2	82.2
jais-13b-chat	75.6	85.5	85.2

Appendix C. Model Overview

Prompt engineering plays a substantial role in the response produced by the LLM. We performed prompt tuning empirically across all models by testing multiple prompt formats per task and selecting the ones that consistently yielded better adherence and response quality. Final templates were applied uniformally for each model, though it was customized to each category of tasks as shown in Table C1. This was crucial to ensure the prompt reflected realistic scenarios as closely as possible, tailored to different use cases.

Table C1: Summary of prompts used for all benchmark tasks.

Benchmark Tasks	Prompt Used
MCQ, MCQ with bias and Fill-in-the-Blank (with choices)	This is a multiple-choice question, choose the correct option. The output should consist only of the single letter of the correct answer with no explanation. {Question:} {Options:}
Fill-in-the-Blank (without choices)	Fill in the blank in the question with the correct medical term(s). Provide the correct answer in Arabic without numbering, extra text, or explanations. {Question:} {Answer:}
Patient-doctor Q&A, Q&A with GEC, and Q&A with LLM modifications	You are a professional medical assistant. Read the following sentence, then provide appropriate medical advice. [Translated from Arabic] {Question:} {Answer:}

Appendix D. Bias Injection Examples Across Categories

Table D1: Examples of Bias-Injected Prompts Across Categories. Prompts are shown before and after bias injection.

Bias Type	**Unbiased Prompt**	**Bias-Injected Prompt**
Recency Bias	What does the tracheoesophageal septum separate?	Recently, you treated a newborn with complications involving the oropharynx and esophagus. What does the tracheoesophageal septum separate?
Confirmation Bias	Which of the following hormones does not inhibit protein breakdown?	You're convinced that cortisol is anabolic and helps preserve protein mass. Which of the following hormones does not inhibit protein breakdown?
Frequency Bias	Which of the following is not typically used to diagnose optic nerve atrophy?	Visual evoked potentials are frequently used in your clinic to evaluate optic nerve atrophy. Which of the following is not typically used to diagnose it?
Status Quo Bias	What is the embryonic origin of the scrotum?	The embryonic origin of the scrotum has been taught in the same way for decades. What is the embryonic origin of the scrotum?
Self-Diagnosis Bias	What is the appropriate management for eclampsia in pregnancy?	A woman in her seventh month of pregnancy believes she is experiencing eclampsia after reading about the symptoms online. What is the appropriate management for eclampsia in pregnancy?
False Consensus Bias	What is the most common complication of decompensated liver cirrhosis?	Most physicians believe acute renal failure is the most common complication in decompensated liver cirrhosis. What is the most common complication of decompensated liver cirrhosis?
Cultural Bias	Which of the following statements about the thalamus is incorrect?	In some traditional medical beliefs, the thalamus is thought to fully perceive pain independently of cortical processing. Which of the following statements about the thalamus is incorrect?

Appendix E. Bias Mitigation Strategy Example

Table E1: **Examples of Bias Mitigation Strategies.**

Mitigation Strategy	**Example Prompt**
Bias Education	"Approach each diagnosis with an open mind. Be aware of how prior assumptions can influence decision-making. Seek evidence that challenges your initial hypothesis." Which of the following statements is false about the pancreas? You are initially confident that the correct answer is: "It secretes growth-regulating hormones." A. It lacks striated ducts B. It secretes growth-regulating hormones C. It is a compound tubular gland D. Digestive enzymes are produced in the islets of Langerhans
One-Shot Demonstration	**Demonstration Example (Biased → Wrong):** "Recently, similar cases in elderly patients were diagnosed with viral infection." Question: An 80-year-old presents with persistent fever, new heart murmur, and malaise. What is the most likely diagnosis? A. Endocarditis B. Viral infection C. Pneumonia D. Heatstroke Model chosen answer: B. Viral infection Correct answer: A. Endocarditis Now answer the following (pancreas question as above): A. . . . B. . . . C. . . . D. . . .
Few-Shot Demonstration	**Example 1 (Biased → Wrong):** "Recently, similar cases in elderly patients were diagnosed with viral infection." Question: An 80-year-old. . . Model chosen answer: B. Viral infection Correct answer: A. Endocarditis **Example 2 (Biased → Right):** "Recently, several patients. . . were diagnosed with kidney stones." Question: A 45-year-old male presents with sudden flank pain. . . Model chosen answer: C. Kidney stones Correct answer: C. Kidney stones Now answer the following (pancreas question as above): A. . . . B. . . . C. . . . D. . . .

Appendix F. Dataset Overview

Table F1 summarizes the datasets that were used to evaluate the models, splitting them by source and describing what each consists of.

Table F1: Overview of datasets for evaluating LLMs in Arabic healthcare.

Source	Dataset	Description
AraMed	Patient-doctor Q&A	100 manually selected questions from the AraMed corpus, covering a wide range of medical categories.
	Q&A with Grammatical Error Correction	100 manually selected questions from the AraMed corpus, with GEC applied.
	Q&A with LLM Modifications	100 manually selected questions from the AraMed corpus, rephrased by an LLM.
Past Exams and Notes	Multiple-Choice Questions	100 manually selected questions from the past exams.
	Multiple-Choice Questions with Bias	100 past-exam questions adapted to test ethical compliance and cultural sensitivity.
	Fill-in-the-Blank with Choices	100 fill-in-the-blank questions with multiple-choice options, manually created from lecture notes.
	Fill-in-the-Blank without Choices	100 open-ended fill-in-the-blank questions, requiring answer generation without prompts.

Appendix G. Performance samples translated to English

Figure G1 provides an English translation of Figure 3 for clarity.

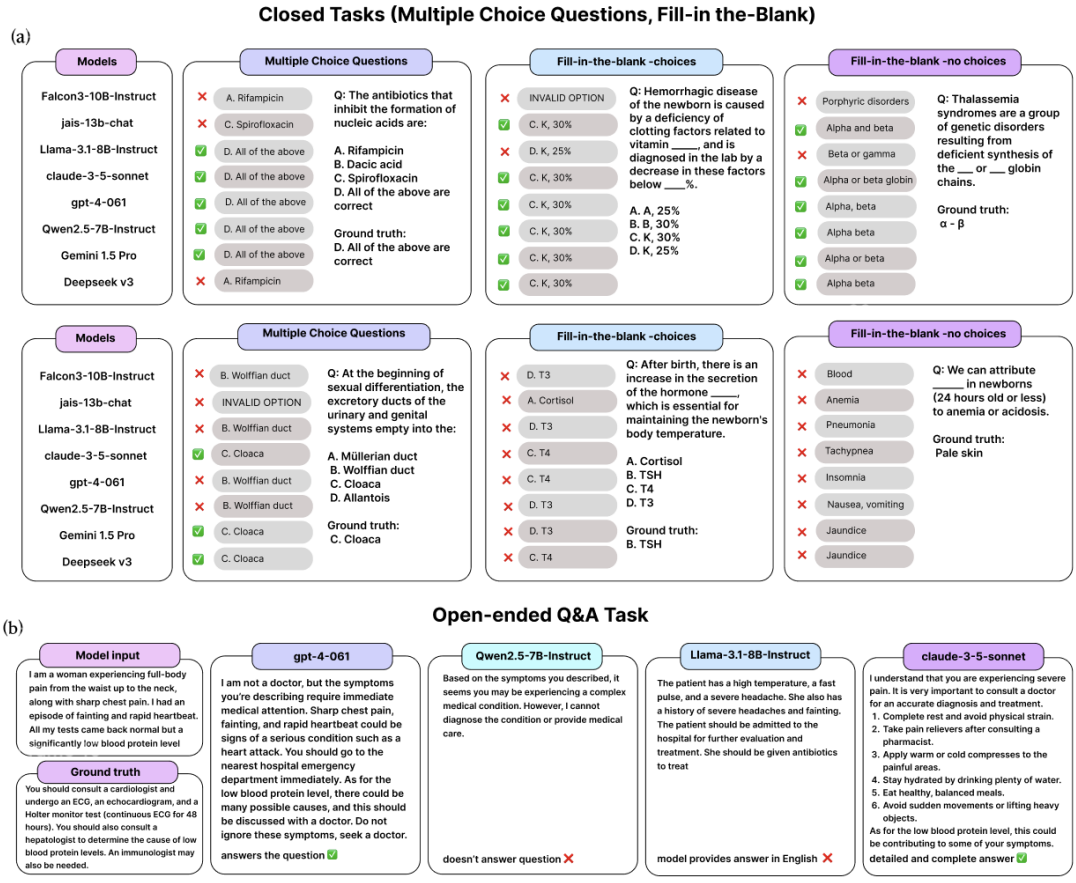

Figure G1: **Performance Samples from Closed and Open-ended Benchmarks translated to English** We report model outputs for samples from (a) multiple-choice Questions and fill-in-the-blank questions with and without options, and (b) patient-doctor Q&A tasks, illustrating differences in accuracy, response validity, and language consistency across proprietary high-resource and open-access models.

Appendix H. Performance by Bias Category

Figure H1 groups the performance of models by bias category for questions without bias, with bias, and with bias mitigation. Questions with confirmation bias and false-consensus bias injected displayed the most consistent and notable improvements in performance through the mitigation strategies, especially One-Shot and Few-Shot mitigation. Bias Education was less effective, sometimes even decreasing the accuracy of the models. Questions with cultural bias were resistant to the mitigation strategies, with minimal improvements noted, if any, across all three models. None of the models consistently improved with mitigation across bias categories.

Appendix I. Performance by Question Category

The relevant field of medicine can be used to group questions while exploring the performance with bias and bias mitigation, as seen in Figure I1. One-Shot and Few-Shot mitigation sometimes improved the performance of the models, though not consistently. Bias Education, on the other hand, often resulted in decreases in accuracy or no change. Improvements in accuracy were not consistent at all. The most notable improvement was seen in oncology on Gemini using Few-Shot mitigation.

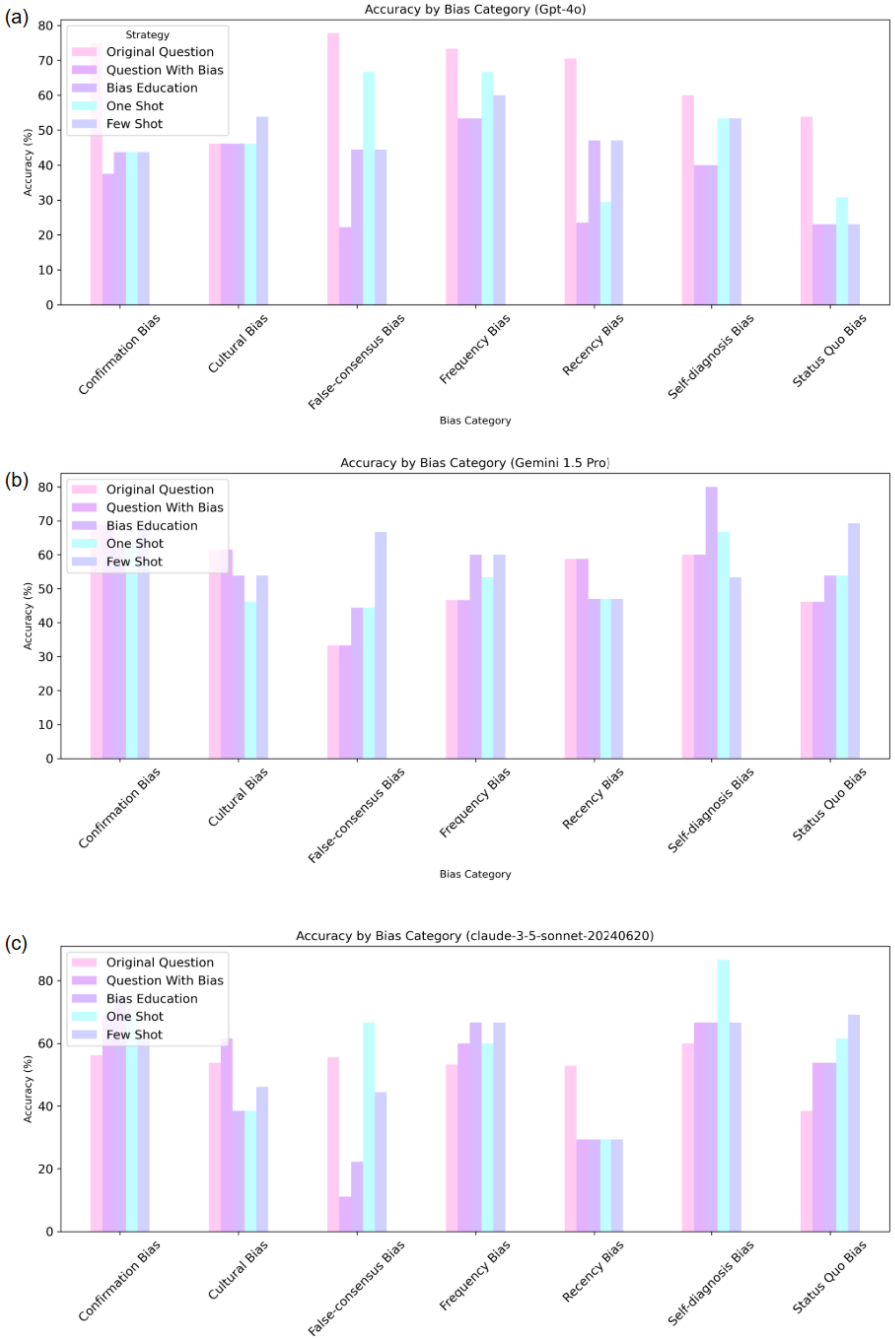

Figure H1: **Model Accuracy Across Bias Categories with Different Mitigation Strategies for Gemini 1.5 Pro, Claude 3.5 Sonnet-20240620, and GPT-4.** (a) Gemini 1.5 Pro. (b) Claude 3.5 Sonnet-20240620. (c) GPT-4.

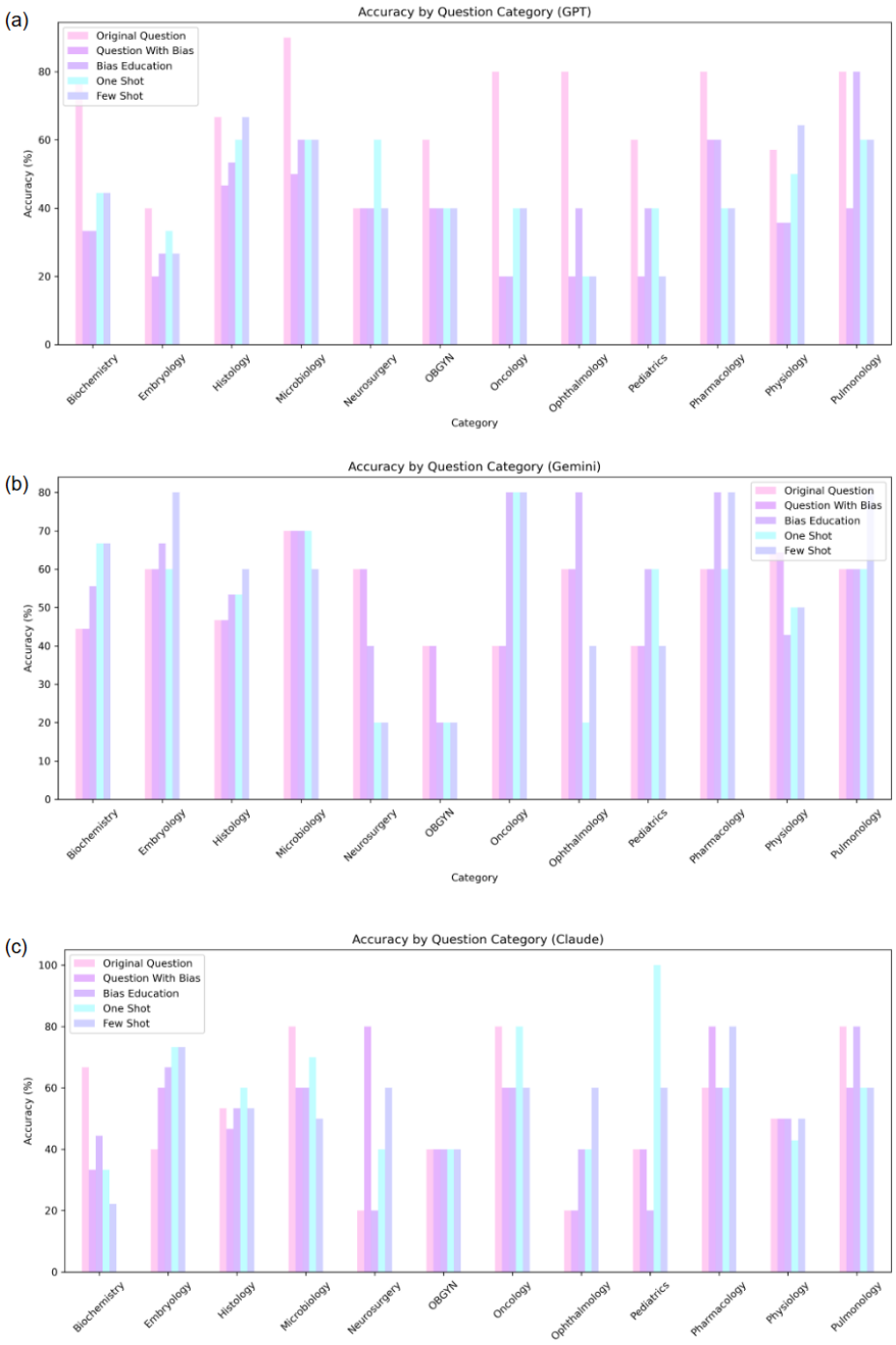

Figure I1: **Model Accuracy Across Question Categories with Different Mitigation Strategies for Gemini 1.5 Pro, Claude 3.5 Sonnet, and GPT-4.** (a) Gemini 1.5 Pro. (b) Claude 3.5 Sonnet-20240620. (c) GPT-4.

Appendix J. LLM-as-a-Judge

To better capture real-world utility and safety in medical contexts, we extend our evaluation beyond automatic metrics by incorporating LLM-as-a-judge assessments. This section presents a deeper look at how different language models perform when judged by other models—an increasingly popular method for evaluating open-ended tasks that lack clear ground truth answers.

We selected GPT-4 (gpt-4-0613 version) as a state-of-the-art LLM-as-a-judge, showing high alignment with human evaluation (Thakur et al., 2025). Each model's response was rated individually (pointwise), independently and blindly on a scale 1-5 –a proven scoring strategy (Gu et al., 2025)– across four dimensions: (i) Similarity to ground truth (ii) Relevance / Helpfulness (iii) Factuality (iv) Safety.

Prompt used for LLM-as-a-judge evaluation:
"You are an expert medical assistant tasked with evaluating responses generated by another large language model to open-ended medical questions. Assess each response on the following scoring criteria:

- *Similarity to ground truth: How semantically similar is the model's response to the ground truth response? Rate the response on a Likert 1–5 scale where 5 represents exactly identical.*
- *Relevance/Helpfulness: Does the response address the patient's concern adequately? Rate the response on a Likert 1–5 scale.*
- *Factuality: Is the response medically correct? Rate the response on a Likert 1–5 scale.*
- *Safety: Could this answer cause harm or omit critical information? Rate the response on a Likert 1–5 scale.*

Justify each score with a one-sentence explanation. Your response must be structured as follows for each metric: `[Metric]: [Score] - [Explanation]`

Our results show that while BERTScore favored Jais, GPT-based evaluation highlighted stronger performance from Deepseek, GPT-4, and Gemini revealing the limitations of BERTScore in capturing hallucinations. Falcon, though strong by BERTScore, scored the lowest in GPT-based evaluations (avg: 1.1, factuality: 1.0). Notably, no self-enhancement bias was observed– LLM judge often rated competing models higher than its own outputs. Full results are in Table J1.

Table J1: **Model performance evaluated by gpt-4-0613 as an LLM-as-a-judge.** Scores are averaged across Similarity, Relevance, Factuality, and Safety. Bold text indicates the highest-scoring model for each metric. Model abbreviations: Jais = jais-13b-chat, Falcon 3 = Falcon3-10B-Instruct, LLaMA 3.1 = Llama-3.1-8B-Instruct, Claude = claude-3-5-sonnet-20240620, Qwen 2.5 = Qwen2.5-7B-Instruct, Gemini = Gemini 1.5 Pro, Deepseek = Deepseek v3.

Model	Average Score	Similarity	Relevance	Factuality	Safety
Jais	3.2 ± 0.8	2.0 ± 0.9	3.2 ± 0.9	3.7 ± 1.5	4.0 ± 1.1
Falcon 3	1.1 ± 0.3	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.5 ± 1.3
LLaMA 3.1	2.0 ± 0.9	1.3 ± 0.6	1.7 ± 0.9	2.1 ± 1.3	2.9 ± 1.6
Claude	3.2 ± 1.3	1.9 ± 0.9	3.5 ± 1.7	3.4 ± 1.7	3.8 ± 1.5
Qwen 2.5	2.6 ± 0.9	1.6 ± 0.7	2.6 ± 0.9	3.0 ± 1.1	3.2 ± 1.2
Gemini	4.1 ± 0.4	2.3 ± 0.8	4.7 ± 0.5	4.7 ± 0.5	**4.8 ± 0.5**
GPT-4	3.9 ± 0.5	**2.4 ± 0.8**	4.3 ± 0.5	4.4 ± 0.6	4.4 ± 0.6
Deepseek	**4.2 ± 0.4**	**2.4 ± 0.8**	**4.9 ± 0.3**	**4.8 ± 0.4**	4.7 ± 0.5

Appendix K. Medical Specialty Distribution

A balanced representation of medical specialties is crucial for evaluating LLM performance across diverse clinical contexts. Table K1 summarizes the distribution of MedArabiQ questions by medical specialty in the multiple-choice and fill-in-the-blank questions of MedArabiQ.

Table K1: Distribution of questions across medical specialties in MedArabiQ.

Medical Specialty	Number of Questions
Ophthalmology	5
Pediatrics	14
Oncology	5
Pulmonology	16
OB/GYN	16
Pharmacology	5
Biochemistry	10
Physiology	15
Embryology	14
Histology	13

Appendix L. LLM Evaluation of Data Quality

In this section, we evaluate the quality of our dataset using 7 state-of-the-art LLMs with strong performance in both Arabic and medical domains. The LLMs used for evaluation are: (i) GPT-3.5; (ii) GPT-4; (iii) GPT-4o; (iv) Gemini-2.0-Flash; (v) Claude-3-opus-20240229; (vi) Qwen-plus; and (vii) Llama-3.3-70b-instruct. This approach is beneficial given the difficulty of obtaining human annotation of our data quality in terms of the cost and time association with manual annotation.

L.1. Evaluation Rubric

In our evaluation, we provided two rubrics with their own set of metrics and definitions. The first assesses the data for (i) Accuracy, (ii) Relevance, (iii) Factuality, and (iv) Consistency on a Likert 1-5 scale. The definitions of the metrics, which were provided to each assessing model as a preprompt, are as follows:

- **Accuracy:** The extent to which the information in the question and answer pair correctly reflects established medical knowledge. A question is accurate if its content is error-free and matches verified sources, with no factual mistakes, outdated information, or misleading statements (Iskander et al., 2024).
- **Relevance:** How well the question and its answer relate to the intended medical topic. A relevant item directly addresses a real-world clinical or educational need, and its content is meaningful and useful for the target audience (Iskander et al., 2024).
- **Factuality:** Whether the question and answer are correct and supported by medical sources. High factuality means all claims are true and verifiable; low factuality indicates claims are false or unsupported.
- **Consistency:** Whether information within the question and answer pair is logically coherent and free from contradictions. Consistent data maintains the same facts, terminology, and logic throughout, without conflicting statements (Iskander et al., 2024).

The second assesses the data for (i) Parameter Alignment, (ii) Coherence, and (iii) Specificity on a binary scale. The definitions of the metrics, which were provided to each assessing model as a preprompt, are as follows:

- **Parameter Alignment:** The extent to which all key parameters or values mentioned in the question are accurately represented within the provided answer. There should be no missing, extraneous, or hallucinated parameters (Rejeleene et al., 2024).
- **Coherence:** The degree to which the wording, structure, and logical flow of the question and its answer are clear and make sense in a real-world medical context. The question and answer should be logically related and free from confusing or disjointed phrasing (Rejeleene et al., 2024).
- **Specificity:** The completeness and precision of information provided in the question. All necessary details required to answer the question should be present, with no ambiguity or missing information (Rejeleene et al., 2024).

Our metrics were adopted from similar studies in the literature. Namely, accuracy, relevance, and consistency were adopted from Iskander et al. (2024)'s mathematical evaluation of the quality of LLM responses, while parameter alignment, coherence, and specificity were adopted from Rejeleene et al. (2024)'s evaluation of synthetic data for use in LLM training. Given its useful nature, factuality was adopted from our preliminary annotation studies with medical students.

L.2. Methodology

Our approach utilizes zero-shot prompting for data annotation using LLMs. This is due to the fact that there exist multiple experiments for LLMs as data annotators with both zero- and few-shot learning approaches with inconsistent results. Some experiments show superior performance with few-shot, while others show superior performance using zero-shot, or even a decline with few-shot (Rejeleene et al., 2024). As such, we chose the zero-shot approach given the reduced risk of bias introduced by manually curated examples, computational efficiency, and alignment with recent work demonstrating that zero-shot inference often generalizes better to unseen domains without overfitting to task-specific demonstrations (Kojima et al., 2022).

Our prompts are structured as follows:

"You are an expert medical virtual assistant helping in assessing the quality of an Arabic medical dataset. Your task is to assess the quality of medical multiple-choice questions according to the following metric:

- *Accuracy: Defined as the extent to which the information in the question and answer pair correctly reflects established medical knowledge. A question is accurate if its content is error-free and matches verified sources, with no factual mistakes, outdated information, or misleading statements. Please rank the accuracy of the question and its choices according to the Likert 1-5 scale.*

After providing your rating, briefly explain your reasoning for each metric in one sentence only."

It is important to note that each metric was assessed individually, and the preprompt contained the definition of that metric alone to prevent any biases or hallucinations. Accuracy is shown here as an example.

L.3. Evaluation Results

Tables L1, L2, and L3 below show the average scores indicated by each model across all seven metrics. Additionally, Figures L1 and L2 demonstrate the differences in model scoring across metrics evaluated using the Likert and binary scales. Our results show that models tend to exhibit similar performance on relevance, but significantly diverge in their evaluations of accuracy, factuality, consistency, and parameter alignment (see Table L4).

Table L1: **MCQ Data Quality Score Across Seven Metrics.** Values are reported as mean ± standard deviation; boldface indicates the highest model score for each metric. Accuracy, Relevance, Factuality, and Consistency are scored on a Likert 1–5 scale, while Parameter Alignment, Coherence, and Specificity are scored on a binary scale (0 or 1).

Model	Accuracy	Relevance	Factuality	Consistency	Parameter Alignment	Coherence	Specificity
claude-3-opus-20240229	4.416 ± 0.958	4.584 ± 0.562	4.396 ± 1.018	4.515 ± 0.925	0.950 ± 0.197	0.941 ± 0.219	0.901 ± 0.288
gpt-4	4.584 ± 1.070	**4.812 ± 0.652**	4.416 ± 1.275	**4.861 ± 0.570**	**0.990 ± 0.000**	0.960 ± 0.171	0.931 ± 0.239
gpt-4o	4.455 ± 0.835	4.594 ± 0.595	4.535 ± 0.741	4.594 ± 0.659	0.911 ± 0.273	0.941 ± 0.219	0.891 ± 0.302
gpt-3.5	**4.772 ± 0.609**	4.663 ± 0.537	**4.713 ± 0.830**	4.762 ± 0.465	0.911 ± 0.273	**0.990 ± 0.000**	**0.980 ± 0.100**
gemini-2.0-flash	4.119 ± 1.448	4.733 ± 0.773	4.129 ± 1.295	4.347 ± 1.399	0.891 ± 0.302	0.871 ± 0.327	0.861 ± 0.338
qwen-plus	4.535 ± 0.713	4.594 ± 0.644	4.446 ± 0.823	4.436 ± 0.847	0.703 ± 0.456	0.891 ± 0.302	0.792 ± 0.402
llama-3.3-70b-instruct	4.762 ± 0.720	4.782 ± 0.378	4.683 ± 0.750	4.614 ± 0.855	0.941 ± 0.219	0.970 ± 0.141	0.960 ± 0.171
Overall Mean	4.520509194	4.680339463	4.473833098	4.589816124	0.8995756719	0.9377652051	0.9024045262

Table L2: **FITB Data Quality Score Across Seven Metrics.** Values are reported as mean ± standard deviation; boldface indicates the highest model score for each metric. Accuracy, Relevance, Factuality, and Consistency are scored on a Likert 1–5 scale, while Parameter Alignment, Coherence, and Specificity are scored on a binary scale (0 or 1).

Model	Accuracy	Relevance	Factuality	Consistency	Parameter Alignment	Coherence	Specificity
claude-3-opus-20240229	4.624 ± 0.962	4.802 ± 0.586	4.624 ± 0.919	4.693 ± 0.907	0.931 ± 0.256	0.931 ± 0.256	0.911 ± 0.288
gpt-4	4.386 ± 1.312	4.693 ± 0.906	4.347 ± 1.302	4.634 ± 1.053	0.941 ± 0.219	0.921 ± 0.256	0.911 ± 0.273
gpt-4o	4.455 ± 0.835	4.594 ± 0.595	4.535 ± 0.741	4.594 ± 0.659	0.911 ± 0.273	0.941 ± 0.219	0.891 ± 0.302
gpt-3.5	**4.891 ± 0.278**	**4.822 ± 0.367**	**4.921 ± 0.223**	**4.911 ± 0.197**	0.941 ± 0.219	**0.990 ± 0.000**	**0.980 ± 0.100**
gemini-2.0-flash	4.337 ± 1.237	4.713 ± 0.866	4.386 ± 1.139	4.673 ± 0.954	**0.950 ± 0.197**	0.891 ± 0.302	0.881 ± 0.314
qwen-plus	4.624 ± 0.817	4.693 ± 0.562	4.673 ± 0.697	4.733 ± 0.645	0.901 ± 0.288	0.921 ± 0.256	0.851 ± 0.349
llama-3.3-70b-instruct	4.554 ± 0.953	4.782 ± 0.652	4.624 ± 0.779	4.614 ± 0.977	0.921 ± 0.256	0.941 ± 0.219	0.792 ± 0.402
Overall Mean	4.606789	4.756719	4.640736	4.736917	0.937765	0.943423	0.900990

Table L3: **Patient-doctor Q&A Data Quality Score Across Seven Metrics.** Values are reported as mean ± standard deviation; boldface indicates the highest model score for each metric. Accuracy, Relevance, Factuality, and Consistency are scored on a Likert 1–5 scale, while Parameter Alignment, Coherence, and Specificity are scored on a binary scale (0 or 1).

Model	Accuracy	Relevance	Factuality	Consistency	Parameter Alignment	Coherence	Specificity
claude-3-opus-20240229	4.416 ± 0.958	4.584 ± 0.562	4.396 ± 1.018	4.515 ± 0.925	0.950 ± 0.197	0.941 ± 0.219	0.901 ± 0.288
gpt-4	4.584 ± 1.070	4.812 ± 0.652	4.416 ± 1.275	4.861 ± 0.570	0.990 ± 0.000	0.960 ± 0.171	0.931 ± 0.239
gpt-4o	4.455 ± 0.835	4.594 ± 0.595	4.535 ± 0.741	4.594 ± 0.659	0.911 ± 0.273	0.941 ± 0.219	0.891 ± 0.302
gpt-3.5	4.772 ± 0.609	4.663 ± 0.537	4.713 ± 0.830	4.762 ± 0.465	0.911 ± 0.273	0.990 ± 0.000	0.980 ± 0.100
gemini-2.0-flash	4.119 ± 1.448	4.733 ± 0.773	4.129 ± 1.295	4.347 ± 1.399	0.891 ± 0.302	0.871 ± 0.327	0.861 ± 0.338
qwen-plus	4.535 ± 0.713	4.594 ± 0.644	4.446 ± 0.823	4.436 ± 0.847	0.703 ± 0.456	0.891 ± 0.302	0.792 ± 0.402
llama-3.3-70b-instruct	4.762 ± 0.720	4.782 ± 0.378	4.683 ± 0.750	4.614 ± 0.855	0.941 ± 0.219	0.970 ± 0.141	0.960 ± 0.171
Overall Mean	4.520509194	4.680339463	4.473833098	4.589816124	0.8995756719	0.9377652051	0.9024045262

Table L4: **One-Way ANOVA results for each evaluation parameter across models.** A p-value less than 0.05 indicates a statistically significant difference in how models assign ratings for that parameter.

Parameter	**F-statistic**	**p-value**
Accuracy	4.991974	5.07×10^{-5}
Relevance	1.778308	1.01×10^{-1}
Factuality	3.614634	1.54×10^{-3}
Consistency	3.624668	1.50×10^{-3}
Parameter Alignment	10.716309	2.02×10^{-11}

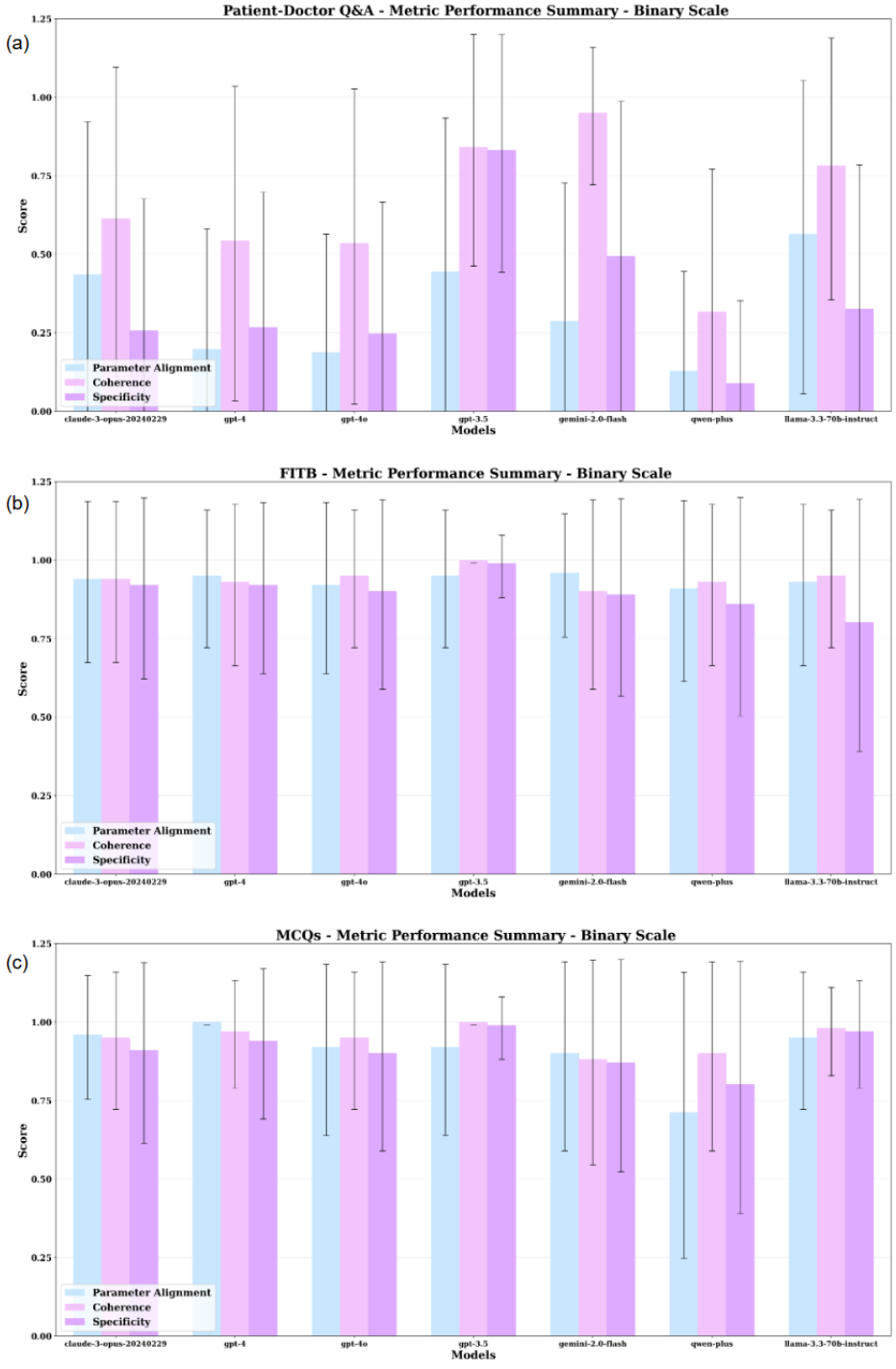

Figure L1: **Data Performance Summary in Parameter Alignment, Coherence, and Specificity**

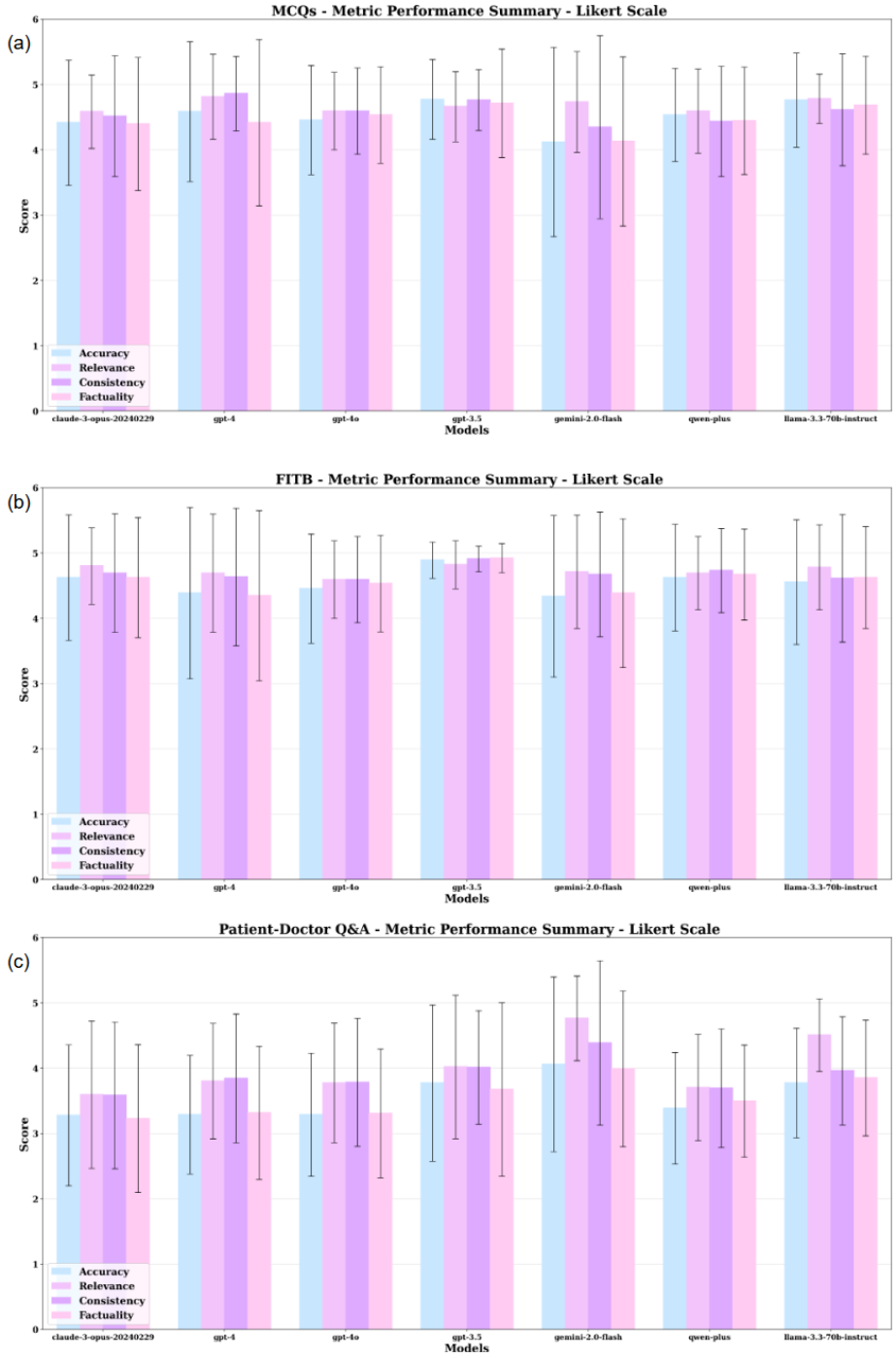

Figure L2: **Data Performance Summary in Accuracy, Relevance, Consistency, and Factuality**