Cross-Lingual Empirical Evaluation of Large Language Models for Arabic Medical Tasks

Chaimae Abouzahir, Congbo Ma, Nizar Habash, Farah E. Shamout

New York University Abu Dhabi
{ca2627,cm7196,nh48,fs999}@nyu.edu

Abstract

In recent years, Large Language Models (LLMs) have become widely used in medical applications, such as clinical decision support, medical education, and medical question answering. Yet, these models are often English-centric, limiting their robustness and reliability for linguistically diverse communities. Recent work has highlighted discrepancies in performance in low-resource languages for various medical tasks, but the underlying causes remain poorly understood. In this study, we conduct a cross-lingual empirical analysis of LLM performance on Arabic & English medical question and answering. Our findings reveal a persistent language-driven performance gap that intensifies with increasing task complexity. Tokenization analysis exposes structural fragmentation in Arabic medical text, while reliability analysis suggests that model-reported confidence and explanations exhibit limited correlation with correctness. Together, these findings underscore the need for language-aware design and evaluation strategies in LLMs for medical tasks.

1 Introduction

Large Language Models (LLMs) have shown remarkable performance on a wide range of medical tasks, including clinical question answering (Singhal et al., 2025), medical reasoning (Chen et al., 2025; Wu et al., 2025), and exam-style benchmarks (Pal et al., 2022), positioning them as as powerful capabilities for advancing healthcare applications. However, these successes are largely demonstrated in English due to limited availability of diverse benchmarks (Singh et al., 2025).

As LLMs move closer to real-world healthcare deployment, their ability to function reliably across languages becomes a critical concern. Recent multilingual evaluations consistently report substantial performance drops when medical LLMs are evaluated outside English, with Arabic as one of the affected languages (Alonso et al., 2024). However, reported results are typically limited to aggregate performance scores (Daoud et al., 2025), providing limited insight into the underlying causes of model underperformance.

Despite growing recognition of performance gaps between Arabic and English, existing explanations remain largely underexplored, often attributing failures to limited pretraining data or domain mismatch (Jin et al., 2024; Qiu et al., 2024). As a result, it remains unclear whether poor Arabic performance stems primarily from linguistic properties of the language, insufficient medical domain adaptation, architectural design choices, or interactions between these factors. This hinders principled adaptation: **without knowing which factors dominate model failure, it is difficult to design effective multilingual training strategies, alignment procedures, or evaluation protocols.** To this end, we present the first systematic study designed to disentangle linguistic, domain-specific, and architectural contributors to LLM performance on Arabic medical tasks. We make the following contributions:

- We design a cross-lingual diagnostic evaluation framework for general-purpose and medical LLMs that enables controlled analysis across languages, output formats, tokenization behavior, and reliability signals.
- We conduct an empirical study on MedAraBench, an Arabic medical question answering dataset, and its English-translated counterpart to isolate language effects while controlling for medical content.
- Our findings show that Arabic performance degradation is driven by interacting representational, alignment, and evaluation factors rather than medical knowledge alone, with gaps amplifying under increased task complexity and free-form generation.

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 158–171
March 28, 2026 ©2026 Association for Computational Linguistics

2 Related Work

2.1 Medical LLMs and Evaluation Benchmarks

LLMs have driven recent progress in clinical NLP, supporting applications including decision support, diagnostic assistance, and clinical text generation. To assess medical reasoning capabilities, several evaluation benchmarks, primarily formulated as question-answering tasks based on medical examinations or curated clinical sources, have been introduced (Zhang et al., 2018; Jin et al., 2019; Pal et al., 2022). Benchmarks such as MedQA and PubMedQA (Jin et al., 2021, 2019) are now widely used to evaluate medical knowledge and reasoning in LLMs.

General-purpose LLMs have achieved strong performance on English medical benchmarks. Notably, GPT-4 exceeded the passing threshold on USMLE-style questions in MedQA, achieving an accuracy of 86.1% (Nori et al., 2023a). This success motivated the development of medical-domain LLMs through domain-specific adaptation. Proprietary models such as Med-PaLM and Med-PaLM 2 (Singhal et al., 2025), as well as GPT-4 Med-Prompt (Nori et al., 2023b), reported substantial gains, with GPT-4 MedPrompt surpassing 90% accuracy on MedQA and achieving significant error reduction.

However, the costs, opacity, and privacy constraints associated with proprietary systems have limited their adoption in real-world clinical settings. In response, several open-access medical LLMs have been proposed, yet their performance on established benchmarks remains limited. Unlike proprietary models, BioMistral only achieves 44.4% accuracy on MedQA, while MedAlpaca and PMC-LLaMA attain 35.4% and 27.6%, respectively (Labrak et al., 2024). These results highlight a persistent performance gap between proprietary and open-source medical LLMs.

2.2 Multilingual Medical Benchmarks and Cross-lingual Generalization

Despite the widespread evaluation of LLMs on English medical benchmarks, their reliability across languages remains limited. Prior work has shown that both general-purpose and medical LLMs are prone to hallucinations (Xiong et al., 2024) and may produce answers based on outdated clinical knowledge (Vladika et al., 2025). Moreover, most medical benchmarks are predominantly English-centric in both their construction and evaluation (Qiu et al., 2024).

Recent multilingual evaluations consistently report substantial performance drops outside English. For instance, significant degradation has been observed on Italian medical QA tasks (Kembu et al., 2025), as well as across a broader range of non-English healthcare queries (Jin et al., 2024). Alonso et al. (2024) further show that both general-purpose and medical LLMs perform markedly worse in Arabic and Hindi than in English. Notably, medical LLMs often underperform the base models from which they are adapted in non-English settings, suggesting that domain adaptation may reduce cross-lingual generalization.

Alonso et al. (2024) further show that medical LLMs often underperform their base models in non-English settings such as Arabic and Hindi, suggesting that domain adaptation may hinder cross-lingual generalization. Complementarily, Jeong et al. (2024) demonstrate that this effect already occurs in English, indicating that specialization alone does not guarantee performance gains even without language mismatch.

To mitigate these disparities, some efforts have focused on developing language-specific medical models. HuatuoGPT (Zhang et al., 2023) is a notable example of a Chinese medical LLM trained on native-language biomedical resources. However, systematic analyses of how domain adaptation interacts with multilingual performance remain limited, particularly for underrepresented languages.

2.3 Challenges in Arabic Medical Language Models

Arabic poses distinct challenges for medical language modeling, including rich morphology, complex tokenization, dialectal variation, and a scarcity of high-quality, domain-specific resources (Farghaly and Shaalan, 2009; Habash, 2010). Although general-purpose Arabic LLMs such as Jais (Sengupta et al., 2023), Fanar (Team et al., 2025) and ALLAM (Bari et al., 2025) have been introduced, the development and evaluation of Arabic medical LLMs remain underexplored.

Existing evaluations report poor performance on Arabic medical tasks (Daoud et al., 2025). However, the underlying causes of these failures are not well understood. It remains unclear whether performance degradation primarily arises from linguistic representation issues, limitations of domain-adaptive training, or their interaction, particularly for medical LLMs adapted from English-centric base models. Moreover, the lack of publicly available Arabic medical benchmarks limits systematic diagnostic analyses comparable to those in English (Alasmari et al., 2024), motivating our investigation into the mechanisms underlying Arabic medical LLM failures beyond aggregate performance.

3 Methodology

To investigate sources of performance degradation in Arabic medical MCQs, we design a targeted evaluation framework probing LLM behavior on several aspects. Rather than introducing a new model, we focus on a set of research questions, which we detail below.

3.1 Research Questions

We compare model accuracy on original Arabic questions and their English-translated counterparts to isolate the role of linguistic representation from medical reasoning.

RQ1: To what extent is performance degradation driven by language rather than medical reasoning? We compare model accuracy on original Arabic questions and their English-translated counterparts to isolate the role of linguistic representation from medical reasoning.

RQ2: How do question-level properties affect model performance? We analyze accuracy as a function of input length, question difficulty, and medical specialty to determine whether linguistic complexity, cognitive demand, or domain-specific content disproportionately affects model outcomes.

RQ3: How do alignment constraints and output formats influence model behavior across languages? We compare soft matching (letter-based option selection) and hard matching (exact answer text generation) to evaluate how instruction following and surface-form generation influence accuracy across languages.

RQ4: Does tokenization behavior contribute to Arabic performance gaps? We examine tokenizer efficiency and fragmentation patterns to understand whether Arabic morphology and segmentation lead to less effective input representations.

RQ5: Are model confidence estimates and generated explanations reliable indicators of correctness? We analyze model-reported confidence and accompanying rationales to assess whether they correlate with accuracy and can be used to diagnose systematic failure modes.

3.2 Dataset

All experiments are conducted on MedAraBench (Abu-Daoud et al., 2026), an Arabic medical question answering benchmark. The questions are originally authored in Modern Standard Arabic (MSA), collected from medical exams, digitized from scanned paper sources, and manually curated to exclude incomplete or ambiguous items. Each question is annotated with the number of answer options (4–6), a difficulty level corresponding to years of medical study (Y1–Y5), and a medical specialty. The dataset covers 19 specialties (e.g., Anatomy, Pathology, Surgery, Pharmacology) and is split into training and test sets using an 80/20 split with matched specialty distributions (19,894 train / 4,989 test). A data sample is shown in Appendix A1.

Models are evaluated on a medical MCQ task in both Arabic and English. English versions are obtained via automatic translation of the original Arabic questions using the Google Translate API and are used solely for controlled cross-lingual analysis. Models are evaluated using accuracy, with a prediction counted as correct if the selected option matches the gold label.

3.3 Evaluated Models

We evaluate several recent open-source large language models as baselines for Arabic medical MCQs. We include recent, large-scale general-purpose LLMs like DeepSeek-V3.2 and LLaMA 3.3 70B as representative contemporary baselines in our evaluation. To examine differences between general language ability and domain-specific modeling, we compare two categories of models:

- **General-purpose LLMs**: DeepSeek-V3.2 (DeepSeek-AI et al., 2025), LLaMA 3.3 70B (Meta AI, 2024), and Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025), all trained on broad multilingual or mixed-domain corpora.

- **Medical-domain LLMs**: Meditron 3 70B (OpenMeditron Initiative, 2024), Med42-70B (Christophe et al., 2024), and MedGemma-27B-text-it (Sellergren et al., 2025), which incorporate domain-adaptive pretraining or fine-tuning on medical data. None explicitly report multilingual medical pretraining, and available documentation indicates predominantly English medical data.

Moreover, our evaluation focuses exclusively on open-source models for both methodological and practical reasons. From a methodological standpoint, open-source models offer full access to embeddings, tokenizers, and intermediate representations, which are essential for both our analysis and our planned cross-lingual adaptation method. For practical concerns, deploying black-box proprietary systems in medical settings poses significant privacy and auditability concerns, underscoring the need for transparent, open-source alternatives.

3.4 Experimental Settings

All models were evaluated using a unified multiple-choice prompting setup implemented via the HuggingFace Transformers API. Inference used greedy decoding (temperature = 0, no sampling, top-p = 1.0, top-k disabled). We fix task-specific maximum generation lengths across languages to ensure comparable inference conditions across models, allowing up to 4 tokens for letter matching, 15 for text matching, and 70 for explanation generation. These limits were chosen to accommodate the required output formats rather than language-specific tokenization characteristics.

The system prompts used follow standardized MCQ templates as shown in Appendix B and are provided in English for all inputs, including Arabic inputs. This is based on preliminary prompt-engineering experiments showing more stable and higher-performing outputs than Arabic prompts. This choice reflects the English-centric instruction-following capabilities of the evaluated models and results in mixed-language inputs for Arabic evaluations.

Due to hardware constraints, all 70B-parameter models were evaluated using 4-bit NF4 quantization with bfloat16 compute (BitsAndBytes) across two 32 GB V100 GPUs. Smaller models, including Medgemma-27B-text-it and Mistral-Small-3.2-24B-Instruct-2506, were evaluated without quantization in full bfloat16 precision on the same hardware. DeepSeek-V3.2 could not be evaluated locally due to its size and was instead accessed via the official DeepSeek API. For cross-lingual analysis, models were additionally evaluated on an English-translated version of the dataset. This setup enables direct comparison between Arabic and English under identical task structures, isolating the effect of language from content.

Models	**Acc (Ar)**	**Acc (En)**	**Δ (En–Ar)**
General-purpose LLMs			
DeepSeek-V3.2	**62.39**	**62.85**	**0.46**
Llama 3.3 70B	42.10	57.61	15.51
Mistral-Small-3.2-24B	50.25	57.75	7.5
Medical-domain LLMs			
Meditron 3 70B	50.51	58.80	8.92
Med42-70B	33.59	53.21	19.62
medgemma-27b-text-it	49.22	52.30	3.08

Table 1: **Results of general-purpose and medical-domain LLMs' Acc(uracy) on the Arabic (Ar) and English (En) datasets.** Bold values indicate the highest accuracy within each column. Mistral-Small-3.2-24B refers to Mistral-Small-3.2-24B-Instruct-250.

4 Empirical Studies and Analyses

4.1 Assessing the Role of Language in LLM Performance

We compare model accuracy on parallel English and Arabic medical benchmarks using a controlled prompting setup (Appendix B2) to isolate the effect of linguistic representation from medical reasoning. As shown in Table 1, accuracy is consistently lower in Arabic than in English across nearly all evaluated models, indicating a systematic language-associated performance gap. DeepSeek-V3.2 is the only model that exhibits comparable performance across languages. Notably, this behavior is not observed uniformly in larger models, indicating that reduced cross-lingual degradation cannot be attributed to model size alone.

For models with comparable parameter sizes ($\leq$ 70B), English consistently outperforms Arabic, indicating that language remains a key factor even under similar capacity constraints. This trend holds for both general-purpose and medical-domain models, suggesting that domain specialization alone does not resolve Arabic performance gaps.

4.2 Effects of Question Length, Difficulty and Medical Specialty

We investigate whether question-level characteristics influence model accuracy by analyzing performance trends with respect to input length, educational difficulty, and medical specialty. We focus our analysis on the best- and worst-performing models overall, DeepSeek-V3.2 and Med42-70B, respectively, and restrict the following experiments to these two models.

Figures 1 (a,b) show accuracy trends as a function of question length for DeepSeek-V3.2 and Med42-70B. Accuracy is relatively stable for

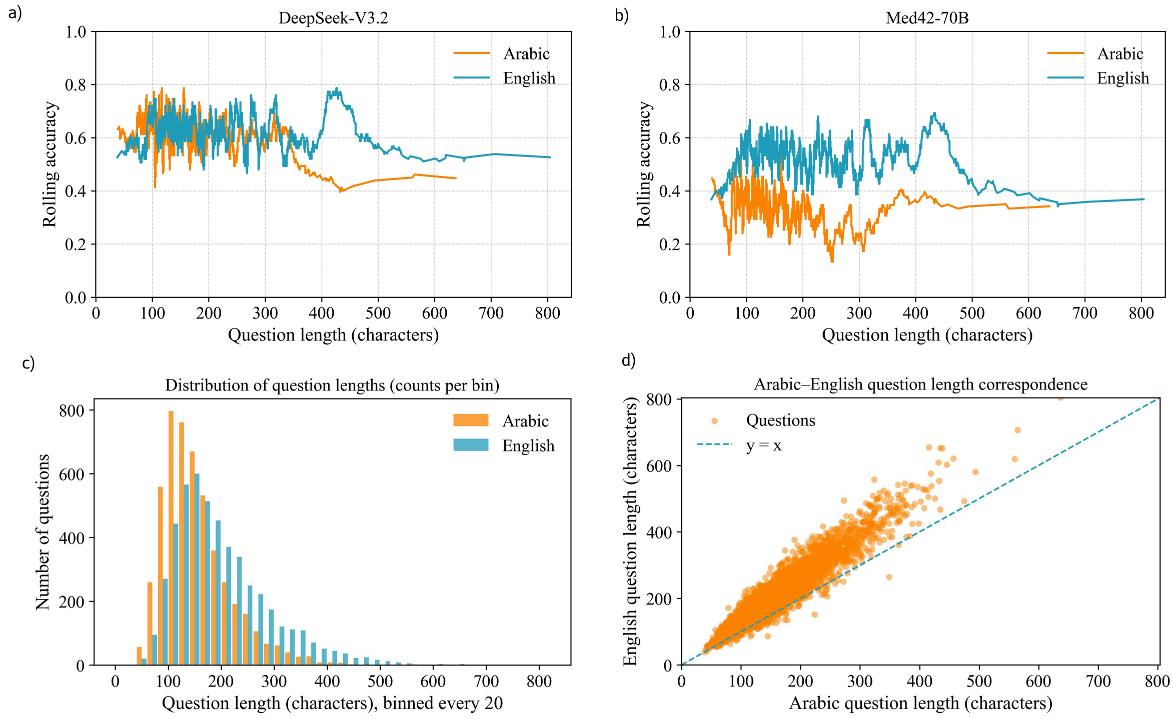


Figure 1: **Effect of question length on accuracy across Arabic and English.** (a–b) Rolling accuracy versus question length for DeepSeek-V3.2 and Med42-70B, respectively. (c) Distribution of question lengths in both languages. (d) Arabic–English length correspondence for aligned question pairs.

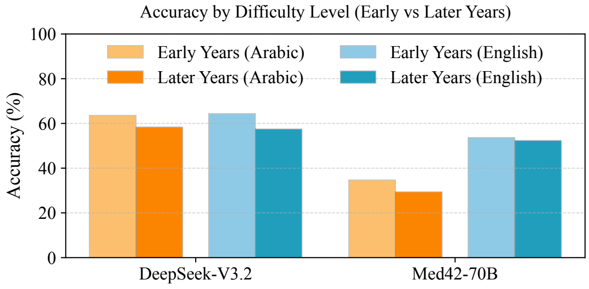


Figure 2: **Accuracy by educational difficulty level (early vs. later years) for DeepSeek-V3.2 and Med42-70B on Arabic and English medical MCQs.**

shorter inputs but degrades as question length increases for Arabic, while English performance remains comparatively stable at longer lengths. Question lengths are strongly correlated across paired Arabic–English items and exhibit overlapping distributions (Figures 1 c,d), indicating that the observed degradation reflects increased sensitivity to input length rather than artifacts of translation or systematic length mismatches.

Figure 2 reports accuracy by educational difficulty level. For both models and languages, accuracy decreases for later years' questions (Y3+) compared to early years' questions (Y1–Y2). The performance drop is consistently larger for Arabic, particularly for Med42-70B.

Figure 3 shows accuracy by medical specialty, revealing substantial variation across domains: performance is higher in clinically oriented fields (e.g., Emergency Medicine, Internal Medicine) and lower in foundational or detail-intensive specialties such as Microbiology and Embryology. English consistently outperforms Arabic across most specialties. This gap is particularly pronounced for Med42-70B, where Arabic performance lags behind English across nearly all specialties, suggesting that language-related performance disparities persist even when controlling for domain.

Overall, these results indicate that input length, difficulty, and domain content systematically affect model performance, and that these effects disproportionately impact Arabic compared to English.

4.3 Alignment Behavior Analysis

To analyze how output format influences model behavior across languages, we evaluate model performance under free-form answer generation using the prompt in Appendix B3. We use token-level

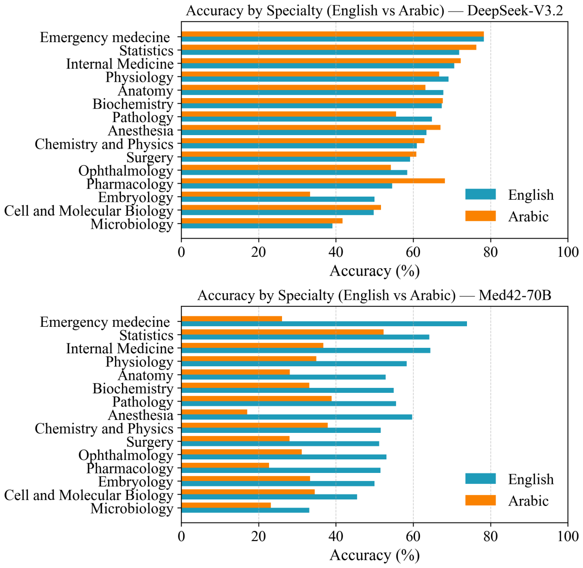


Figure 3: **Accuracy by medical specialty for DeepSeek-V3.2 (top) and Med42-70B (bottom) on Arabic and English medical MCQs.**

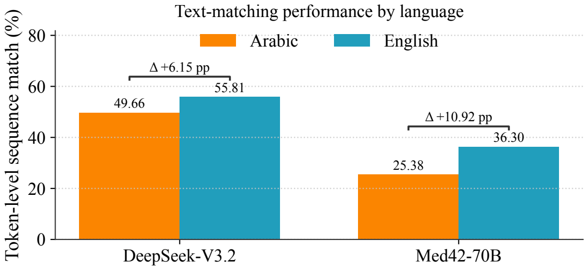


Figure 4: **Token-level sequence-match accuracy (%) for text-matching evaluation in Arabic and English across two models.**

sequence similarity between predicted and ground-truth answer texts computed with the Sequence-Matcher algorithm (Munk and Feitelson, 2022). Figure 4 shows that across both models, surface-form similarity is consistently lower for Arabic than for English, with a larger gap of 10.92 percentage points for Med42-70B compared to 6.15 pp for DeepSeek-V3.2. These discrepancies indicate that, when models are required to generate answer text explicitly, Arabic outputs diverge more substantially from reference answers at the surface-form level. The magnitude of these gaps is also larger than that observed under letter-based option selection reported in Table 1, suggesting that free-form generation amplifies language-specific difficulties beyond those captured by standard MCQ accuracy.

Tokenizer	**Tok/Word**	**Char/Tok**	**Single Char**
Model-native tokenizers			
DeepSeek-V3.2	2.39	2.33	32%
Llama 3.3 70B	2.42	2.27	32%
Meditron 3 70B	2.45	2.27	32%
Mistral-Small-3.2-24B	2.06	2.72	0%
Med42-70B	2.42	2.27	32%
Multilingual-efficient tokenizer			
Gemma-3-4B-it	2.30	2.43	35%
Arabic-focused tokenizer			
CAMeLBERT-MSA	1.76	3.21	36%

(a) Arabic dataset.

Tokenizer	**Tok/Word**	**Char/Tok**	**Single Char**
Model-native tokenizers			
DeepSeek-V3.2	1.52	4.01	28%
Llama 3.3 70B	1.60	3.82	27%
Meditron 3 70B	1.60	3.82	27%
Mistral-Small-3.2-24B	1.56	3.93	0%
Med42-70B	1.60	3.82	27%
Multilingual-efficient tokenizer			
Gemma-3-4B-it	1.57	3.90	35%
Arabic-focused tokenizer			
CAMeLBERT-MSA	2.82	2.13	44%

(b) English dataset.

Table 2: **Tokenization fragmentation statistics for Arabic and English inputs**. We report average tokens per word (subword splitting), average characters per token (compactness), and single-character tokens (reflecting extreme fragmentation).

4.4 Tokenization Efficiency Analysis

We analyze tokenization efficiency and fragmentation to assess whether language-specific tokenization patterns are associated with downstream performance gaps. We report average tokens per word, characters per token, and the proportion of single-character tokens in Table 2. For Arabic, word counts are computed using a linguistically informed tokenizer from CAMeL Tools (Obeid et al., 2020), while English uses whitespace-based word segmentation. Higher tokens per word and single character rates, together with lower characters per token, indicate more fragmented representations.

Across model-native tokenizers, Arabic is consistently more fragmented than English, with approximately 2.4 tokens per word compared to 1.5–1.6 for English. Similar trends hold for the multilingual tokenizer. This increased fragmentation leads to higher token usage for Arabic inputs, which may plausibly contribute to sharper performance degradation as input length and complexity

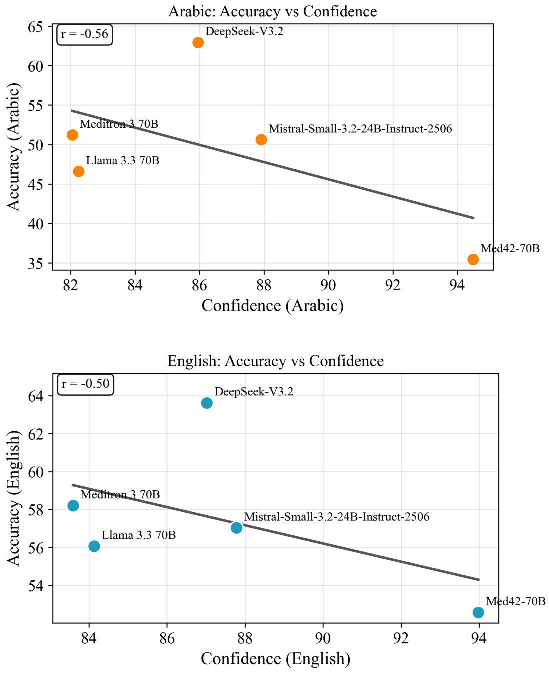


Figure 5: **Relationship between model-reported confidence and accuracy for Arabic (top) and English (bottom) medical MCQ.**

increase. However, this effect is not uniform across models: despite higher token counts for Arabic, some models (e.g., DeepSeek-V3.2) exhibit more stable performance, suggesting that tokenization alone does not fully explain the observed degradation. In contrast, the Arabic-focused CAMeLBERT tokenizer (Inoue et al., 2021) substantially reduces fragmentation for Arabic while increasing fragmentation for English, illustrating that tokenizer efficiency is language-dependent.

4.5 The Role of Confidence Estimates and Explanations

We analyze the relationship between model-reported confidence (prompt in Appendix B4) and accuracy in medical question answering (Figure 5). We observe a moderate negative correlation between confidence and accuracy in both Arabic ($r = -0.56$) and English ($r = -0.50$), indicating that higher confidence predictions are, on average, less accurate. This pattern is consistent across model families and languages, suggesting a general miscalibration of confidence in medical settings. Accordingly, model-reported confidence should not be treated as a reliable proxy for correctness. MedGemma-27B-text-it is excluded due to repeated noncompliance with the required confidence-reporting format under zero-shot prompting, often producing free-form text without a valid answer label. As this reflects instruction-following issues rather than task performance, we omit it from this analysis.

Model	Base^{Ar}	Base^{En}	Exp^{Ar}	Exp^{En}
DeepSeek-V3.2	**62.39**	**62.85**	**63.56**	45.85
Llama-3.3-70B	42.10	57.61	46.27	57.91
Mistral-Small-3.2-24B	50.25	57.75	49.50	**57.99**
Meditron-3-70B	51.05	58.45	28.65	33.65
Med42-70B	33.59	53.21	28.99	33.41

Table 3: **Accuracy with Explanation prompting (Exp) compared to the baseline (Base, no explanation prompt), for Arabic and English**. Bold values denote the highest accuracy for each column.

To examine whether explicit reasoning improves performance, we prompt models to generate a natural language explanation before answer selection, following a chain-of-thought–style prompting strategy (Wei et al., 2022) using the prompt shown in Appendix B5. MedGemma-27B-text-it is excluded for severe instruction non-compliance, consistent with the confidence-based analysis. As shown in Table 3, explanation prompting yields mixed and often detrimental effects across models. While some models exhibit modest gains in Arabic accuracy (e.g., DeepSeek-V3.2: 62.39 → 63.56), explanation prompting often leads to degraded performance in English and, for several models, substantial drops in both languages (e.g., Meditron-3-70B and Med42-70B).

Upon qualitative inspection, we find that in many cases, the model produces medically plausible or partially correct explanations while selecting an incorrect option label. Representative examples illustrating these reasoning–label misalignments are provided in Appendix C6. Requiring explanations appears to encourage verbose reasoning and reinterpretation, which can decouple reasoning quality from discrete multiple-choice selection. While explanation-conditioned prompting reduces accuracy for some models, this setting also surfaces cases where letter-only evaluation may overestimate performance by rewarding option matching despite stem–option inconsistencies. In such cases, explanation prompting exposes reasoning–label misalignment rather than pure knowledge errors.

5 Discussion

5.1 Language as a Source of Degradation

Across nearly all evaluated models, performance on Arabic medical MCQs is consistently lower than on their English-translated counterparts, despite identical medical content, indicating a persistent language-related performance gap beyond medical knowledge alone. DeepSeek-V3.2 is a notable exception, achieving near-parity between Arabic and English, demonstrating that strong cross-lingual performance in Arabic medical QA is achievable in open models.

This robustness cannot be explained by model scale alone, as public estimates suggest that DeepSeek-V3.2 and LLaMA 3.3 70B are trained at comparable orders of magnitude in compute and training tokens (Epoch AI, 2024), yet only DeepSeek-V3.2 maintains high Arabic performance. This suggests that language robustness depends on specific design and training choices beyond scale, including data curation, language balance, and post-training procedures.

5.2 Interaction between Task Complexity and Language

Performance across all evaluated models is systematically higher for shorter questions and for lower-difficulty (early-year) items, regardless of language. However, the rate of performance degradation differs substantially between Arabic and English. As question length increases and as questions progress from early to later years' material, accuracy declines more sharply for Arabic than for English, as shown in Figures 1 (a-b) and 2.

The relatively strong performance on short and early-year Arabic questions indicates that models can successfully answer simpler medical queries in Arabic, meaning that basic medical knowledge is present. The decline observed for longer and more advanced questions points to a reduced robustness of Arabic representations as task complexity increases, rather than a lack of medical understanding. A similar pattern emerges across medical specialties in Figure 3, where the gap between Arabic and English is larger in specialties that involve finer-grained distinctions, reinforcing the interaction between language effects and task complexity.

5.3 Alignment Constraints and Evaluation Sensitivity

The alignment analysis shows that output format influences how language-specific performance differences manifest. Because letter-based MCQ accuracy and token-level text similarity measure distinct aspects of model behavior, we restrict our analysis to within-format comparisons between Arabic and English. Under free-form answer generation, models consistently achieve lower token-level similarity in Arabic than in English, yielding larger cross-lingual gaps than those observed with constrained option selection. This suggests that removing output constraints introduces additional language-dependent variability not reflected by letter-based evaluation. In particular, free-form generation places greater demands on lexical choice and morphological realization, which are more challenging in Arabic. Since token-level similarity measures surface-form overlap, lower scores primarily reflect increased variation in answer expression rather than incorrect medical reasoning.

5.4 Tokenization as a Structural Constraint

Compared to English, Arabic medical text is consistently more fragmented under model-native tokenizers, with words split into more subword units and a higher prevalence of single-character tokens. This reflects a mismatch between subword tokenizers, optimized for frequent training forms, and Arabic medical vocabulary, which combines rich morphology with low-frequency, variable domain-specific terms, leading to unstable subword representations and finer-grained splits.

Multilingual tokenizers partially mitigate this effect by covering broader lexical distributions, while Arabic-focused tokenizers further reduce fragmentation by explicitly modeling Arabic morphology. This shows how tokenizer training implicitly prioritizes certain linguistic distributions in ways that disadvantage underrepresented languages. Such fragmentation imposes a structural constraint on downstream processing: longer effective input sequences reduce usable context and increase sensitivity to question length, offering a plausible explanation for the sharper performance degradation observed for Arabic as task complexity increases.

5.5 Reliability of Confidence and Explanations

The negative relationship between model-reported confidence and correctness suggests that self-assessed confidence reflects surface-level fluency rather than medical correctness. In multiple-choice settings, this can lead to confident selection of plausible but incorrect options, limiting the usefulness of confidence as an indicator of output reliability. While this effect appears across languages, it is particularly problematic in low-resource settings, where lower baseline accuracy increases the risk of over-trusting incorrect outputs.

Our explanation prompting results caution against treating generated rationales as a reliable remedy. Rather than improving outcomes, explanations induce a model- and language-dependent behavioral shift that reallocates generation toward coherent justifications instead of answer selection. These findings show that self-reported confidence and free-form explanations are insufficient as stand-alone reliability signals for multilingual medical QA, motivating evaluation and calibration approaches beyond model-internal self-assessments.

6 Conclusion and Future Work

Our findings underscore the need for language-aware adaptation across the entire modeling pipeline. At the representation level, tokenization must better capture morphological and domain-specific structure; at evaluation, alignment constraints should avoid conflating surface-form variation with reasoning errors; and at deployment, stronger calibration is required, as model confidence and explanations are unreliable in multilingual medical settings. Overall, these results suggest that improving medical LLM performance in underrepresented languages requires coordinated design choices rather than isolated model scaling or domain specialization.

More broadly, we present a diagnostic evaluation framework that combines controlled cross-lingual comparisons, question-level analysis, and reliability assessment to expose systematic weaknesses obscured by aggregate accuracy metrics. Although our study focuses on Arabic–English medical tasks, the methodology is language-agnostic and applicable to other low-resource or multilingual settings. We hope this work encourages evaluation protocols that explicitly account for linguistic structure, robustness, and reliability when developing medical AI systems for diverse clinical populations.

Limitations

This study examines Arabic and English as a controlled language pair, with Arabic representing a widely spoken yet underrepresented language in medical NLP. While Arabic presents distinct linguistic and morphological challenges, it does not reflect the full diversity of low-resource or typologically distant languages; therefore, the generalizability of our findings beyond this pairing remains an open question.

Our analysis is diagnostic rather than causal. We identify systematic performance patterns across language, task complexity, tokenization behavior, and reliability signals, but do not isolate the effects of specific architectural choices, pre-training strategies, or data composition. This limitation is exacerbated by limited transparency in recent LLM training pipelines, which precludes controlled comparisons between adaptation paradigms such as instruction fine-tuning and continued pre-training.

Several design choices may also influence the results. Large models (70B) are evaluated using 4-bit quantization, which may introduce size-dependent effects but is required for evaluation at scale. In the explanation generation experiment, fixed generation budgets may disproportionately constrain Arabic outputs due to higher tokenization fragmentation; exploring language-adaptive generation limits is left to future work.

Finally, Arabic evaluations involve mixed-language prompts. While fully Arabic prompting was preliminarily tested and yielded lower performance, a systematic comparison of prompt-language strategies was out of scope. English versions of the dataset were obtained via automatic translation and were not manually validated; although the goal is cross-lingual comparison rather than translation quality assessment, translation noise may affect English performance.

Acknowledgements

This work was supported by the Meem Foundation and the New York University Abu Dhabi (NYUAD) Center for Interdisciplinary Data Science and AI (CIDSAI), funded by Tamkeen under the NYUAD Research Institute Award CG016. The research was carried out on NYUAD's High Performance Computing resources (Jubail).

Ethical Considerations

This work evaluates LLMs for medical question answering, which carries inherent risks if such systems are deployed without appropriate safeguards. Our study is strictly evaluative and does not advocate the use of LLMs as standalone clinical decision-making tools. The dataset used in this study consists of non-patient-specific medical questions and does not involve real clinical records or personal health information. Furthermore, by highlighting systematic performance disparities across languages, this work aims to support more equitable evaluation and development of medical AI systems.

References

Mouath Abu-Daoud, Leen Kharouf, Omar El Hajj, Dana El Samad, Mariam Al-Omari, Jihad Mallat, Khaled Saleh, Nizar Habash, and Farah E. Shamout. 2026. Medarabench: Large-scale arabic medical question answering dataset and benchmark. *Preprint*, arXiv:2602.01714.

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. 2024. 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) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 50–56, Torino, Italia. ELRA and ICCL.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering. *Artificial Intelligence in Medicine*, 155:102938.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLaM: Large Language Models for Arabic and English. In *The Thirteenth International Conference on Learning Representations*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards Medical Complex Reasoning with LLMs through Medical Verifiable Problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A Suite of Clinical LLMs. *Preprint*, arXiv:2408.06142.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E. Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks. *Preprint*, arXiv:2505.03427.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the Frontier of Open Large Language Models. *Preprint*, arXiv:2512.02556.

Epoch AI. 2024. Data on Notable AI Models. `https://epoch.ai/data/notable-ai-models`. Accessed: 2025-05-11.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Sale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024.

Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Vignesh Kumar Kembu, Pierandrea Morandini, Marta Bianca Maria Ranzini, and Antonino Nocera. 2025. Are LLMs Truly Multilingual? Exploring Zero-Shot Multilingual Capability of LLMs for Information Retrieval: An Italian Healthcare Use Case. *Preprint*, arXiv:2512.04834.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.3 70b instruct. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`. Hugging Face model card.

Mistral AI. 2025. Mistral-small-3.2-24b-instruct-2506. `https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506`. Hugging Face model card.

Moshe Munk and Dror G. Feitelson. 2022. When Are Names Similar Or the Same? Introducing the Code Names Matcher Library. *Preprint*, arXiv:2209.03198.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of GPT-4 on Medical Challenge Problems. *Preprint*, arXiv:2303.13375.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023b. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *Preprint*, arXiv:2311.16452.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. 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, Marseille, France. European Language Resources Association.

OpenMeditron Initiative. 2024. Llama-3.1 meditron-3 [70b]. `https://huggingface.co/OpenMeditron/Meditron3-70B`. Hugging Face model card; publication forthcoming.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. MedGemma Technical Report. *Preprint*, arXiv:2507.05201.

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, and 13 others. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *Preprint*, arXiv:2308.16149.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward Expert-Level Medical Question Answering with Large Language Models. *Nature Medicine*, 31(3):943–950.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An Arabic-Centric

Multimodal Generative AI Platform. *Preprint*, arXiv:2501.13944.

Juraj Vladika, Mahdi Dhaini, and Florian Matthes. 2025. Facts Fade Fast: Evaluating Memorization of Outdated Medical Knowledge in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9161–9174, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs. *Preprint*, arXiv:2504.00993.

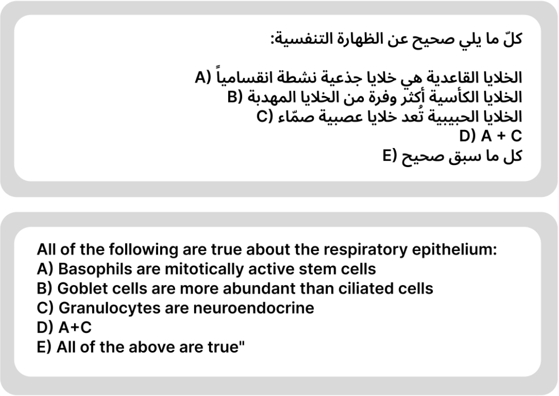
Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical Exam Question Answering with Large-Scale Reading Comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

A Data Samples

Figure A1 shows representative examples from the dataset, including the original Arabic question and its corresponding English translation, to illustrate the structure and content of the bilingual data used in our experiments.



كلّ ما يلي صحيح عن الظهارة التنفسية:

A) الخلايا القاعدية هي خلايا جذعية نشطة انقساماً
B) الخلايا الكأسية أكثر وفرة من الخلايا المهدبة
C) الخلايا الحبيبية تُعد خلايا عصبية صمّاء
D) A + C
E) كل ما سبق صحيح

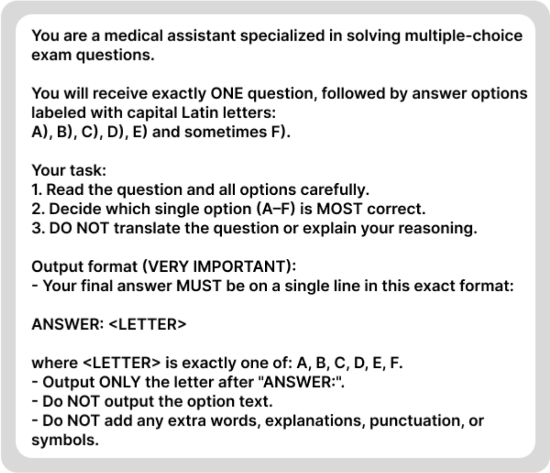
All of the following are true about the respiratory epithelium:
A) Basophils are mitotically active stem cells
B) Goblet cells are more abundant than ciliated cells
C) Granulocytes are neuroendocrine
D) A+C
E) All of the above are true"

Figure A1: **Example dataset entry showing an Arabic MCQ (top) and its English translation (bottom).**

B Prompts Used

B.1 Letter-Based Prompting

Figure B2 shows the exact prompt template used for letter-based multiple-choice question answering, where the model is instructed to return only a single answer option (A–F) without additional explanation.

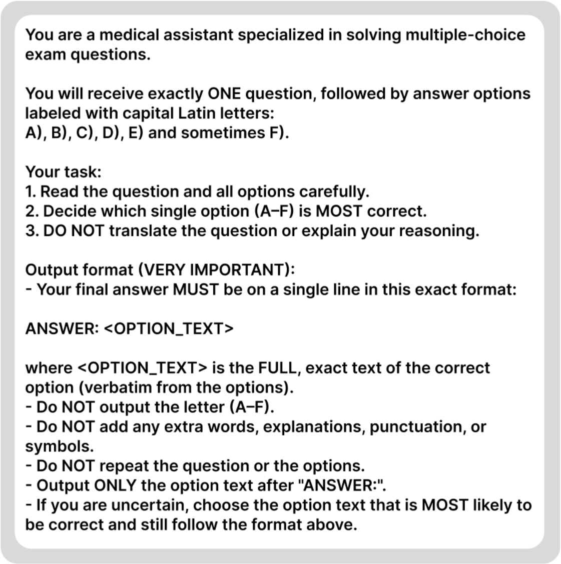


```
You are a medical assistant specialized in solving multiple-choice
exam questions.

You will receive exactly ONE question, followed by answer options
labeled with capital Latin letters:
A), B), C), D), E) and sometimes F).

Your task:
1. Read the question and all options carefully.
2. Decide which single option (A–F) is MOST correct.
3. DO NOT translate the question or explain your reasoning.

Output format (VERY IMPORTANT):
- Your final answer MUST be on a single line in this exact format:

ANSWER: <LETTER>

where <LETTER> is exactly one of: A, B, C, D, E, F.
- Output ONLY the letter after "ANSWER:".
- Do NOT output the option text.
- Do NOT add any extra words, explanations, punctuation, or
symbols.
```

Figure B2: **Prompt template used for letter-based MCQ answering.**

B.2 Text Generation Prompting

Figure B3 shows the prompt template used for exact text generation matching, where the model is instructed to return the exact text sequence corresponding to the correct answer option without additional explanation.

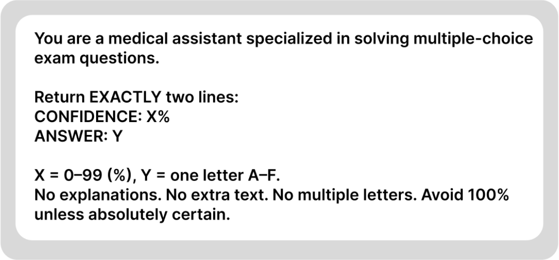


```
You are a medical assistant specialized in solving multiple-choice
exam questions.

You will receive exactly ONE question, followed by answer options
labeled with capital Latin letters:
A), B), C), D), E) and sometimes F).

Your task:
1. Read the question and all options carefully.
2. Decide which single option (A–F) is MOST correct.
3. DO NOT translate the question or explain your reasoning.

Output format (VERY IMPORTANT):
- Your final answer MUST be on a single line in this exact format:

ANSWER: <OPTION_TEXT>

where <OPTION_TEXT> is the FULL, exact text of the correct
option (verbatim from the options).
- Do NOT output the letter (A–F).
- Do NOT add any extra words, explanations, punctuation, or
symbols.
- Do NOT repeat the question or the options.
- Output ONLY the option text after "ANSWER:".
- If you are uncertain, choose the option text that is MOST likely to
be correct and still follow the format above.
```

Figure B3: **Prompt template used for text generation MCQ answering.**

B.3 Confidence Generation Prompting

Figure B4 shows the exact prompt template used for confidence-aware MCQ answering, where the model is instructed to report an explicit confidence estimate alongside its selected answer using a strictly defined output format, without providing any additional explanation.

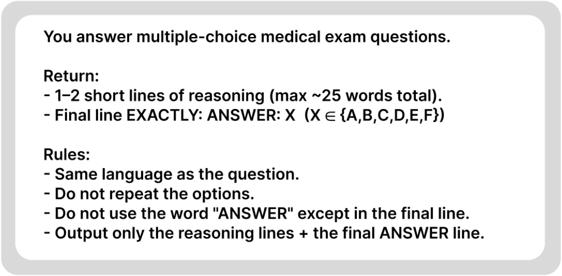


```
You are a medical assistant specialized in solving multiple-choice
exam questions.

Return EXACTLY two lines:
CONFIDENCE: X%
ANSWER: Y

X = 0–99 (%), Y = one letter A–F.
No explanations. No extra text. No multiple letters. Avoid 100%
unless absolutely certain.
```

Figure B4: **Prompt template used for confidence generation MCQ answering.**

B.4 Explanation Generation Prompting

Figure B5 shows the exact prompt template used for explanation-based MCQ answering, where the model is instructed to generate a brief medical rationale followed by a final answer selection, enabling analysis of whether generated explanations correlate with answer correctness.



You answer multiple-choice medical exam questions.

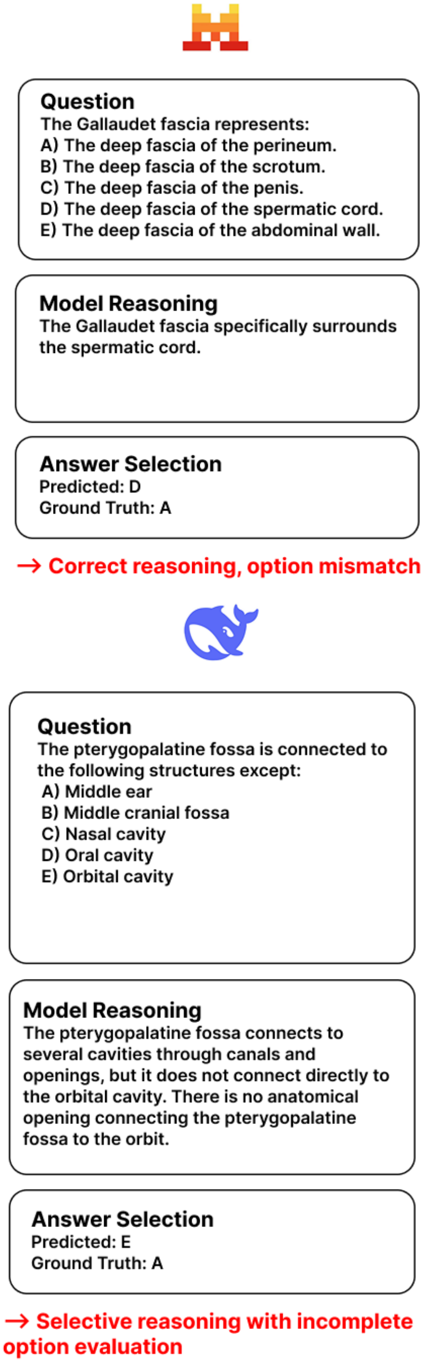
Return:
- 1–2 short lines of reasoning (max ~25 words total).
- Final line EXACTLY: ANSWER: X (X ∈ {A,B,C,D,E,F})

Rules:
- Same language as the question.
- Do not repeat the options.
- Do not use the word "ANSWER" except in the final line.
- Output only the reasoning lines + the final ANSWER line.

Figure B5: **Prompt template used for explanation-based MCQ answering.**

C Examples of Failure Modes Under Explanation Prompting

Figure C6 shows representative examples of failure modes observed under explanation-conditioned prompting. In these cases, models generate medically plausible or partially correct explanations but select an incorrect answer option, revealing misalignment between reasoning and final answer selection.



Question
The Gallaudet fascia represents:
A) The deep fascia of the perineum.
B) The deep fascia of the scrotum.
C) The deep fascia of the penis.
D) The deep fascia of the spermatic cord.
E) The deep fascia of the abdominal wall.

Model Reasoning
The Gallaudet fascia specifically surrounds the spermatic cord.

Answer Selection
Predicted: D
Ground Truth: A

→ Correct reasoning, option mismatch

Question
The pterygopalatine fossa is connected to the following structures except:
A) Middle ear
B) Middle cranial fossa
C) Nasal cavity
D) Oral cavity
E) Orbital cavity

Model Reasoning
The pterygopalatine fossa connects to several cavities through canals and openings, but it does not connect directly to the orbital cavity. There is no anatomical opening connecting the pterygopalatine fossa to the orbit.

Answer Selection
Predicted: E
Ground Truth: A

→ Selective reasoning with incomplete option evaluation

Figure C6: **Examples of reasoning–label misalignment under explanation prompting.** Models may produce correct or salient medical reasoning while selecting an incorrect option due to option mismatch or incomplete evaluation of alternatives.