PersianMedQA: Language-Centric Evaluation of LLMs in the Persian Medical Domain

Anonymous EMNLP submission

Abstract

Large Language Models (LLMs) have achieved remarkable performance on a wide range of NLP benchmarks, often surpassing human-level accuracy. However, their reliability in high-stakes domains such as medicine, particularly in low-resource languages, remains underexplored. In this work, we introduce **PersianMedQA**, a large-scale, expert-validated dataset of multiple-choice Persian medical questions, designed to evaluate LLMs across both Persian and English. We benchmark over 40 state-of-the-art models, including general-purpose, Persian fine-tuned, and medical LLMs, in zero-shot and chain-of-thought (CoT) settings. Our results show that closed-source general models (e.g., GPT-4.1) consistently outperform all other categories, achieving **83.1%** accuracy in Persian and **83.3%** in English, while Persian fine-tuned models such as Dorna underperform significantly (e.g., **35.9%** in Persian), often struggling with both instruction-following and domain reasoning. We also analyze the impact of translation, showing that while English performance is generally higher, Persian responses are sometimes more accurate due to cultural and clinical contextual cues. Finally, we demonstrate that model size alone is insufficient for robust performance without strong domain or language adaptation. PersianMedQA provides a foundation for evaluating multilingual and culturally grounded medical reasoning in LLMs.

Medical Examples

Clinical:

A 48-year-old man has been brought to the emergency room with chest pain that started 4 hours ago. In the ECG, ST-segment elevation is evident in the anterior leads. On examination, the patient has sweating, blood pressure of 90/60 mmHg, distended neck veins, and rales heard at the base of the lungs. What is the most effective treatment?
Options:
1. Administer fibrinolytic and if necessary, emergency angioplasty
2. Administer fibrinolytic
3. Emergency angioplasty
4. Administer fibrinolytic and angioplasty 48 hours later

Answer: 3

Non-Clinical:

All of the following can be causes of acute retinal necrosis, except:
Options:
1. Cytomegalovirus
2. Herpes simplex type 1
3. Toxoplasmosis
4. Varicella Zoster

Answer: 3

Figure 1: A translated medical question example from the dataset.

1 Introduction

LLMs have become the go-to solution for many tasks, showcasing promising results on standard benchmarks, potentially replacing humans across various domains (Brown et al., 2020; OpenAI, 2023). However, their reliability in tasks that require real attention to detail, such as tasks that directly impact human life, remains concerning (Bommasani and et al., 2021; Zhang and et al., 2023). Medical tasks like clinical decision-making represent a critical domain where experts must possess comprehensive knowledge in cultural contexts, medical principles, pharmaceutical information, and numerous other specialized areas in healthcare (Liu and et al., 2023; Lee and et al., 2023).

Although recent works have demonstrated that LLMs may achieve accuracy rates exceeding 90% on English medical question-answering tasks (Singhal et al., 2022; Nori and et al., 2023), their performance falls off significantly in other languages other than English (Lee and et al., 2023; AlGhanem and et al., 2023). Importantly, simply translating questions is inadequate, as subtle cultural cues and localized standards of care often vanish in the process (Joshi and et al., 2020; Liu and et al., 2023). These nuances can decisively alter diagnoses, treatment plans, and ultimately pa-

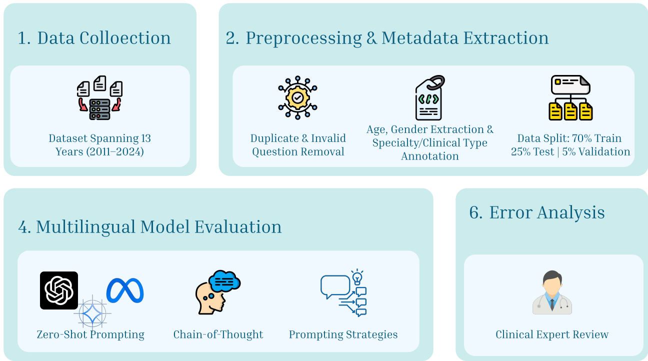

Figure 2: Overview of the PersianMedQA dataset construction process, including data collection, cleaning, annotation, and partitioning steps.

tient outcomes (Vilares and et al., 2023; Min and et al., 2023).

For low-resource languages, the evidence base is even thinner. Limited research has investigated the specific factors that mislead LLMs in medical contexts, in multilingual and low-resource language settings like Persian. A deeper investigation on the medical sub-fields in which LLMs excel or underperform is essential for identifying suitable use cases and implementing necessary safeguards (Bommasani and et al., 2021). Such insights are crucial for the responsible deployment of LLMs in clinical environments, where errors carry substantial clinical risk (Zhang and et al., 2023; Liu and et al., 2023).

To fill this gap, we introduce **PersianMedQA**, a large-scale, expert-annotated dataset covering 23 medical specialties. The dataset includes a comprehensive bilingual dictionary of Persian and medical terms to support both evaluation and model adaptation. As a benchmark, we evaluate a range of state-of-the-art models, including general-purpose models, Persian fine-tuned models, and medical models in both Persian and English. Our experiments uncover a pronounced language gap between Persian and English: closed-source models such as GPT-4.1 significantly outperform open-source counterparts. Notably, Persian fine-tuned models exhibited minimal domain knowledge and performed the worst, while medical fine-tuned models showed only modest improvements and failed to generalize effectively to Persian clinical data. Figure 2 illustrates the overall workflow of our study, including dataset collection, model evaluation, and analysis steps.

In Section 2, we review prior work on medical QA benchmarks, multilingual LLM evaluation, and Persian language models. Section 3 describes the PersianMedQA dataset construction, including data collection, cleaning, and annotation. Section 4 presents our experimental setup and zero-shot, CoT, translation-impact, and ensembling evaluations, and the analysis of the results across medical subfields, model sizes, and artifact reliance. Finally, Section 5 concludes with key findings, limitations, and directions for future research.

2 Related Works

Medical Question Answering (QA). Medical QA has long been used as a benchmark for machine reasoning in high-stakes domains. Progress accelerated with domain-specific language pretraining: BIOBERT (Lee et al., 2020) and PUBMEDBERT (Gu et al., 2021) each delivered sizable gains on benchmark datasets including PUBMEDQA (Jin et al., 2019), MEDQA (Jin et al., 2021), and MEDMCQA (Pal et al., 2022). Recent retrieval-augmented generation approaches (RAG) (Lewis et al., 2020) attempt to ground LLM outputs in trusted sources, yet factual consistency remains a challenge (Singhal et al., 2022). Multilingual coverage is also expanding: the CBM benchmark (Zhang et al., 2023) introduces a comprehensive suite of Chinese medical QA tasks, underscoring the fields rising attention to multilingual healthcare evaluation.

LLMs in Medical Practice. Medical-specific LLMs such as MED-PALM and MED-PALM 2 have shown that combining domain-specific pre-training, instruction tuning, and CoT prompting can enhance performance on the United

States Medical Licensing Examination (USMLE) well above the passing threshold (Singhal et al., 2022, 2023; Wei et al., 2022). Despite these advances, most published evaluations remain English-centric. While General-purpose models like GPT-4 achieve strong zero-shot results on a variety of medical QA benchmarks (OpenAI, 2023; Nori and et al., 2023; Lee and et al., 2023), their behavior in multilingual clinical settings, especially for low-resource languages like Persian, has not been systematically explored.

Multilingual Medical QA and the Limits of Translation. A common workaround for evaluating medical QA in low-resource languages is to translate questions into English. Yet recent work shows that this *translation-first* pipeline can strip away critical terminology and distort local clinical guidelines, ultimately hurting accuracy and safety. At the same time, several multilingual, or at least non-English medical QA benchmarks have appeared, including MedQA (Chinese) (Jin et al., 2021), MedMCQA (Hindi) (Pal et al., 2022), CBM (Chinese) (Zhang et al., 2023), and the aggregated MultiMedQA suite (Singhal et al., 2022). MedExpQA further augments these resources with cross-lingual explanations (Alonso et al., 2024). Despite this progress, no publicly available dataset targets Persian medical QA, leaving its distinct clinical context completely unrepresented. These gaps highlight the need for language-aware medical LLMs and benchmarks rather than one-size-fits-all translation strategies.

Persian Language Models and QA. Efforts in Persian NLP have produced strong monolingual models such as ParsBERT (Farahani et al., 2020), Dorna (PartAI, 2024). These models outperform multilingual baselines on tasks like sentiment analysis and classification. However, few are trained or evaluated in the medical domain. SINA-BERT (Taghizadeh et al., 2021) mark early attempts to address this gap, yet focus on document classification or conversational QA.

3 PersianMedQA Construction

The PersianMedQA dataset was developed by collecting 14 years of multiple-choice questions from the Iranian residency and pre-residency medical exams. Each item includes the question text, four answer options, and the correct answer key. Figure 1 presents representative examples of clinical

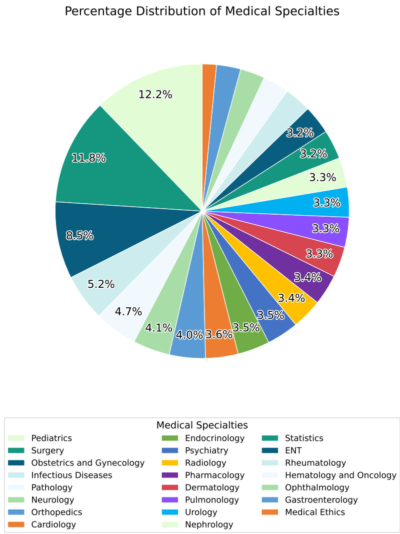

Figure 3: Distribution of medical fields in the dataset.

and non-clinical questions. The raw dataset underwent a rigorous preprocessing pipeline to ensure quality, consistency, and relevance for multilingual medical QA evaluation. The following steps summarize the construction process:

3.1 Data Cleaning and Filtering

In order to eliminate noise and redundancy, we ran a three-step cleaning pipeline:

- **Duplicate Removal:** Automatically prune exact and near-duplicate questions using string matching and sentence-embedding similarity to maintain diversity.
- **Image Dependent Exclusion:** Discard any question that relies on medical images (e.g., radiographs, histology slides) so the benchmark remains purely text-based.
- **Answer Key Verification:** Conduct a review to remove items with missing, conflicting, or implausible answer keys.

3.2 Annotation and Categorization

To enhance interpretability and analysis, the cleaned dataset was annotated as follows:

- **Subject Verification:** Partnered with medical experts to confirm and correct each questions subject tag.
- **Domain Classification:** Labeled questions as *clinical* (patient cases and diagnosis) or

Table 1: Train/Validation/Test split in PersianMedQA

Subset	Number of Questions
Training	14,549
Validation	1,000
Test	5,236
Total	**20,785**

non-clinical (basic sciences and theoretical concepts).

- **Demographic Extraction:** Utilized Gemini to automatically extract patient attributes (e.g., age, gender) for every question.

3.3 Dataset Overview

The final **PersianMedQA** dataset comprises 20,785 unique, expert-validated multiple-choice medical questions, collected over 14 years from Iranian national residency, pre-residency, and board exam archives. Approximately 70% of the questions are classified as clinical, with the remaining 30% labeled non-clinical. The items span 23 medical specialties and cover both patient-case and theoretical knowledge domains. We also include and analyze additional metadata (e.g., patient gender, age, and other demographic attributes). A full breakdown of these distributions appears in Appendix A.

The dataset is randomly partitioned into 14,549 training examples, 1,000 validation examples, and 5,236 test examples to support robust model development and evaluation (see Table 1). Figure 3 summarizes the distribution of questions across medical domains.

4 Experiments

4.1 Zero-shot Scenario

We conducted zero-shot evaluations on the PersianMedQA dataset using a wide range of state-of-the-art open-source and closed-source LLMs in both Persian and English. All models were prompted using identical instructions (provided in the C), with temperature set to 0 and a sufficiently large generation length. Prompts were issued in English across both language settings to control for instruction comprehension.

The overall accuracy of a samples of evaluated models on Persian and English test sets is presented in Figure 4. Among all models, the closed-source GPT-4.1 achieved the highest zero-shot accuracy in both languages, scoring 83.09% in Persian and 83.34% in English. Notably, the best-performing open-source model, LLaMA 3.1-405B Instruct, achieved a strong 69.25% in Persian and 75.83% in English. In terms of medical-tuned models, Meditron3-8B scored only 39.70% in Persian and 51.64% in English, revealing substantial room for improvement in domain adaptation for Persian.

Persian fine-tuned models significantly underperformed across the board; some of them suffered greatly from not being able to follow instructions. PersianMind-1.0 achieved only 23.98% in Persian (roughly equivalent to random guessing) and 25.90% in English, suggesting limited medical knowledge and insufficient generalization capability in clinical domains. Similarly, Dorna2-LLaMA-3.1-8B-Instruct, another Persian fine-tuned model, scored just 35.96% in Persian and 53.10% in English, indicating slightly better instruction following but still poor domain alignment in the Persian medical setting.

Overall, closed-source models consistently outperformed both open-source and fine-tuned medical models, particularly in Persian. While most models exhibited performance degradation when evaluated in Persian compared to English, some top-tier models, such as GPT-4.1 and Gemini 2.5-Flash, showed minimal to no drop, indicating stronger crosslingual transfer capabilities.

We further analyze model performance across different medical specialties. Figure 5 presents a heatmap of accuracy scores for each model across all medical fields in the PersianMedQA dataset.

Several factors shaped model performance across medical subfields. For example, pharmacology questions, which hinge on factual recall rather than complex clinical reasoning, yielded the highest accuracies for most models. Likewise, non-clinical items (theoretical or basic-science questions) tended to be answered more accurately than clinical case scenarios, reflecting their relatively straightforward nature.

In contrast, performance was dropped sharply in subfields such as surgery and medical statistics, which require complex reasoning, quantitative interpretation, and a deeper understanding of language-specific clinical guidelines and protocols. These findings show that factual recall alone is insufficient: robust medical QA calls for deeper reasoning and cultural grounding across subfields.

Translation Impact. English dominates both the web-scale corpora that power modern LLMs and the medical literature on which they are

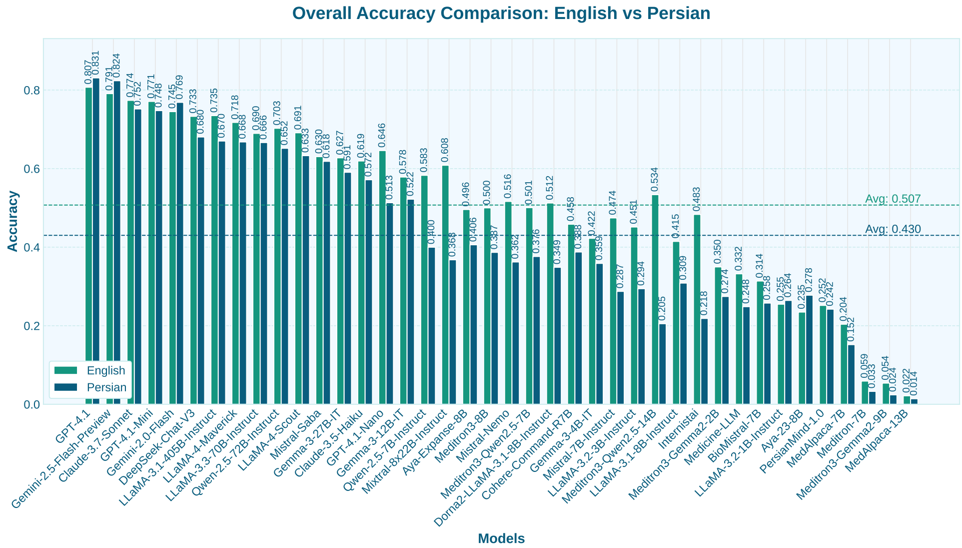

Figure 4: Overall accuracy of models on Persian and English test sets.

trained. To assess the effect of language, we translated the PersianMedQA dataset into English and compared model performance on the original Persian versus the translated English questions.

We generated translations using three methods: Google Translate, the GPT-4.1 API, and the Gemini-2.5-Flash API, and evaluated them for fluency and domain fidelity. Both GPT-4.1 and Gemini-2.5-Flash produced more accurate, natural translations than Google Translate. Due to its combination of quality and accessibility, we use Gemini-2.5-Flash translations as our default in all subsequent experiments.

To better understand model behavior across languages, we categorized every question into three mutually exclusive categories based on the LLM's correctness in Persian, in English, or in both.

1. **Correct in Both Languages:** Questions answered correctly in both Persian and translated English.

2. **Correct Only After Translation:** Questions solved only after the question's translation indicate a boost from the models stronger English competence.

3. **Correct Only in Persian:** Questions answered correctly only in Persian, suggests that language- or culture-specific cues outweigh any gains from translation.

This categorization revealed two consistent patterns:

- Most models were trained predominantly on English medical data, and thus benefited from translation due to stronger representation and alignment with English-language knowledge bases.

- However, a non-negligible number of questions were only correctly answered in Persian. Upon further analysis, these questions often involved region-specific clinical guidelines and protocols that are more prevalent in the Iranian medical system. In such cases, translation introduced semantic drift or failed to preserve culturally grounded medical knowledge, leading to incorrect answers in English.

Impact of Model Size. We further analyzed whether model size correlates with performance across different model types. Figure 6 illustrates the relationship between model size and accuracy for some of the evaluated models. While larger models generally show better performance, this trend is not consistent across all categories:

- For general-purpose models, increased scale appears beneficial GPT-4.1 (the largest) leads with over 83% accuracy, while smaller GPT variants (e.g., GPT-4.1-Nano) fall to the 50-60% range.

- For medical fine-tuned models, larger size does not guarantee better performance. Despite their size, MedAlpaca-13B, Meditron3, Gemma-9B, and MedAlpaca-7B all scored very low in both Persian and English,

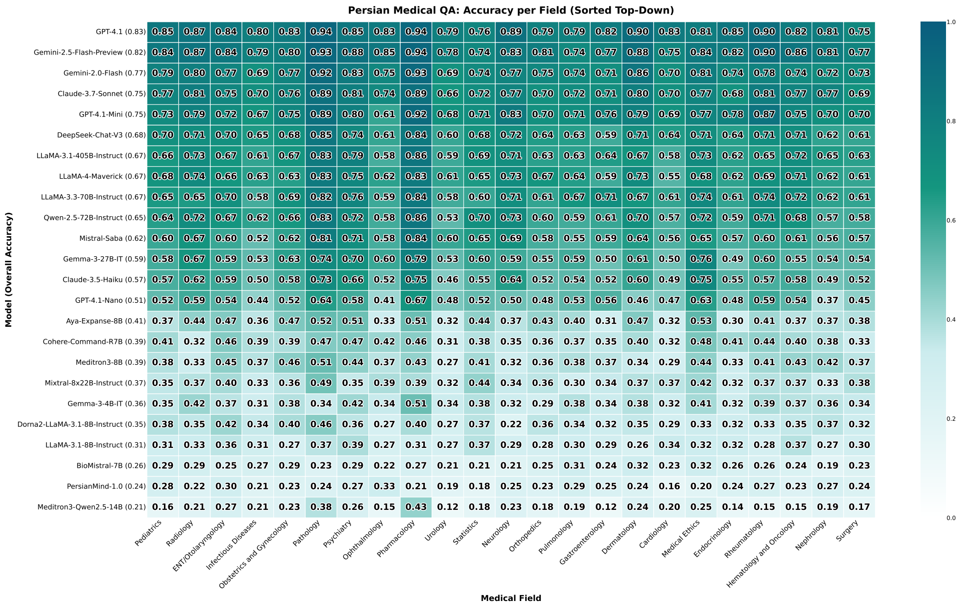

Figure 5: Heatmap showing the accuracy of each model across all medical specialties in the PersianMedQA dataset. Each cell represents the accuracy for a particular model-field pair.

performing far below even small general-purpose models.

- Persian fine-tuned models struggle regardless of scale. Even relatively large models like Dorna2-LLaMA-3.1-8B perform poorly (35.96%), likely due to limited training data or weak domain alignment.

These results indicate that model scale is beneficial only when accompanied by sufficient high-quality training data and domain coverage.

4.2 Prompting Strategies and Few-shot Learning

We experimented with various prompting strategies and few-shot learning approaches; the results are summarized below.

Role-based prompting, where the model was instructed to act as a specialist based on the medical field of the question (e.g., "You are a cardiologist..."), resulted in slightly improved performance, but the gains were marginal.

Few-shot learning For every test question we drew the in-context examples exclusively from the PersianMedQA training split (up to k = 5 per query). We experimented with several retrieval schemes for picking those training examples, LaBSE cosine similarity, TF-IDF, and random selection, but none of them produced consistent gains over the zero-shot baseline. A plausible reason is the absence of high-quality embedding models tailored to Persian medical text, which makes it difficult to retrieve truly helpful training examples.

We also experimented with augmenting each question with a medical dictionary, extracted by a larger, more capable model (Gemini-2.5-Flash), that provided both translations and concise definitions of key terms. This dictionary (see F) was released alongside the dataset to help smaller models interpret domain-specific terminology. However, we found that this augmentation had a negligible effect on overall performance, especially for weaker or instruction-tuned models.

4.3 Answer-Only Evaluation of LLM Medical Reasoning

To test whether LLMs genuinely understand medical questions, or merely exploit memorized patterns and statistical regularities in the answer choices, we adopted the *partial-input* protocol of Balepur et al. (2024). Each model was informed that it would answer a medical question but received *only* the four answer options, never the question stem. Accuracy noticeably above the 25 % random-guess baseline, therefore signals dependence on answer-choice artifacts rather than true comprehension.

The key findings were that in the answer-only setting, bigger LLMs like GEMINI still outperformed their smaller counterparts. Performance varied markedly by specialty: Knowledge-

Table 2: Majority-vote ensembles. "Δ_{best} is the gain over the best single model in the group.

Ensemble / Baseline	**Acc.**	**Avg. Acc.**	$\mathbf{\Delta_{best}}$
Top-3 Overall	0.834	0.808	+0.003
Top-5 Overall	0.831	0.790	-0.001
Top-3 GPT Family	0.803	0.704	-0.028
Top-3 Google Family	0.795	0.728	-0.029
Top-3 Claude Family	0.777	0.684	-0.001
Top-5 Open Sources	**0.737**	0.679	**+0.033**
Human Baseline	0.75		

Table 3: Selective answering performance (sample of models)

Model	Orig. Acc.	Sel. Acc.	Improvement	Coverage
GPT-4.1	0.8309	0.8524	+0.0215	59.2%
Claude-3.7-Sonnet	0.7519	0.7817	+0.0298	42.6%
Gemini-2.0-Flash	0.7686	0.7941	+0.0254	45.1%
Gemma-3-27B-IT	0.5906	0.6413	+0.0507	21.7%
Gemini-2.5-Flash-Preview	0.8237	0.8420	+0.0183	56.8%

heavy specialties like Pharmacology, Radiology, and Nephrology stayed near random, whereas principle-driven areas such as Medical Ethics yielded noticeably higher scores.

Manual inspection revealed that models exploit three recurrent answer-choice artifacts: (i) *logically exclusive options*, where an implausible or self-contradictory choice can be discarded without the context; (ii) *hierarchical cues*, in which an ordered sequence (e.g., steps in a protocol) reveals the correct rank; and (iii) *linguistic or formatting cues*, where options with precise terminology, numeric specificity, or textbook phrasing that signals the right answer.

Running the same experiment on the English translations produced similar patterns, with a slight overall accuracy gain. These results warn that current medical MCQ benchmarks may overstate LLM reasoning abilities by permitting exploitation of answer-choice artifacts instead of requiring genuine medical understanding.

4.4 Model Ensembling

Analysis of the confusion matrices reveals that different models exhibit varying strengths across different medical subjects, and models from distinct families often demonstrate differing agreement patterns on answers (see Appendix for visualizations). This suggests that ensembling models, especially those from diverse families, may yield higher overall accuracy.

Additionally, some of the highest-performing models, such as GPT and Gemini, are not open source, limiting their potential for future development and adaptation. Therefore, leveraging open-source models in ensemble methods remains highly valuable.

Our evaluation of various ensemble configurations shows that combining models from the same family does not necessarily lead to substantial accuracy gains. In contrast, ensembles that mix models from different families achieve higher performance, with the top-5 ensemble reaching an accuracy of 0.831. Notably, an ensemble of five open-source models achieves 73.7% accuracy, which is comparable to the top-performing closed models and highlights the potential of open models for future research and development.

4.5 Selective Answering

In high-stakes domains like medicine, it is preferable to abstain rather than to provide incorrect answers. To that end, we implement a confidence-based selective answering strategy. We embed each question with LaBSE, assign a pseudo-confidence equal to the mean accuracy of its three cosine-nearest neighbors, and answer only when this score exceeds a tunable threshold. We evaluate performance in terms of *accuracy* (on the answered subset) and *coverage* (the fraction of questions answered). Aggregating across all models, selective answering yields substantial accuracy gains when partial coverage is acceptable. Table 3 shows a small sample of models under this regime.

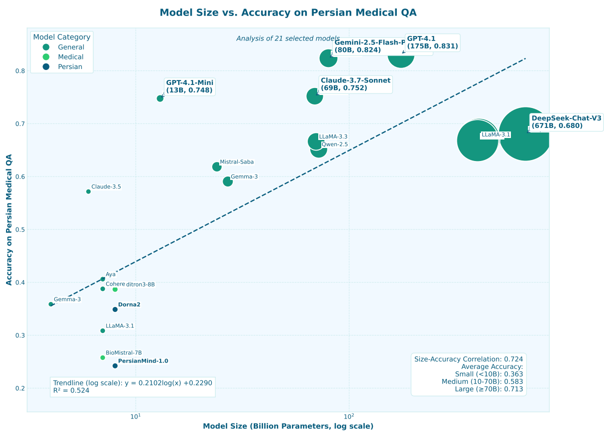

Figure 6: Relationship between model size and accuracy across different model types.

4.6 Chain-of-Thought Evaluation.

To evaluate the impact of CoT reasoning on LLMs, we applied CoT prompting to four different LLMs, including two general-purpose models (GPT-4.1 and Gemini-2.5-Flash), one medical model (Meditron3), and one Persian fine-tuned model (Dorna).

Performance Gains. For the top-performing models (GPT-4.1 and Gemini-2.5-Flash), CoT prompting led to an average accuracy improvement of approximately 2%, suggesting that explicit reasoning instructions can enhance even highly capable models when addressing complex medical questions. In contrast, the effect small models was modest, with little to no observed improvement, likely due to their more limited Persian language understanding and reasoning abilities compared to the larger models. Notably, we also observed that CoT prompting yielded greater accuracy gains on clinical questions in the large models, highlighting that clinical scenarios particularly benefit from explicit reasoning steps.

Expert analysis of CoT. A board-certified clinician manually reviewed samples of GPT-4.1 CoT responses and identified four recurring error modes: (i) *Contextual Mismatch:* Some English answers were grounded in protocols not aligned with Iranian clinical practices, resulting in incorrect reasoning chains despite accurate general knowledge. (ii) *Ambiguity in Options:* GPT-4.1 often failed when faced with highly similar or subtly misleading answer choices. In these cases, the CoT outputs reflected confusion or overconfidence in selecting between near-identical options. (iii) *Reasoning Failures:* A subset of errors was attributed to incomplete or logically inconsistent reasoning, even when the model possessed the necessary knowledge. This highlights a gap between knowledge representation and reliable inference. (iv) *Knowledge Gaps:* Some mistakes were traced to an outright lack of factual information where CoT prompting could not compensate for missing knowledge. Illustrative examples for each error category are provided in the Appendix (Section B).

5 Conclusion

In this study, we present PersianMedQA, a large-scale question-answer collection designed to analyze how well todays language models grasp medical content in Persian. We benchmarked a range of open-source and closed-source LLMs on both the original Persian questions and on English translations.

The results expose a wide performance analysis: only a handful of top-tier LLMs, such as GPT-4.1 and Gemini-2.5-Flash, handled Persian questions as well as, or better than, their English versions. Most other models performed better on the English set, underscoring the persistent barriers to truly multilingual medical AI. This gap was most acute in the smaller models, indicating that simply scaling parameters is not a sufficient recipe for robust cross-lingual medical reasoning.

Future work should (1) build retrieval-augmented or knowledge-grounded LLMs that can query authoritative Persian and English medical sources, (2) create large, domain-specific Persian medical models, and (3) expand benchmarks to other specialties (e.g., dentistry) and multimodal inputs (text plus medical images) to produce clinically reliable AI.

Limitations

Several factors constrained this study. (i) *API restrictions:* cost and rate limits for commercial LLMs (e.g., GPT-4) reduced the number of evaluation runs and chain-of-thought variants we could conduct. (ii) *Licensing barriers:* copyright restrictions prevented us from using larger multilingual biomedical corpora, limiting the scope of our experiments. As a result, our reported scores should be considered conservative lower bounds; broader data access and greater computational resources would enable a more exhaustive evaluation.

Ethics Statement

This study involved the analysis and evaluation of LLMs on publicly available or previously released medical examination data. No private, identifiable, or patient-specific information was used. All data is de-identified and non-sensitive, originating from official Iranian medical entrance and licensing examinations.

Our findings and evaluations aim to improve the responsible deployment of language models in healthcare, especially for underrepresented languages. Also, we emphasize that the models tested are not certified for clinical use and should not be deployed in real-world healthcare settings without strict oversight. We advocate for continued expert-in-the-loop development and further inclusion of diverse linguistic and cultural considerations in medical AI research.

References

Nouran AlGhanem and et al. 2023. Multilingual evaluation of medical language models. *arXiv preprint arXiv:2305.15035*.

Miguel Alonso, Aarthi Balachandran, Akanksha Singh, Shikhar Vashishth, and et al. 2024. Medexpqa: A multilingual benchmark for medical question answering with expert explanations. *arXiv preprint arXiv:2403.05789*.

Aditya Balepur, Rohan Anil, Ronan Le Bras, Yejin Choi, Ashish Sabharwal, and Oyvind Tafjord. 2024. Artifacts in multiple choice question answering: More than just clever hans. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Rishi Bommasani and et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *arXiv preprint arXiv:2005.12515*.

Yu Gu, Robert Tinn, Hao Cheng, Jesse Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Di Jin, Yue Pan, Xiang Ouyang, and Zhiyuan Liu. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6201–6218.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Pratik Joshi and et al. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of ACL*.

Jinhyuk Lee and et al. 2023. Benefits and risks of cross-lingual transfer in biomedical question answering. In *Proceedings of ACL*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Haoran Liu and et al. 2023. Trustworthy multilingual medical ai systems. *arXiv preprint arXiv:2306.07207*.

Sewon Min and et al. 2023. Multilingual evaluation of generative medical question answering. *arXiv preprint arXiv:2305.13552*.

Harsha Nori and et al. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, Vikram Dwivedi, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *arXiv preprint arXiv:2203.14746*.

PartAI. 2024. Dorna2-llama3.1-8b-instruct. `https://huggingface.co/PartAI/Dorna2-Llama3.1-8B-Instruct`. Accessed: 2025-05-17.

Karan Singhal, Shekoofeh Azizi, Talia Tu, Sina Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, and et al. 2022. Large language models encode clinical knowledge. In *arXiv preprint arXiv:2302.12347*.

Karan Singhal, Talia Tu, Sina Mahdavi, Jason Wei, and et al. 2023. Towards expert-level medical question answering with gpt-4. *arXiv preprint arXiv:2305.09617*.

Nasrin Taghizadeh, Ehsan Doostmohammadi, Elham Seifossadat, Hamid R. Rabiee, and Maedeh S. Tahaei. 2021. Sina-bert: A pre-trained language model for analysis of medical texts in persian. *arXiv preprint arXiv:2104.07613*.

David Vilares and et al. 2023. Evaluation of multilingual medical qa: Beyond english and high-resource languages. In *Findings of ACL*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Haoran Zhang and et al. 2023. An audit of large language models for medical applications: Bias, hallucinations, and fairness. *arXiv preprint arXiv:2304.14454*.

Yong Zhang, Jiahua Zhang, Yifan Zhao, et al. 2023. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

A Demographic distributions

We present additional statistics on the demographic metadata present in the PersianMedQA dataset. To extract this information, we experimented with both regular expressions and a LLM. The LLM-based extraction demonstrated consistently high accuracy on this task, outperforming the regex approach in terms of precision and recall.

A.1 Gender distribution

Table 4: Distribution of patient gender across questions.

Gender	Count
Unspecified	9,361
Female	5,831
Male	5,590

A.2 Age category distribution

Table 5: Distribution of patient age categories across questions.

Age Category	Count
Adult (18+)	10,241
Unspecified	6,765
Child (2–17)	2,675
Infant (0–1)	1,101

A.3 Clinical vs. non-clinical distribution

Table 6: Distribution of clinical vs. non-clinical questions.

Category	Count
Clinical (1.0)	14,724
Non-Clinical (0.0)	6,061

B Examples of CoT Error Patterns

This section presents representative error patterns identified in model-generated CoT outputs, as annotated by clinical experts. For each example, we highlight the clinical context, the correct answer, the model's response, and a summary of the expert's evaluation. These cases illustrate the most common types of reasoning failures observed in our analysis, which were further corroborated by the structured expert review.

1. Contextual Mismatch

Question: What is the next step in an immunocompromised patient with nasal congestion and suspected invasive fungal sinusitis?
Correct: Endoscopy and biopsy
Model: Imaging (MRI) is needed before biopsy.
Expert Evaluation: *Incorrect evaluation.* The model follows a Western protocol; however, local clinical practice requires urgent biopsy due to high mortality risk.

2. Ambiguity in Options

Question: What is the most common malignant neoplasm of the liver?
Correct: Hepatocellular carcinoma (HCC)
Model: Metastasis is more common overall, so we choose that.
Expert Evaluation: *Incomplete question.* The model selected a technically true but contextually incorrect answer; expert notes ambiguity in phrasing and clinical intent.

3. Reasoning Failure

Question: What is the correct order of action in a 25-year-old with lymphoma and meningitis signs but no neurologic deficits?
Correct: Blood culture → Lumbar puncture → Empiric antibiotics
Model: CT scan should be done first due to immunosuppression.
Expert Evaluation: *Incorrect conclusion.* The expert highlights that the patients immunosuppression requires a different clinical approach, which the model failed to identify.

4. Knowledge Gap

Question: Which drug works via motilin receptor stimulation for gastroparesis?
Correct: Erythromycin
Model: Metoclopramide is commonly used for gastroparesis, so we choose that.
Expert Evaluation: *Knowledge gap.* Model lacks pharmacologic mechanism knowledge and defaults to common treatments.

C Zero-shot evaluation prompt

Zero-shot Prompt

You are a medical expert tasked with answering multiple-choice medical questions.
Question format

```
Question: [Medical question text]
1: [Option 1]
2: [Option 2]
3: [Option 3]
4: [Option 4]
```

Important notes

[nosep]Select the best answer from the provided choices. Your output must be **only the option number** (1, 2, 3, or 4). Do **not** add explanations or extra text. Base your answers on authoritative medical knowledge.

D CoT reasoning prompt

CoT Prompt

You are a medical expert taking a medical board examination.

For each question, please

[itemsep=0pt]Read and understand the question carefully. Analyze the options (14) systematically. Apply your medical knowledge step by step. Show your chain-of-thought (CoT) reasoning clearly. Explain why each incorrect option is wrong and the chosen one is correct. Explicitly state which option (1, 2, 3, or 4) is your final answer.

Response format (JSON)

[itemsep=0pt]`"CoT"` your step-by-step reasoning. `"Final_Answer"` the option number (1 | 2 | 3 | 4). `"Reasoning"` a concise justification of the answer.

Be methodical, precise, and thorough in your analysis, just as you would in a medical examination. Your expertise as `{english_specialty}` is critical for answering these specialized questions correctly.

E User interfaces

To facilitate expert interaction throughout various phases of our study, we developed multiple user interfaces, primarily implemented as Telegram bots, to streamline collaboration with medical professionals.

E.1 Subject annotation interface

We created a Telegram-based annotation bot to support subject-level classification. Experts could review ambiguous or unclassified questions and select the most appropriate medical field from a predefined list of 23 specialties.

E.2 CoT reasoning interface

To analyze the reasoning behind model outputs, we designed an interface that presented experts with a curated 200-question subset of the dataset. For each question, experts were asked to:

- Select whether a predefined reasoning category applied (e.g., domain knowledge, commonsense, causal inference).
- Optionally assign a new category if the reasoning did not fit existing labels.
- Provide a brief explanation justifying the correct answer.

Figure 7: Telegram interface for expert subject classification of ambiguous questions.

F Persian Medical Dictionary

We present an extracted Persian medical dictionary derived from the dataset. Table 7 summarizes, for each category file, final number of unique medical terms extracted.

G Agreement of Different Model Families

Figure 9 shows the pairwise agreement rates between model families. As expected, each family agrees most with itself, while cross-family agreement ranges roughly from 40% to 80%.

Figure 8: Telegram interface for expert annotation of reasoning categories and explanations.

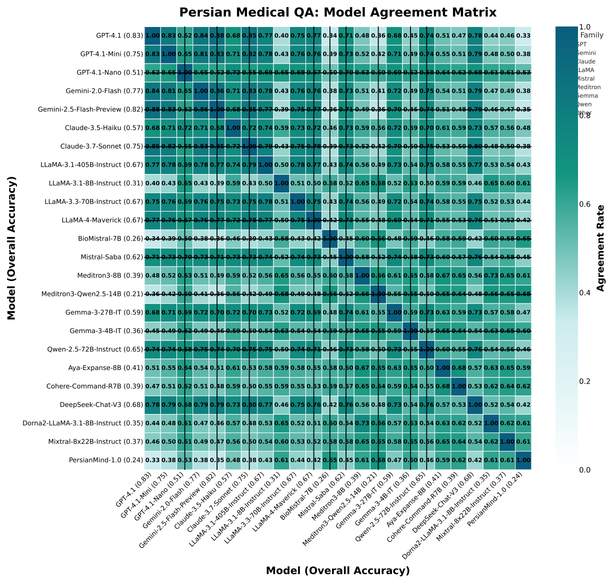

Figure 9: Pairwise agreement matrix of model families. Each cell indicates the percentage of test items for which both families produced the same answer.

Table 7: Distribution of extracted Persian medical terms

Category	Unique Terms
Medical Devices	866
Medical Specialties	273
Lab Tests	6 410
Medical Abbreviations	2 596
Traditional Medicine Terms	64
Procedures	9 632
Anatomical Terms	8 120
Symptoms	14 397
Medications	5 905
Diseases	16 400