PersianMedQA: Evaluating Large Language Models on a Persian-English Bilingual Medical Question Answering Benchmark

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved remarkable performance on a wide range of Natural Language Processing (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 dataset of 20,785 expert-validated multiple-choice Persian medical questions from 14 years of Iranian national medical exams, spanning 23 medical specialties and designed to evaluate LLMs in both Persian and English. We benchmark 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 83.09% accuracy in Persian and 80.7% in English, while Persian fine-tuned models such as Dorna underperform significantly (e.g., 34.9% 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, 3-10% of questions can only be answered correctly in Persian due to cultural and clinical contextual cues that are lost in translation. Finally, we demonstrate that model size alone is insufficient for robust performance without strong domain or language adaptation. PersianMedQA provides a foundation for evaluating bilingual 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. For seeing an original example of dataset check A.

1 Introduction

Large Language Models (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; Team, 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., 2022). Medical tasks, such as 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 within healthcare. In other words, clinical excellence requires more than just biomedical knowledge (Campinha-Bacote, 2002).

Although recent works have demonstrated that LLMs may achieve high accuracy on English medical question-answering tasks (Sing-

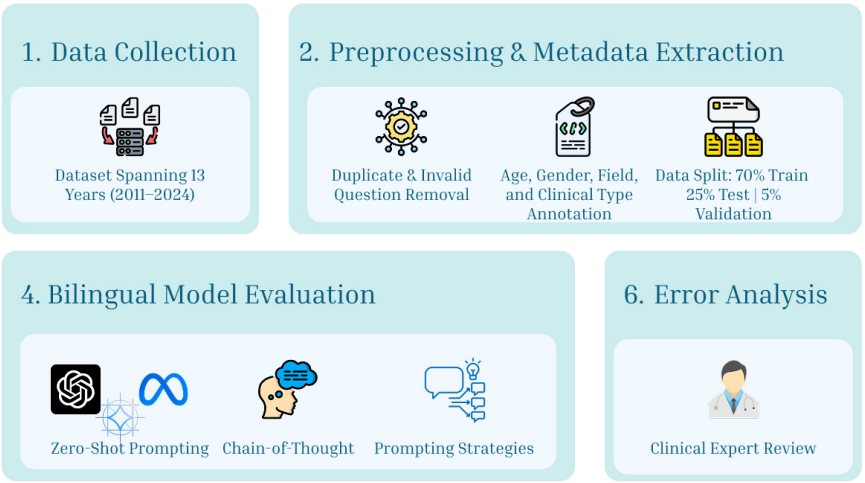

Figure 2: Overview of the PersianMedQA dataset construction process, including data collection, cleaning, annotation, and partitioning steps.

hal et al., 2022; Saab et al., 2024), their performance falls off significantly in other languages (Qin et al., 2025; Alonso et al., 2024). This gap is particularly pronounced in medicine, where high-quality corpora are centered on English, restricting the models' applicability in global healthcare settings. Importantly, simply translating questions is inadequate, as such pipelines can strip away critical terminology, subtle cultural cues, and localized standards of care, potentially leading to life-threatening consequences in clinical practice (Mehandru et al., 2022).

Medical practice is inherently shaped by contextual factors, including sociocultural, socioeconomic, regional, and healthcare system variables that extend beyond language translation (Kleinman, 1978; Betancourt et al., 2003). Clinical decision-making protocols and symptom interpretation vary significantly across healthcare systems and populations due to genetic variations, dietary patterns, climate-related health risks, and socioeconomic determinants, with the same clinical presentation potentially indicating different underlying pathologies across ethnic groups (Kirmayer, 2001; Wennberg, 2002; Risch et al., 2002; Zborowski, 1952). Additionally, vaccination schedules, drug availability, and standard-of-care protocols differ markedly between regions, making direct translation of medical guidelines inefficient. These considerations highlight why medical AI systems cannot rely solely on linguistic translation but must incorporate understanding of regional medical practices and population-specific health patterns.

These contextual complexities are particularly pronounced in low-resource language settings, where the intersection of linguistic barriers and distinct medical practices creates compounded challenges for AI evaluation. Limited research has investigated the specific factors that mislead LLMs in medical contexts, particularly in multilingual and low-resource language settings like Persian. A deeper investigation into the medical sub-fields in which LLMs excel or underperform is essential for identifying suitable use cases and implementing necessary safeguards.

To fill this gap, we introduce PersianMedQA, a large-scale, expert-annotated dataset covering 23 medical specialties. Given the scarcity of standardized Persian medical terminology resources, the dataset includes a comprehensive bilingual dictionary of Persian medical terms for consistent terminology usage during evaluation and model adaptation. As a benchmark, we evaluate state-of-the-art models, including general-purpose models, Persian fine-tuned models, and medical fine-tuned models on both original Persian questions and their English translations. Our experiments uncover a huge language gap: closed-source models such as GPT-4.1 significantly outperform open-source counterparts. Notably, Persian fine-tuned models exhibited minimal un-

derstanding of the Persian medical field 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.

Section 2 reviews prior work on medical QA benchmarks and Persian language models. Section 3 describes the PersianMedQA dataset construction. Section 4 presents our experimental setup and evaluations. Section 5 concludes with key findings and future research directions.

2 Related Works

Medical Question Answering Datasets and Multilingual Challenges. Medical question answering has emerged as a critical benchmark for evaluating machine reasoning capabilities in high-stakes healthcare domains. The field evolved from early information retrieval benchmarks (Athenikos and Han, 2010; Cao et al., 2011) to standardized datasets such as PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2020), and MedMCQA (Pal et al., 2022), driving domain-specific model development like BioBERT (Lee et al., 2019) and PubMedBERT (Gu et al., 2021). However, most benchmarks focus exclusively on English, creating significant evaluation gaps. While native-language datasets have emerged—including CMB (Wang et al., 2024), Huatuo-26M (Li et al., 2023), MedQA-SWE (Hertzberg and Lokrantz, 2024), FrenchMedMCQA (Labrak et al., 2022), and HeadQA (Vilares and Gómez-Rodríguez, 2019)—many frameworks rely on problematic "translate-test" methodologies that distort clinical terminology and miss culturally-specific practices (Jin et al., 2023; Liu et al., 2025). Recent multilingual efforts like MedExpQA (Alonso et al., 2024) show around 10% accuracy drops for non-English languages, with critical gaps remaining for low-resource languages requiring native evaluation approaches.

Large Language Models in Medical Practice. Large language models have transformed medical AI applications, with specialized models demonstrating remarkable capabilities on standardized medical examinations. Med-PaLM 2 (Singhal et al., 2023) achieved groundbreaking performance on the USMLE, while general-purpose models like GPT-4 (Team, 2023) showed impressive zero-shot performance across medical QA benchmarks (Nori et al., 2023). Recent advances include open-source models such as MEDITRON-70B (Chen et al., 2023), multilingual approaches like MMed-Llama 3 (Qiu et al., 2024) covering six languages, and specialized Chinese models such as TCM-Chat (Dai et al., 2024) and BianCang (Wei et al., 2024). However, systematic evaluation across diverse languages and clinical settings remains limited, particularly for morphologically rich and low-resource languages such as Persian.

Persian Language Models and Medical Applications. Persian natural language processing has witnessed significant progress with the development of robust monolingual models. ParsBERT (Farahani et al., 2021) established strong baselines for various Persian NLP tasks, consistently outperforming multilingual alternatives on sentiment analysis and text classification benchmarks. Recent advances include Dorna (Team, 2024b), a large-scale Persian language model. In the medical domain, SINA-BERT (Taghizadeh et al., 2021) represents an early attempt at Persian medical NLP, utilizing pre-training on large-scale medical corpora including both formal and informal medical texts from diverse online resources. Furthermore, existing Persian medical NLP efforts lack the expert validation and standardized evaluation protocols necessary for reliable clinical assessment, highlighting the need for comprehensive Persian medical QA benchmarks with rigorous validation procedures.

3 PersianMedQA Construction

The PersianMedQA dataset was developed by collecting 14 years of multiple-choice questions from the official Iranian residency and pre-residency medical exams, administered by Sanjeshp. Each exam was created by the official Iranian medical board and reflects real-world, high-stakes evaluation standards. Each item includes the question text, four answer

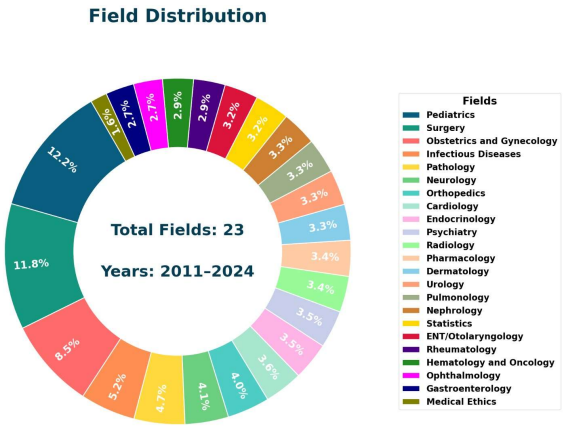

Figure 3: Distribution of medical fields in the dataset.

options, the correct answer key, and the medical field to which the question belongs. Figure 6 presents representative examples of clinical and non-clinical questions. The raw dataset underwent a rigorous preprocessing pipeline to ensure quality, consistency, and relevance for multilingual medical QA evaluation.

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 from the Language-agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2020) 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:** Most questions already contained subject tags from the original examination. For questions lacking subject tags, both medical specialists and Gemini 2.5-Flash independently classified them, achieving over 90% agreement. Final subject labels were determined through expert medical review to ensure high accuracy. The subject annotation interface used by the specialists is provided in Appendix H.
- **Domain Classification:** Questions were labeled as *clinical* (patient cases and diagnosis) or *non-clinical* (basic sciences and theoretical concepts). This classification was performed using Gemini 2.5-Flash and validated by a medical specialist.
- **Demographic Extraction:** Utilized Gemini 2.5-Flash to automatically extract patient attributes (e.g., age, gender) for every question, motivated by analyzing data distribution to ensure comprehensive representation across patient demographics and enable future research on potential LLM performance gaps in specific demographic subgroups.

Detailed information about the medical expert's qualifications and the complete demographic distributions is provided in Appendix B, Appendix C, and Appendix C.3.

3.3 Dataset Overview

The **PersianMedQA** dataset comprises 20,785 unique, expert-validated multiple-choice medical questions, collected over 14 years from Iranian national residency and pre-residency exams. Approximately 70% of the questions are classified as clinical, with the remaining 30% labeled as non-clinical. The items span 23 medical specialties, covering a broad range of topics relevant to medical education and practice.

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. Figure 3 summarizes the distribution of questions across medical domains.

3.4 Data Contamination and Evaluation Integrity

To ensure the reliability of our medical evaluation, we implemented multiple safeguards

against data contamination and memorization artifacts:

Secure Sourcing: The dataset is not from easily crawlable, free public websites. Questions come from official Iranian medical residency exams administered by Sanjeshp in PDF format, providing an additional layer of protection against training data leakage.

Exact Search: We conducted exact match searches on a limited sample (under 50 questions), including both full question stems and key medical terms. The analysis revealed **minimal overlap** with publicly available online sources, indicating **limited or no data leakage** into LLM training corpora.

Temporal Analysis: We conducted an accuracy analysis across examination years (2011-2024). Performance remained consistent even for 2024 and 2023 questions, which post-date training cutoffs for most models(see Appendix J).

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 (the complete list of models is available in the K). All models were prompted using identical instructions (provided in the F), 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. Figure 4 presents the overall accuracy of the evaluated models on both Persian and English test sets. Among all models, the closed-source GPT-4.1 achieved the highest zero-shot accuracy in both languages, scoring 83.09% in Persian and 80.71% in English. Notably, the best-performing open-source model, LLaMA-3.1-405B-Instruct, achieved a strong 67.02% in Persian and 73.49% in English. In terms of medical-tuned models, Meditron3-8B scored only 38.67% in Persian and 50.00% 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 24.22% in Persian (roughly equivalent to random guessing) and 25.17% 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 34.87% in Persian and 51.24% 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-Preview, showed minimal to no drop, indicating stronger cross-lingual 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 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 trained. This bilingual evaluation is crucial for understanding a key trade-off in multilingual medical AI: while translating questions into English may align them better with a model's core knowledge base, it risks erasing subtle clinical guidelines and cul-

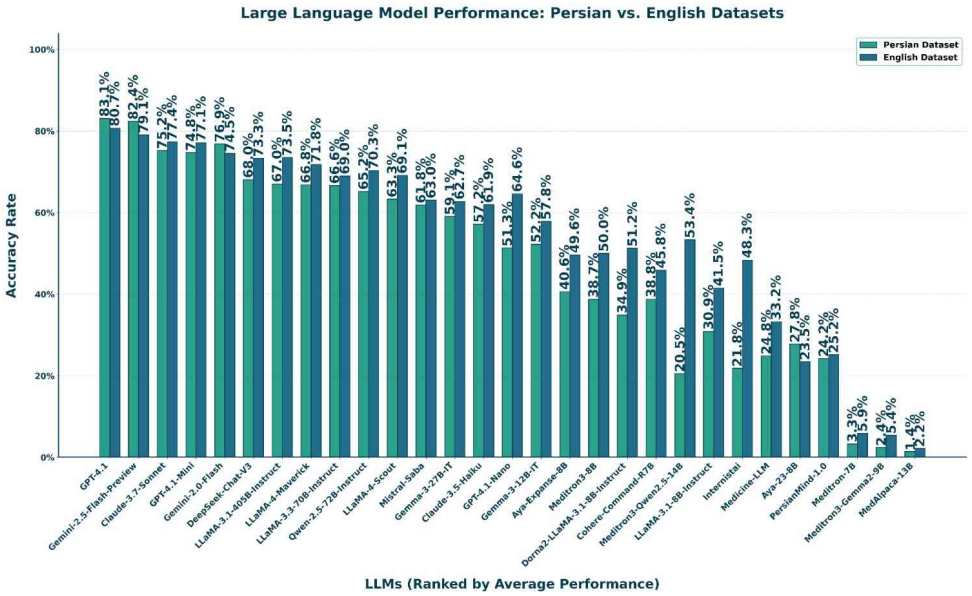

Figure 4: Overall accuracy of models on Persian and English test sets.

tural contexts unique to local practice. To quantify this effect, 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 analyzed performance based on whether questions were answered correctly in only Persian, only English, or both. As expected, most models performed better on the English translations—even those fine-tuned on Persian—reflecting their predominant exposure to English medical data during training. However, a substantial subset of questions (ranging from 3-10% across models) were answered correctly only in the original Persian. Manual analysis showed these cases often involve crucial local context that is lost in translation. This includes healthcare system-specific protocols where Iranian clinical guidelines differ from Western standards, population-specific clinical considerations like regional disease prevalence, and semantic drift where the precise meaning of Persian medical terms is altered. Such translation errors were most damaging in fields like pharmacology and surgery, where imprecise terminology led to incorrect answers even when the underlying reasoning was sound. Appendix M provides detailed examples illustrating these patterns.

Impact of Model Size. Our analysis of model size versus performance reveals that scale is not a universal solution. While larger general-purpose models like GPT-4.1 (83% accuracy) clearly outperformed their smaller counterparts, increased size offered no advantage for specialized models. Both large medical-specific (e.g., MedAlpaca-13B) and Persian-tuned (e.g., Dorna2-LLaMA-3.1-8B) models struggled significantly, often scoring below smaller general-purpose models. These results underscore that model scale must be paired with high-quality, domain-relevant training data to achieve strong performance. For a detailed visualization, see Appendix L.

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

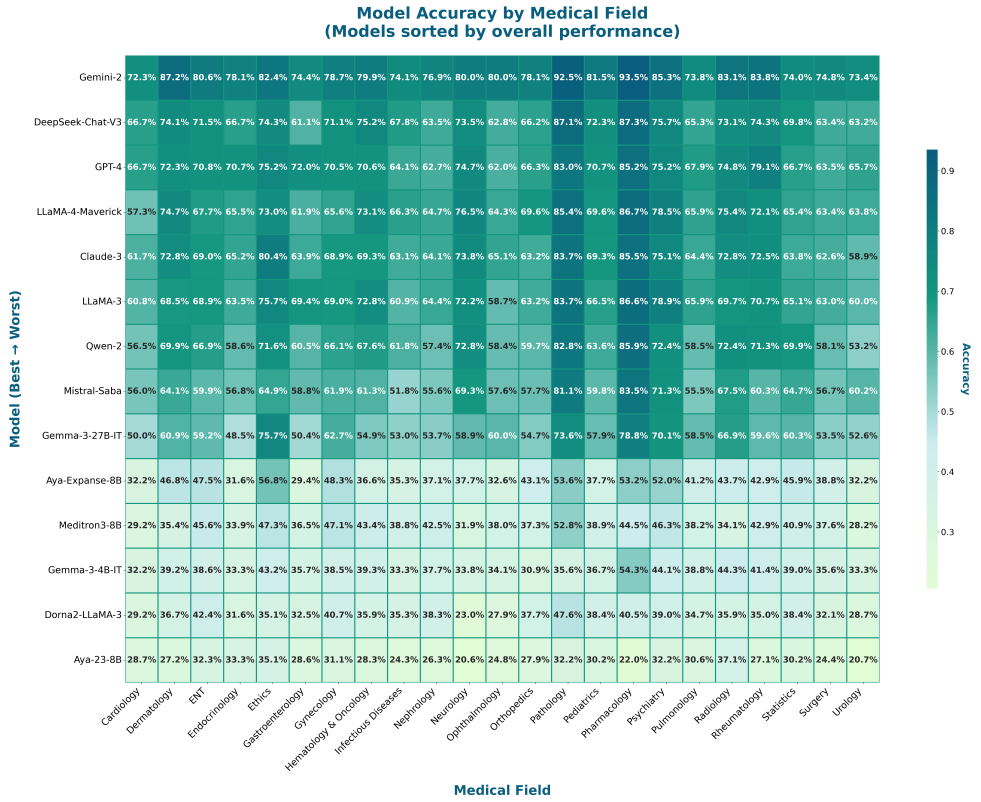

Figure 5: Heatmap showing the accuracy of each model across all medical specialties in the PersianMedQA dataset. Each cell represents the accuracy for a particular model-field pair. The overall performance of all models is available in Appendix K.

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 I) 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 statistical patterns in answer choices, we adopted the *partial-input* protocol of Balepur et al. (2024). Each model received *only* the four answer options without the question stem. Gemini-2.5-Flash-Preview achieved 35.60% accuracy, substantially outperforming random guessing (25%).

Manual inspection revealed that models exploit recurrent answer-choice artifacts, particularly evident in medical ethics—the highest-performing field at 46.8% accuracy. In ethics questions, models can infer correct answers through: (i) *hierarchical ethical principles*, where options containing phrases like "patient autonomy," "informed consent," or "professional disclosure" signal standard bioethical frameworks; (ii) *logically exclusive options*, where choices violating fundamental medical ethics (e.g., "withhold information from patient") can be discarded; and (iii) *linguistic*

cues, where options with formal ethical terminology indicate textbook-correct responses. These patterns suggest that high performance in ethics may reflect recognition of moral vocabularies rather than genuine ethical reasoning, warning that medical MCQ benchmarks may overstate LLM capabilities by permitting exploitation of answer-choice artifacts.

4.4 Model Ensembling

Different models exhibit varied strengths across medical subjects, suggesting that ensembling diverse models can improve accuracy. Since top-performing models like GPT and Gemini are not open source, developing open-source ensembles remains highly valuable.

Table 1: Majority-vote ensembles. "Δ_{best}" is the gain over the best single model in the group.

Ensemble / Baseline	Acc.	Avg. Acc.	Δ_{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	—	—

As shown in Table 1, ensembles of diverse model families outperform single models, whereas same-family ensembles offer little benefit. Notably, an ensemble of five open-source models achieved 73.7% accuracy, a significant gain over the best individual model in that group (+0.033).

4.5 Chain-of-Thought Evaluation

We evaluated the impact of Chain-of-Thought (CoT) prompting on four models: two large (GPT-4.1, Gemini-2.5-Flash), one medical (Meditron3), and one Persian (Dorna).

Performance Gains. For large models, CoT improved accuracy by approximately. 2%, with the greatest gains on clinical questions, suggesting it aids complex medical reasoning. Smaller models showed negligible improvement, likely due to weaker language and reasoning capabilities.

Expert Analysis of CoT Errors. A clinician's review of GPT-4.1's CoT responses identified four primary error types:

- **Contextual Mismatch:** Applying reasoning based on non-Iranian clinical protocols.
- **Ambiguity in Options:** Failing to distinguish between very similar or misleading answer choices.
- **Reasoning Failures:** Exhibiting illogical or incomplete reasoning despite possessing the required knowledge.
- **Knowledge Gaps:** Lacking the necessary factual information to answer correctly.

Examples for each category are available in Appendix E.

5 Conclusion

In this study, we present PersianMedQA, a dataset of 20,785 expert-validated Persian medical questions from 14 years of Iranian national medical exams, designed to evaluate how well current language models understand medical content across Persian and English contexts. Our evaluation of 40 models reveals a significant performance hierarchy: closed-source models like GPT-4.1 (83.1% Persian, 80.7% English) substantially outperform open-source alternatives, with the best open-source model (LLaMA-3.1-405B) achieving 67.0% in Persian. Persian fine-tuned models performed poorly (Dorna: 34.9%), while medical fine-tuned models showed only modest improvements over general models. Critically, our cross-linguistic analysis revealed that 3-10% of questions require Persian-specific cultural and clinical knowledge, demonstrating that translation-based evaluation approaches are insufficient for medical AI in non-English contexts. Future work should focus on developing Persian medical corpora for domain-specific training, creating retrieval-augmented systems that can access culturally appropriate medical guidelines, and expanding evaluation to other low-resource languages and multimodal medical contexts. This work establishes a foundation for culturally grounded medical AI evaluation beyond English-centric benchmarks.

Limitations

Several factors constrained this study. (i) *API restrictions:* cost and rate limits for commercial LLMs (e.g., GPT-4.1) 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

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model.

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.

Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Technical Report*.

Sofia J. Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question?

Joseph R. Betancourt, Alexander R. Green, J. Emilio Carrillo, and Owusu Ananeh-Firempong. 2003. Defining cultural competence: a practical framework for addressing racial/ethnic disparities in health and health care. *Public Health Reports*, 118(4):293–302.

Rishi Bommasani and et al. 2022. On the opportunities and risks of foundation models.

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

Josepha Campinha-Bacote. 2002. The process of cultural competence in the delivery of healthcare services: a model of care. *Journal of Transcultural Nursing*, 13(3):181–201.

Yu Guan Cao, Fei Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. In *Journal of biomedical informatics*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

Yizheng Dai, Xin Shao, Jinlu Zhang, Yulong Chen, Qian Chen, Jie Liao, Fei Chi, Junhua Zhang, and Xiaohui Fan. 2024. Tcmchat: A generative large language model for traditional chinese medicine. *Pharmacological Research*, 210:107530.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle,

Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Yu Gu, Robert Tinn, Hao Cheng, Michael 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*, 3(1):1–23.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2025. Medalpaca – an open-source collection of medical conversational ai models and training data.

Niclas Hertzberg and Anna Lokrantz. 2024. MedQA-SWE - a clinical question & answer dataset for Swedish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186, Torino, Italia. ELRA and ICCL.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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

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

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries.

Laurence J. Kirmayer. 2001. Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment. *The Journal of Clinical Psychiatry*, 62 Suppl 13:22–30.

Arthur Kleinman. 1978. Concepts and a model for the comparison of medical systems as cultural systems. *Social Science Medicine. Part B: Medical Anthropology*, 12:85–93.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2016–2025, New York, NY, USA. Association for Computing Machinery.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023.

Capabilities of gpt-4 on medical challenge problems.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

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.

Neil Risch, Esteban Burchard, Elad Ziv, and Hua Tang. 2002. Categorization of humans in biomedical research: genes, race and disease. *Genome Biology*, 3:comment2007.1.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

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.

DeepSeek-AI Team. 2024a. Deepseek llm: Scaling open-source language models with longtermism.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemma Team and Google. 2025. Gemma 3 technical report.

OpenAI Team. 2023. Gpt-4 technical report.

PartAI Research Team. 2024b. Dorna: A large-scale persian language model. Hugging Face Model Hub. Available at: `https://huggingface.co/PartAI/Dorna_Llama3.1_8B_Instruct`.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. Cmb: A comprehensive medical benchmark in chinese.

Sibo Wei, Xueping Peng, Yi fei Wang, Jiasheng Si, Weiyu Zhang, Wenpeng Lu, Xiaoming Wu, and Yinglong Wang. 2024. Biancang: A traditional chinese medicine large language model.

John E. Wennberg. 2002. Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ (Clinical Research Ed.)*, 325(7370):961–964.

Mark Zborowski. 1952. Cultural components in response to pain. *Journal of Social Issues*, 8(4):16–30.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction fine-tuned open-access multilingual language model.

Medical Question

Persian Question:

خانم 53 ساله به دلیل زخم ایجاد شده در ناحیه مچ پای چپ، مراجعه کرده است. وی اظهار میکند از چند سال قبل پیگمانتاسیون روی قوزک داخلی مچ پای چپ داشته و به مرور زخم بدون درد از دو ماه قبل در همین ناحیه ایجاد شده است. سابقه بیماری دیگری ندارد. او معلم است و اکثر اوقات در حالت ایستاده کار می کند. در معاینه، نبض اندامهای تحتانی به خوبی لمس میشوند. عروق واریسی سطحی در ساق پای چپ وجود دارد و در بیوپسی از محل ضایعه، به جز التهاب و noitareclu نکته دیگری ندارد. آنتی بیوتیک در بهبود زخم اثربخش نبوده است. کدام اقدام تشخیصی زیر مناسب تر است؟

English Translation: A 35-year-old woman presents with a wound in the left ankle area. She reports having had pigmentation over the medial malleolus of the left ankle for several years, and a painless wound gradually developed in the same area two months ago. She has no other medical history. She is a teacher and mostly works in a standing position. On examination, lower limb pulses are well palpable. There are superficial varicose veins in the left leg, and biopsy from the lesion site shows only inflammation and ulceration with no other findings. Antibiotics have not been effective in wound healing. Which of the following diagnostic measures is more appropriate?

Answer Options:

اسکن هسته‌ای با گلبول سفید نشاندار
Translation: (Nuclear scan with labeled white blood cells)
سی‌تی‌اسکن از محل ضایعه
Translation: (CT scan of the lesion site)
ام‌آرآی از محل ضایعه
Translation: (MRI of the lesion site)
سونوگرافی داپلر وریدی
Translation: (Venous Doppler ultrasound)

Question Attributes:

Specialty: (Surgery)
Question Year: 1,401 (2022)
Correct Answer: 4

Figure 6: Medical question sample from the dataset.

A Original Dataset

An example of a Persian dataset (Sanjesh exam questions) is provided here in Persian along with its translation. You can see it in Figure 6.

B Medical Specialist Background

The medical specialist involved in this study is a 30-year-old female internal medicine expert. She graduated from the University of Tehran with a degree in general medicine and completed her specialty training in internal medicine at Shahid Beheshti University of Medical Sciences. She is a board-certified clinician with 5 years of clinical practice and has successfully passed the Iranian national medical licensing examinations. Her contributions to this work include validating subject classification, checking translation accuracy, overseeing the dataset curation process, and performing chain-of-thought (CoT) error analysis.

C Data Verification and Quality Assurance

C.1 Answer Verification Process

All questions in our dataset undergo rigorous multi-level verification by the National Center for Medical Education Assessment (Sanjesh): (1) Initial expert committee review by board-certified medical professionals, (2) Public comment period where medical students and practitioners can report concerns or discrepancies, (3) Final review and correction process incorporating feedback from the medical community. While we do not have traditional inter-annotator agreement scores, this established verification process has ensured high accuracy across 14 years of national medical examinations.

C.2 Subject Classification Validation and Error Analysis

Subject classification was validated by one medical specialist (board-certified clinician) with 5 years of medical practice experience who has studied and successfully passed these Iranian medical licensing examinations. The specialist reviewed ambiguous cases using our Telegram interface, achieving greater than 90% agreement with Gemini 2.5-Flash classifications over a sample of 200 questions.

Error Analysis: The same medical specialist conducted a comprehensive error analysis of Chain-of-Thought (CoT) reasoning outputs from top-performing models. Through manual review of GPT-4.1 responses, the specialist identified four primary error categories: (1) Contextual mismatch where answers reflected non-Iranian clinical protocols, (2) Ambiguity handling failures when faced with similar answer choices, (3) Logical reasoning inconsistencies despite adequate knowledge, and (4) Fundamental knowledge gaps where CoT could not compensate for missing information. This expert analysis provided critical insights into model limitations and reasoning patterns in Persian medical contexts.

C.3 Question Diversity Distribution

The National Center for Medical Education Assessment (Sanjesh) maintains natural diversity across medical specialties following standardized distributions mandated by the Min-

istry of Health. The typical distribution includes: Internal Medicine (40 questions), Pediatrics (25 questions), Obstetrics/Gynecology (20 questions), Surgery (20 questions), with minor specialties including Neurology (9 questions), Pathology (9 questions), Pharmacology (6 questions), and other specialties following official guidelines. This distribution reflects the emphasis placed on different medical fields in Iranian medical education and practice.

D Demographic Distributions

Motivated by analyzing the data distribution to ensure sufficient and comprehensive representation across patient demographics, we systematically extracted demographic metadata from all questions in the dataset. This demographic analysis enables future research to identify potential LLM performance gaps in specific demographic subgroups and opens avenues for studying bias and fairness in medical AI systems. To extract this information, we experimented with both regular expressions and LLM-based approaches. The LLM-based extraction demonstrated consistently high accuracy on this task, outperforming the regex approach in terms of precision and recall across all demographic categories.

D.1 Gender distribution

Table 2: Distribution of patient gender across questions.

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

D.2 Age category distribution

Table 3: 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

D.3 Clinical vs. non-clinical distribution

Table 4: Distribution of clinical vs. non-clinical questions.

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

E 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 patient's 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.The model lacks pharmacologic mechanism knowledge and defaults to common treatments.*

F 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

- 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.

G CoT reasoning prompt

CoT Prompt

You are a medical expert taking a medical board examination.
For each question, please

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

Response format (JSON)

- `"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.

H 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.

H.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.

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

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

I Persian Medical Dictionary

We present an extracted Persian medical dictionary derived from the dataset. Table 5 summarizes, for each category file, the final number of unique medical terms extracted.

J Temporal Analysis

To ensure the absence of data leakage, we performed a temporal analysis across different years of exam data. Accuracy remained stable over time, especially after 2022, which marks the temporal cutoff used during data curation and evaluation.

However, we observed a noticeable dip in performance for some models on the 2020 and 2021 exams. This decline is likely attributable to the COVID-19 pandemic period, during which students had more time to study and exams were reportedly more challenging. This suggests that model performance variations may, in part, reflect changes in exam difficulty rather than data inconsistencies.

Table 5: 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

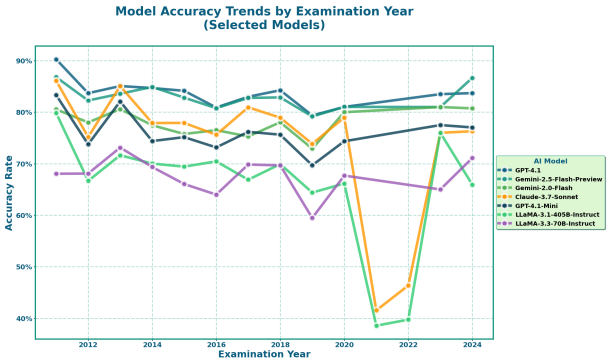

Figure 9: LLM Performance Across Exam Years

K Overall Performance Comparison

Table 6 shows performance on the original Persian questions, the translated English questions, and the average of the two scores. Models are sorted by their average performance. The five lowest-performing models struggled significantly with following instructions, resulting in accuracy scores worse than random guessing.

Table 6: Detailed zero-shot accuracy comparison of all evaluated models on the PersianMedQA test set.

Model	Persian (%)	English (%)	Average (%)
GPT-4.1	83.09	80.71	81.90
Gemini-2.5-Flash-Preview(Team, 2025)	82.37	79.09	80.73
Claude-3.7-Sonnet(Anthropic, 2024)	75.19	77.37	76.28
GPT-4.1-Mini	74.76	77.10	75.93
Gemini-2.0-Flash	76.86	74.50	75.68
DeepSeek-Chat-V3(Team, 2024a)	68.05	73.30	70.67
LLaMA-3.1-405B-Instruct(Dubey et al., 2024)	67.02	73.49	70.25
LLaMA-4-Maverick	66.79	71.75	69.27
LLaMA-3.3-70B-Instruct	66.63	68.96	67.80
Qwen-2.5-72B-Instruct(Bai et al., 2023)	65.17	70.26	67.72
LLaMA-4-Scout	63.29	69.12	66.21
Mistral-Saba	61.85	63.04	62.45
Gemma-3-27B-IT(Team and Google, 2025)	59.06	62.74	60.90
Claude-3.5-Haiku	57.16	61.94	59.55
GPT-4.1-Nano	51.32	64.59	57.95
Gemma-3-12B-IT	52.22	57.85	55.03
Qwen-2.5-7B-Instruct	39.99	58.29	49.14
Mixtral-8x22B-Instruct(Jiang et al., 2024)	36.78	60.85	48.82
Aya-Expanse-8B(Üstün et al., 2024)	40.60	49.58	45.09
Meditron3-8B	38.67	50.00	44.34
Mistral-Nemo	36.23	51.64	43.94
Meditron3-Qwen2.5-7B	37.62	50.06	43.84
Dorna2-LLaMA-3.1-8B-Instruct	34.87	51.24	43.06
Cohere-Command-R7B	38.77	45.84	42.30
Gemma-3-4B-IT	35.87	42.25	39.06
Mistral-7B-Instruct(Jiang et al., 2023)	28.74	47.44	38.09
LLaMA-3.2-3B-Instruct	29.43	45.13	37.28
Meditron3-Qwen2.5-14B	20.51	53.36	36.94
LLaMA-3.1-8B-Instruct	30.85	41.46	36.16
Internistai	21.85	48.34	35.09
Meditron3-Gemma2-2B	27.44	34.97	31.21
Medicine-LLM	24.85	33.21	29.03
BioMistral-7B(Labrak et al., 2024)	25.76	31.38	28.57
LLaMA-3.2-1B-Instruct	26.44	25.48	25.96
Aya-23-8B	27.77	23.47	25.62
PersianMind-1.0(Rostami et al., 2024)	24.22	25.17	24.69
MedAlpaca-7B(Han et al., 2025)	15.18	20.38	17.78
Meditron-7B	3.28	5.90	4.59
Meditron3-Gemma2-9B	2.41	5.39	3.90
MedAlpaca-13B	1.41	2.16	1.79
PersianLLaMA-13B(Abbasi et al., 2023)	0.00	0.00	0.00

L Model Size vs. Accuracy

We have plotted accuracy versus model size (some model sizes are estimated, as they are not publicly available); see Figure 10.

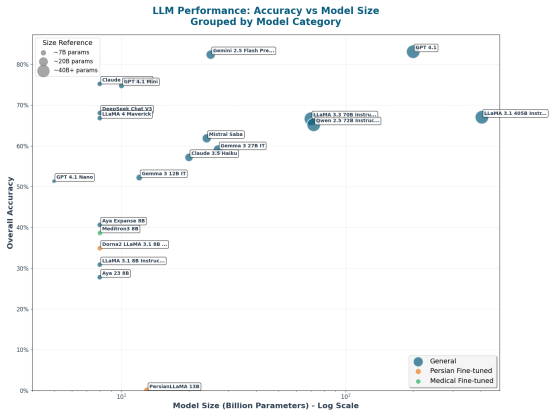

Figure 10: Relationship between model size and accuracy across different model types

M Cross-Linguistic Performance Analysis Examples

Representative Question Categories

Our cross-linguistic analysis revealed three distinct performance patterns, each reflecting different aspects of medical knowledge representation and cultural context preservation.

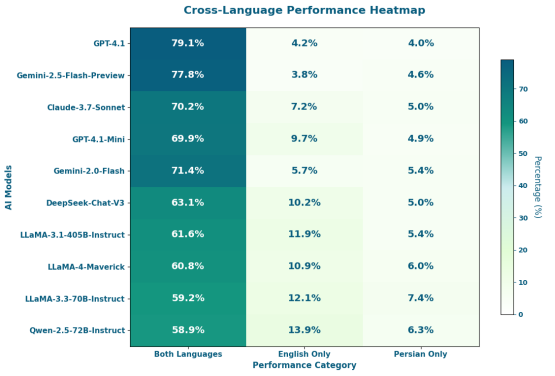

Figure 11: Cross-linguistic performance distribution for top 10 models.

Category 1: Correct in Both Languages

These questions demonstrate robust medical knowledge that transfers seamlessly across languages, typically involving standardized clinical protocols and universal pathophysiological concepts.

Example: Emergency Management Protocol

سوال فارسی: بیمار آقای مسنی به‌دلیل پانکراتیت صفراوی حاد بستری است. بعد از 42 ساعت درمان همچنان بیمار lli است. درد بیمار بهبود نیافته است. علائم انسداد صفراوی مشهود است. بهترین اقدام در مورد این بیمار کدام است؟

English Translation: "An elderly male patient is hospitalized due to acute biliary pancreatitis. After 24 hours of treatment, the patient is still ill. The patient's pain has not improved. Symptoms of biliary obstruction are evident. What is the best course of action for this patient?"

Analysis: Emergency biliary obstruction management follows universal protocols (ERCP, surgical intervention) consistent across healthcare systems.

Category 2: Correct Only After Translation (English-Only)

These questions benefit significantly from models' superior English medical training, particularly in highly specialized terminology and advanced clinical knowledge domains.

Example: Specialized Anatomical Pathology

سوال فارسی: خطر گسترش تنوسینوویت عفونی تاندون فلکسور کدام انگشت دست به ساعد وجود دارد؟

English Translation: "Which finger's flexor tendon sheath infection (infectious tenosynovitis) is at risk of spreading to the forearm?"

Analysis: Specialized anatomical terminology and hand surgery concepts are predominantly represented in English medical literature, improving performance post-translation.

Example: Population-Specific Clinical Considerations

سوال فارسی: نوجوان 71 ساله ای به علت تصادف به اورژانس آورده شده است. در معاینه له شدگی نسج زخم همراه با آلودگی با خاک مشاهده می شود. در سابقه اعلام می دارد که واکسیناسیون روتین کشوری را انجام داده است.

English Translation: "A 17-year-old adolescent has been brought to the emergency department due to an accident. Examination reveals crushed tissue with soil contamination in the wound. History indicates routine national vaccination has been performed."

Analysis: Iranian vaccination schedules and tetanus prophylaxis protocols differ from Western standards in timing and risk stratification.

Category 3: Correct Only in Persian

These questions involve Iran-specific medical practices, regional treatment protocols, or clinical contexts that are altered or lost during translation.

Example: Regional Treatment Protocols

سوال فارسی: موتورسوار 03 ساله ای بعد از تصادف در جاده با زخم حدود 01 سانتی متر روی استخوان ساق پای راست به همراه شکستگی خرد شده تیبیا به اورژانس مراجعه می کند. در درمان این بیمار از کدام آنتی بیوتیک و برای چه مدت استفاده می کنید؟

English Translation: "A 30-year-old motorcyclist presents to the emergency department after a road accident with a wound approximately 10 cm over the right tibia bone along with a comminuted tibia fracture. Which antibiotic and for how long would you use in the treatment of this patient?"

Analysis: Iranian trauma protocols reflect regional resistance patterns and drug availability, differing from Western guidelines in antibiotic selection and duration.