

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

Mohammad Javad Ranjbar Kalahroodi¹, Sepehr Karimi^{1*}, Amirhossein Sheikholeslami^{1*},
Sepideh Ranjbar Kalahroodi², Hesham Faili¹, Azadeh Shakery^{1,3}

¹School of Electrical and Computer Engineering, University of Tehran, Iran

²Shahid Beheshti University of Medical Sciences, Iran

³Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

{mohammadjranjbar, ah.sheikh, sepehrkarimi, hfaili, shakery}@ut.ac.ir

*These authors contributed equally to this work and are listed in random order.

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 41 state-of-the-art models, including general-purpose, Persian, and medical LLMs, in zero-shot and chain-of-thought (CoT) settings. Our results show that closed-weight 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 LLMs 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. The dataset, along with a bilingual medical dictionary, is publicly available at [PersianMedQA](#).

Keywords: Medical Question Answering, Persian Language Models, Low-Resource Languages, Multilingual Evaluation, Clinical NLP

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 those directly impacting human life, remains concerning (Bommasani 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 (Singhal et al., 2022; Saab et al., 2024), their performance drops 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, sim-

Medical Examples

Clinical:

A 48-year-old man presents with chest pain (4h), anterior ST-elevation, sweating, BP 90/60 mmHg, distended neck veins, and basal rales. Most effective treatment?

1. Fibrinolytic + emergency angioplasty if needed
2. Fibrinolytic only
3. Emergency angioplasty
4. Fibrinolytic + angioplasty after 48h

Answer: 3

Non-Clinical:

All of the following can cause acute retinal necrosis, *except*:

1. Cytomegalovirus
2. Herpes simplex type 1
3. Toxoplasmosis
4. Varicella Zoster

Answer: 3

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

ply 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 con-

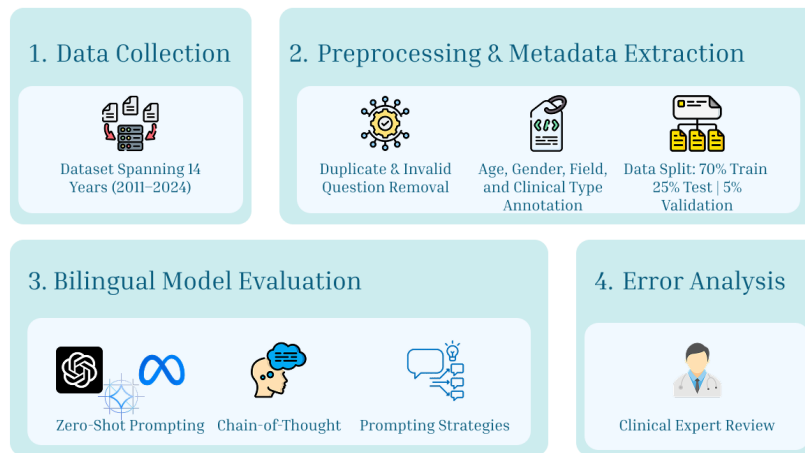


Figure 2: Overview of the PersianMedQA dataset construction process, including data collection, cleaning, annotation, and partitioning steps.

textual 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 an 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 termi-

nology usage during evaluation and model adaptation. As a benchmark, we evaluate state-of-the-art models, including general-purpose models, Persian LLMs, and medical LLMs on both original Persian questions and their English translations. Throughout this work, we use *medical LLMs* to refer to models that have undergone domain adaptation for medicine, either via domain-adaptive pre-training on biomedical corpora or supervised instruction fine-tuning on medical tasks. Similarly, *Persian LLMs* refers to models domain-adapted for Persian, via continued pre-training or Persian instruction fine-tuning. Our experiments uncover a substantial language gap: closed-weight models such as GPT-4.1 significantly outperform open-weight counterparts. Notably, Persian LLMs exhibited minimal understanding of the Persian medical field and performed the worst, while medical LLMs 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 Work

2.1. 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 has evolved from early in-

formation 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 (Jin et al., 2023; Liu et al., 2025) (the practice of machine-translating non-English inputs into English before evaluation, rather than evaluating in the source language) 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.

2.2. LLMs in Medical Practice

LLMs 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-weight 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.

2.3. Persian Language Models and Medical Applications

Persian natural language processing has witnessed significant progress with the development of robust monolingual models. PARSBERT (Fara-hani et al., 2021) established strong baselines for various Persian NLP tasks, consistently outperforming multilingual alternatives on sentiment analysis and text classification benchmarks. Recent

Field Distribution (2011-2024)

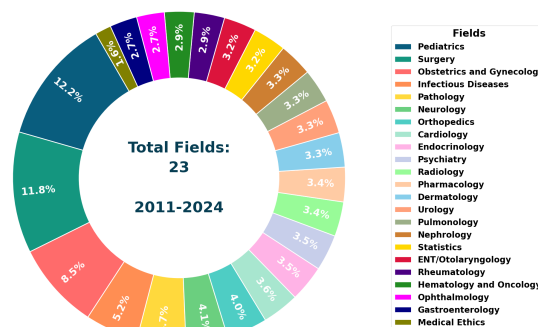


Figure 3: Distribution of medical fields in the dataset.

advances include DORNA (Team, 2024), 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 medical residency and pre-residency examinations administered by the Medical Education Assessment Center (Sanjeshp) under the Iranian Ministry of Health, the governmental body responsible for national standardized examinations in the medical domain, analogous to organizations such as the National Board of Medical Examiners (NBME) in the United States. These exams serve as the mandatory licensing gateway for medical graduates seeking specialist training in Iran, ensuring that all questions reflect real-world, high-stakes clinical evaluation standards at the national level. 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 options, the correct answer key, and the medical field to which the question belongs. Figure 1 presents representative examples of clinical and non-clinical questions. The raw dataset underwent a rigorous pre-processing 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:** All reference answers were provided by the official Iranian Medical Board and underwent a rigorous three-level verification process by the National Center for Medical Education Assessment (Sanjesh): (1) Initial review by board-certified medical professionals, (2) Public comment period where students and practitioners could report concerns or discrepancies, and (3) Final review and correction incorporating community feedback. In addition to this official validation, we manually removed questions with incorrect answer formatting (e.g., missing or mismatched option numbers) and those with multiple plausible correct answers to ensure evaluation reliability.

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 (i.e., the medical specialty labels shown in Figure 3, such as cardiology, surgery, or pediatrics) 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.
- **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:** We utilized Gemini 2.5-Flash to automatically extract patient attributes (e.g., age, gender) for every question, motivated by the need to analyze data distribution to ensure comprehensive representation across patient demographics and enable future research on potential LLM performance gaps in specific demographic subgroups.

All medical annotations and validations were performed by a board-certified internal medicine specialist with 5 years of clinical practice (see Appendix 8 for full details).

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.

We partitioned PersianMedQA into training (14,549), validation (1,000), and test (5,236) splits using stratified sampling based on year and field to facilitate future research, including potential fine-tuning studies, and to establish standardized evaluation protocols. Figure 3 summarizes the distribution of questions across medical domains.

Demographic Coverage: To ensure comprehensive representation and enable future bias analysis, we systematically extracted patient demographics using LLM-based extraction validated by medical experts. Table 1 presents the distribution across gender, age categories, and question types. The dataset demonstrates balanced representation across demographics, with substantial coverage of both male and female patients, diverse age groups from infants to adults, and a clinically-relevant mix of case-based and theoretical questions.

Table 1: Demographic distribution in Persian-MedQA dataset.

Category	Subcategory	Count
Gender	Male	5,590
	Female	5,831
	Unspecified	9,361
Age	Infant (0–1)	1,101
	Child (2–17)	2,675
	Adult (18+)	10,241
	Unspecified	6,765
Question Type	Clinical	14,724
	Non-Clinical	6,061

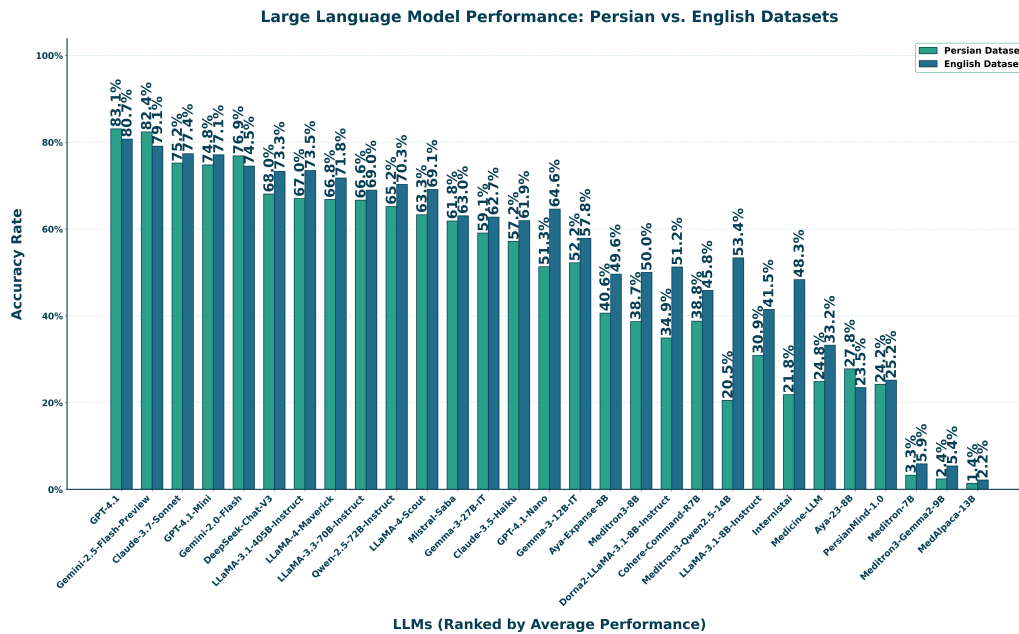


Figure 4: Overall accuracy of models on Persian and English test sets.

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 randomly sampled subset of under 50 questions, querying both full question stems and isolated key medical terms across publicly accessible web sources. The analysis revealed **minimal overlap**, with the vast majority of sampled questions returning no verbatim matches on any publicly accessible platform, indicating **limited or no data leakage** into LLM training corpora.

Temporal Analysis: To empirically verify the absence of data leakage, we performed a year-by-year accuracy analysis across examinations from 2011-2024 (Figure 5). Model performance remained consistent even on 2023-2024 questions that post-date training cutoffs for most evaluated models, strongly suggesting our dataset was not present in training corpora. The observed performance decrease in 2020-2021 likely reflects increased exam difficulty during the COVID-19 pandemic period rather than data inconsistencies, as students had extended study time and examiners reportedly created more challenging assessments.

3.5. English Translation

To enable bilingual evaluation, we generated English translations of the dataset using three methods: Google Translate, the GPT-4.1 API, and the Gemini-2.5-Flash API. All translations were assessed by our board-certified medical expert for accuracy of medical terminology and preservation of clinical meaning. Both GPT-4.1 and Gemini-2.5-Flash produced more accurate, natural translations than Google Translate, with expert validation confirming superior preservation of medical concepts and terminology. Due to its combination of quality, accessibility, and expert-validated accuracy, we use Gemini-2.5-Flash translations as our default in all subsequent experiments.

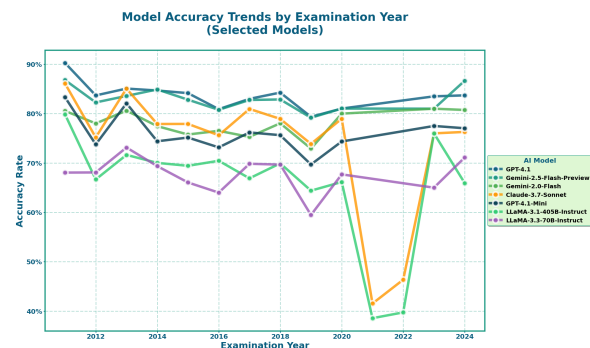


Figure 5: LLM performance across exam years (2011-2024).

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-weight and closed-weight LLMs in both Persian and English. All models were evaluated with temperature set to 0 and allowed to generate up to their maximum generation length. We employed identical prompts and evaluation protocols across all models to ensure fair comparison. Figure 6 shows the complete prompt template used for all zero-shot evaluations. We used English prompts across both Persian and English evaluation settings, as most models demonstrated superior instruction-following capabilities in English compared to Persian. All results reported in this study are based on the 5,236-question test set to facilitate reproducible evaluation and reduce computational costs. We validated this approach by running experiments on multiple models using the full dataset, confirming that results were consistent with the smaller test set.

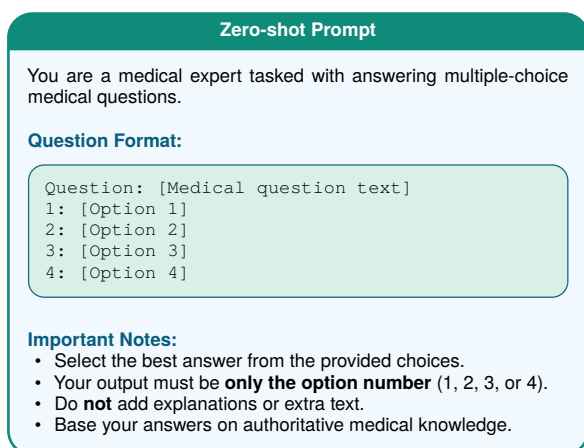


Figure 6: The zero-shot prompt used for evaluation

Figure 4 presents the overall accuracy of the evaluated models on both Persian and English test sets. Among all models, the closed-weight 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-weight model, DeepSeek-Chat-V3, achieved a strong 68.05% in Persian and 73.30% in English, followed closely by LLaMA-3.1-405B-Instruct with 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 LLMs significantly underperformed across the board; some of them suffered

greatly from poor instruction-following ability. PersianMind-1.0 achieved only 24.22% in Persian (roughly equivalent to random guessing) and 25.17% in English, suggesting limited medical knowledge and inadequate generalization capability in clinical domains. Similarly, Dorna2-LLaMA-3.1-8B-Instruct, another Persian LLM, 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-weight models consistently outperformed both open-weight and medical LLMs, 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 (83.09% Persian vs. 80.71% English) and Gemini-2.5-Flash-Preview (82.37% Persian vs. 79.09% English) — actually scored higher on the original Persian questions. This is consistent with our translation analysis: the English set is machine-translated from Persian, and even high-quality translation introduces subtle semantic drift and loss of clinical context that can slightly disadvantage models on the translated version.

We further analyzed model performance across different medical specialties. Figure 7 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.

4.2. 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 cultural contexts unique to lo-

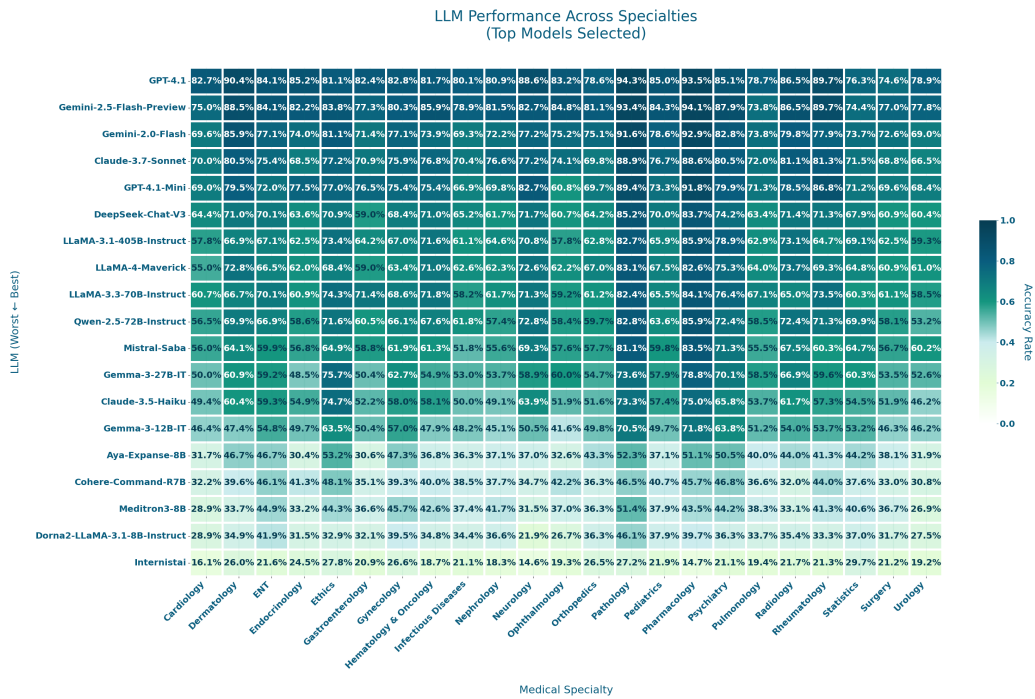


Figure 7: Heatmap showing the accuracy of each model across all medical specialties in the Persian-MedQA dataset. Each cell represents the accuracy for a particular model-field pair. The full model list is available in Appendix 15.

cal 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.

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 Persian LLMs—reflecting their predominant exposure to English medical data during training. However, a substantial subset of questions (ranging from 3-10% across models) was answered correctly only in the original Persian.

Manual analysis showed these cases often involved crucial local context that is lost in translation. These included 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. Figure 8 illustrates a representative case where Iran-specific antibiotic protocols lead to correct answers only when questions are presented in Persian, as translation obscures regional resistance patterns and drug availability that differ from Western guidelines.

Such translation errors were most damaging in fields like surgery, where imprecise terminology

led to incorrect answers even when the underlying reasoning was sound (see Appendix 16 for further examples).

4.3. Impact of Model Size

Our analysis of model size versus performance revealed 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.

4.4. 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.

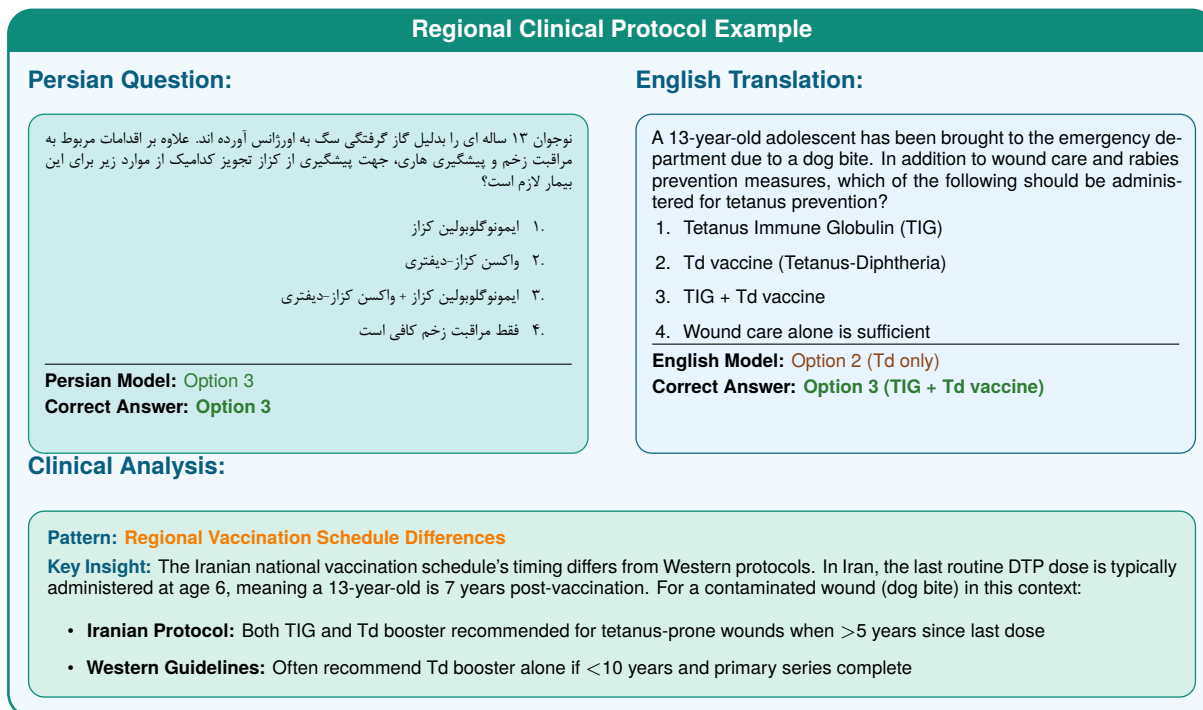


Figure 8: Example demonstrating regional vaccination protocol differences affecting tetanus post-exposure prophylaxis decisions.

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 is 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.

4.5. 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 exploited 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, suggesting that medical MCQ benchmarks may overstate LLM capabilities by permitting exploitation of answer-choice artifacts.

4.6. 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-weight, developing open-weight ensembles remains highly valuable.

Table 2: 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-weights	0.737	0.679	+0.033
Human Baseline	0.75	—	—

As shown in Table 2, majority-vote ensembles of diverse model families outperformed single models, whereas same-family ensembles offer little benefit. Notably, an ensemble of five open-weight models achieved 73.7% accuracy, a significant gain over the best individual model in that group (+0.033).

4.7. Chain-of-Thought Evaluation

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

Performance Gains: For large general-purpose models, CoT improved accuracy by approximately 2%, with the greatest gains on clinical questions, highlighting that clinical scenarios particularly benefit from explicit reasoning steps. Smaller models showed negligible improvement, likely due to weaker language and reasoning capabilities.

Expert Analysis of CoT Errors: A clinical expert reviewed GPT-4.1’s CoT responses and 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 10.

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 lan-

guage models understand medical content across Persian and English contexts.

Our evaluation of 41 models revealed a significant performance hierarchy: closed-weight models like GPT-4.1 (83.1% Persian, 80.7% English) substantially outperform open-weight alternatives, with the best open-weight model (LLaMA-3.1-405B) achieving 67.0% in Persian. Persian LLMs performed poorly (Dorna: 34.9%), while medical LLMs 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 simple translation-based evaluation approaches are inadequate 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. Additionally, the training split of PersianMedQA remains underutilized in this study; future work should explore fine-tuning open-weight models directly on this split and compare against the prompt-based baselines established here. This work establishes a foundation for culturally grounded medical AI evaluation beyond English-centric benchmarks.

6. Limitations

Several factors constrained this study. (i) *API restrictions and computational resources:* 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. Additionally, the lack of access to large-scale GPU infrastructure precluded fine-tuning experiments on the training split, which we leave for future work. (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. (iii) *Generalization scope:* The exam-based format of PersianMedQA is best suited for evaluating physician-level clinical reasoning. Results may not generalize to patient-focused question answering, such as informal health queries collected from online forums or patient communities, which represent a distinct and complementary evaluation setting.

7. Ethics Statement

This study involved the analysis and evaluation of LLMs on publicly available medical examination data. The dataset consists exclusively of multiple-choice examination questions from official Iranian medical licensing exams and contains no patient records, clinical notes, or any personally identifiable information. No anonymization was required, as the data is entirely de-identified by nature — all questions are impersonal exam items testing medical knowledge, not individual cases.

Regarding intellectual property, the questions are sourced from official Iranian national medical residency examinations administered by Sanjeshp (the Medical Education Assessment Center), a governmental body under the Iranian Ministry of Health. These examination papers are publicly released after each examination cycle and have been widely circulated by medical education institutions, test-preparation resources, and commercial study materials for over a decade. The dataset is used exclusively for non-commercial academic research purposes, consistent with established practices in the Persian NLP research community.

PersianMedQA is publicly available to support reproducible research and further development of Persian medical NLP. The dataset and bilingual medical dictionary are available at [PersianMedQA](#).

Our findings aim to improve the responsible deployment of language models in healthcare, especially for underrepresented languages. 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

- Iñigo Alonso, Maite Oronoz, and Rodrigo Aggeri. 2024. [Medexpqa: Multilingual benchmarking of large language models for medical question answering](#). *Artificial Intelligence in Medicine*, 155:102938.
- Sofia J. Athenikos and Hyoil Han. 2010. [Biomedical question answering: A survey](#). *Computer Methods and Programs in Biomedicine*, 99(1):1–24.
- 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, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khat-tab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Muniyikwa, Suraj Nair, Avaniika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Nieves, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 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.
- 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.
- 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.
- 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 health-care 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.
- 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 Malaikanan 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.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutarō Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaeckermann, 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 Schaeckermann, 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](#).
- OpenAI Team. 2023. [Gpt-4 technical report](#).
- PartAI Research Team. 2024. [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.

8. Medical Specialist Background

The medical specialist involved in this study is a board-certified internal medicine physician. 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 has 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.

9. Data Verification and Quality Assurance

9.1. Answer Verification Process

As described in Section 3, all questions underwent a rigorous three-level verification process 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, and (3) Final review and correction process incorporating community feedback. This established process has ensured high accuracy across 14 years of national medical examinations, though formal inter-annotator agreement scores in the traditional sense were not applicable given the institutional verification structure.

9.2. Subject Classification Validation

Subject classification was validated by the board-certified medical specialist described in Appendix 8. The specialist reviewed ambiguous cases using a custom Telegram-based annotation interface (see Appendix 13), achieving greater than 90% agreement with Gemini 2.5-Flash classifications over a randomly selected sample of 200 questions. We acknowledge that the absence of a second human annotator means formal inter-annotator agreement scores could not be computed; future work should include dual annotation on a representative subset to formally assess metadata quality.

9.3. Question Diversity Distribution

The National Center for Medical Education Assessment maintains natural diversity across medical specialties following standardized distributions mandated by the Ministry of Health. Each exam

is dominated by the core clinical specialties: Pediatrics (~26 questions), Surgery (~24 questions), and Obstetrics/Gynecology (~18 questions), followed by Infectious Disease (~10 questions) and Pathology (~9 questions). The remaining 18 specialties—including Cardiology, Neurology, Psychiatry, Orthopedics, Endocrinology, Dermatology, and others—each contribute between 2 and 8 questions per exam, following official Ministry of Health guidelines.

10. Examples of CoT Error Patterns

This section presents representative error patterns identified in model-generated CoT outputs, as annotated by our clinical expert. For each example, we highlight the clinical context, the correct answer, the model's response, and a summary of the expert's evaluation.

1. Contextual Mismatch

Question: What is the next step in an immunocompromised patient with nasal congestion and suspected invasive fungal sinusitis?

Correct Answer: Endoscopy and biopsy

Model Answer: Imaging (MRI) is needed before biopsy.

Expert 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 Answer: Hepatocellular carcinoma (HCC)

Model Answer: Metastasis is more common overall, so we choose that.

Expert Evaluation: 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 Answer: Blood culture → Lumbar puncture → Empiric antibiotics

Model Answer: CT scan should be done first due to immunosuppression.

Expert Evaluation: 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 Answer: Erythromycin

Model Answer: Metoclopramide is commonly used for gastroparesis, so we chose that.

Expert Evaluation: Model lacks pharmacologic mechanism knowledge and defaults to common treatments.

11. Few-shot Evaluation Prompt

In-context examples are drawn from the Persian-MedQA training split using LaBSE cosine similarity, TF-IDF, and random selection (up to $k = 5$).

Few-shot Prompt

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

In-context Examples:

```
Question: [Example question 1]
1: [Option 1] 2: [Option 2]
3: [Option 3] 4: [Option 4]
Answer: [Correct option number]
...
Question: [Example question k]
1: [Option 1] 2: [Option 2]
3: [Option 3] 4: [Option 4]
Answer: [Correct option number]
```

Now answer this question:

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

Important Notes:

- 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.

12. CoT Reasoning Prompt

CoT Prompt

You are a medical expert taking a medical board examination.

For each question, please:

- Read and understand the question carefully.
- Analyze the options (1–4) 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):

```
{
  "CoT": "your step-by-step reasoning",
  "Final_Answer": 1 | 2 | 3 | 4,
  "Reasoning": "concise justification"
}
```

Be methodical, precise, and thorough in your analysis. Your expertise as {english_specialty} is critical for answering these specialized questions correctly.

13. User Interfaces

To facilitate expert interaction throughout various phases of our study, we developed multiple Telegram-based interfaces to streamline collaboration with our medical specialist.

13.1. Subject Annotation Interface

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



Figure 9: Telegram interface for expert subject classification of ambiguous questions.

13.2. CoT Reasoning Interface

To analyze the reasoning behind model outputs, we designed an interface that presented the expert with a curated 200-question subset. For each question, the expert was asked to: (1) select whether a predefined reasoning category applied (e.g., domain knowledge, commonsense, causal inference), (2) optionally assign a new category if none fit, and (3) provide a brief explanation justifying the correct answer.



Figure 10: Telegram interface for expert annotation of reasoning categories and explanations.

14. Persian Medical Dictionary

Table 3 summarizes the number of unique medical terms extracted per category in the bilingual Persian medical dictionary released alongside the dataset.

Table 3: 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

15. Overall Performance Comparison

Table 4 shows the zero-shot accuracy of all 41 evaluated models on the original Persian questions, the translated English questions, and their average. Models are sorted by average performance. The five lowest-performing models struggled significantly with instruction-following, resulting in accuracy scores at or below random guessing.

Table 4: Zero-shot accuracy of all evaluated models. ^cGeneral Closed-Weight, ^oGeneral Open-Weight, ^mMedical, ^pPersian.

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

16. Cross-Linguistic Performance Analysis

Our cross-linguistic analysis revealed three distinct performance patterns across models. Representative examples for each category are provided below.

Category 1: Correct in Both Languages

These questions involve standardized clinical protocols and universal pathophysiological concepts that transfer seamlessly across languages.

Category 1: Emergency Management Protocol

English:	Persian:
An elderly male patient is hospitalized due to acute biliary pancreatitis. After 24 hours, the patient is still ill. Symptoms of biliary obstruction are evident. What is the best course of action?	بیمار آقای مسنی به دلیل پانکراتیت صفراوی حاد بستری است. بعد از ۲۴ ساعت درمان همچنان حال بیمار وخیم است. علامت انسداد صفراوی مشهود است. بهترین اقدام کدام است؟

Pattern: Universal Protocol — Emergency biliary obstruction management is consistent across healthcare systems, transferring correctly in both languages.

Figure 11: Category 1: correct in both languages.

Category 2: Correct Only After Translation

These questions benefit from the model's stronger English medical training, particularly in specialized terminology.

Category 2: Specialized Anatomical Pathology

English:	Persian:
Which finger's flexor tendon sheath infection (infectious tenosynovitis) is at risk of spreading to the forearm?	خطر گسترش تنوسینوویت عفونی تاندون فلکسور کدام انگشت دست به ساعد وجود دارد؟

Pattern: English Literature Dominance — Specialized anatomical terminology is predominantly represented in English medical literature, improving performance post-translation.

Figure 12: Category 2: correct only after translation.

Category 3: Correct Only in Persian

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

Category 3: Iran-Specific Clinical Protocols

Example A: Regional Antibiotic Protocols	
A 30-year-old motorcyclist presents with a ~10 cm wound over the right tibia with a comminuted fracture. Which antibiotic and for how long?	موتورسوار ۳۰ ساله ای با زخم ۱۰ سانتی متری روی ساق پای راست به همراه شکستگی خرد شده تبسیا به اورژانس مراجعه می کند. از کدام آنتی بیوتیک و برای چه مدت؟
Example B: Vaccination Schedule Differences	
A 17-year-old presents with a contaminated wound. History indicates routine national vaccination has been performed.	نوجوان ۱۷ ساله با زخم آلوده به خاک به اورژانس آمده. در سابقه واکسیناسیون روتین کشوری انجام داده است.

Pattern: Regional Protocol Differences — Iranian antibiotic and vaccination protocols differ from Western guidelines. Critical clinical context is lost in translation, leading to incorrect answers on the English version.

Figure 13: Category 3: correct only in Persian due to Iran-specific protocols.