PerMedCQA: Benchmarking Large Language Models on Medical Consumer Question Answering in Persian Language

Naghmeh Jamali[1] Milad Mohammadi[2] Danial Baledi[2] Zahra Rezvani[3] Hesham Faili[2]
[1]School of Computer Science, Institute for Research in Fundamental Sciences, Tehran, Iran.
[2]School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran.
[3] Department of Computer Science, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran.
naghme.jamali.ai@gmail.com, miladmohammadi@ut.ac.ir,
baledi.danial@gmail.com, z.rezvani@gmail.com, hfaili@ut.ac.ir

Abstract

Medical consumer question answering (CQA) is crucial for empowering patients by providing personalized and reliable health information. Despite recent advances in large language models (LLMs) for medical QA, consumer-oriented and multilingual resources—particularly in low-resource languages like Persian—remain sparse. To bridge this gap, we present **PerMedCQA**, the first Persian-language benchmark for evaluating LLMs on real-world, consumer-generated medical questions. Curated from a large medical QA forum, PerMedCQA contains 68,138 question-answer pairs, refined through careful data cleaning from an initial set of 87,780 raw entries. We evaluate several state-of-the-art multilingual and instruction-tuned LLMs, utilizing **MedJudge**, a novel rubric-based evaluation framework driven by an LLM grader, validated against expert human annotators. Our results highlight key challenges in multilingual medical QA and provide valuable insights for developing more accurate and context-aware medical assistance systems. The data is publicly available on https://huggingface.co/datasets/NaghmehAI/PerMedCQA

1 Introduction

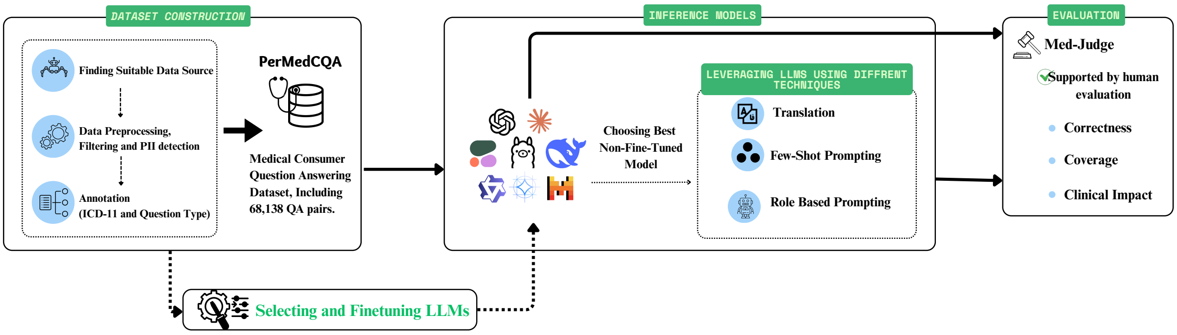
Recent advances in large language models (LLMs) have significantly enhanced the capabilities of Medical Question Answering (MQA) systems, facilitating rapid and reliable access to healthcare information and supporting clinical decision-making (Wang et al., 2024; He et al., 2025; Zheng et al., 2025). These systems have demonstrated impressive performance on standardized, exam-style questions predominantly within English-language contexts, significantly aiding clinicians and patients alike by providing timely, accurate medical knowledge (Meng et al., 2024; Shi et al., 2024; Tong et al., 2025). However, existing datasets and benchmarks primarily target structured, multiple-choice, or short-form question answering, often failing to capture the complexities and nuanced nature of real-world patient inquiries (Pal et al., 2022a; Manes et al., 2024; Kim et al., 2024). Additionally, the emphasis on high-resource languages, particularly English, leaves substantial gaps in linguistic and cultural diversity, posing significant limitations for truly inclusive and equitable healthcare AI systems (Tian et al., 2019; Daoud et al., 2025; Sviridova et al., 2024).

Addressing these gaps necessitates the development of consumer-oriented Medical QA datasets that authentically reflect informal, culturally specific, and real-world patient questions encountered in everyday healthcare interactions (Nguyen et al., 2023; Hosseini et al., 2024). Such resources are particularly scarce for low-resource languages like Persian (Farsi). To the best of our knowledge, no existing studies or datasets explicitly focus on Persian medical consumer question answering, highlighting a critical barrier to the advancement of equitable and culturally-aligned healthcare AI systems in this language.

To bridge this critical gap, we introduce **PerMedCQA**, the first large-scale Persian-language benchmark specifically designed for consumer-oriented medical question answering. PerMedCQA comprises 68,138 carefully curated question-answer pairs derived from real-world interactions collected from four prominent Persian online medical forums. These interactions involve authentic consumer-generated health queries answered by licensed medical professionals, encompassing a wide range of medical specialties and enriched with detailed metadata such as patient demographics, physician specialty, and platform source.

To ensure the dataset's reliability, clinical utility, and compliance with privacy standards, we implemented a rigorous two-stage data cleaning process. Initially, rule-based heuristics were applied to remove invalid entries, extremely short

Figure 1: Overview of PerMedCQA


DATASET CONSTRUCTION
Finding Suitable Data Source
Data Preprocessing, Filtering and PII detection
Annotation (ICD-11 and Question Type)
PerMedCQA
Medical Consumer Question Answering Dataset, Including 68,138 QA pairs.
INFERENCE MODELS
Choosing Best Non-Fine-Tuned Model
LEVERAGING LLMS USING DIFFRENT TECHNIQUES
Translation
Few-Shot Prompting
Role Based Prompting
EVALUATION
Med-Judge
Supported by human evaluation
Correctness
Coverage
Clinical Impact
Selecting and Finetuning LLMs

interactions, duplicates, and non-textual content, significantly enhancing the data's structural quality. Subsequently, we employed a LLM to label instances containing personally identifiable information (PII) such as names, phone numbers, addresses, and emails, further ensuring user privacy and ethical compliance. This comprehensive preprocessing pipeline resulted in a high-quality, deidentified dataset suitable for robust medical AI research.

To facilitate rigorous benchmarking and fine-grained analyses, we systematically annotated each QA pair along two critical dimensions: (1) disease categorization using the International Classification of Diseases, 11th Revision (ICD-11) (Khoury et al., 2017), comprising 28 distinct medical categories; and (2) classification into one of 25 standardized question types that capture the structural and semantic intent of patient inquiries (Abacha et al., 2019a). This extensive annotation significantly enhances PerMedCQA's utility for comprehensive evaluation and analysis of medical LLM capabilities.

Given the inherent challenges in evaluating open-ended medical QA tasks, where clinically acceptable responses often vary significantly in phrasing, we adopt an innovative evaluation framework utilizing a large language model (**Med-Judge**) as an automated grader. Validated through expert physician assessments, this rubric-driven system systematically compares model-generated responses against expert-provided answers, enabling nuanced evaluations beyond traditional lexical similarity metrics like BLEU or ROUGE.

We conducted extensive benchmarking experiments across diverse state-of-the-art language models, including both proprietary and prominent open-source variants. Robust baseline performance was established through zero-shot inference, followed by exploration of several inference-time enhancements—such as role-based conditioning and pivot translation—to improve response quality without parameter updates. Additionally, supervised fine-tuning experiments using parameter-efficient methods (LoRA) were performed on selected smaller models to assess the learnability and practical utility of PerMedCQA as a training resource.

Our key contributions can be summarized as follows:

- We introduce **PerMedCQA**, the first large-scale, real-world Persian medical QA benchmark, meticulously constructed through a rigorous two-step data cleaning process—rule-based filtering and LLM-based PII detection—and annotated comprehensively with ICD-11 categories and standardized question types, substantially addressing the resource gap for low-resource language medical QA while ensuring data quality, privacy, and ethical compliance.

- We present an automated yet clinically informed evaluation framework (**Med-Judge**), validated by expert physician reviews, providing reliable and nuanced assessments of open-ended medical question answering quality.

- We comprehensively benchmark a variety of proprietary and open-source LLMs, identifying substantial performance variations and demonstrating the effectiveness of prompt-based techniques for enhancing model outputs.

- We evaluate supervised fine-tuning strategies using PerMedCQA, highlighting their potential and limitations for improving smaller-scale model performance and clinical reliability.

An overview of the complete PerMedCQA workflow—from data collection and annotation through to model evaluation—is provided in Figure 1. Through these contributions, PerMedCQA establishes a critical foundation for future research aimed at developing trustworthy, culturally-sensitive, and linguistically inclusive medical AI systems specifically tailored to Persian-speaking populations.

2 Related Work

The rapid development of Large Language Models (LLMs) has significantly advanced the field of Medical Question Answering (Medical QA). While models such as GPT-4 (Nori et al., 2023), PaLM (Chowdhery et al., 2023), Mistral (Chaplot et al., 2023), and LLaMA (Touvron et al., 2023) have achieved impressive results on English-language Medical QA benchmarks, the progress of LLMs in non-English and consumer-focused medical domains remains underexplored. In this section, we review key datasets and modeling approaches that have shaped current research in Medical QA.

2.1 Medical Question Answering Dataset

The progress in Medical QA is closely tied to the availability of high-quality and diverse datasets. To evaluate clinical accuracy and factual consistency, LLMs have been tested on exam-style multiple-choice benchmarks such as MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019). These benchmarks feature questions framed in the style of medical licensing exams, focusing on factual recall. However, multiple-choice formats often fail to reflect the complexity and nuance of real-world medical inquiries, as they constrain responses to predefined options and limit models' ability to generate explanatory or contextual answers (Welivita and Pu, 2023). To address these limitations, recent benchmarks such as Medical Long-Form QA (Hosseini et al., 2024) and MedRedQA (Nguyen et al., 2023) prioritize practical utility and open-ended responses, aligning more closely with consumer-oriented healthcare needs.

While most Medical QA datasets are in English, efforts have been made to extend coverage to other languages. For Chinese, ChiMed (Tian et al., 2019), built from large online medical forums, serves as a benchmark for QA in Chinese. MedDialog (Zeng et al., 2020), comprising real-world, open-ended medical dialogues in both Chinese and English, has enabled progress in conversational medical systems via transfer learning. Huatuo-26M (Li et al., 2023a), a large-scale dataset with 26 million QA pairs sourced from Chinese encyclopedias, knowledge bases, and online consultations, has further boosted model training. More recently, MMedC (Qiu et al., 2024) was introduced as a multilingual medical corpus containing approximately 25.5 billion tokens across six languages, aiming to support the development of more capable and generalizable medical LLMs, particularly for low-resource languages. For Arabic, AraMed (Alasmari et al., 2024) provides QA pairs extracted from Al-Tibbi, an online doctor-patient discussion platform.

Several foundational datasets have been consolidated into comprehensive benchmarks for evaluating LLMs in medical QA. (Jin et al., 2019; Ben Abacha et al., 2019; Jin et al., 2021) MultiMedQA (Singhal et al., 2023b), for instance, aggregates six multiple-choice datasets—MedQA (USMLE), MedMCQA (Pal et al., 2022b), PubMedQA, LiveQA (Abacha et al., 2019b), MedicationQA (Abacha et al., 2019a), and MMLU clinical topics (Hendrycks et al., 2020)—along with the HealthSearchQA (Singhal et al., 2023a) dataset, to support a wide range of evaluation tasks. In the context of Retrieval-Augmented Generation (RAG), MIRAGE (Xiong et al., 2024) offers a standardized benchmark that combines subsets of MedQA, MedMCQA, PubMedQA, MMLU clinical topics, and BioASQ-YN (Tsatsaronis et al., 2015), enabling systematic evaluation of RAG techniques in medical QA.

2.2 Medical Large Language Models (LLMs in healthcare)

Recent years have seen the emergence of several domain-specific LLMs for healthcare, such as Med-Gemini (Saab et al., 2024), Med-PaLM (Singhal et al., 2023b), MedPaLM-2 (Singhal et al., 2025), BioMistral (Labrak et al., 2024a), PMC-LLaMa (Wu et al., 2024) and MMed-Llama 3 (Qiu et al., 2024), each demonstrating superior performance in medical reasoning and generation tasks. These advancements necessitate the development of more robust and nuanced benchmarks to thoroughly assess model capabilities.

The field has also expanded with fine-tuned medical LLMs such as HuatuoGPT (Zhang et al., 2023), BianQue (Chen et al., 2023), ClinicalGPT (Wang et al., 2023), DoctorGLM (Xiong et al., 2023), Chatdoctor (Li et al., 2023b), Baize-healthcare (Xu et al., 2023), zhongjing (Yang et al., 2024), Clinical Camel (Toma et al., 2023), and Me-LLaMA (Xie et al., 2024), many of which were trained on real-world clinical notes and patient-doctor dialogues. In parallel, multilingual medical models such as CareBot (Zhao et al., 2025), Medical-mT5 (García-Ferrero et al., 2024), ChiMed-GPT (Tian et al., 2024), and BiMediX (Pieri et al., 2024) have extended support for cross-lingual medical applications, increasing the global accessibility of medical AI systems.

Despite these advancements, There is a gap in the literature when it comes to real-world datasets and most open-ended Medical QA datasets remain limited to English, leaving a substantial gap in multilingual evaluation—especially for low-resource languages.

3 PerMedCQA Dataset Construction

To support the development and evaluation of medical consumer question answering systems in Persian, we introduce PerMedCQA — a large-scale dataset of real-world consumer health questions and expert answers, sourced from verified specialists across multiple public Persian forums. This section details the data collection process, preprocessing pipeline, automatic annotation, and benchmark construction.

3.1 Data Sources and Raw Collection

The initial dataset comprises 87,780 question-answer pairs collected from four major Persian-language health Q&A platforms: (DrYab), (HiSalamat), (GetZoop), and (Mavara-e-Teb). Each platform hosts a diverse set of verified physicians, collectively covering over 100 medical specialties. All data was collected from publicly accessible pages between November 10, 2022, and April 2, 2024, in compliance with ethical standards for public web data usage.

Each QA instance includes metadata fields such as `Title`, `Category` (user-assigned), `Physician Specialty`, `Age`, and `Sex`. All metadata across sources were standardized to a unified format, ensuring consistency. The dataset includes various QA interaction structures: (1) single-turn dialogues (DrYab, HiSalamat), (2) multi-turn conversations (GetZoop), and (3) cases with multiple expert responses (Mavara-e-Teb).

3.2 Preprocessing and Cleaning

We employed a two-stage data cleaning pipeline: rule-based preprocessing and large language model (LLM)–based processing (filtering and tagging).

Rule-based Filtering. Briefly, the following heuristics were applied across all data sources: (1) Entries without a valid user or assistant message were removed. (2) QA pairs where either message had fewer than three words, images, videos, or URLs were discarded. (3) Duplicate (user, assistant) pairs across different files were eliminated. These rule-based filters reduced the dataset from 87,780 to 73,416 instances, removing a total of 14,364 entries.

PII Detection. Given the sensitive nature of healthcare data, we further removed QA pairs containing personally identifiable information (PII), such as names, phone numbers, addresses, and emails. GPT-4o-mini was employed to detect any records containing PII with high accuracy. This stage resulted in the removal of 5,278 additional instances. The final cleaned dataset contains 68,138 QA pairs, resulting the final numbers QA pairs in PerMedCQA. Table 1 shows the distribution of QA pairs in terms of their data source.

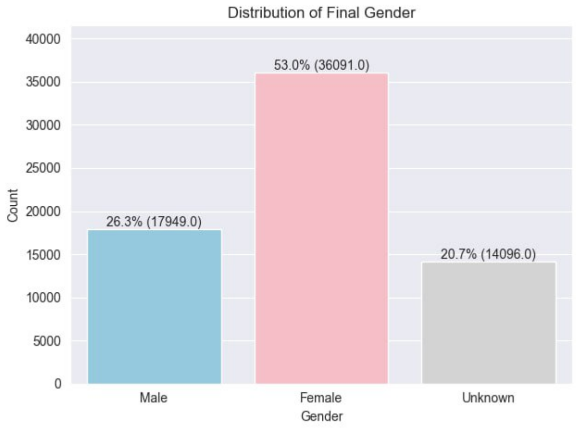
Table 1: Distribution of QA pairs in the data sources

Data Source	#QA Pairs	Percentage
DrYab	34427	50.5%
GetZoop	24006	35.2%
HiSalamat	5011	7.4%
Mavara-E-Teb	4692	6.9%

3.3 Annotation and Benchmark Splitting

ICD-11 Tagging. (Khoury et al., 2017) To facilitate medically meaningful categorization, the International Classification of Diseases 11th Revision (ICD-11) that consists of 28 categories, assigned to each QA pair by employing GPT-4o-mini. The prompt for ICD-11 classification and PII tagging showed in Figure 3 in Appendix A. This process yielded 28 distinct ICD-11 classes. Unlike user-assigned categories, ICD-11 provides a consistent and standardized taxonomy for disease classification. Figure 5 shows the distribution of ICD-11 categories in PerMedCQA.

Figure 2: Gender Distribution in PerMedCQA



Question Type Tagging. We categorized each QA pair into one of 25 predefined question types (Abacha et al., 2019a), enabling a deeper structural understanding of the dataset. The definition of each category illustrates in Table 3 along with some examples and the prompt of question type tagging is shown in Figure 7. Figure 6 shows the distribution of ICD-11 categories in PerMedCQA.

Extra Analysis based on Gender. Culture affects shaping the distribution of questions across ICD-11 disease categories 5, question types 6, and the gender of questioners. As shown in Figure 2, the gender-based distribution indicates that women and unknown constitute the majority of users who post questions on Persian Forums. The highest number of questions are allocated to the categories of sexual health (9,266 questions), digestive systems (6,868 questions), and Skin care and disease (5,478 questions), comprising roughly one-third of the entire dataset. In all three categories, the dominant "gender" of the questioners is "women", and the "question type" is predominantly "information". This observation highlights the substantial cultural influence on the prominence of these categories. For example, the tendency among women to ask about their sexual health concerns anonymously, seek home remedies for digestive issues, and obtain skincare advice without in-person consultation reflects cultural norms, explaining the higher participation of women in Persian-language medical forums.

Benchmark Split. To support model training and evaluation, PerMedCQA were split according to appropriate percent subsets for training, evaluation and test, partitioning into: Train set contains 64,280 instances, Eval set includes 345 instances (15 per ICD-11 category), and Test set comprises 3,513 instances (150 per ICD-11 category). Evaluation and test splits were stratified by ICD-11 category to ensure balanced representation across medical domains.

4 Experiments

This section outlines the experimental setup used to evaluate models on the PerMedCQA dataset. Due to the long-form (LF) nature of the task, we adopt a structured evaluation approach using a large language model as an automatic judge (Zheng et al., 2023), referred to as Med-Judge, used to assess model performance on the PerMedCQA dataset. To validate the reliability of this automatic evaluation protocol, we also conducted a human expert evaluation like (Hosseini et al., 2024). We then benchmarked a range of state-of-the-art language models using zero-shot inference to establish robust baselines. Based on this initial evaluation, the best-performing model in Persian for further analysis was chosen to apply a variety of advanced inference techniques (e.g, prompt-based strategy). Furthermore, to assess the quality and learnability of the PerMedCQA dataset, supervised fine-tuning (SFT) was applied to a set of language models.

4.1 LLM–based Evaluation with `Med-Judge`

Traditional evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) penalize legitimate paraphrases and thus fail for open-ended medical QA, where multiple correct answers may differ significantly in wording while still conveying clinically equivalent information. To address this limitation, we adopted **`Med-Judge`**, a LLM (Gemini-Flash-2.5) (Team et al., 2024)) prompted to evaluate model answers against expert references using a predefined rubric. Med-Judge is based on the criteria (Hosseini et al., 2024), comprising Correctness (Correct, Partially correct, Incorrect, Contradictory), Coverage (equal, model_subset, expert_subset, overlap_none), and Critical_Impact (Negligible, Moderate, Significant, Critical), explaining further details in A.1. Our full prompt of Med-Judge is shown in Figure 11 and 12 in section A.

4.2 Baseline Model Benchmarking

To establish robust performance benchmarks on the PerMedCQA dataset, we conducted zero-shot evaluations as an initial evaluation across a diverse suite of both proprietary and open-source language models. In total, we tested 16 models spanning a range of scales and training paradigms, primarily multilingual general-purpose models, along with a single biomedical-tuned variant. All evaluations were conducted on the fixed PerMedCQA test set, allowing for consistent comparisons across methods. The full list of evaluated models is summarized in Table 2.

Table 2: The list of LLM models in our experiments

Model Name	Affiliation
GPT-4.1, 4.1-mini, 4.1-nano (Achiam et al., 2023)	OpenAI
Claude 3.5 Haiku (Claude3.5)	Anthropic
Claude 3.7 Sonnet (Claude3.7)	Anthropic
Mistral-Saba (mistral-saba)	Mistral
Command A- 111B (Cohere et al., 2025)	Cohere
BioMistral-7B (Labrak et al., 2024b)	Mistral
DeepSeek-V3-670B (Liu et al., 2024)	DeepSeek
LLaMA 4 Scout-109B (LLaMA4Scout)	Meta
LLaMA 3.3 70B	Meta
LLaMA 3.1 8B (Grattafiori et al., 2024)	Meta
Gemma3 (Team, 2025) 4B, 12B, 27B	DeepMind
qwen3-14b (Yang et al., 2025)	Qwen

All models baseline experiments were evaluated under a consistent zero-shot setting using the same prompt. This prompt instructed the model to adopt the persona of a professional medical doctor answering in fluent Persian. The instructions emphasized direct, precise, and actionable guidance, discouraging excessive elaboration or default referrals unless medically necessary. Prompt structure and a representative example of the prompt is shown in Figure 8 in Appendix A.

4.3 Prompt-based Enhancement Methods

In addition to the baseline zero-shot evaluations, inference-time and non-parametric strategies were explored to aim the improvement of model performance without gradient updates or fine-tuning. These techniques were designed to enhance answer quality by modifying the input context or interaction style, while keeping the model parameters fixed. The three main techniques considered are described in the following paragraphs:

Pivot Translation for LLM Processing. (Tanaka et al., 2024) To leverage the strength of LLMs in English, we translated the Persian input question into English, requested an English-language response from the model, and then performed back-translation the answer to Persian. GPT-4.1 was employed for both directions, the prompt shows in Figure 10 in A. This method introduces latency but often enhances fluency, completeness, and structured reasoning, potentially compensating for model weaknesses in low-resource language handling.

Role-based Prompting. (Grabb, 2023) Each question was prepended with a system-level role prompt tailored to its ICD-11 category. For example, if the ICD-11 tag was related to mental disorders (tag 6), the model received the instruction: *"You are an experienced 'Psychiatrist' providing reliable..."*. This approach aimed to inject domain-specific priors and guide the model toward more specialized, context-aware answers. The prompt shows in Figure 9 in A.

Few-shot Prompting. (Maharjan et al., 2024) To test few-shot prompting, where each model was shown five randomly selected QA pairs from the PerMedCQA training set before answering the target question. The exemplars were selected without regard to topic simila rity to preserve generalizability. This strategy aimed to improve format consistency, response completeness, and adherence to clinical style by offering implicit demonstration of expected output structure. These techniques were applied to the best-performing model from the baseline evaluation, selected based on MedJudge scores. Their effects discussed in Section 5.

4.4 Supervised Fine-Tuning

To assess the learnability and utility of the PerMedCQA dataset, supervised fine-tuning (SFT) experiments were conducted on three language models, comprising Gemma 4B, LLaMA 3.1 8B, and BioMistral 7B. In terms of accessibility, architectural diversity, and suitability for efficient instruction tuning in resource-constrained environments, these models were selected. The training was conducted on the full PerMedCQA training set for 1 epoch, comprising 64,279 QA pairs. Moreover, despite lacking native support for Persian, BioMistral was included due to its specialization in biomedical and clinical domains. We employed LoRA-based (Low-Rank Adaptation) parameter-efficient fine-tuning (Hu et al., 2022), using LLaMaFactory framework (Zheng et al., 2024). LoRA configuration included a rank of 8 and $\alpha = 16$, with a learning rate of 2×10^{-5} and context length of 2048 tokens.

Rather than aiming for state-of-the-art performance, these fine-tuning experiments were primarily designed to evaluate the effectiveness of PerMedCQA as a training resource, and to explore how much medical QA capability could be learned by modestly sized models through supervised instruction tuning.

5 Results and Analysis and Conclusion

This section is currently under development and will be included in a future revision.

Limitations

Despite the comprehensive scale and rigorous cleaning of PerMedCQA, our study faces several limitations. The dataset, while large, is derived from a limited set of public Persian medical forums, which may introduce topical and demographic biases and restrict generalizability beyond Persian-speaking communities or to clinical domains not well represented in our sources. Furthermore, while we implemented robust quality control—including automated PII removal and human expert validation—our human evaluation was limited by the availability of clinical experts, restricting the depth and diversity of manual review. Finally, the dataset focuses on text-only QA, excluding cases requiring visual information, which are common in real-world healthcare settings and present important avenues for future work.

Ethical Considerations

Throughout this work, we prioritized privacy, fairness, and legal compliance. All data were collected exclusively from publicly accessible forums, and a thorough multi-stage pipeline was applied to remove any personally identifiable information (PII), ensuring no private patient or clinician data is present in the released dataset. We confirmed that data collection practices respected the terms and conditions of the source websites, and we will respond to takedown requests as needed. While releasing PerMedCQA, we recognize the risk of perpetuating cultural or topical biases inherent in consumer-generated content, as well as the potential for LLMs to generate incorrect or misleading medical advice. As such, both the dataset and any models fine-tuned with it are intended strictly for research purposes and should not be used as a substitute for professional medical care. We strongly recommend that downstream applications include clear disclaimers, robust safety measures, and, when possible, human expert oversight to prevent harm from inaccurate or inappropriate responses.

References

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019a. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 25–29. IOS Press.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019b. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. 2024. AraMed: Arabic medical question answering using pretrained transformer language models. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 50–56, Torino, Italia. ELRA and ICCL.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

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, and 1 others. 2023. 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, william el sayed. *arXiv preprint arXiv:2310.06825*.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, and 1 others. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Claude3.5. 2025. https://www.anthropic.com/news/claude-3-family.

Claude3.7. 2025. https://api.semanticscholar.org/CorpusID:276612236.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, and 1 others. 2025. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

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

DrYab. 2025. https://doctor-yab.ir/faq/.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, and 1 others. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. In *LREC-COLING 2024-2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Gemini-Flash-2.5. 2025. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash.

GetZoop. 2025. https://getzoop.com/doctors/online.

Declan Grabb. 2023. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*, 6.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

HiSalamat. 2025. https://www.hisalamat.com.

Pedram Hosseini, Jessica M Sin, Bing Ren, Bryceton G Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. *arXiv preprint arXiv:2411.09834*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Brigitte Khoury, Cary Kogan, and Sariah Daouk. 2017. *International Classification of Diseases 11th Edition (ICD-11)*, pages 1–6. Springer International Publishing, Cham.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. MedExQA: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024a. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard

Dufour. 2024b. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

LLaMA4Scout. 2025. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-qa: A real-world medical q&a benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294.

Mavara-e-Teb. 2025. `https://mavarateb.com`.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, and 1 others. 2024. The application of large language models in medicine: A scoping review. *Iscience*, 27(5).

mistral-saba. 2025. `https://mistral.ai/news/mistral-saba`.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022a. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022b. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual medical mixture of experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002, Miami, Florida, USA. Association for Computational Linguistics.

PII. 2025. `https://www.dol.gov/general/ppii`.

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

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023b. Publisher correction: Large language models encode clinical knowledge. *Nature*, 620(7973):19–19.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. Casimedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures. *arXiv preprint arXiv:2410.05235*.

Yudai Tanaka, Takuto Nakata, Ko Aiga, Takahide Etani, Ryota Muramatsu, Shun Katagiri, Hiroyuki Kawai, Fumiya Higashino, Masahiro Enomoto, Masao Noda, and 1 others. 2024. Performance of generative pretrained transformer on the national medical licensing examination in japan. *PLOS Digital Health*, 3(1):e0000433.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team. 2025. Gemma 3.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. ChiMed-GPT: A Chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173, Bangkok, Thailand. Association for Computational Linguistics.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. ChiMed: A Chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, Florence, Italy. Association for Computational Linguistics.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Ran Tong, Ting Xu, Xinxin Ju, and Lanruo Wang. 2025. Progress in medical ai: Reviewing large language models and multimodal systems for diagonosis. *AI Med*, 1(1):165–186.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Jinqiang Wang, Huansheng Ning, Yi Peng, Qikai Wei, Daniel Tesfai, Wenwei Mao, Tao Zhu, and Runhe Huang. 2024. A survey on large language models from general purpose to medical applications: Datasets, methodologies, and evaluations. *arXiv preprint arXiv:2406.10303*.

Anuradha Welivita and Pearl Pu. 2023. A survey of consumer health question answering systems. *Ai Magazine*, 44(4):482–507.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 others. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Lulu Zhao, Weihao Zeng, Xiaofeng Shi, and Hua Zhou. 2025. Carebot: A pioneering full-process open-source medical language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26039–26047.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. 2025. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, 16(2):1015–1040.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Med_Jugde

For every test instance, the MedJudge LLM receives the user's question, the expert-provided "gold" answer, and the model-generated answer. The LLM is strictly instructed to evaluate based solely on the expert reference, without drawing from external sources or its own medical knowledge. The output is a structured JSON object with the following fields:

- **Brief_analysis**: A concise comparison highlighting key similarities or differences between expert and model answers.
- **Key_missing_facts**: Important medical facts present in the expert answer but missing from the model answer (in Persian).
- **Key_extra_facts**: Additional details present in the model answer but absent from the expert answer (in Persian).
- **Correctness**: Assessment of factual and clinical consistency: *Correct*, *Partially_correct*, *Incorrect*, or *Contradictory*.
- **Coverage**: Degree of factual overlap: *Equal*, *Model_subset*, *Expert_subset*, or *Overlap_none*.
- **Clinical_impact**: Estimated clinical significance of any discrepancies: *Negligible*, *Moderate*, *Significant*, or *Critical*.
- **Judge_confidence**: The LLM's self-reported confidence (*High*, *Medium*, *Low*).

A.2 Label Definitions, Scoring, and Reliability

Label Definitions Each output field is defined as follows:

- **Correctness**:
  1. *Correct*: Clinically equivalent; no meaningful differences.
  2. *Partially_correct*: Minor deviations, no significant clinical impact.
  3. *Incorrect*: Substantial differences that affect accuracy or completeness.
  4. *Contradictory*: Model advice directly conflicts with the expert reference.
- **Coverage**:
  1. *Equal*: Both answers contain the same key facts.
  2. *Model_subset*: Model omits critical facts present in the expert answer.
  3. *Expert_subset*: Model introduces relevant facts not found in the expert answer.
  4. *Overlap_none*: No substantial factual overlap.
- **Clinical_impact**:

1. *Negligible*: No effect on care or understanding.
2. *Moderate*: Slight effect on treatment or comprehension.
3. *Significant*: Likely to affect recommendations or outcomes.
4. *Critical*: May result in unsafe or harmful guidance.

Reliability To assess the reliability of MedJudge, we compared its labels against blinded ratings from board-certified physicians on a 100-item subset. Agreement on the primary dimension—*correctness*—was 75% (collapsed to "acceptable" vs. "problematic"), with quadratic Cohen's $\kappa = 0.42$ (95% CI 0.19–0.58) and $F_1 = 0.82$. Ordinal rank-correlation was significant ($\tau = 0.41$, $p < 0.001$), indicating that higher human scores were mirrored by the LLM. For *clinical-impact*, MedJudge achieved 78% accuracy and detected one-third of truly high-impact discrepancies ($\kappa = 0.07$; prevalence-adjusted).

A.3 Prompt Engineering

Prompt engineering is considered crucial for the evaluation of LLMs. In this section, the prompts used at various stages of the work are presented. Figures 3 and 4 illustrate the task instructions provided for ICD-11 classification and PII tagging during the MedPerCQA stage. As shown in Figure 7, question type tagging across 25 categories was performed using a dedicated prompt. In addition, a wide range of prompts was employed in extensive experiments, as illustrated in Figures 8, 9, and 10. Finally, the structured output schema used in the Med-Judge evaluation pipeline is shown in Figure 13, enabling reliable parsing and analysis of model responses; the prompt design incorporates a chain-of-thought strategy to guide the LLM through consistent reasoning steps.

Figure 3: Task instructions for ICD-11 classification and PII tagging.

ICD-11 and PII Tagging System Prompt

Task Description
You are a medical expert tasked with classifying and analyzing patient–expert dialogue. Your task has two main parts:

1) ICD-11 Classification
Classify the content based on ICD-11 categories using the catalogue provided in the next page. Return only the integer (1–28) that best corresponds to the core subject matter.

2) PII Detection
Check if any personal information of the patient or expert is exposed in the messages. Personal information includes examples such as real name, address, phone number, email, etc. If such information exists, set `"identity"` to `true`; otherwise, set it to `false`.

Figure 4: Standardized ICD-11 classification codes used for QA annotation.

ICD-11 Category Catalogue (28-class taxonomy)

1. (1A00–1H0Z) Certain infectious or parasitic diseases
2. (2A00–2F9Z) Neoplasms
3. (3A00–3C0Z) Diseases of the blood or blood-forming organs
4. (4A00–4B4Z) Diseases of the immune system
5. (5A00–5D46) Endocrine, nutritional or metabolic diseases
6. (6A00–6E8Z) Mental, behavioural or neurodevelopmental disorders
7. (7A00–7B2Z) Sleep-wake disorders
8. (8A00–8E7Z) Diseases of the nervous system
9. (9A00–9E1Z) Diseases of the visual system
10. (AA00–AC0Z) Diseases of the ear or mastoid process
11. (BA00–BE2Z) Diseases of the circulatory system
12. (CA00–CB7Z) Diseases of the respiratory system
13. (DA00–DE2Z) Diseases of the digestive system
14. (EA00–EM0Z) Diseases of the skin
15. (FA00–FC0Z) Diseases of the musculoskeletal system or connective tissue
16. (GA00–GC8Z) Diseases of the genitourinary system
17. (HA00–HA8Z) Conditions related to sexual health
18. (JA00–JB6Z) Pregnancy, childbirth or the puerperium
19. (KA00–KD5Z) Certain conditions originating in the perinatal period
20. (LA00–LD9Z) Developmental anomalies
21. (MA00–MH2Y) Symptoms, signs or clinical findings, not elsewhere classified
22. (NA00–NF2Z) Injury, poisoning or other consequences of external causes
23. (PA00–PL2Z) External causes of morbidity or mortality
24. (QA00–QF4Z) Factors influencing health status or contact with health services
25. (RA00–RA26) Codes for special purposes
26. (SA00–SJ3Z) Traditional Medicine Conditions – Module I
27. (VA00–VC50) Functioning assessment
28. (XA0060–XY9U) Extension Codes

Table 3: Question Type Categories.

Question Type	**Definition**	**Example**
Information	Asks for general identification or classification of a drug.	What type of drug is amphetamine?
Dose	Queries recommended or safe dosage.	What is a daily amount of prednisolone eye drops to take?
Usage	Seeks instructions on how to take/administer a drug.	How to self inject enoxaparin sodium?
Side Effects	Asks about adverse reactions.	Does benazepril aggravate hepatitis?
Indication	Asks why/for which condition the drug is prescribed.	Why is pyridostigmine prescribed?
Interaction	Concerns compatibility with another drug/-substance.	Can I drink cataflam when I drink medrol?
Action	Mechanism of action or physiological effect.	How does xarelto affect homeostasis?
Appearance	Asks about physical look (colour, shape, imprint).	What color is 30 mg prednisone?
Usage/Time	Best time of day to take the medicine.	When is the best time to take lotensin?
Stopping/Tapering	How to discontinue or taper.	How to come off citalopram?
Ingredient	Active ingredient(s) contained in a product.	What opioid is in the bupropion patch?
Action/Time	Onset/duration of drug effect.	How soon does losartan affect blood pressure?
Storage and Disposal	Proper storage temperature or disposal method.	In how much temp should BCG vaccine be stored?
Comparison	Compares two therapies/drugs.	Why is losartan prescribed rather than a calcium channel blocker?
Contraindication	Whether the drug is safe given allergy/condition.	If I am allergic to sulfa can I take glipizide?
Overdose	Consequences of taking too much.	What happens if your child ate a Tylenol tablet?
Alternatives	Asks for substitute medications.	What medicine besides statins lowers cholesterol?
Usage/Duration	How long treatment should continue.	How long should I take dutasteride?
Time (Other)	Other time-related effectiveness/protection questions.	How long are you protected after the Hep B vaccine?
Brand Names	Asks for commercial brand names.	What is brand name of acetaminophen?
Combination	How to combine two treatments in one regimen.	How to combine dapagliflozin with metformin?
Pronunciation	How to pronounce a drug name.	How do you pronounce Humira?
Manufacturer	Asks who makes or markets the drug.	Who makes nitrofurantoin?
Availability	Whether the drug is still on the market/shortages.	Has lisinopril been taken off the market?
Long-term-Consequences	Long-term effects of prolonged use.	What are the long-term consequences of using nicotine?

Figure 5: Distribution of ICD-11 Categories in PerMedCQA

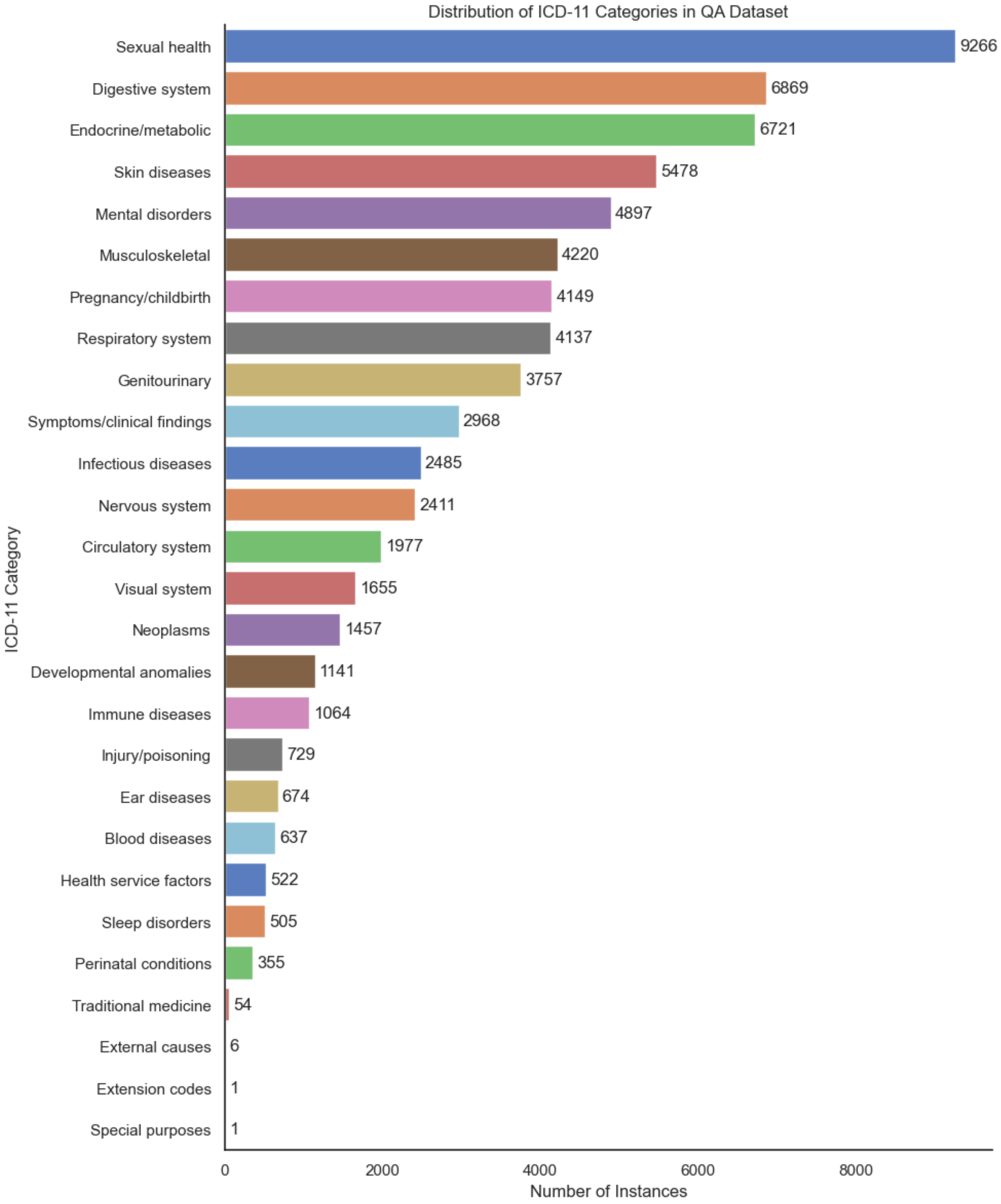


Figure 6: Distribution of Question Type in PerMedCQA

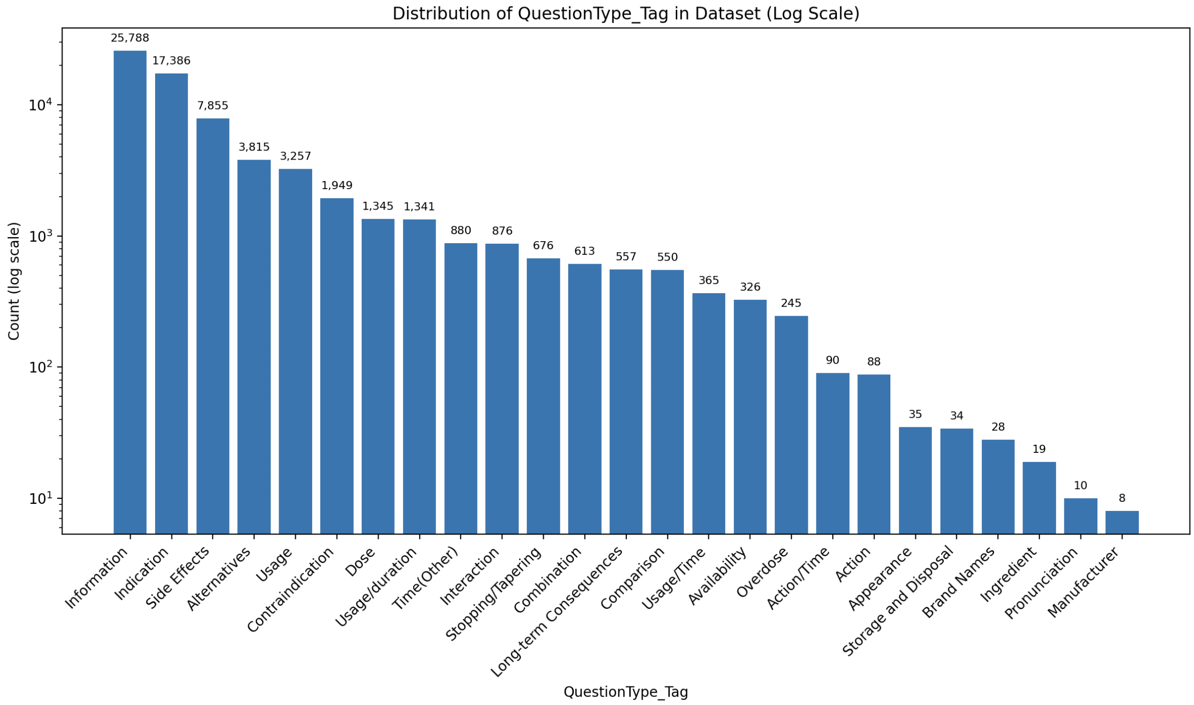


Figure 7: Prompt description for the Question Type Tagging Task.

Question Type Tagging System Prompt

Task Description
You are a medical NLP annotator. Label every consumer medical question with:

- `question_type_explanation` – a concise and brief (1 sentence) justification of **why** the chosen `question_type` is appropriate, citing question wording and any domain knowledge you considered.
- `question_type` – **exactly one** of the 25 categories listed below in the table (use the label verbatim).

Figure 8: Baseline system prompt used for all models in the default setting.

Baseline System Prompt (Main Prompt)

You are an experienced professional medical doctor providing reliable medical advice in fluent Persian. Users are reaching out through an online medical forum seeking clear, concise, and accurate answers to their health-related questions. Respond directly, briefly, and precisely to their inquiries. Always provide practical and actionable medical guidance. Refer users to consult a doctor only when it is strictly necessary and impossible to safely or accurately address their concerns online. Prioritize addressing their concerns clearly and thoroughly without unnecessary elaboration or hesitation.

Figure 9: Role-based prompting strategy using ICD-11-conditioned specialist titles (e.g., `Psychiatrist`, `Gynecologist`).

Role-Based Prompt (ICD-11 Conditioned)

You are an experienced `{role}` providing reliable medical advice in fluent Persian. Users are reaching out through an online medical forum seeking clear, concise, and accurate answers to their health-related questions. Respond directly, briefly, and precisely to their inquiries. Always provide practical and actionable medical guidance. Refer users to consult a doctor only when it is strictly necessary and impossible to safely or accurately address their concerns online. Prioritize addressing their concerns clearly and thoroughly without unnecessary elaboration or hesitation.

Figure 10: Translation pipeline (separate LLM inference) for leveraging English LLMs with Persian input and output.

Translation-Based Prompting (Three-Step Pipeline)

Step 1: Translate Question (Persian → English)
You are a professional medical translator. Translate the following Persian medical question into clear, neutral English. Only output the translation.

Step 2: Answer in English
You are an experienced professional medical doctor providing reliable advice in English. Answer directly, briefly and precisely. Provide practical guidance and advise an in-person visit only when absolutely necessary.

Step 3: Translate Answer (English → Persian)
You are a professional medical translator. Translate the following English answer into fluent Persian suitable for the patient. Only output the translation.

MedJudge System Prompt

You are "Med-Judge", an objective medical QA grader responsible for evaluating how closely a Model-generated answer aligns with an Expert-provided answer (the gold-standard).

Strict Guidelines for Evaluation:

1. **Exclusivity:** Base your judgment **only on the Expert's provided answer**, and never rely on your own medical knowledge or external resources.
2. **Step-by-step and Explainable Evaluation:**
   - Provide a brief and concise comparison analysis (`brief_analysis`), clearly explaining your evaluation.
   - Clearly identify the key medical facts, statements, or recommendations in both Expert and Model answers.
   - List explicitly any critical facts appearing **only** in the Expert's answer (`key_missing_facts`) or **only** in the Model's answer (`key_extra_facts`). These should be keywords in Persian (Farsi).
3. **Labeling Correctness:** Categorize the answer strictly according to these definitions:
   - `correct` – Answers have essentially identical meaning with no clinically meaningful differences.
   - `partially_correct` – Answers differ slightly without critical differences affecting clinical understanding or advice.
   - `incorrect` – Answers substantially differ, significantly impacting clinical meaning or completeness.
   - `contradictory` – Model explicitly contradicts Expert's answer or provides opposite medical advice.
4. **Coverage Analysis:** Evaluate factual coverage using:
   - `equal` – Both answers contain exactly the same key facts.
   - `model_subset` – Model answer misses one or more critical facts present in the Expert answer.
   - `expert_subset` – Model answer includes additional key facts not found in the Expert's answer.
   - `overlap_none` – No meaningful shared key facts between answers.
5. **Clinical Impact Estimation:** Judge how the differences could clinically affect a patient's safety or care using:
   - `negligible`, `moderate`, `significant`, `critical`
6. **Confidence Rating:** Express your certainty level:
   - `high`, `medium`, `low`

Attention:

- Be strict and literal in comparing the two answers.
- Do not assume or infer correctness beyond the Expert answer.
- Be concise and structured. Avoid vague commentary.

Output Format: Return **only** a valid JSON object matching the schema.

Figure 11: MedJudge system prompt with structured grading rubric for model evaluation.

MedJudge User Prompt Template

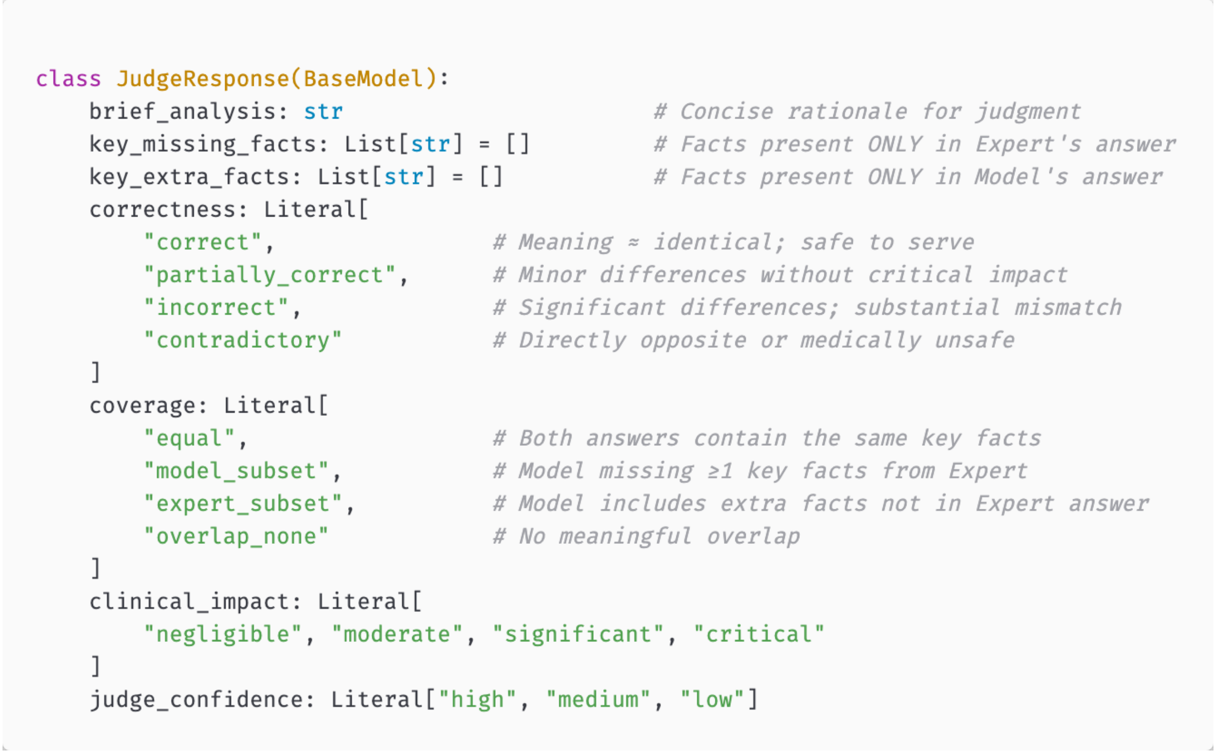
Question: `{question}`

Expert Answer: `{expert_answer}`

Model Answer: `{model_answer}`

Figure 12: Template used for submitting evaluation tasks to MedJudge. Placeholders are replaced at runtime.

Figure 13: MedJude Structured Output Schema



```
class JudgeResponse(BaseModel):
    brief_analysis: str                        # Concise rationale for judgment
    key_missing_facts: List[str] = []          # Facts present ONLY in Expert's answer
    key_extra_facts: List[str] = []            # Facts present ONLY in Model's answer
    correctness: Literal[
        "correct",                  # Meaning ≈ identical; safe to serve
        "partially_correct",        # Minor differences without critical impact
        "incorrect",                # Significant differences; substantial mismatch
        "contradictory"             # Directly opposite or medically unsafe
    ]
    coverage: Literal[
        "equal",                    # Both answers contain the same key facts
        "model_subset",             # Model missing ≥1 key facts from Expert
        "expert_subset",            # Model includes extra facts not in Expert answer
        "overlap_none"              # No meaningful overlap
    ]
    clinical_impact: Literal[
        "negligible", "moderate", "significant", "critical"
    ]
    judge_confidence: Literal["high", "medium", "low"]
```

Figure 14: Human Evaluation

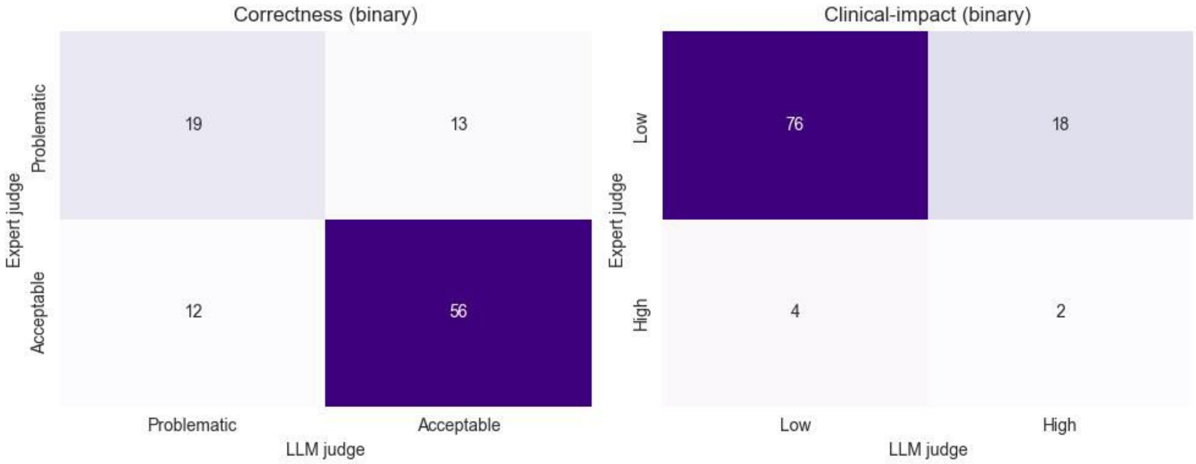


Figure 15: Correctness for baselines and advanced Prompting

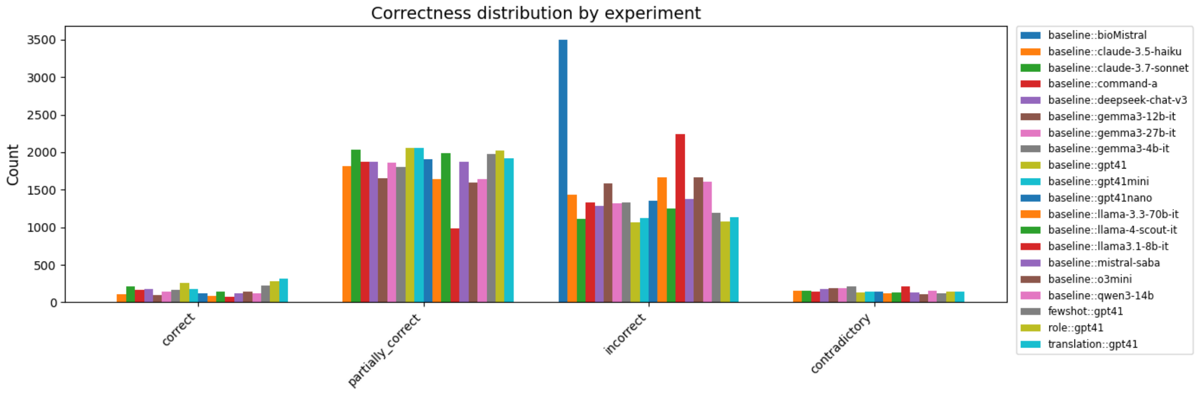


Figure 16: Coverage for baselines and advanced Prompting

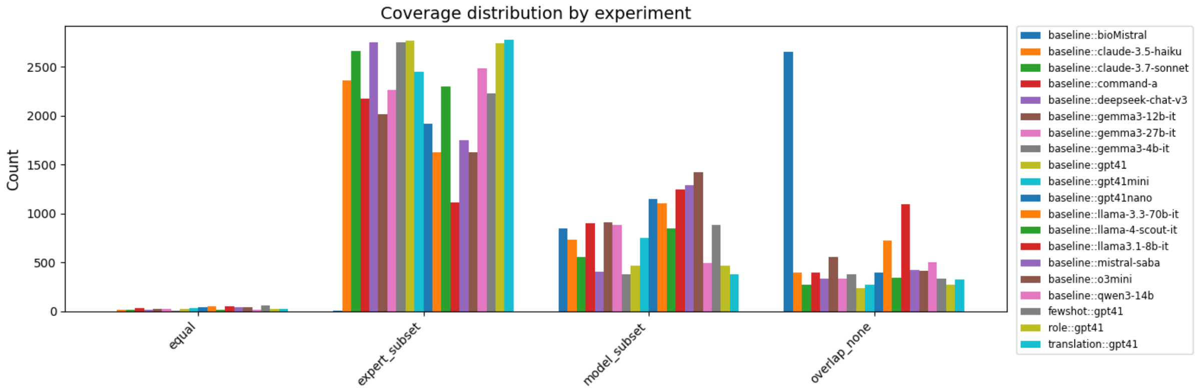
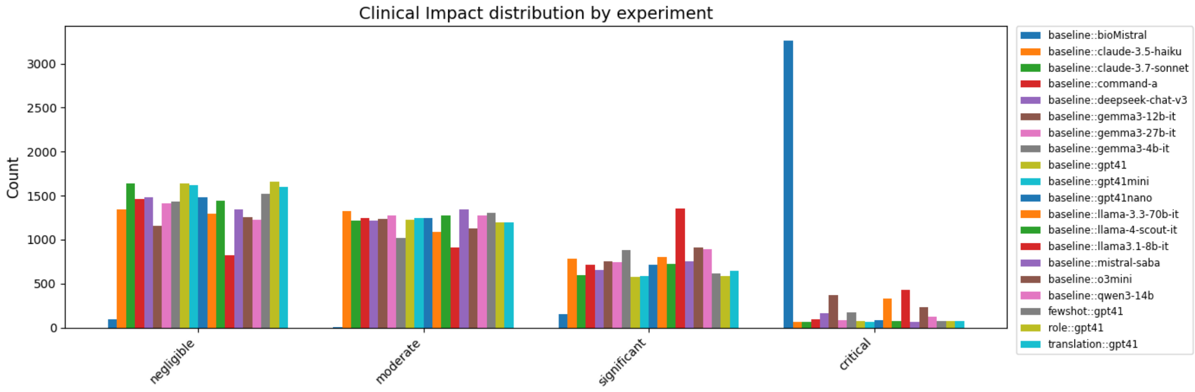


Figure 17: Critical impact for baselines and advanced Prompting


Clinical Impact distribution by experiment
Count
negligible
moderate
significant
critical
baseline::bioMistral
baseline::claude-3.5-haiku
baseline::claude-3.7-sonnet
baseline::command-a
baseline::deepseek-chat-v3
baseline::gemma3-12b-it
baseline::gemma3-27b-it
baseline::gemma3-4b-it
baseline::gpt41
baseline::gpt41mini
baseline::gpt41nano
baseline::llama-3.3-70b-it
baseline::llama-4-scout-it
baseline::llama3.1-8b-it
baseline::mistral-saba
baseline::o3mini
baseline::qwen3-14b
fewshot::gpt41
role::gpt41
translation::gpt41