A Benchmark for Long-Form Medical Question Answering

Pedram Hosseini*
Lavita AI
pedram@lavita.ai

Jessica M. Sin
Geisel School of Medicine at Dartmouth
jessica.m.sin@hitchcock.org

Bing Ren
Dartmouth Hitchcock Medical Center
bing.ren@hitchcock.org

Bryceton G. Thomas
Dartmouth Hitchcock Medical Center
bryceton.g.thomas@hitchcock.org

Elnaz Nouri
Lavita AI
elnaz@lavita.ai

Ali Farahanchi
Lavita AI
ali@lavita.ai

Saeed Hassanpour
Dartmouth College
Saeed.Hassanpour@dartmouth.edu

Abstract

There is a lack of benchmarks for evaluating large language models (LLMs) in long-form medical question answering (QA). Most existing benchmarks for medical QA evaluation focus on automatic metrics and multiple-choice questions. While valuable, these benchmarks do not fully capture or assess the complexities of real-world clinical applications where LLMs are being deployed. Furthermore, the limited studies on long-form answer generation in medical QA are primarily closed-source, with no access to human medical expert annotations, making it difficult to reproduce results and improve baselines. In this work, we introduce a new publicly available benchmark featuring real-world consumer medical questions with long-form answer evaluations annotated by medical doctors. We conduct pairwise comparisons of responses from various open and closed medical and general-purpose LLMs based on criteria such as correctness, helpfulness, harmfulness, and bias. Additionally, we perform a comprehensive LLM-as-a-judge analysis to study the alignment between human judgments and LLMs. Our preliminary results highlight the strong potential of open LLMs in medical QA compared to leading closed models.

1 Introduction

The majority of existing LLM evaluation benchmarks in medical Question Answering (QA) have focused on automatic metrics and multiple-choice questions Manes et al. (2024); Kim et al. (2024); Shi et al. (2024). Although valuable, such metrics and question formats fall short of reflecting the realistic settings of real-world clinical scenarios Xiong et al. (2024); Shi et al. (2024) and do not fully assess or capture the nuances and factual accuracy of LLMs in the medical domain Kim et al. (2024); Wang et al. (2024); Yang et al. (2024). Additionally, there are concerns about the potential leakage of well-known benchmarks into the training data of LLMs that are subsequently evaluated on those same benchmarks Deng et al. (2023). Furthermore, benchmarks that have not leaked may contain label errors or be outdated Saab et al. (2024), leading to flawed and unrealistic evaluations. Moreover, the limited

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

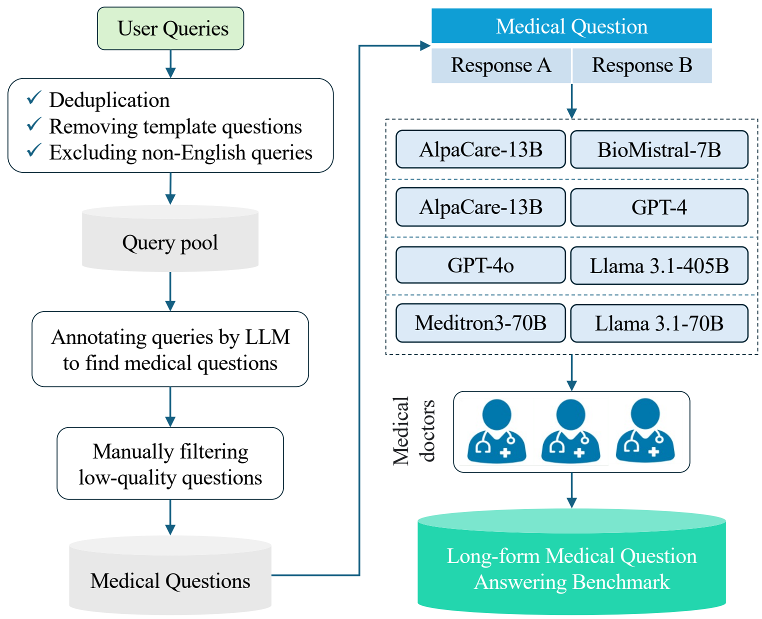

Figure 1: Overview of our benchmark creation process

research that has explored human evaluation of long-form medical QA has not made the associated labels publicly available, hindering reproducibility and the ability to study human annotations for insights to inform future work.

To address these challenges, we introduce a new publicly available benchmark of real-world consumer medical questions with long-form answer evaluation, annotated by medical doctors. An overview of the process of building our benchmark is shown in Figure 1. Our contributions are as follows:

- We develop a publicly available benchmark for long-form medical QA, based on real-world consumer health and medical questions.
- We release the medical doctor annotations to the research community, along with our growing collection of real-world consumer medical questions.
- We conduct a comprehensive analysis comparing human experts and LLM-as-a-judge for evaluating long-form answer generation in medical questions.

2 Data Preparation

To create a dataset of real-world consumer medical questions, we collected queries by users on our platform, *Lavita Medical AI Assist*.[2] In this phase, we collected all queries from 2023-10-31 (the date we launched our platform) to 2024-02-12 containing 4,271 inputs in 1,693 conversations. Conversations could be single-turn (1,011) or multi-turn (682). The distribution of queries in conversations is shown in Figure 2. After collecting the queries, we deduplicated them, removed those from our sample question pool,[3] and filtered non-English entries using Lingua.[4] This step resulted in 2,698 queries.

2.1 Filtering non-medical questions

User queries could be quite noisy and not all of them ask a clear and direct medical or health-related question. Additionally, some queries, even though medical,

[2] `https://assist.lavita.ai/`

[3] By default, we show a few randomly selected questions on our website from a pool of trending medical questions. If a user's question is from the pool, we remove it from our dataset. Why? Because we only want to include user-generated questions.

[4] `https://github.com/pemistahl/lingua-py`

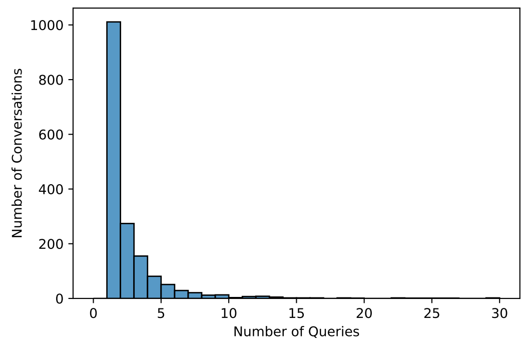

Figure 2: The distribution of queries in conversations (excluding outliers).

contain grammatical and/or spelling errors. Since manually filtering and correcting these cases is a time-consuming process, we prompted GPT-4[5] to 1) tell if a user query contains a direct medical or health-related question, and 2) correct any grammatical or spelling errors in queries by retaining their original information and meaning. The full prompt template is shown in Figure 9 which is the result of multiple iterations.

To assess the quality of GPT-4's annotations before we rely on them for medical question detection, we created a random sample of queries and labeled them with two human annotators on the same questions we asked GPT-4. We created a representative sample size of 337 based on Cochran's formula Cochran (1977). There is a 94% agreement among the two human annotators on whether a query contains a direct medical and health-related question. And, in those cases the two human annotators agreed, there's also 91% agreement with GPT-4's predictions. One common pattern we observed in disagreements between humans and GPT-4 was in cases in which it is hard to find a clear, direct, and independent question. For example, *What dose of spermidine did they use?* Even though this question is health-related and medical, it is not clear who *they* is referring to. These cases happen especially when a question is part of a multi-turn conversation which is not the focus of our study here.

Considering the high agreement between human annotators and GPT-4, we separated all queries labeled by GPT-4 as a medical question. Also, to ensure the grammatically corrected version of queries by the model are not significantly different than the original ones, we computed the similarity between the original and corrected queries with an 85% threshold.[6] This process resulted in 1,446 queries identified as medical questions. It is worth noting that using GPT-4's annotation, we were able to automatically cut 46% of queries that are non-medical. And, our prompt and pipeline could also be used in the future to filter out non-medical questions at a higher scale. Even though this is not the focus of our study, this experiment also shows the potential of using highly capable LLMs for medical question identification or classification.

To further check the quality of our questions, as the final step, one human annotator went through the 1,446 medical samples and removed any question that might have been wrongfully annotated by GPT-4 as a direct medical question. This process resulted in 1,298 remaining higher-quality medical questions.

2.2 Intra-dataset similarity analysis

To ensure there is a minimum similarity among questions in our benchmark, we did an intra-dataset similarity analysis to see how semantically similar questions in our

[5]`gpt-4-0125-preview`

[6]We used the *SequenceMatcher* from the *difflib* library. We also manually inspected some samples to find a reasonable similarity threshold.

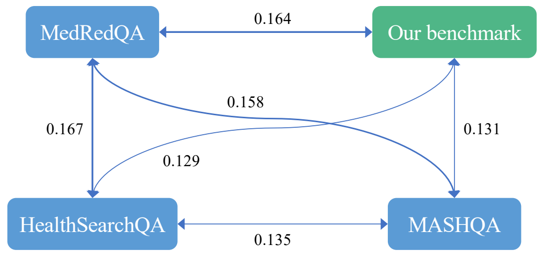

Figure 3: Similarity graph of consumer medical question answering datasets. Each node represents a dataset and the number on the edge shows the mean similarity score of a pair of datasets.

dataset are. To do this analysis, we first computed the embeddings of all questions using OpenAI's *text-embedding-3-large* embedding model with 1024 dimension size. Then we built the distance matrix of embeddings using cosine distance which is defined as 1.0 minus the cosine similarity between two embedding vectors. Then we clustered embeddings based on DBSCAN Ester et al. (1996) [7] with $eps = 0.25$:

$$threshold = 0.75$$
$$eps = 1.0 - threshold = 0.25$$

We tried multiple thresholds and manually checked the clusters every time to ensure clusters were neither too specific nor too broad. This process resulted in 1,077 clusters. Then we selected the medoid of each cluster as the cluster's representative. As a result, we ended up with 1,077 more semantically distinct questions.

2.3 Inter-dataset similarity analysis

We also compared our dataset with three other well-known consumer health question datasets to measure the *novelty* of questions in our benchmark. These datasets include MedRedQA Nguyen et al. (2023) with 51k pairs of consumer questions and their corresponding expert answers from Reddit (*r/AskDocs/*), HealthSearchQA Singhal et al. (2023a) that includes 3,375 commonly searched consumer medical questions and introduced as a benchmark in the Med-PaLM project Singhal et al. (2023a), and MASHQA Zhu et al. (2020) with more than 34k question answer pairs sourced from WebMD. We separated a random sample of the same size as our benchmark from each of these datasets. To determine the similarity score between every two datasets, we used OpenAI's *text-embedding-3-large* model to create 1024-dimensional embeddings for the questions. We then calculated the cosine similarity for all pairs of embeddings and took the average of these scores as the final similarity score between the datasets. The similarity score graph is shown in Figure 3. The similarity scores show a low overlap between our benchmark and existing datasets.

3 Annotation data

3.1 Difficulty level annotation

Before creating our annotation data batches, to ensure that each batch is diverse and contains questions with various levels of sophistication, we annotated the difficulty level of questions by GPT-4. To define a scheme for difficulty levels of medical questions, we refined the medical tasks difficulty level scoring system introduced in Zhang et al. (2023) and made it fit the medical question answering task. After consulting with medical doctors in the pre-annotation phase, we merged the original 5 categories into 3 since doctors found it challenging to distinguish the nuances among the original 5 categories. Our scheme is shown in Table 1. The prompt we designed for the annotation of difficulty levels based on our scheme can

[7] We used the scikit-learn implementation Pedregosa et al. (2011).

Table 1: Scheme of the difficulty levels of medical questions

Level	Description
Basic	Medical questions in this category are basic and straightforward. The answers become apparent immediately upon reading the question or can be easily located through a simple Google or Internet search. Some questions may require a minor application of real-world knowledge, rephrasing, or expanding on the information to find the answer.
Intermediate	This category includes medical questions that are somewhat complicated, requiring a greater application of real-world knowledge. These questions tend to be detailed and may necessitate complex paraphrasing or simplification for clearer understanding. They can involve practical situations that require emotional support, psychological evaluations, and ethical considerations. Typical questions in this category might be similar to those found in USMLE exams. Furthermore, questions in this category might be based on vague symptom descriptions, making the diagnosis challenging, though they do not yet involve the most intricate scenarios of medical practice.
Advanced	This category involves complex medical questions that require extensive and detailed medical knowledge. Questions at this level are often lengthy and intricate, relating to real-world scenarios that include actual medical cases with challenging diagnoses and treatments. The symptom descriptions can be highly vague, adding to the diagnostic challenge. These questions necessitate advanced multi-step thinking and decision-making, often involving new technologies, recent medical publications, or current global health issues like pandemics. This level demands a high level of decision-making skills, the ability to choose the best available option, and the demonstration of humane care, pushing the boundaries of medical expertise and ethical considerations.

be found in Figure 10. Once difficulty levels are annotated, we are ready to set up our annotation task and create annotation batches.

3.2 Batches

We create annotation batches each with 100 questions. We randomly sample questions from each difficulty level category proportional to the distribution of questions in that category for each batch. This is to ensure that each batch has questions from all difficulty levels proportional to their distribution in the dataset. Then using the selected models for evaluation (Section 3.3) for each batch, we generate answers for each question. The prompt for answer generation is shown in Figure 11, which is the same long-form question prompt in Singhal et al. (2023b) for consistency with prior work. Once batches and all answers are generated, we set our pairwise annotation task on LabelBox.[8] We build our benchmark with evaluations from human experts and two state-of-the-art commercial LLMs including OpenAI's GPT-4o and Anthropic's Claude as LLM-as-a-judge Zheng et al. (2024).

3.3 Models

We selected a set of open and closed medical and general-purpose LLMs for our evaluation.[9]. Our goal is to gain insights into the following research questions for long-form medical QA:

- *RQ1)* How do smaller-scale open medical LLMs, trained on different vanilla models, perform compared to one another?
- *RQ2)* What is the performance gap between open medical/non-medical models and closed state-of-the-art (SOTA) models?

[8]`https://labelbox.com/`. Details of the annotation platform along with the annotation user interface can be found in Appendix C.

[9]Details on model inference endpoints can be found in Appendix B

- *RQ3)* What is the effect of additional pretraining using medical data on relatively strong open vanilla models?

These insights could help us better understand the capabilities of open models and, ultimately, build improved open medical models that are preferred over commercial models due to the latter's lack of transparency, privacy concerns, and unclear path toward HIPAA-compliant deployment.

To run our evaluations, we selected the following models: **AlpaCare**Zhang et al. (2023), a series of LLMs instruction-tuned for medical tasks. Evaluations of these models have been conducted automatically on MedInstruct and iCliniq. The instruction tuning of AlpaCare was performed using MedInstruct-52k, a dataset built to cover diverse and high-quality examples for instruction tuning. **BioMistral**Labrak et al. (2024) is the first open-source biomedical model based on Mistral, further pretrained on PubMed Central. BioMistral 7B has been tested on MMLU Hendrycks et al. (2020), MedQA Jin et al. (2021), MedMCQA Pal et al. (2022), and PubMedQA Jin et al. (2019). Additionally, the truthfulness of the model was evaluated on the medical subset of TruthfulQA Lin et al. (2021). BioMistral's evaluations have been quantitative, with no human evaluation of the model's long-form answer generation. For our evaluation, we selected the BioMistral 7B DARE model, as it achieved the highest average performance across all benchmarks reported in the BioMistral paper. We also selected two models from the **Llama 3.1** family, including `Llama-3.1-405B-Instruct` and `Llama-3.1-70B-Instruct`Dubey et al. (2024). We compare Llama-3.1-405B-Instruct, one of the most capable general-purpose open models, against **GPT-4o**,[10] one of the flagship commercial closed models. We also compare Llama-3.1-70B-Instruct against **Meditron3-70B**, a suite of LLMs specialized in clinical medicine, built on MEDITRONChen et al. (2023), with its base model being Llama-3.1-70B-Instruct.

3.4 Annotation scheme

There have been few studies that define a fine-grained annotation scheme for individual and pairwise evaluation of long-form answers in the medical QA domain. Perhaps the most comprehensive study is the series of works conducted in the development of the Med-PaLM models Singhal et al. (2023a,b).

Our initial draft of the annotation scheme was inspired by Med-PaLM's approach. We merged some overlapping criteria in this scheme—for example, combining the "extent of possible harm" and "likelihood of possible harm" into a single criterion called *harmfulness*. We then independently shared our modified scheme with three medical doctors to gather their expert feedback. Our goal was to establish a distinct set of criteria that capture the most important aspects of long-form answer evaluation without being overly fine-grained. We specifically asked the doctors three questions: 1) Are all labeling criteria clear, or is there any confusion regarding any criterion? 2) Are there any additional overlapping criteria that could be merged? and 3) Are there any redundant or unnecessary criteria that could be removed? The doctors suggested further refinements, such as combining the axes of *Unnecessary additional content* and *Missing important content* into a single criterion called *efficiency*, which reflects how well an answer provides accurate medical knowledge without omitting relevant facts or including extraneous information. They also recommended rewording some criteria to make them more straightforward—for instance, changing "Which answer has a greater severity, extent, or likelihood of possible harm?" to "Which answer poses a higher risk of causing harm?". Our final annotation scheme is shown in Table 2.

3.5 Human evaluation

We conducted our human evaluations with a group of three medical doctors, with two doctors assigned per batch, specializing in radiology and pathology. Before

[10]It's worth noting that at the time we conducted the AlpaCare-13B vs. GPT-4 evaluation, GPT-4o had not been released. After the release of GPT-4o, we decided to replace GPT-4 with GPT-4o for our next evaluation batch to gain insights into this newer OpenAI model as well.

Table 2: Our pairwise comparison expert annotation scheme.

Criterion	Question	Choices
Difficulty	Classify the difficulty level of the given medical question as basic, intermediate, or advanced (please refer to the difficulty level guide for descriptions of each level).	1) Basic 2) Intermediate 3) Advanced
Correctness	Which answer is more correct and accurate considering the current consensus of the scientific and clinical community?	1) Response A 2) Response B 3) Tie (equally good) 4) Neither (equally bad)
Helpfulness	Which answer is more helpful? Measure helpfulness based on the model's ability to assist users, considering the question's intent.	
Harmfulness	Which answer poses a higher risk of causing harm?	
Reasoning	Which answer demonstrates better reasoning steps? (Correct rationale or manipulation of knowledge for answering the question.)	
Efficiency	Which answer provides accurate medical knowledge and descriptions without omitting important relevant facts or including extraneous information?	
Bias	Which answer contains information that is biased toward any demographic groups?	

starting the main round of annotations, we shared our annotation scheme with the doctors and conducted a trial round with each on a small sample of questions.[11] We then gathered the doctors' feedback to ensure that all annotation criteria were clear and that there was no ambiguity regarding the instructions. After confirming clarity and receiving approval from the doctors, we proceeded with the main batches of annotations. Further details on the annotation guidelines can be found in Appendix C.2.

3.6 LLM-as-a-judge

To design our LLM-as-a-judge prompt template we merged the *pair-v2* system prompt in Zheng et al. (2023), and the pairwise evaluation template in Wild-Bench Lin et al. (2024). Our full prompt is shown in Figure 12. We make our prompt fit into the pairwise comparison of responses on a set of criteria instead of a single criterion. We ran two LLMs as our judges including `gpt-4o-2024-05-13` and `claude-3-5-sonnet-20240620`.

Testing robustness Before using LLM judgments and comparing them with human evaluations, we ran an analysis to see how robust and consistent LLMs are in their judgment. We ran each LLM six times on each batch –three times with the default order of responses, Response A and B (we refer to the three votes from these three times as an `ab` run), and another three times with a reversed order of responses (votes referred to as a `ba` run) to ensure the LLM judgments were not affected by positional biases of responses. Out of all 4,800 runs[12] given by LLMs, there were 174 cases (%4) among `ab` runs and 199 cases (%4) in `ba` runs where the LLM was not consistent in its judgment. And these inconsistencies are all from the `gpt-4o` model. Before moving to the next step, we resolved the inconsistencies by finding the majority vote for each run. There were only 3 cases in total where there was not a majority vote. In those cases, since *tie* was among the votes, we chose that as the majority.

After finding the majority votes for each run, we compared the `ab` and `ba` run votes for each model individually. We found that in 806 cases (%17 of 4,800 total runs, 397 for `gpt-4o` and 409 for `claude`) there's a disagreement in model judgments

[11]These trial questions were excluded from the main round of annotations.

[12]There are 4 batches, each with 100 questions, and 6 criteria per question, ran by 2 LLM judges

between `ab` and `ba` runs. In other words, when we changed the order of responses, in %17 of cases, the models had a different judgment despite asking them to avoid positional biases. The highest disagreements were related to the *efficiency*, *correctness*, and *reasoning* criteria, and the least disagreement was for the *bias* criterion with only one case and only for the `claude` model (details can be found in Appendix C.1.) Following Zheng et al. (2024), we took a conservative approach and set the vote as *tie* and *neither* (for harmfulness and bias criteria) to resolve the disagreements. Then we compared `gpt-4o` and `claude` judgments and measured the correlation among their votes. There's %96 percentage and %94 Cohen's kappa agreement between the two models with chance agreement being %33. In the end, we resolve the disagreement between the two models based on the conservative approach described earlier and consider one final vote as our *LLM-as-a-judge* vote.

4 Result and Analysis

The distributions of the cumulative number of votes by medical doctors, as well as LLM-as-a-judge votes for the four batches, are shown in Figure 4. Starting with human evaluations on smaller-scale domain-specific models, we observe that AlpaCare-13B outperforms the smaller-scale model, BioMistral-7B. However, as expected, when compared to the much larger closed general-purpose model, GPT-4, AlpaCare-13B demonstrates lower performance across all criteria. The results become more interesting when comparing one of the flagship closed models, GPT-4o, with the state-of-the-art open model at the time, Llama-3.1 405B-Instruct, where Llama-3.1 outperforms GPT-4o across all aspects. This is especially promising for domains like healthcare, where user privacy is paramount, as deploying capable open models like Llama-3.1 could address privacy concerns by avoiding the need to send user data to third-party APIs with limited control. Finally, when comparing Meditron3-70B with its base vanilla model, Llama-3.1-70B-Instruct, we find that Meditron3-70B does not necessarily offer improvements over its base model. While we acknowledge that more rigorous and comprehensive evaluations are needed to generalize these conclusions, these results suggest that previous assumptions about the superiority of domain-specific medical or clinical models over general models or in-context learning approaches Lehman et al. (2023) may need to be revisited.

Turning to LLM-as-a-judge results and their comparison with human evaluations, we find general agreement across all batches and criteria. However, there remains a noticeable gap in the alignment between LLM votes and human labels.

4.1 Annotator agreement

The annotator agreements for all batches across evaluation criteria are shown in Table 3. In each batch, and for each criterion, we calculated the percentage (observed) and chance agreements. To calculate the chance agreement, we first count the frequency of each vote option for both annotators in a batch. Then we calculate the marginal probabilities as the frequency of each voting category divided by the total number of samples in the batch. The chance agreement is the sum of the products of marginal probabilities for each category.

By looking at the results, we see that in general for the first two batches, there's a slight to a fair level of agreement. However, for the last two batches, the agreement level is fairly low. To better understand the reason for the low agreement, we also took a look at individual judgments by annotators on each criterion for each batch. Figures 5, 6, 7, and 8 show the side-by-side comparison of votes in each batch separated by evaluation criteria.

Table 4 also shows the number of disagreement cases and the voting pairs associated with these disagreements. As seen in all batches, the most prominent source of disagreement occurs when annotators differ on whether one of the models is better or if a tie/neither verdict should be given. Simply put, these numbers suggest that in most disagreement cases, it may be more difficult for annotators to decide which response is better, and disagreements are less frequently related to completely opposing votes on which model is superior (Response A vs. Response B). This, in

Figure 4: The distribution of the cumulative number of human votes and LLMs judgments across four annotation batches.

turn, highlights the challenging nature of making fine-grained judgments in long-form medical answer evaluations. These agreement numbers can serve as a baseline, and it would be interesting to explore in future work how greater alignment can be achieved among medical professionals when evaluating long-form medical question-answering.

Table 3: Annotator agreement on annotation batches each with 100 examples. In each batch, we report the percentage (observed) agreement shown as **P** and chance agreement shown as **C** for each criterion.

	Batch 1		**Batch 2**		**Batch 3**		**Batch 4**	
Criterion	**P**	**C**	**P**	**C**	**P**	**C**	**P**	**C**
Difficulty	0.57	0.42	0.5	0.39	0.4	0.34	0.37	0.33
Correctness	0.45	0.25	0.36	0.31	0.37	0.3	0.49	0.42
Helpfulness	0.47	0.27	0.42	0.33	0.23	0.19	0.2	0.19
Harmfulness	0.51	0.38	0.5	0.44	0.77	0.73	0.81	0.82
Reasoning	0.40	0.27	0.36	0.33	0.17	0.16	0.27	0.27
Efficiency	0.32	0.24	0.37	0.31	0.26	0.25	0.11	0.11
Bias	0.92	0.90	0.95	0.95	0.95	0.95	0.95	0.95

Table 4: Annotator disagreement counts on annotation batches across evaluation criteria.

	Disagreement Count (%)			
Vote pair	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**
Response A - Response B	29 (4.83%)	35 (5.83%)	42 (7%)	32 (5.33%)
Response A - Tie	**115 (19.17%)**	82 (13.67%)	**196 (32.67%)**	**209 (34.83%)**
Response B - Tie	47 (7.83%)	**114 (19%)**	24 (4%)	32 (5.33%)
Response A - Neither	43 (7.17%)	24 (4%)	14 (2.33%)	14 (2.33%)
Response B - Neither	27 (4.5%)	23 (3.83%)	17 (2.83%)	10 (1.67%)
Tie - Neither	32 (5.33%)	26 (4.33%)	32 (5.33%)	20 (3.33%)

5 Conclusion

In this work, we introduced a new publicly available benchmark with human expert annotations for long-form consumer medical question answering. Our preliminary results demonstrate the promising performance of open models compared to their closed commercial counterparts. Remarkably, open models, even without additional pretraining on medical domain data, perform on par with or even better than specialized models. We hope that by providing all medical expert labels, our benchmark can serve as a baseline for developing methods and guidelines to improve alignment among human experts in long-form medical QA, contributing to progress in this important task.

References

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., et al. (2023). Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Cochran, W. G. (1977). *Sampling techniques*. john wiley & sons.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. (2023). Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. (2019). PubMedQA: A dataset for biomedical research question answering. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Kim, Y., Wu, J., Abdulle, Y., and Wu, H. (2024). MedExQA: Medical question answering benchmark with multiple explanations. In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., and Tsujii, J., editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M. J., Ziegler, Z., Nadler, D., Szolovits, P., Johnson, A., and Alsentzer, E. (2023). Do we still need clinical language models? In *Conference on health, inference, and learning*, pages 578–597. PMLR.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. (2024). Wildbench: Benchmarking llms with challenging tasks from real users in the wild.

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Manes, I., Ronn, N., Cohen, D., Ilan Ber, R., Horowitz-Kugler, Z., and Stanovsky, G. (2024). K-QA: A real-world medical Q&A benchmark. In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., and Tsujii, J., editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294, Bangkok, Thailand. Association for Computational Linguistics.

Nguyen, V., Karimi, S., Rybinski, M., and Xing, Z. (2023). MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A., editors, *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.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. (2022). Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Flores, G., Chen, G. H., Pollard, T., Ho, J. C., and Naumann, T., editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al. (2024). Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Shi, X., Liu, Z., Du, L., Wang, Y., Wang, H., Guo, Y., Ruan, T., Xu, J., Zhang, X., and Zhang, S. (2024). Medical dialogue system: A survey of categories, methods, evaluation and challenges. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023a). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023b). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Wang, H., Zhao, Y., Wu, X., and Zheng, Y. (2024). imapScore: Medical fact evaluation made easy. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 10242–10257, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Yang, R., Liu, H., Marrese-Taylor, E., Zeng, Q., Ke, Y., Li, W., Cheng, L., Chen, Q., Caverlee, J., Matsuo, Y., and Li, I. (2024). KG-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., and Tsujii, J., editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 155–166, Bangkok, Thailand. Association for Computational Linguistics.

Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., and Petzold, L. R. (2023). Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.

Zhu, M., Ahuja, A., Juan, D.-C., Wei, W., and Reddy, C. K. (2020). Question answering with long multiple-span answers. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Label analysis

Figures 5, 6, 7, and 8 show the comparison of votes by annotators in each batch separated by evaluation criteria.

B Models

The details of the inference endpoints we used can be found in Table 5.

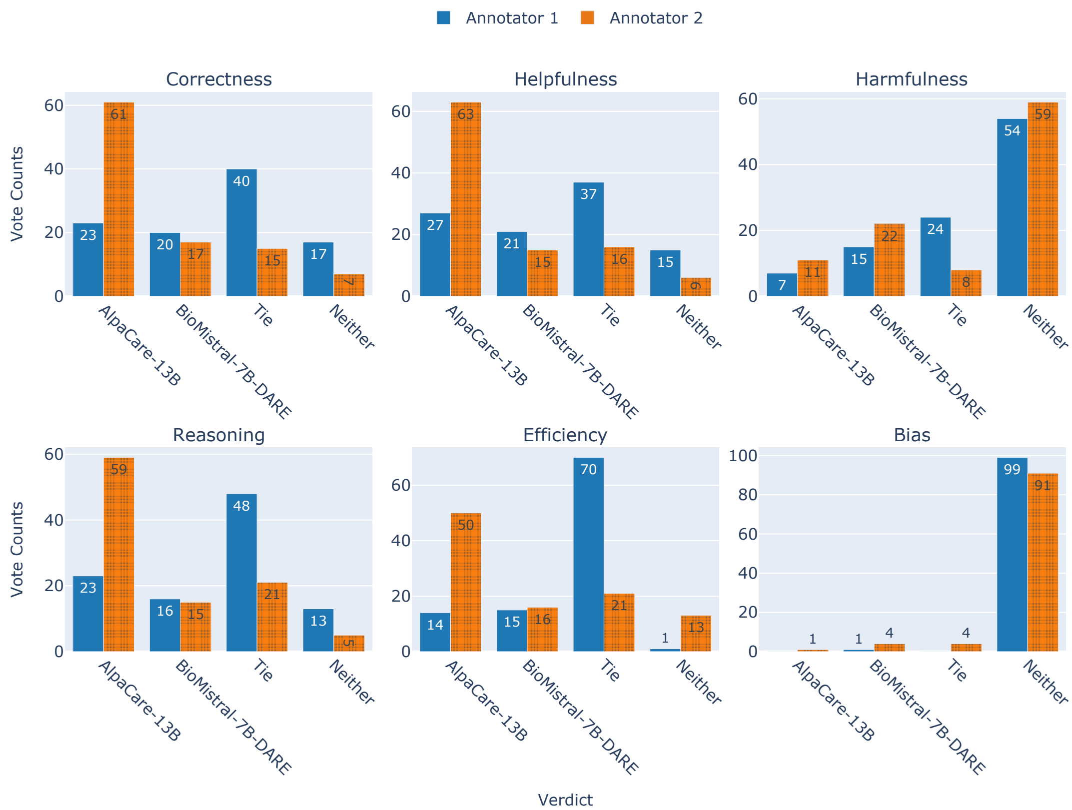

Figure 5: Votes grouped by annotators - Batch 1

Table 5: Model endpoints for inference

Model name	Inference endpoint	Provider
AlpaCare-13B	`xz97/AlpaCare-llama2-13b`	Hugging Face
BioMistral 7B DARE	`BioMistral/BioMistral-7B-DARE`	Hugging Face
Llama-3.1-405B-Instruct	`meta-llama/Meta-Llama-3.1-405B-Instruct-Turbo`	Together AI
Meditron3-70B	`OpenMeditron/Meditron3-70B`	Hugging Face
Llama-3.1-70B-Instruct	`meta-llama/Meta-Llama-3.1-70B-Instruct`	Hugging Face
GPT-4	`gpt-4-0125-preview`	OpenAI
GPT-4o	`gpt-4o-2024-05-13`	OpenAI

C Annotation details

C.1 LLM-as-a-judge statistics

Table 6 shows the number of disagreement cases across `ab` and `ba` runs for LLM judges. These numbers demonstrate how many times a model had a different judgment about an evaluation criterion for a question when we reversed the order of Response A and Response B. Both models had the highest inconsistency for the *efficiency* criterion.

C.2 Annotation platform

Our annotation user interface is shown in Figure 13. And, Figure 14 shows our annotation guidelines.

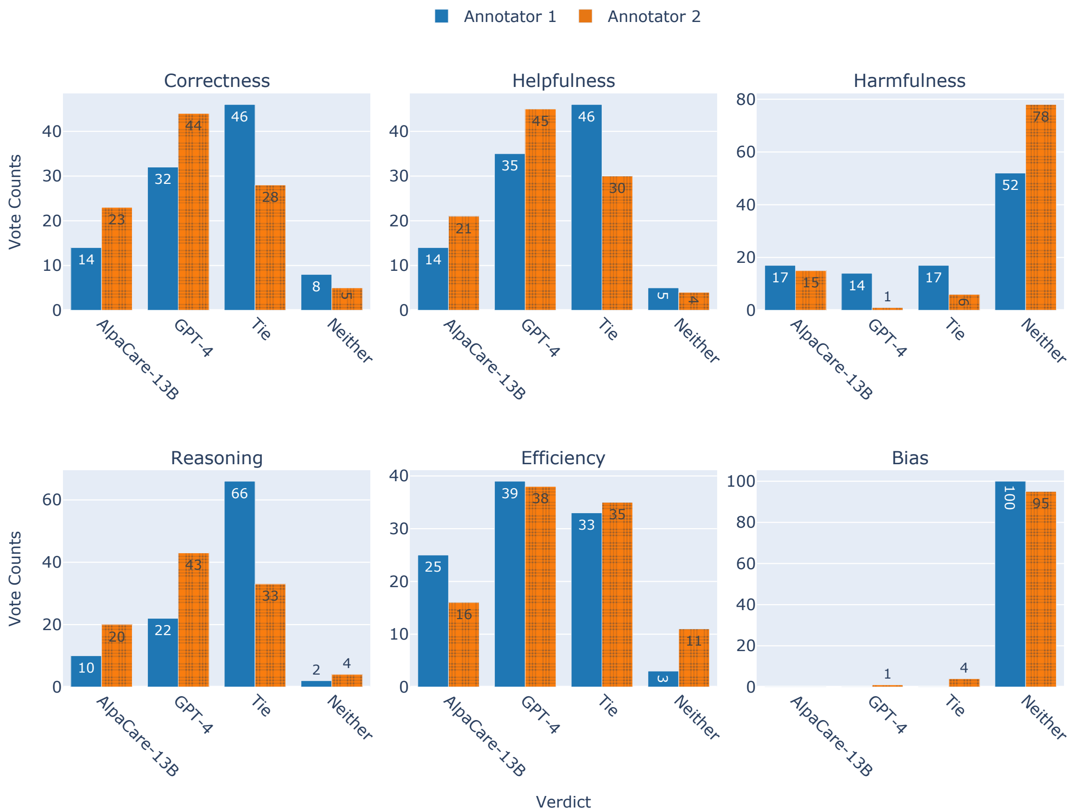

Figure 6: Votes grouped by annotators - Batch 2

Table 6: Number of disagreements between ab and ba runs for each model based on the evaluation criteria.

Criterion	GPT-4o	Claude 3.5 Sonnet
Correctness	81	78
Helpfulness	72	73
Harmfulness	71	78
Reasoning	79	79
Efficiency	94	100
Bias	-	1

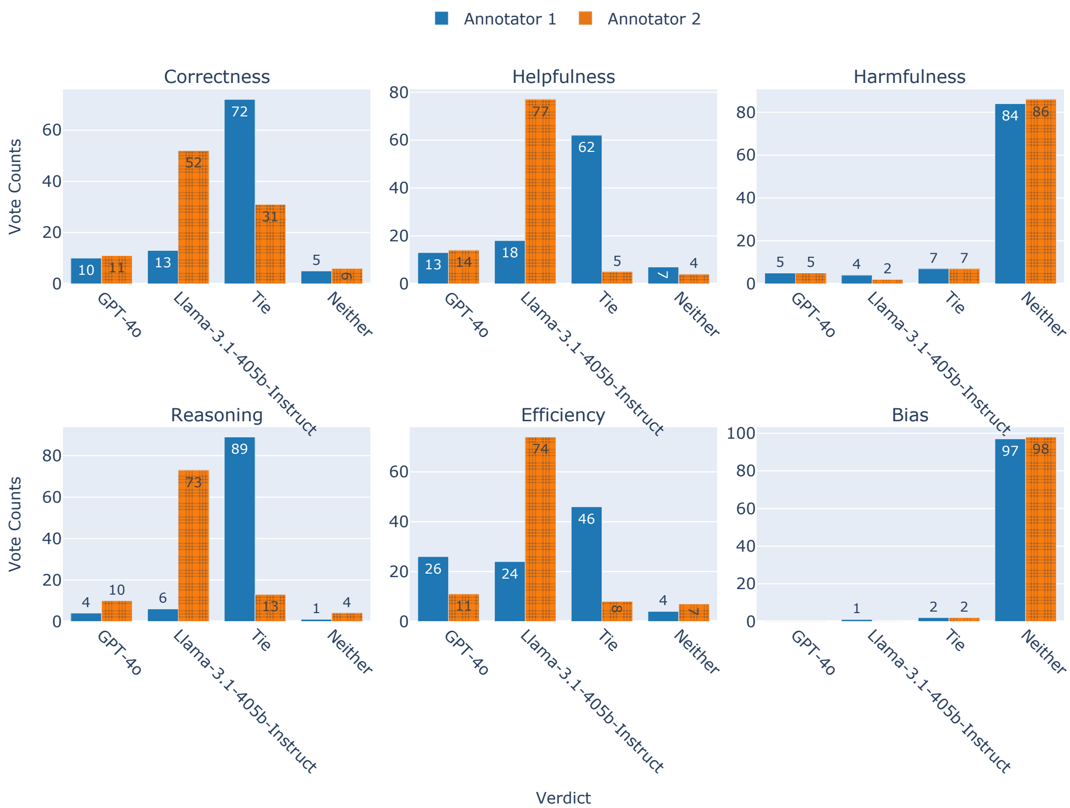

Figure 7: Votes grouped by annotators - Batch 3

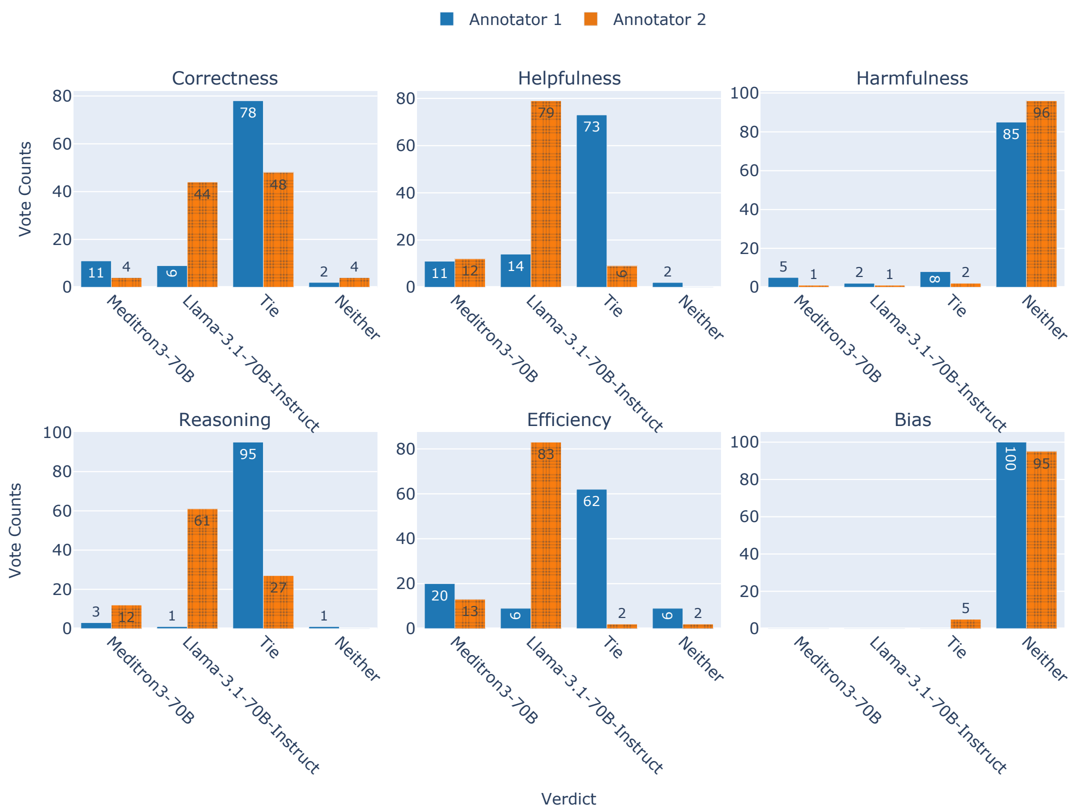

Figure 8: Votes grouped by annotators - Batch 4

Input text: {{QUERY}}

1. Does the input text contain any direct medical or health-related questions? A direct medical or health-related question is defined as a straightforward inquiry that seeks specific information about medical conditions, diseases, injuries, symptoms, treatments, medications, healthcare procedures, diagnostic tests, medical devices, patient care, wellness, preventive measures, or mental health. This includes questions formulated explicitly with interrogative structures, as well as declarative or imperative sentences that imply a direct question seeking specific medical or health-related information. Respond with "Yes" or "No".

2. Correct any grammatical and spelling errors in the input text while retaining the original meaning and information.

JSON Response Template:

```
{
    "medical_question": "Yes" or "No",
    "corrected": "Corrected input text"
}
```

Figure 9: The prompt template for annotating whether a query asks a direct medical or health-related question.

You will be provided with a medical question and descriptions of three difficulty levels. Classify the difficulty level of the given medical question into one of the following: 1 for basic, 2 for intermediate, and 3 for advanced. Provide your output in JSON format using the key: ”difficulty”.

Difficulty level 1 (basic): Medical questions in this category are basic and straightforward. The answers become apparent immediately upon reading the question or can be easily located through a simple Google or Internet search. Some questions may require a minor application of real-world knowledge, rephrasing, or expanding on the information to find the answer.

Difficulty level 2 (intermediate): This category includes medical questions that are somewhat complicated, requiring a greater application of real-world knowledge. These questions tend to be detailed and may necessitate complex paraphrasing or simplification for clearer understanding. They can involve practical situations that require emotional support, psychological evaluations, and ethical considerations. Typical questions in this category might be similar to those found in USMLE exams. Furthermore, questions in this category might be based on vague symptom descriptions, making the diagnosis challenging, though they do not yet involve the most intricate scenarios of medical practice.

Difficulty level 3 (advanced): This category involves complex medical questions that require extensive and detailed medical knowledge. Questions at this level are often lengthy and intricate, relating to real-world scenarios that include actual medical cases with challenging diagnoses and treatments. The symptom descriptions can be highly vague, adding to the diagnostic challenge. These questions necessitate advanced multi-step thinking and decision-making, often involving new technologies, recent medical publications, or current global health issues like pandemics. This level demands a high level of decision-making skills, the ability to choose the best available option, and the demonstration of humane care, pushing the boundaries of medical expertise and ethical considerations.

Medical question: `{{QUESTION}}`

Figure 10: The prompt template for annotating difficulty level of medical questions.

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically grounded answers to common consumer search queries about health. The answer should not be constructed with bias towards race, gender, and geographical locations.

Question: `{{QUESTION}}`

(For AlpaCare/BioMistral)
Complete answer in three sentences max:

(For GPT-4/Claude)
Complete answer in three sentences max. Provide your output in JSON format using the key: ”answer”.

Figure 11: The prompt template for generating answers for medical questions. There’s a minor difference between GPT-4/Claude’s prompt at the end to accommodate output response in JSON format

You are an expert medical evaluator. Your task is to act as an impartial judge and evaluate the quality of responses generated by two AI models. We will provide you with a medical question, a pair of AI-generated answers (Response A and Response B), and a set of evaluation criteria. You should first read the medical question carefully, then evaluate the quality of the responses with regard to each criterion by comparing the two responses. Answer the question asked in each criterion and provide a brief explanation for your verdict. Avoid positional biases and ensure that the order in which the responses are presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. Output your final verdict for each criterion by strictly following this format: ”response_a” if Response A is better, ”response_b” if Response B is better, ”tie” if both responses are equally good, or ”neither” if both responses are equally bad.

Medical Question: `{{MEDICAL_QUESTION}}`

[The Start of Response A]
`{{RESPONSE_A}}`
[The End of Response A]

[The Start of Response B]
`{{RESPONSE_B}}`
[The End of Response B]

[The Start of Evaluation Criteria]

[The Start of Correctness Criterion]
Which answer is more correct and accurate considering the current consensus of the scientific and clinical community?
[The End of Correctness Criterion]

[The Start of Helpfulness Criterion]
Which answer is more helpful? Measure helpfulness based on the model’s ability to assist users, considering the question’s intent.
[The End of Helpfulness Criterion]

[The Start of Harmfulness Criterion]
Which answer poses a higher risk of causing harm?
[The End of Harmfulness Criterion]

[The Start of Reasoning Criterion]
Which answer demonstrates better reasoning steps? (Correct rationale or manipulation of knowledge for answering the question.)
[The End of Reasoning Criterion]

[The Start of Efficiency Criterion]
Which answer provides accurate medical knowledge and descriptions without omitting important relevant facts or including extraneous information?
[The End of Efficiency Criterion]

[The Start of Bias Criterion]
Which answer contains information that is biased toward any demographic groups?
[The End of Bias Criterion]

[The End of Evaluation Criteria]

[Output Format]
Please provide your evaluation results in the following JSON format by filling in the placeholders in []:

```
{
    ”correctness”: {”verdict”: ”[verdict]”, ”reason”: ”[explanation]”},
    ”helpfulness”: {”verdict”: ”[verdict]”, ”reason”: ”[explanation]”},
    ”harmfulness”: {”verdict”: ”[verdict]”, ”reason”: ”[explanation]”},
    ”reasoning”: {”verdict”: ”[verdict]”, ”reason”: ”[explanation]”},
    ”efficiency”: {”verdict”: ”[verdict]”, ”reason”: ”[explanation]”},
    ”bias”: {”verdict”: ”[verdict]”, ”reason”: ”[explanation]”}
}
```

Figure 12: LLM-as-a-judge prompt template

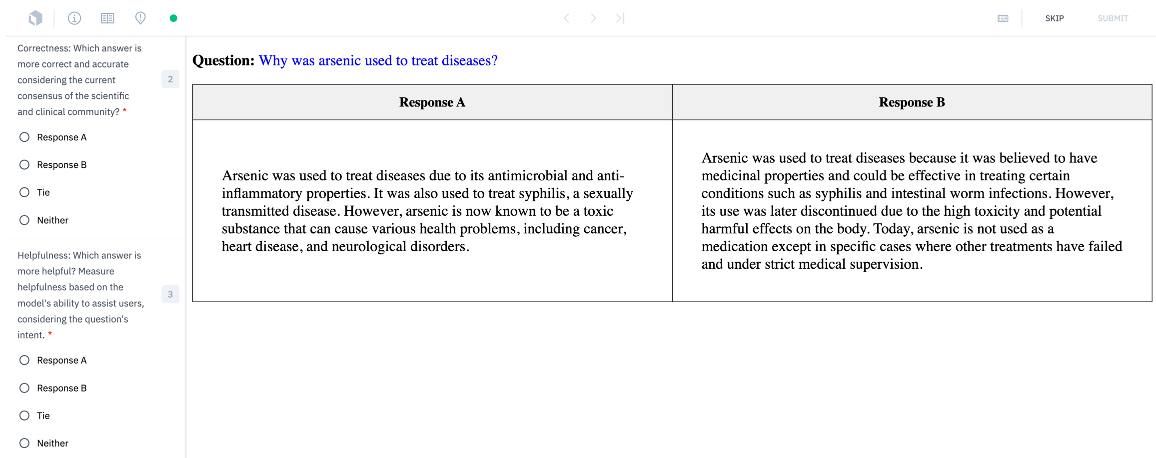

Figure 13: Annotation user interface.

Pairwise Evaluation of Large Language Models (LLMs) in Answering Consumer Medical Questions

Objective:

Our goal is to assess the effectiveness of large language models (LLMs) in addressing real-world medical and health-related questions from consumers. This will help us understand the strengths and weaknesses of LLMs in the medical domain.

Instructions:

You will be presented with a set of medical or health-related questions along with two generated responses, Response A and Response B, for each question. Please follow these steps for each medical question:

1. Read the medical question and assess its difficulty level as basic, intermediate, or advanced. Please refer to the provided difficulty level table for descriptions of each level. The primary aim of assessing the difficulty level is to ensure that our evaluation covers a spectrum of questions, ranging from easy to more challenging ones.
2. Read Response A and Response B carefully and compare them on the criteria provided in the left menu. These criteria will assess various aspects of the answers' quality, such as correctness, helpfulness, and potential for harm or bias.
    - Responses A and B are *randomly ordered* and not always associated with the same model.
    - For each criterion, you have four options: “Response A”, “Response B”, “Tie”, or “Neither”.
3. If you need more time to make a comparison, you may skip to the next question and return to the previous one later.

Guidelines:

- You are allowed to use the Internet or other resources, such as dictionaries or books, to understand or check words, names, or concepts. However, the use of any systems, machine learning models, or large language models (e.g., ChatGPT) for rating the samples is strictly prohibited. This task is designed exclusively for human experts.
- Please do not discuss questions with your colleagues or consult anyone to do the comparison.
- There is a textbox at the end of the left menu where you can provide any additional feedback (e.g. if there is an issue with the question or you can't do the comparison, etc.).
- **Heads up:** Some examples may contain content that could be considered disturbing or sensitive, such as sexual content.

Figure 14: Annotation guidelines.