CQA-EVAL: Designing Reliable Evaluations of Multi-paragraph Clinical QA under Resource Constraints

Federica Bologna[1] **Tiffany Pan**[1] **Matthew Wilkens**[1] **Yue Guo**[2] **Lucy Lu Wang**[3]
[1]Cornell University [2]University of Illinois Urbana-Champaign [3]University of Washington
Correspondence: fb265@cornell.edu

Abstract

Evaluating multi-paragraph clinical question answering (QA) systems is resource-intensive and challenging: accurate judgments require medical expertise and achieving consistent human judgments over multi-paragraph text is difficult. We introduce CQA-EVAL, an evaluation framework and set of evaluation recommendations for limited-resource and high-expertise settings. Based on physician annotations of 300 answers to 100 real patient questions answered by physicians and LLMs, we compare coarse answer-level versus fine-grained sentence-level evaluation over the dimensions of correctness, relevance, and risk disclosure. We find that inter-annotator agreement (IAA) varies by dimension: fine-grained annotation improves agreement on correctness, coarse improves agreement on relevance, and judgments on risk disclosures remain inconsistent. Additionally, annotating only a small subset of sentences can provide reliability comparable to full annotations, reducing cost and effort.[1]

1 Introduction

Rising healthcare costs and limited provider availability make it difficult for patients to receive timely answers to their clinical questions. Generative models integrated into electronic health record (EHR) systems could help, but evaluating their responses requires medical expertise (Ayers et al., 2023; Singhal et al., 2023). Expert annotators are costly, scarce, and often disagree on what constitutes a good answer, leading to low inter-annotator agreement (IAA) (Hosseini et al., 2024; Leonardelli et al., 2021; Khasentino et al., 2025). Simplifying the annotation task is thus fundamental to reducing annotation fatigue and improving annotation reliability.

Although fine-grained sentence-level annotations improve IAA on faithfulness in summarization (Krishna et al., 2023), it remains unclear whether these benefits extend to other dimensions, particularly in high-stakes domains where disagreement is common (Leonardelli et al., 2021; Mostafazadeh Davani et al., 2022; Hosseini et al., 2024). In this work, we validate these guidelines in the high-expertise, high-consequence setting of consumer health QA and examine their generalizability to annotation facets beyond faithfulness.

We introduce CQA-EVAL, a framework for evaluating multi-paragraph clinical QA across the dimensions of *correctness*, *relevance*, and *communicates-risks* (whether the answer discloses risks). We conduct a randomized human annotation study on pairs of patient questions and clinical answers, tasking annotators with evaluating answers under two annotation designs: a *coarse* design in which raters assess the entire answer, and a *fine-grained* design in which they evaluate individual sentences. This allows us to assess whether specific annotation designs can reduce annotator fatigue while maintaining or improving annotation quality. We also assess the impact of annotation design on system ratings, the viability of partial annotations, and whether our annotation designs can be adapted for LLM-as-judge settings.

Based on our annotation study findings, we derive a set of recommendations for achieving reproducible annotations under resource and expertise constraints. We find, for example, that fine-grained annotations improve IAA primarily for factual dimensions such as correctness, while coarse annotations are sufficient or sometimes better for dimensions like *relevance* and *communicates-risks*. At the same time, partial fine-grained annotations—evaluating only a few sentences per answer—can achieve ratings and variance comparable to full fine-grained ratings while reducing annotation cost. These results suggest that annotation design should be tailored to the evaluation dimension: fine-grained schemes for factuality-based dimensions,

[1]Data and code available under a CC-BY license at https://github.com/federicabologna/cqa-eval.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 32694–32707
October 24-29, 2026 ©2026 Association for Computational Linguistics

Dataset (author, year)	# QA pairs	Question source	Reference source	Number raters	Number criteria	IAA	Expert ratings	Codebook	LLM-as-judge
Ben Abacha et al. 2019	2,479	Web	Extracted	3	1	✓	✓	✗	✗
Ayers et al. 2023	100	Reddit	Extracted	3	3	✗	✓	✗	✗
Singhal et al. 2023	140	Mix	Experts	1	12	✗	✗	✗	✗
Singhal et al. 2025	1,446	Mix	Experts	15	12	✓	✗	✗	✗
Our study	300	Health app	Experts	6	3	✓	✓	✓	✓

Table 1: Comparison of clinical QA evaluation datasets and annotation practices.

coarse ones for context-based ones.

Our LLM-as-judge experiments show that fine-grained instructions are less effective at improving LLM-expert agreement than expert-expert agreement. Nonetheless, our annotation instructions yield LLM-expert agreement comparable to expert-expert agreement for correctness and relevance. In addition, we find that fine-grained evaluation may help counteract length-related biases of coarse assessments.

In practice, fine-grained ratings for correctness and coarse ratings for relevance provide reliable, efficient metrics for assessing LLM-generated clinical answers. In sum, our contributions include:

- A **dataset** of 300 QA pairs with annotations for answer correctness, relevance, and risk disclosure collected from six independent medical experts, with each pair evaluated by three experts;
- The CQA-EVAL **annotation framework** for both coarse and fine-grained evaluation of answers to clinical questions;
- A **comparison study** of coarse and fine-grained annotations through a randomized human annotation study of ground-truth clinician answers and those generated by two LLMs widely used for this task;
- A set of **recommendations** that helps developers of clinical LLMs find the best annotation design for their goals;
- An **analysis** of how our annotation framework generalizes to LLM-as-judge settings.

2 Related Work

Annotation Rubrics Human evaluation is the gold standard for assessing clinical QA, as it captures qualities such as factual correctness, safety, and contextual appropriateness that automated metrics often miss (Singhal et al., 2023, 2025; Yim et al., 2025; Khasentino et al., 2025). Recent benchmarks incorporate human evaluation but rely on broad categorical rubrics without detailed annotation guidance: MultiMedQA (Singhal et al., 2023) and MedQA (Jin et al., 2021) use three-level scales for correctness and harm, while HealthBench (Arora et al., 2025) and MEDIC (Kanithi et al., 2024a) apply general Likert-style ratings at the answer level. Such underspecified schemes reduce clarity, inter-annotator agreement, and reproducibility. Moreover, annotator agreement is seldom reported: among 14 expert-annotated studies, only four report IAA (Singhal et al., 2025; Ben Abacha and Demner-Fushman, 2019; Johnson et al., 2023; Hosseini et al., 2024), and just one (Ayers et al., 2023) releases annotations. This lack of transparency hinders the development of automatic metrics and systematic comparison between human and LLM judgments. To address these shortcomings, we introduce CQA-EVAL, a structured framework with explicit definitions, sub-criteria, and examples, enabling more consistent and actionable evaluation of clinical QA. In addition, we release the human and LLM-generated annotations from our study, along with the corresponding IAA.

Annotation Granularity Most prior work in clinical QA uses answer-level evaluation, where annotators assign a single score to the entire response (Singhal et al., 2023; Arora et al., 2025; Kanithi et al., 2024a; Guo et al., 2024; Yim et al., 2025; Mallinar et al., 2025). While simple, this approach obscures mixed-quality content. Long answers may include both correct and incorrect statements, safe and unsafe advice, or relevant and irrelevant details. Collapsing these into one score makes error localization difficult, increases subjectivity, and reduces inter-annotator agreement (Castilho, 2020; Popović, 2021; Krishna et al., 2023). In contrast, sentence-level evaluation assesses each unit independently, offering clearer error signals, higher reliability, and more actionable feedback for system development.

3 Annotation Study Design

To evaluate reliable designs for multi-paragraph clinical QA, we augment a dataset of real patient

questions with LLM-generated answers (§3.1), define evaluation dimensions and granularities (§3.2), and recruit experienced physicians to provide annotations (§3.3).

3.1 Dataset

We conduct experiments on the K-QA dataset (Manes et al., 2024), which consists of real patient questions submitted to K Health, an AI-driven clinical platform with over 8 million users. These consumer health questions span general primary care topics, from medication safety to disease prevention, covering hundreds of medical conditions from neurology to reproductive health. Reference answers in the dataset are curated by a team of medical doctors.

We randomly sample 100 questions and their corresponding physician answers from K-QA for our annotation study. We augment these QA pairs by generating answers to the patient questions using GPT-4 (OpenAI et al., 2024) and Llama-3.1-Instruct-405B (Grattafiori et al., 2024). We focus on these two models because they are widely deployed and have shown strong performance on medical tasks in previous studies (Kanithi et al., 2024a). The models are prompted using five-shot in-context learning (ICL) (Brown et al., 2020) and chain-of-thought reasoning (Wei et al., 2022). The full prompt is available on Github. In each case, we instruct the model to limit its answers to 270 words, the maximum length of physician answers in the K-QA dataset, and set the max number of output tokens to 300. Four answers generated by Llama-3.1-Instruct-405B still exceeded this limit and are returned with truncated sentences; we remove these sentences from the analysis. Table 2 shows answer length distributions; both models produce longer replies on average than the original physicians. We provide example model and physician replies in Appendix A.

Each of the 100 sampled patient questions is paired with three answers (physician, GPT-4, Llama), yielding 300 total QA pairs.

3.2 Annotation Task Formulation

There is no consensus on which dimensions of clinical answers should be assessed. After carefully reviewing codebooks and rubrics of previous clinical evaluation studies (Singhal et al., 2023, 2025; Hosseini et al., 2024; Ayers et al., 2023; Ben Abacha and Demner-Fushman, 2019; Ben Abacha et al., 2023; Kanithi et al., 2024b), we identify three dimensions:

Answer type	Mean tokens ± SD
GPT-4	124 ± 50
Llama-3.1-Instruct-405B	170 ± 52
Physician (ground-truth)	106 ± 54

Table 2: Average number of tokens and standard deviation for each answer type in our dataset.

- **Correctness**: the answer aligns with current medical knowledge
- **Relevance**: the answer addresses the specific medical question asked
- **Communicates-risks**: the answer communicates contraindications or risks

We decide to evaluate risk disclosure rather than possibility of harm because it is complementary to correctness. Both of these dimensions (omitting mentions of potential risks and providing incorrect medical information) have the potential to harm the reader. More detailed and extensive definitions of these dimensions can be found in Appendix B.

In our study, annotators evaluate multi-paragraph answers to patient questions along these three dimensions under two annotation settings:

- **Coarse-grained**: annotators are shown a question and an answer, and are asked to consider the *answer as a whole*; they rate each dimension of the answer on a five-point Likert scale—*Disagree, Partially Disagree, Neutral, Partially Agree, Agree*.
- **Fine-grained**: annotators are shown a question and an answer with one sentence highlighted, and are asked to consider the *sentence* in the context of the whole answer; they assess each dimension of the sentence on the same five-point Likert scale.

Prior work on fine-grained annotations explored a variety of units, such as entities, numbers, clauses, or sentences (Krishna et al., 2023). We choose to conduct fine-grained evaluation at the sentence-level because answers in the K-QA dataset are generally concise and annotating entities would be prohibitively expensive. As 75% of answers consist of fewer than eight sentences, we randomly sample six sentences from each answer to be evaluated, balancing annotation cost with coverage. We hypothesize that dimensions where evidence is localized (e.g., correctness of a factual claim) would

benefit most from fine-grained annotations.

We develop two codebooks, one for each setting, providing annotators with detailed definitions and criteria, as well as instructions and real-world examples. Information in our codebooks is drawn from reputable sources such as Mayo Clinic, Cleveland Clinic, and Harvard Medical School. The full codebooks can be found on Github. Annotators have access to the codebook at all times while performing the annotation task.

3.3 Experimental Design

Annotators We hire six physicians from the freelance recruiting platform Upwork following previous work (Yim et al., 2025; Jin et al., 2023). We select candidates based on their credentials and their recent jobs on the platform. After the participants provide their consent, we further determine their clinical expertise through two pilot studies where they are asked to provide motivations in case of large disagreements with other annotators.

All the selected physicians have first-hand experience with patient care. They report specialties in General Medicine, Internal Medicine, Pathology, Radiology, or Surgery and had between 3 and 15 years of patient-care experience. Annotators are compensated at a rate of 25 USD per hour for their work.

Annotation Design We divide the six annotators into two groups, where *Group A* is assigned all answers (from the physician and two LLMs) for 50 randomly sampled patient questions and *Group B* is assigned all answers for the remaining 50 patient questions. In other words, each QA pair is evaluated by three of the six physicians, and each physician evaluates 150 QA pairs.

For each annotator, half of these assigned QA pairs are annotated following a coarse annotation design, and half following a fine-grained annotation design. To implement this, each group's 150 pairs are split into 17 coarse batches and 17 fine-grained batches, with up to 9 pairs per batch except for batch 17, which contains 6 pairs.

The order of questions within each batch is randomized to control for confounding due to fatigue. For each annotator, we also randomize in which batch each of the 50 questions appears to control for temporal effects in annotation decisions. However, to reduce context switching, all three possible answers for each question are grouped and always assigned to the same batch.

To minimize bias introduced by the sequence in which the two annotation designs are performed, we alternate coarse and fine-grained annotation tasks. Annotators first complete nine coarse batches, followed by eight fine-grained batches with entirely new pairs. They then evaluate nine fine-grained batches on pairs they had previously annotated under the coarse scheme, and conclude with the remaining eight coarse batches, which include pairs they had already evaluated in the fine-grained setting.

Lastly, annotators repeat annotations for 6 QA pairs in both coarse and fine-grained settings that they had evaluated at the beginning of the study, allowing us to measure intra-rater reliability (IRR) for each annotator.

Additional Measures To measure *annotator effort*, we record the time elapsed between when annotators are shown a new QA pair and when they proceed to the next one. For each QA pair, we also measure *annotator confidence*. We ask annotators to respond to the question "How confident do you feel about your evaluation of the answers?" on a five-point Likert scale ranging from *Not confident at all* to *Very confident*.

At the end of each annotation batch (9 QA pairs), we collect feedback on the clarity of the codebook instructions. Annotators rate "How easy to follow are the annotation instructions?" on a five-point Likert scale ranging from *Very difficult* to *Very easy*.

Post-annotation Survey At the end of the annotation study, annotators complete a two-part feedback survey structured after the NASA Task Load Index (TLX) (Hart and Staveland, 1988). The first part asks about their experience in the coarse annotation setting, while the second repeats the same questions for the fine-grained setting. Questions are reproduced in Appendix C; annotators answer each question on a 7-point scale. Lastly, we collect qualitative feedback by asking the open-ended question "Which annotation design do you prefer?"

3.4 LLM-as-judge

Limited evidence exists regarding the ability of LLMs to provide human-aligned evaluations of model-generated clinical answers. Jin et al. (2023) show that expert annotators agree 75% of the time with GPT-4o regarding the correctness of clinical answers. However, Szymanski et al. (2025) find a much lower agreement rate in the context of dietetics and mental health. We fill this gap by assessing

our evaluation dimensions with LLM-judges, and reporting the IAA between the LLM-judge and our expert annotators. We investigate whether LLM-as-judge agreement with expert raters is impacted by coarse- versus fine-grained prompting strategies.

Following prior work, we select GPT-4o as our judge model (Chen et al., 2025; Hosseini et al., 2024; He et al., 2024; Kanithi et al., 2024b; Szymanski et al., 2025). Although this choice may bias LLM-judge ratings in favor of GPT-4-generated answers because the models belong to the same family, we use GPT-4o because it is the most commonly used judge model in prior clinical work, allowing for a more direct comparison between our results and prior work. We prompt the model with our two codebooks, each minimally adapted from the human versions. Full prompts can be found on our Github page.

3.5 Measures of Reliability

To ensure comparability across settings, we collapse coarse annotations from a 5-point to a 3-point Likert scale. Subtle distinctions (e.g., Agree vs. Partially Agree) introduced variability without added value, consistent with prior work showing that collapsing categories improves reliability in subjective tasks (Artstein and Poesio, 2008; Davani et al., 2022). We map ratings as follows: `+1` for Agree/Partially Agree, `0` for Neutral, and `-1` for Disagree/Partially Disagree.

For fine-grained annotations, we consider two schemes. In Fine 3pt, we collapse 5-point ratings to 3-point as above at the sentence level. In Fine Binary, we aggregate sentence-level ratings into a passage-level judgment using dimension-specific rules. For correctness, an answer is assigned `0` if any sampled sentence is rated incorrect and `1` otherwise. For relevance and communicates-risks, an answer is assigned `1` if at least one sampled sentence is rated as relevant or as communicating risks, respectively, and `0` otherwise. This binary scheme provides an interpretable answer-level aggregation of the fine-grained annotations that can be compared against coarse annotations and across systems.

We use IAA as our primary reliability measure and report Randolph's κ following prior work on expert evaluation of clinical QA (Singhal et al., 2025). For comparisons between coarse and fine-grained annotation settings, we compute and report Randolph's κ by pooling ratings from all six annotators for each annotation setting. Error bars show the 95% CI from 100 bootstrap resamples. For comparisons of system-level performance, we report average annotator ratings and p-values computed using the Mann-Whitney U test (Mann and Whitney, 1947). For repeated annotations, we measure IRR as percent self-agreement. We also report annotators' self-rated confidence and ease (average Likert scores), and effort (average seconds per QA pair, aggregated across sentences in the fine-grained setting).

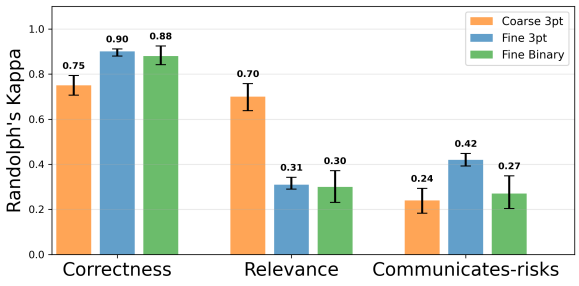


Figure 1: Average inter-annotator agreement (IAA) (Randolph's κ) for expert annotators across all annotation groups for the three dimensions: correctness, relevance, and communicates-risks. Error bars show the 95% CI from 100 bootstrap resamples.

4 Results & Recommendations

Fine-grained annotations achieve higher IAA for correctness but not for other dimensions. For correctness, agreement is substantially higher in the fine-grained settings (0.90, 0.88) compared to coarse 3-point (0.75) (Figure 1). In contrast, in the case of relevance, annotators reach higher agreement in the coarse condition (0.71) than in either fine-grained condition (0.31, 0.30). Communicates-risks shows low agreement in all settings, with an increase in the fine-grained conditions (0.42, 0.27) compared to coarse 3pt (0.24). Because IAA improvement in fine-grained settings varies based on the aggregation strategy, we cannot reliably conclude that fine-grained annotations improve IAA for communicates-risks.

The first part of our results confirms findings from Krishna et al. (2023) that IAA increases with fine-grained annotations. However, we find this only holds true for specific dimensions, notably correctness. We hypothesize that dimensions that rely on factuality and the alignment of a single piece of text against a reference—such as faithfulness (as assessed in Krishna et al. (2023)) and correctness—are more suited to a fine-grained annotation design. Dimensions that are constructed across mul-

tiple pieces of texts (such as relevance), or that rely on identifying missing information (such as communicates-risks) may not be efficiently captured by fine-grained annotations.

This hypothesis is supported by annotator responses to our feedback survey. They reported that, while fine-grained annotations helped them provide more accurate ratings, those annotations were more difficult when context was important. A sample of quotes from annotator responses is available in Appendix D. Furthermore, responses on the NASA-TLX (Table 3) show that annotators, on average, found coarse annotations to be easier and less mentally and physically demanding. However, they felt more successful and less discouraged when completing fine-grained annotations.

In addition to comparing IAA across the two annotation settings, we also measure IRR in both coarse and fine-grained settings on 6 QA pairs experts evaluated at the beginning and end of the study. We report these results in Appendix E. Given the small sample size, these results did not achieve statistical significance and do not offer strong evidence for recommendations. Future work should investigate whether a fine-grained design improves IRR across dimensions.

Recommendation: Researchers should choose an annotation design based on the dimensions being assessed. Developing a detailed codebook, and collecting annotator feedback through pilot studies, surveys, or interviews before starting the evaluation can help researchers decide whether coarse or fine-grained annotations are better suited for the dimension of interest.

Partial fine-grained annotations achieve comparable results at half the cost. Calculating Spearman's rank correlation coefficient between partial and full fine-grained annotations, we find that ratings based on assessing only three sentences per answer already correlate strongly (correlation coefficient above 0.8) with six-sentence annotations (Figure 2 left). Furthermore, variance in fine-grained ratings decreases rapidly with additional sentences: with only three sentences, variance is lower than coarse annotations for correctness and communicates-risks, and only slightly higher for relevance (Figure 2 right).

Although coarse annotations are significantly faster—taking an average of 239.3 seconds per answer compared to 459.8 seconds for full fine-grained annotations (Table 3; p<0.001, effect size of -0.454)—sampling three sentences reduces the annotation time and cost to a level comparable with coarse annotations, while maintaining the benefits in IAA. In terms of confidence, annotators report similar ratings for coarse (mean = 3.37) and fine-grained (mean = 3.40) annotations, with no significant difference (p = 0.54).

Measure	**Coarse**	**Fine**	p**-value**
Confidence	3.37	3.40	$= .54$
Ease	3.05	2.56	$< .001$
Time spent (sec)	239.3	459.8	$< .001$

Question	**Coarse**	**Fine**
Mental demand	4.00	4.50
Physical demand	3.67	4.33
Temporal demand	3.83	3.50
Performance	6.00	6.33
Effort	4.00	4.83
Frustration	2.50	1.50

Table 3: Above: Comparison of mean annotator self-reported confidence, ease, and time taken to complete tasks between coarse and fine-grained annotation settings. Below: Mean NASA-TLX ratings for coarse and fine-grained tasks. Ratings are reported on a 7-point scale where higher indicates more of the thing in question. Statistical significance was computed with Mann Whitney U.

Recommendation: Sampling a portion of the answer can be a cost-effective strategy for fine-grained evaluation. We find that this preserves reliability and variance relative to larger samples, while keeping annotation costs and annotator effort comparable to coarse annotations.

Fine-grained annotations may mitigate biases related to response length when computing system-level ratings and rankings Fine-grained annotations result in comparable system-level ratings (similar range of ratings) for the three annotation dimensions, but produces different system rankings (Figure 3). GPT-4's and Llama-3.1-Instruct-405B's answers receive near-perfect ratings for correctness in both the coarse (0.92 and 0.93) and fine-grained (0.98 and 0.93) settings. However, where this minimal difference in performance between the two system is not statistically meaningful in the coarse setting, it is significant in the fine-grained one, where GPT-4 shows better performance. Furthermore, physician answers receive much lower

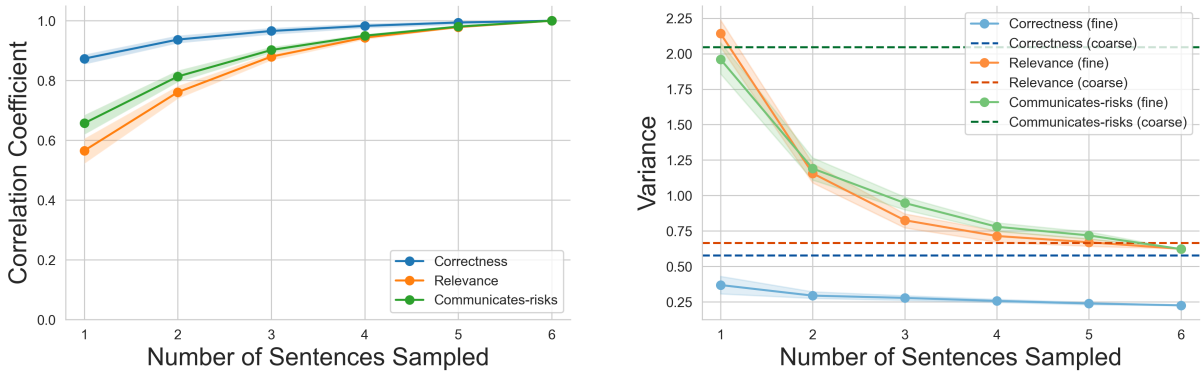


Figure 2: Correlation between partial fine-grained annotations and full fine-grained annotations and variance when partially annotating an answer (left). Inter-annotator variance when partially annotating an answer compared against coarse annotations. Confidence intervals shown are 95% and computed across 100 random subsets.

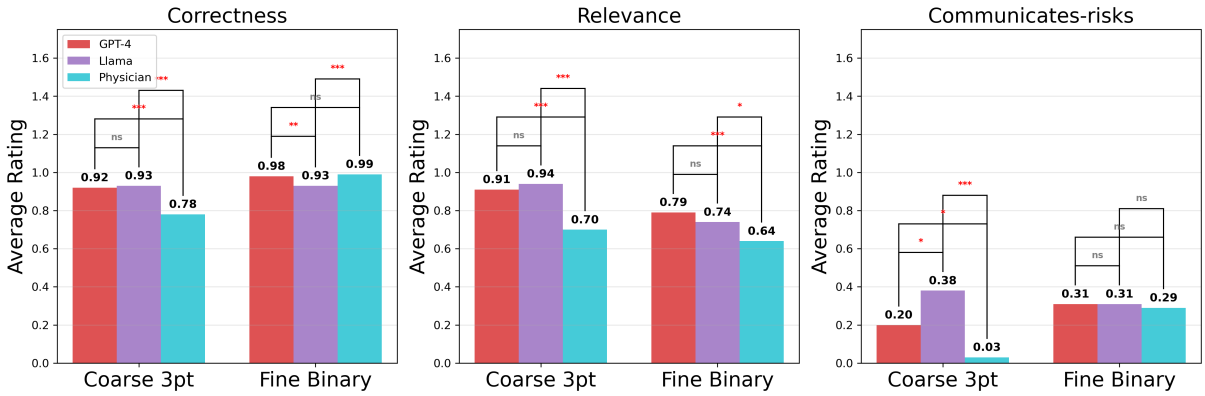


Figure 3: Comparison of system-level average coarse versus fine-grained ratings for correctness, relevance, and communicates-risks. Scores range from 0 to 1, with 1 indicating optimal performance. Statistical significance was computed using Mann-Whitney U.

correctness ratings in the coarse condition (0.78) than in the fine-grained one (0.99).

Physician answers are consistently shorter and more to the point than model-generated answers, offering limited elaboration (Table 2). We hypothesize that this difference biases annotators toward rating physician-written answers lower, despite comparable correctness. Length may be conflated with informational content, and the style of information presentation used by models may be preferred by users. We speculate that by constraining annotators to evaluate one sentence at a time, fine-grained annotations help annotators separate characteristics related to correctness (e.g., accuracy) from those related to relevance (e.g., elaboration), leading to fairer ratings.

For relevance, coarse and fine-grained designs produce different system-level ratings, but the relative system ranking is consistent between the two. All systems receive higher relevance scores in the coarse annotation setting compared to the fine-grained one (0.91, 0.94, 0.70 vs. 0.79, 0.74, 0.64). However, in both annotation conditions, GPT-4 and Llama-3.1-Instruct-405B answers are rated statistically similarly, while physician answers are rated significantly lower.

For communicates-risks, Llama's ratings are similar across coarse and fine-grained settings (0.38 vs. 0.31), whereas GPT-4 and physician outputs show substantially lower communicates-risks ratings for coarse annotations (0.20 and 0.03) than for fine-grained annotations (0.31 and 0.29). Similar to correctness, the lower ratings for physician answers disappears in the fine-grained annotation setting. Overall, however, all models and physicians are rated poorly in terms of communicates-risks across both settings.

Our ratings are consistent with prior work showing that state-of-the-art LLMs can perform at or above physician level on measures of correctness and relevance, while struggling with communicates-risks. (Ayers et al., 2023; Kanithi

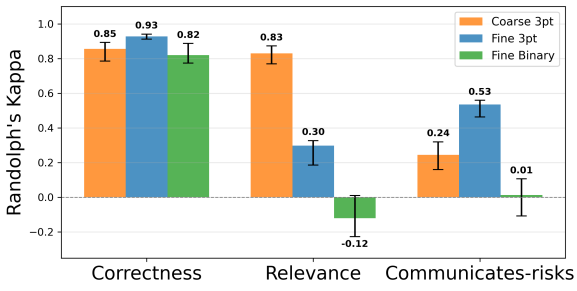


Figure 4: Average inter-annotator agreement (Randolph's κ) between average expert ratings and LLM-judge ratings for correctness, relevance, and communicates-risks. Error bars show the 95% CI from 100 bootstrap resamples.

et al., 2024b; Hosseini et al., 2024).

Recommendation: Our findings suggest that fine-grained annotations should generally be preferred for producing system ratings and rankings, particularly for correctness, because they may reduce biases introduced by answer length and presentation style yielding fairer comparisons across systems. Coarse-grained annotations are adequate for dimensions such as relevance. More work is needed to assess whether model-generated answers provide satisfactory risk information.

Prompting with fine-grained annotation instructions does not consistently improve agreement between LLM-as-judge and expert annotators. Figure 4 shows IAA between GPT-4o (our judge model) and expert annotators in the coarse and fine-grained evaluation settings. We observe that prompting with fine-grained instructions does not consistently improve human-LLM agreement across rating schemes. Fine-grained 3-point ratings enhance IAA for correctness and communicates-risks, while decreasing it for relevance. In contrast, fine-grained binary ratings reduce IAA across all dimensions. Even so, when prompted with our annotation instructions, LLM-as-judge can achieve agreement with experts that is comparable to—or even higher than—expert-expert agreement for correctness and relevance.

Recommendations: We do not propose LLM-as-judge as a replacement for expert evaluation, but as a supplement when expert resources are constrained. If leveraging LLM-as-judge to evaluate multi-paragraph clinical QA, fine-grained annotation instructions can improve agreement with expert raters for factual dimensions such as correctness on a 3-point scale, while coarse annotation instructions remain more suitable for dimensions such as relevance.

In our experiments, we also find that prompting GPT-4o with our annotations instructions yield LLM-expert IAA for correctness and relevance that is comparable to, and in some cases exceeds, expert-expert IAA. These findings suggest that LLM-as-judge could be effectively used to supplement expert judgments when evaluating these dimensions in annotation studies of clinical QA with constrained resources.

5 Discussion & Conclusion

In this work, we introduce CQA-EVAL, a framework for evaluating multi-paragraph clinical question answering using both coarse (answer-level) and fine-grained (sentence-level) annotation settings. Through a randomized human annotation study of 300 QA pairs of real patient questions, we find that annotation design matters by dimension: fine-grained annotations improve IAA for factual dimensions like correctness, while coarse annotations remain more reliable for context-dependent dimensions such as relevance. However, our framework did not improve IAA for communicates-risks, which remains low in both fine and coarse annotation settings. This result may be diagnostic of the intrinsic difficulty of this dimension: identifying missing or insufficient risk disclosures requires annotators to reason about what should have been stated in the answer, rather than what it contains. The low agreement suggests that annotation design alone is insufficient for achieving reliable judgments for this type of dimension; future work should explore alternative strategies (e.g., counterfactual prompts, risk checklists).

We further demonstrate that partial fine-grained evaluation (annotating only three sentences per answer) closely approximates agreement compared to full fine-grained annotations while reducing annotation effort and cost, making it a practical alterative for resource-constrained studies, particularly when evaluating correctness.

Fine-grained annotations may also reduce sensitivity to differences in answer length and presentation style in system-level evaluations. In our study, shorter physician answers received substantially lower correctness ratings under coarse evaluation than fine-grained evaluation, suggesting that passage-level judgments may be conflating correctness with other characteristics of the response.

Across designs, communicates-risks remains a persistently challenging dimension, with low ratings and low inter-annotator agreement. This highlights the need for targeted evaluation and improvement methods for evaluating appropriate risk disclosure in consumer health QA.

Overall, our findings provide actionable guidance for designing multi-paragraph clinical QA evaluation studies. We recommend employing partial fine-grained annotations to achieve high IAA for correctness while keeping annotator workloads manageable. For relevance, we recommend retaining coarse-level annotations to preserve context. These insights can help researchers and practitioners balance annotation quality, cost, and reliability, ultimately supporting the safe and effective deployment of LLM-powered question-answering systems in patient care.

Limitations

Our dataset is relatively small and consists of only questions on primary care topics, which may limit the generalizability of our findings and their applicability to clinical questions about conditions that require specialized care, or to diverse patient populations who might experience such conditions. While the dataset we use in this study is smaller than other datasets with free-text response answers, we chose this dataset because it contains real patient questions and reference answers curated by a team of physicians, providing a high-quality testbed for studying evaluation methods in a realistic consumer health QA scenario. It is also easily accessible, allowing for the reproducibility of our study. To this set we add LLM-generated answers, a codebook for evaluation, expert ratings collected in two settings, IAA measurements, and LLM-as-judge ratings, providing a helpful expansion to the existing literature. We provide a comprehensive comparison to prior work in Table 1.

Second, expert annotators were recruited on UpWork. While we did request medical credentials and verify them to the best of our ability, there is still a possibility that the credentials our annotators provided are outdated, overstated, or falsified.

Third, we recruited only six expert annotators, which constrains the variability in perspectives and may affect the robustness of inter-annotator agreement (IAA) and intra-rater reliability (IRR) estimates. This choice reflects the practical constraints of working with scarce and costly medical expertise. Our annotation study employs more annotators than other comparable evaluation studies in the medical QA domain (Ben Abacha and Demner-Fushman, 2019; Jin et al., 2023; Ayers et al., 2023; Singhal et al., 2023; Khasentino et al., 2025; Yim et al., 2025), and the results we report achieve statistical significance. Our study design also mirrors the resource-limited conditions under which clinical QA systems are often evaluated in practice, and our recommendations are intended to be useful in such settings.

Fourth, sentence-level annotations may not guarantee coverage of incorrectness between sentences (e.g. dosages), although our findings show that sentence-level annotations produce safety ratings equivalent to passage-level ratings.

Fifth, the sample collected to measure IRR is small, potentially limiting the precision of reliability assessments across dimensions. Among this small sample, we find that fine-grained annotations do not achieve higher IRR across all dimensions. We provide further details and interpretation in Appendix E. This result highlights the need for larger follow-up studies to confirm whether the trends we observe in IAA extend to individual annotator self-consistency.

Finally, our evaluation focuses on only two widely used language models, GPT-4 and Llama-3.1-Instruct-405B. These models were chosen because they have been documented as performing well on clinical QA, and we limit the number of models in our study in order to obtain more annotations over a larger set of questions. Nonetheless, our findings may not extend to other models or to future versions of these models with different capabilities. Future work should investigate larger datasets, more annotators, and a broader range of models to validate and extend these findings.

Ethical Considerations

The design of our human annotation study was submitted to our university's Institutional Review Board (IRB), which determined that no IRB approval or exemption was required, as the study does not meet the regulatory definition of human subjects research. All annotators recruited through Upwork were fairly compensated at a rate of 25 USD per hour, consistent with their suggested hourly wage. We provided participants with sufficient time to complete their annotations, offered flexible deadlines to accommodate personal or health-related needs, and remained available to address feedback and questions throughout the study.

Acknowledgments

This work was supported by the University of Washington eScience Institute's AWS cloud credits program, and gift funding from Google and the Allen Institute for AI.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 183(6):589–596.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An Investigation of Evaluation Methods in Automatic Medical Note Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *NeurIPS*.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. Association for Computational Linguistics (ACL).

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. Appls: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 9194.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland. ISSN: 0166-4115.

Zhe He, Balu Bhasuran, Qiao Jin, Shubo Tian, Karim Hanna, Cindy Shavor, Lisbeth Garcia Arguello, Patrick Murray, and Zhiyong Lu. 2024. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. *Journal of Medical Internet Research*, 26(1):e56655. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

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 [cs].

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.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. *arXiv preprint*. ArXiv:2310.13132 [cs].

Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, Elizabeth Scoville, Tyler Reese, Debra Friedman, Julie Bastarache, Yuri van der Heijden, Jordan Wright,

Nicholas Carter, Matthew Alexander, Jennifer Choe, and 15 others. 2023. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Research Square*, pages rs.3.rs–2566942.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024a. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Praveen K. Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024b. MEDIC: Towards a Comprehensive Framework for Evaluating LLMs in Clinical Applications. *arXiv preprint*. ArXiv:2409.07314 [cs].

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A. Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, Robby Bryant, Ryan G. Gomes, Allen Jiang, Roy Lee, Yun Liu, Javier Perez, Jameson K. Rogers, Cathy Speed, Shyam Tailor, and 15 others. 2025. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, pages 1–10. Publisher: Nature Publishing Group.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neil Mallinar, A. Ali Heydari, Xin Liu, Anthony Z. Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, Javier L. Prieto, Daniel McDuff, and Ahmed A. Metwally. 2025. A Scalable Framework for Evaluating Health Language Models. *arXiv preprint*. ArXiv:2503.23339 [cs].

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, Bangkok, Thailand. Association for Computational Linguistics.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110. Place: Cambridge, MA Publisher: MIT Press.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Publisher: Nature Publishing Group.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966, Cagliari Italy. ACM.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *NeurIPS*.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. MORQA: Benchmarking Evaluation Metrics for Medical Open-Ended Question Answering. *arXiv preprint*. ArXiv:2509.12405 [cs].

A Example Questions from K-QA

Table 4 presents representative consumer health questions from the K-QA dataset alongside corresponding answers provided by the physician (from the dataset), GPT-4, and Llama-3.1-Instruct-405B. These examples illustrate a diversity of question types and differences in style, length, and content among human- and model-generated responses. **These examples are provided solely for research and discussion purposes and should not be interpreted as medical advice.**

B Annotation Dimensions

Extended definitions for our evaluation dimensions are as follows:

1. The answer aligns with current medical knowledge. This dimension considers whether the answer:
   - Provides information based on existing scientific evidence
   - Expresses certainty in line with the current state of medical knowledge
   - Does not contain contradictions or logical flaws
2. The answer addresses the specific medical question. This dimension considers whether the answer:
   - Directly addresses all the concerns included in the question
   - Provides requested details and context
   - Does not contain information unrelated to the question
3. The answer communicates contraindications or risks. This dimension considers whether the answer:
   - Explains risks in clear, accessible language
   - Details all potential consequences

C Feedback Survey (modeled after NASA-TLX)

Annotators complete a feedback survey at the end of our study. The following questions are adapted from the NASA-TLX and asked for each of the coarse- versus fine-grained annotation settings. Each question is answered on a 7-point scale:

- How mentally demanding was the task?
- How physically demanding was the task?
- How hurried or rushed was the pace of the task?
- How successful were you in accomplishing what you were asked to do?
- How hard did you have to work to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed, and annoyed were you?

D Feedback Survey Responses

These are some quotes from annotators' open-text responses to the question "Which annotation design do you prefer?" Fine-grained annotations are referred to as sentence-level in the survey.

> **Quote from Annotator A:** *Presenting the response in a well-structured paragraph ensures that all relevant information is covered comprehensively.*

> **Quote from Annotator B:** *Sentence level or "highlighted text" was kind of tricky because you need to put the highlighted text in context as well.*

> **Quote from Annotator C:** *I will prefer sentence-level annotation. Though it was more mentally demanding for me but there were more depth to the annotation and i felt i was able to annotate questions better [...]*

> **Quote from Annotator D:** *it was easier to interpret levels [Likert-scale values] in sentence type especially the third question [communicates-risks].*

E Intra-rater reliability (IRR)

Fine-grained annotations achieve higher IRR for only one of three dimensions (Figure 5). For correctness and relevance, IRR is lower in the fine-grained settings (86.4 and 63.9, respectively) compared to the coarse 3-point settings (97.2 and 86.1). Communicates-risks shows a higher IRR in the fine-grained setting (66.7) compared to the coarse setting (47.2).

These findings are based on a limited sample of question–answer pairs (only six repeated QA pairs per annotator) and therefore have substantial uncertainty. Larger studies are needed to determine whether these descriptive patterns are robust.

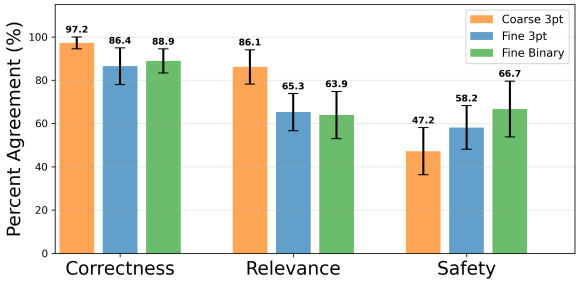


Figure 5: Average intra-rater reliability (IRR) (percent agreement) across annotators for correctness, relevance, and communicates-risks.

Question	*Is accutane like an antibiotic and would it affect a baby if I got pregnant?*
Physician	Isotretinoin (Accutane) is a form of vitamin A that helps to reduce the amount of oil produced by your skin's oil glands, but it's not an antibiotic. This drug is used to treat severe acne that has not responded to other treatments. It's important to note that this medication can cause severe birth defects or even fetus death if taken during pregnancy. It's crucial to have a negative pregnancy test before starting Isotretinoin and to use two forms of birth control throughout the treatment. If you're considering becoming pregnant or are already pregnant, you should speak to your healthcare provider about alternatives.
GPT-4	Accutane, also known as isotretinoin, is not an antibiotic. It is a form of vitamin A that helps the skin renew itself more rapidly and is used primarily to treat severe nodular acne. Unlike antibiotics, which eliminate bacteria causing infections, isotretinoin works by reducing the amount of oil released by oil glands in the skin and helping the skin renew itself. Isotretinoin can have serious side effects and is known to cause severe birth defects if taken during pregnancy. It is critical to avoid pregnancy while taking this medication. Women of childbearing age are required to use two forms of birth control and undergo regular pregnancy tests before, during, and after the treatment with isotretinoin. It is vital to consult with a healthcare provider to discuss the risks and necessary precautions while on this medication.
Llama-3.1	Accutane, also known as isotretinoin, is not an antibiotic but a powerful oral retinoid medication used to treat severe acne, including nodular and conglobate acne, which are resistant to other treatments. Accutane can cause severe birth defects if taken during pregnancy. It is contraindicated in pregnant women, and the FDA mandates the iPLEDGE program to monitor and prevent use during pregnancy. This program requires patients to commit to using two forms of birth control and undergo regular pregnancy tests to prevent pregnancy while taking Accutane.If you are taking Accutane and become pregnant, you should immediately stop taking the medication and inform your doctor. Due to the high risk of birth defects, your doctor may discuss options for ending the pregnancy. It is essential to carefully consider the risks and benefits of taking Accutane, especially if you are a woman of childbearing age.
Question	*what causes ringworm ?*
Physician	Ringworm is a contagious skin infection caused by fungi. It is characterized by a circular, red, and itchy rash and can affect the skin, hair, and nails. The infection can be transmitted through contact with people or pets, as well as through shared spaces like locker rooms and shower stalls.Treatment usually consists of topical antifungal drugs.
GPT-4	Ringworm, also known as tinea, is a fungal infection of the skin caused primarily by dermatophytes such as Trichophyton, Microsporum, and Epidermophyton. These fungi thrive in warm, moist environments, making the skin a favorable site for infection. It spreads through direct contact with infected individuals, animals, or contaminated objects like towels, clothing, and surfaces. The infection is characterized by a circular, red, itchy rash that may be scaly or have a raised border, resembling a ring. Proper hygiene and avoidance of sharing personal items can help prevent ringworm. Treatment typically involves antifungal medications applied topically or taken orally depending on the severity and location of the infection.
Llama-3.1	Ringworm, also known as tinea, is a fungal infection that affects the skin, hair, and nails. It is caused by a group of fungi called dermatophytes, which include Trichophyton, Microsporum, and Epidermophyton. These fungi thrive in warm, moist environments and can infect humans through direct contact with an infected person, animal, or contaminated object. Ringworm can also be spread through indirect contact, such as walking barefoot in areas where the fungus is present, or by sharing personal care items like towels or clothing. In some cases, ringworm can be contracted from infected animals, such as cats or dogs. The most common causes of ringworm include: • Direct contact with an infected person or animal • Indirect contact with contaminated objects or surfaces • Sharing personal care items • Walking barefoot in areas where the fungus is present • Poor hygiene and skin care Symptoms of ringworm typically appear within 4-10 days of exposure and can include a circular, itchy rash, redness, and scaling. Treatment for ringworm usually involves antifungal medications, either topical or oral, and good hygiene practices to prevent further spread of the infection.

Table 4: Example patient questions and answers from physicians, GPT-4, and Llama-3.1-Instruct-405B. Text is lightly edited for formatting. These examples are provided for research and should not be taken as medical advice.