K-QA: A Real-World Medical Q&A Benchmark

Itay Manes[1], **Naama Ronn**[1], **David Cohen**[1], **Ran Ilan Ber**[1],
Zehavi Horowitz-Kugler[1], **Gabriel Stanovsky**[2]
[1]K Health Inc, New York, NY
[2]School of Computer Science, The Hebrew University of Jerusalem
{itay.manes,david.cohen,ran.ilanber}@khealth.com

Abstract

Ensuring the accuracy of responses provided by large language models (LLMs) is crucial, particularly in clinical settings where incorrect information may directly impact patient health. To address this challenge, we construct K-QA, a dataset containing 1,212 patient questions originating from real-world conversations held on K Health (an AI-driven clinical platform). We employ a panel of in-house physicians to answer and manually decompose a subset of K-QA into self-contained statements. Additionally, we formulate two NLI-based evaluation metrics approximating recall and precision: (1) *comprehensiveness*, measuring the percentage of essential clinical information in the generated answer and (2) *hallucination rate*, measuring the number of statements from the physician-curated response contradicted by the LLM answer. Finally, we use K-QA along with these metrics to evaluate several state-of-the-art models, as well as the effect of in-context learning and medically-oriented augmented retrieval schemes developed by the authors. Our findings indicate that in-context learning improves the comprehensiveness of the models, and augmented retrieval is effective in reducing hallucinations. We will make K-QA available to to the community to spur research into medically accurate NLP applications.[1]

1 Introduction

Recent advancements in large language models (LLMs) have led to a growing interest in their use in the medical domain in patient-facing applications, where LLMs hold the promise of providing laypersons with high-quality advice at a relatively low cost (Singhal et al., 2023; Han et al., 2023). For instance, in response to the question *"What's good for muscular pain?"*, a good patient-facing response may include, in addition to medical information (the name of a muscle relaxant), also the advice *"Seek medical attention if you have numbness or tingling in limbs"*.

However, there is a lack of benchmarks reflecting user needs and corresponding medically-accurate answers to test these models under real-world conditions. Most existing benchmarks assume textbook questions with multiple-choice or span-based answers (Tsatsaronis et al., 2015; Ben Abacha et al., 2017; Jin et al., 2019). In contrast, real-world questions, like *"Is there any way I can get some medicine for cold sore and ulcer that is killing me?"*, often include various interacting medical conditions (*"cold sore"*, *"ulcer"*), use ambiguous, non-medical jargon (*"that is killing me"*), and require long-form, nuanced answers.

In this work, we present K-QA, a medical QA benchmark containing 1,212 deidentified questions asked by real users on K Health,[2] an AI-driven clinical platform with over 8 million unique users. The questions in K-QA were curated from K Health's vast database of patient-physician interactions, aiming to capture stand-alone medical questions. These can be answered solely based on the information provided in the question, and do not require any prior knowledge about the patient's history or demographics. The resulting corpus is diverse and challenging, spanning over 100 different medical conditions (see examples in Figure 1).

To evaluate state-of-the-art models against K-QA, a team of 12 in-house medical doctors invested more than 400 person hours rigorously answering 201 questions from the dataset in a free-text format. Doctors consulted credible medical sources, such as UpToDate[3] and PubMed[4] to provide accurate and scientifically-backed answers. Their answers

[1]The data and the evaluation script are available at https://github.com/Itaymanes/K-QA. Results and model comparisons can be viewed at https://huggingface.co/spaces/Itaykhealth/K-QA

[2]https://khealth.com

[3]https://www.wolterskluwer.com/en/solutions/uptodate

[4]https://pubmed.ncbi.nlm.nih.gov/

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 277–294
August 16, 2024. ©2024 Association for Computational Linguistics

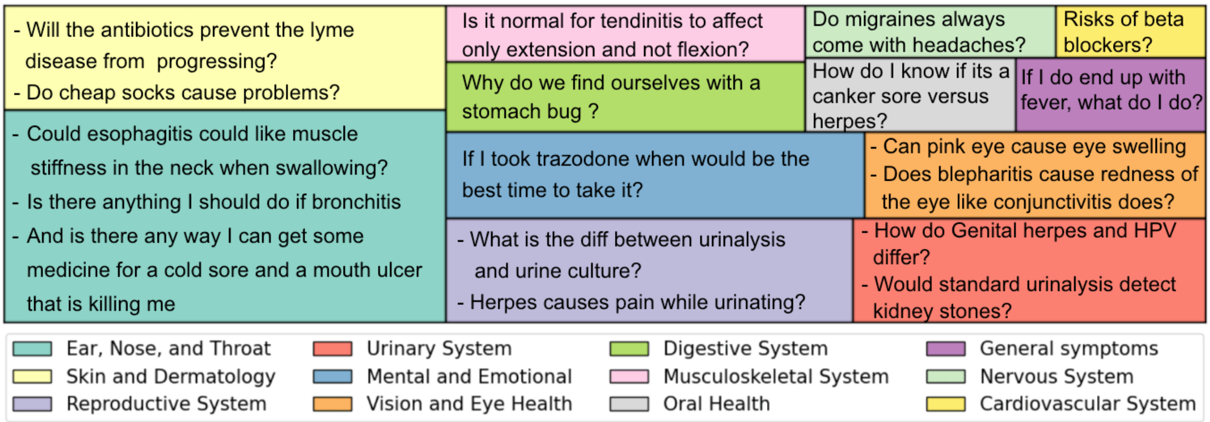

Figure 1: Visualization of K-QA, with box sizes indicating the distribution of patients' reported chief complaints across a wide range of healthcare topics. The questions are open-ended and diverse.

were further reviewed by an experienced overseeing physician. The remaining portion of roughly 1K questions present opportunities for expanding the benchmark through various augmentation techniques, as already demonstrated by Cherian et al. (2024).

To allow fine-grained evaluation, doctors decomposed each answer into an average of roughly 8 minimal semantic content units (Nenkova et al., 2007), resulting in over 1.5K individual statements. In addition, the importance of each statement was manually marked as either (1) *Must Have*, indicating that a model must include this statement in order to be medically accurate (e.g., providing all contraindications for a drug) , or (2) *Nice to Have*, indicating the statement is supplemental in nature (e.g., providing additional conditions where this drug may be helpful).

Following recent work on evaluation of text generation, we use the decomposed ground-truth answers in a natural language inference (NLI)-based evaluation of predicted answers (Honovich et al., 2021; Laban et al., 2022; Aharoni et al., 2023). Concretely, we define two complementing evaluation metrics. First, *comprehensiveness*, which is similar to recall, measures the percentage of ground-truth statements conveyed in the predicted answer. In order to excel in this metric, a model must cover all of the *Must Have* statements annotated by doctors. Second, *hallucination rate*, which is similar to precision, measures how many of all ground-truth statements (either *Must Have* or *Nice to Have*) contradict the predicted answer. To excel in this metric models must not produce *any* medically-inaccurate statements. We find that recent LLMs, like GPT-4, are able to approximate both comprehensiveness and hallucination rate, nearing human assessment of both metrics.

Finally, we evaluate various state-of-the-art LLM-based architectures on K-QA, spanning a wide range of families, including open- and closed-source models, zero-shot vs. in-context learning, and retrieval-augmented generation. We find that all models struggle on *comprehensiveness*, with the best performing model covering only 67.7% of medically-important statements, and while hallucinations seem to decrease with model size and augmented generation, all models still provide medically-dangerous advice in subtle ways which are especially risky for lay users.

We hope that future work adopts K-QA and accompanying metrics as a valuable benchmark to produce medically-accurate NLP applications which can be safely deployed in real-world scenarios.

2 Background

We review existing medical NLP datasets and other recent challenging benchmarks. We highlight key comparisons with K-QA in Table 1.

Medical QA benchmarks. Several diverse health-related question-answering datasets have been compiled, including over biomedical scientific literature (Tsatsaronis et al., 2015; Jin et al., 2019) and medical examinations (Zhang et al., 2018; Pal et al., 2022). The majority of these datasets rely on multiple-choice or span extraction (Jin et al., 2022), which simplify the evaluation process but do not reflect complexity of free-form responses which are often needed in real-world situations (Gehrmann et al., 2023).

Dataset	Consumer Health	Open Domain	Patient-Physician Interaction	Answer Decomposition	Answer Format
BioASQ (Tsatsaronis et al., 2015)	✗	✗	✗	✗	span-based & binary
MedQA (Jin et al., 2021)	✗	✓	✗	✗	multiple choice
LiveMedQA (Ben Abacha et al., 2017)	✓	✗	✗	✓	retrieval
MedicationQA (Abacha et al., 2017)	✓	✗	✗	✗	retrieval
MedQuAD (Ben Abacha et al., 2019)	✓	✗	✗	✗	retrieval
MEDIQA-AnS (Savery et al., 2020)	✓	✗	✗	✗	ranking
HealthSearchQA (Singhal et al., 2023)	✓	✓	✗	✗	-
K-QA (Ours)	✓	✓	✓	✓	long-form generation

Table 1: Comparison between K-QA and previous benchmarks in the field of medical question-answering.

In the context of consumer health questions, our dataset is different from existing benchmarks like MEDIQA-AnS (Savery et al., 2020), LiveMedQA (Ben Abacha et al., 2017) and MedicationQA (Ben Abacha et al., 2019) in several additional key ways. While these datasets source their questions from users searching healthcare websites via the ChiQA system (Demner-Fushman et al., 2020) and retrieve answers through keyword matching, ours originates from authentic patient-physician interactions, ensuring genuine medical inquiries. Furthermore, our dataset includes free-form open-domain responses carefully curated by medical professionals. In addition, the answers in K-QA are segmented into finer atomic statements, enabling fine-grained evaluation.

Challenging LLM benchmarks. Our work joins a recent line of test sets which are challenging for state-of-the-art LLMs, thus enabling further development and experimentation. For example, the Bamboogle benchmark consists of 125 multi-hop questions which stump popular search engines (Press et al., 2022), while the GPQA benchmark contains 445 graduate-level questions in various domains (Rein et al., 2023). K-QA consists of 1,212 questions, as well as a subset of 201 answers, specially curated by in-house physicians.

3 The K-QA Benchmark

In this section, we describe curation and annotation the K-QA dataset, depicted in Figure 2. K-QA consists of two portions - a medium-scale corpus of diverse real-world medical inquiries written by patients on an online platform (Section 3.1) and a subset of rigorous and granular answers, annotated by a team of in-house medical experts (Section 3.2). In Section 3.3, we present an analysis of the dataset, illustrating its medical and linguistic diversity.

3.1 Curating Questions from Real-World Patient-Physician Conversations

All of the questions in K-QA originate from de-identified real-world text-based conversations in English held on a proprietary online medical platform. These conversations contain a wide variety of user intents, such as billing inquiries or prescription renewals, alongside a wealth of queries on varied medical subjects (see Figure 1).

Our goal in creating K-QA is to extract from this large and noisy corpus a diverse dataset of *medical* questions which can be used to test automated models' ability to provide factual and comprehensive medical answers. In particular, we aim for the extracted questions to be as *stand-alone* as possible, without relying on the patient's medical record or the context of the medical discourse. For example, K-QA includes questions such as *"How do Genital herpes and HPV differ?"* (adapted from Figure 1), while we omit questions such as *"Can this allergic reaction be related to my age?"* which assumes prior knowledge about the patient and their previous symptoms.

To achieve this, we performed a rigorous manual annotation, aided by a preliminary automatic preprocessing step. First, we used an open-source BERT-based classifier (Devlin et al., 2018), fine-tuned for distinguishing questions from statements, such as *"sounds like hives to me"*.[5] Next, we ap-

[5] https://huggingface.co/mrsinghania/

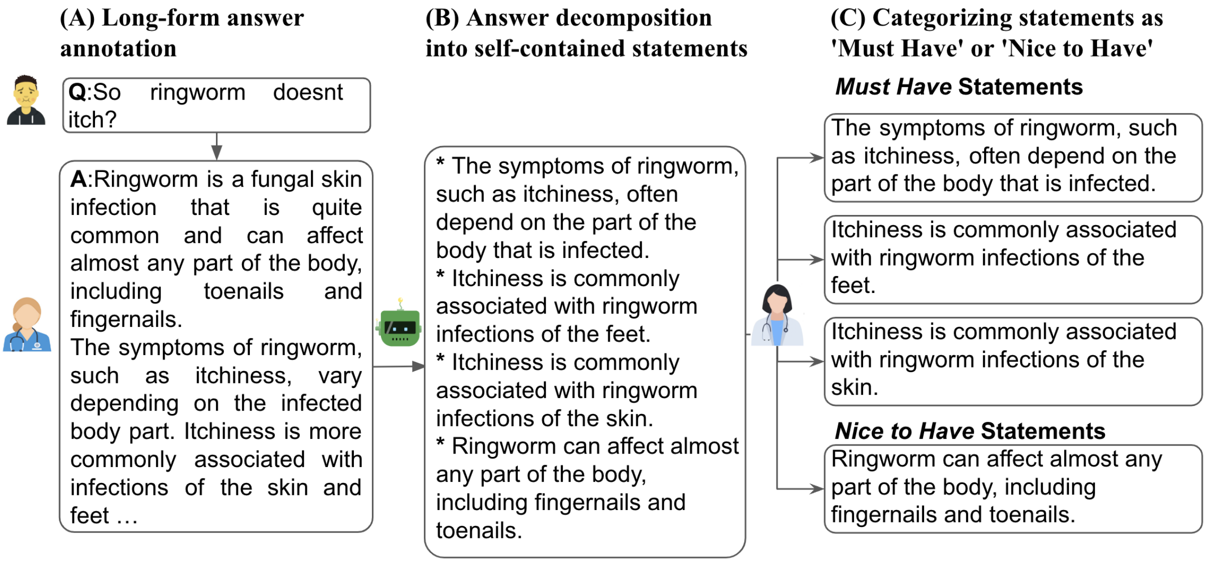

Figure 2: High-level description of the annotation of the K-QA dataset, starting with a physician's considered offline response to an actual patient query obtained from patient-physician interactions. We then use LLMs to decompose the response into self-contained statements, subsequently reviewed and categorized by a panel of medical experts as *Must Have* or *Nice to Have*. The example was simplified for presentation purposes.

plied regular expressions to filter questions about logistics (e.g., billing or delivery instructions). This preprocessing yielded roughly 26K questions, each individually assessed by a medical professional to identify those suitable as stand-alone questions. The dataset comprises diverse questions, each paired with the medical condition diagnosed by the physician at the end of their interaction with the patient, according to ICD-10 conventions (WHO, 1993). For example, the question in Figure 1 is classified as *"Dermatitis, unspecified"*.

3.2 Annotating Granular Physician Answers

We provide comprehensive and granular answers for a diverse subset of K-QA questions, annotated in three steps by a team of 12 in-house medical doctors. This subset enables us to automatically compare different LLMs against high-quality expert answers.

Step 1: Long-form answer annotation. In the first annotation step, exemplified in Figure 2(A), six medical physicians were tasked with providing free-form responses to different sets of questions from K-QA, while an additional physician reviewed their answers and advised where needed. Overall, the first step required roughly 400 skilled person hours (at a cost of roughly 26K USD, based on average physician hourly pay in the U.S.), during which 201 questions from K-QA were answered. Each physician was granted unlimited time and access to reputable medical resources such as UpToDate and PubMed for referencing purposes. Figure 3 depicts the distribution of used sources. Notably, they were explicitly instructed to *avoid* using any generative language models or services. To best emulate the requirements from a user-facing model in the medical domain, annotators were further instructed to write answers tailored for a lay audience seeking consumer-health information. For example, note how the answer in Figure 2 regarding ringworm strays from medical jargon.

Step 2: Answer decomposition into self-contained statements. Following literature on the evaluation of text generation via minimal semantic content units (Nenkova et al., 2007; Liu, 2022), we guided annotators to decompose answers into self-contained statements. Each statement is expected to capture a distinct fact and include sufficient context for independent evaluation. Answer decomposition is presented in Figure 2(B), illustrating the decomposition of a natural answer into atomic statements.

This step was carried out by a panel of 6 medical doctors (distinct from the annotators in the first step) who deconstructed each answer into individual statements. To assist in this process, the panel utilized GPT-4 with a few-shot prompt suggesting potential answer decompositions. The full prompt is provided in Appendix D.1. The annotators could

asr-question-detection

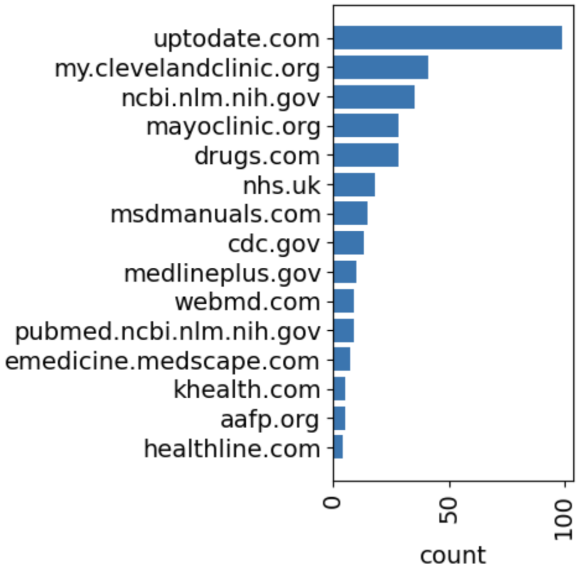

Figure 3: The 15 most used medical resources by the annotators during the curation of the long-form answers.

amend or remove noisy statements, as well as add any missing statements, which they did for 6.86% of the automatically generated statements. In total, this process yielded 1,589 annotated statements, averaging roughly 7.9 statements per answer. The completion of this phase required a total of approximately 30 hours at the cost of approximately 2K USD.

Step 3: Categorizing statements as *Must Have* or *Nice to Have*. At the last step, we asked the same group of medical professionals from step 2 to classify each statement into one of two categories: (1) *Must Have* – facets of the answer which are crucial to convey to a patient when providing medical advice; or (2) *Nice to Have* – statements which are supplemental or informative, but not clinically crucial. For example, as can be seen in Figure 2(C), the long-form answer regarding the itchiness of ringworm was decomposed into four statements, three of which were deemed as *Must Have*, while a statement which provided information about ringworm which is not related to its itchiness was deemed as *Nice to Have*.

A more complex example of medically oriented statement decomposition is presented below. In response to a question about treating hypertension in diabetic patients, the physician recommends ACE inhibitors or ARBs, either alone or in combination with other drugs like calcium channel blockers and thiazides. This scenario illustrates an exclusive OR relation often observed in medical contexts, where multiple treatments are optional but not advised together. To address the dependency between treatment options, we use our statement classification into *Must Have* and *Nice to Have* to preserve the physician's intention, emphasizing the importance of taking either ACE or ARBs and suggesting an additional optional treatment for each. This results in one *Must Have* statement: *"A recommended treatment includes either ACE or ARBs, but not both."*, and two *Nice to Have* statements: (1) *"ARBs can be taken alone or with other medications, such as calcium channel blockers and thiazides."*; and (2) *"ACE can be taken alone or with other medications, such as calcium channel blockers and thiazides."* The full annotation guidelines and more examples are provided in Appendix B.1.

This process was carried out collaboratively, facilitating discussions within the medical group to collectively assess and reach consensus regarding the perceived levels of importance, amounting to approximately 20 person hours, at the cost of roughly 1.5K USD.

The categorization into *Must Have* and *Nice to Have* represents a discrete approach to assigning importance to statements. However, in a broader context, this method can be extended to assign various weighted scores to each statement and each metric.

	Count	# Words (avg.)
Questions	1,212	10.06
Answers	201	88.52
Statements		
Must Have	892	14.9
Nice to Have	697	13.74

Table 2: Statistics of the K-QA benchmark.

3.3 Dataset Statistics

K-QA is derived from a diverse group of 1,055 unique users featuring 1,212 questions, including 201 answers meticulously curated by physicians. Table 2 shows detailed statistics on statements and word counts, while information on the distribution of age and biological sex among users can be found in Table 3.

The questions in our dataset address a wide array of health concerns, as evidenced in Figure 1, covering 172 different medical conditions, according to the ICD-10 system. The diversity in questions is highlighted in Figure 4, showing the top five

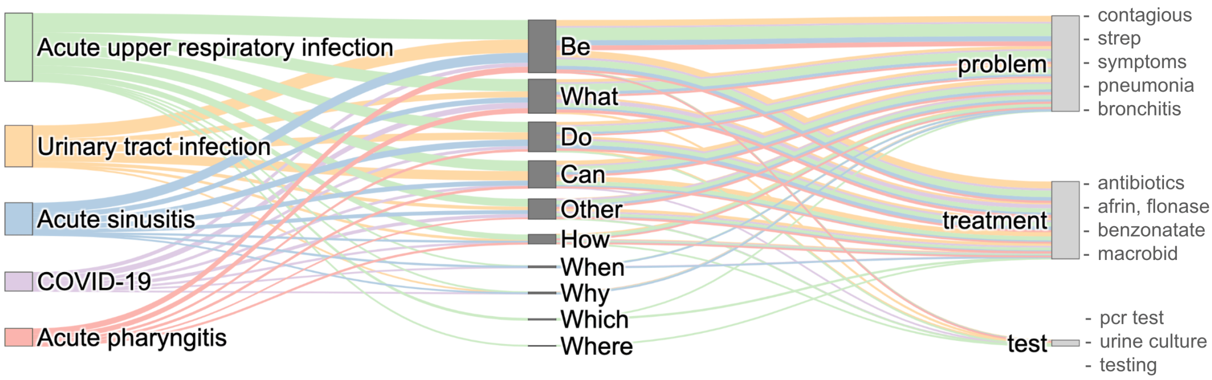

Figure 4: Distribution of the top 5 most prevalent medical conditions, the types of questions related to each condition, and the frequencies of clinical entities within those questions. On the far right, the text most frequently matched with the clinical entities is displayed.

Age Group	Sex (% of users)	
	Female	Male
18-25	9.09	6.67
26-45	36.97	30.30
46-60	9.70	3.64
60+	2.42	1.21

Table 3: Distribution of the users in K-QA by age group and biological sex.

most prevalent medical conditions and their distributions across various question types, such as Be, WH-questions and other forms, as well as various clinical categories (*problem*, *treatment*, and *test*) classified by an open-source, fine-tuned named entity recognition BERT-based model.[6] For example, a question like *"Is it usually normal for someone to have side effects when first starting vitamins?"* is labeled with both *problem* ("side effects") and *treatment* ("vitamins").

4 Evaluation Metrics for K-QA

To evaluate models against K-QA, we propose a natural language inference (NLI; Dagan et al., 2005; Bowman et al., 2015) framework, following recent text generation evaluation (Honovich et al., 2021; Laban et al., 2022; Aharoni et al., 2023).

Our evaluation framework is inspired by FActScore (Min et al., 2023), a metric that measures factual *precision* by computing the percentage of *atomic facts* in a generated answer supported by a reliable external source. Unlike FActScore, which automatically generates statements and assigns them equal importance, our approach involves predefined medical statements with varying clinical significance. This modification enables us to extend the framework and establish a proxy metric for factual *recall* as well.

We consider a predicted answer as a *premise* and each ground-truth statement derived from an annotated answer as an *hypothesis*. Intuitively, a correctly predicted answer should entail every ground-truth statement. This formulation aims to quantify the extent to which the model's answer captures the meaning of the gold answer, abstracting over the wording chosen by a particular expert annotator.

As formulated below, we devise two NLI-based metrics: *comprehensiveness* and *hallucination rate*. These adapt the evaluation of text generation to the medical domain by taking into account K-QA's annotation of *Must Have*, i.e., clinically crucial facets of information, and *Nice to Have* statements, which are supplemental in nature. Both metrics were aggregated across all assessed questions, where higher values of the comprehensiveness metric and lower values of hallucination rates indicate better performance. Figure 5 provides an example illustrating the complete process of evaluating a generated answer and deriving these metrics.

Formally, let $\hat{P}$ denote the model's predicted answer, *Must Have* represents the set of ground-truth statements marked as crucial, *Nice to Have* represents the set of ground-truth statements marked as supplemental, and $S = \textit{Must Have} \cup \textit{Nice to Have}$ is the set of all statements in the gold reference answer.

Comprehensiveness metric. This metric measures how many of the clinically crucial claims are

[6] https://huggingface.co/samrawal/bert-base-uncased_clinical-ner

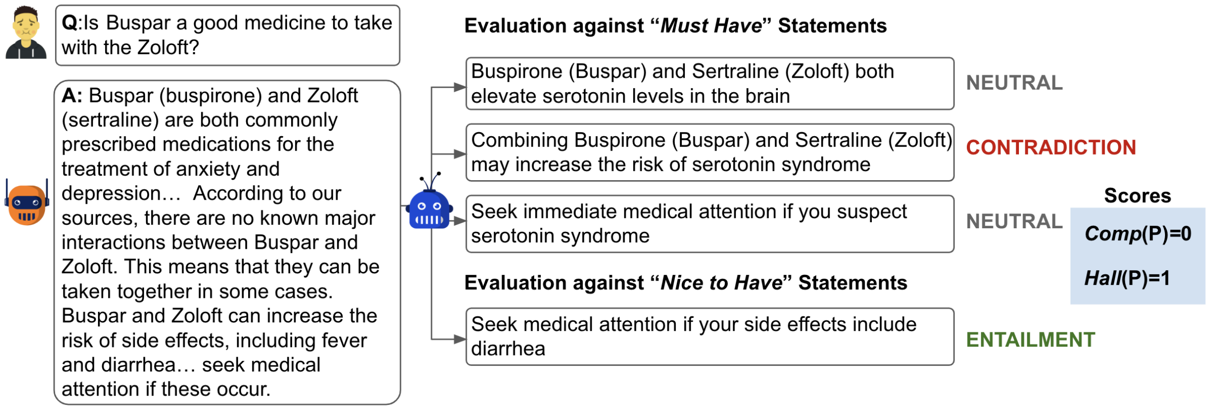

Figure 5: An example of the evaluation procedure, starting with a patient's question and a generated answer from a language model. Each statement is then automatically tested in an NLI framework to determine its relationship to the generated answer. Finally, our metrics are computed, where $Hall(\hat{P})$ counts the number of contraditions (1), and $Comp(\hat{P})$ is equal to 0, because none of the *Must Have* statements were entailed. Different robot symbols signify different models, and the example was simplified for presentation.

included in the predicted answer.

$$Comp(\hat{P}) = \frac{|\{x \in Must\_Have | \hat{P} \text{ entails } x\}|}{|Must\_Have|} \tag{1}$$

I.e., similarly to recall, $0 \leq Comp(\hat{P}) \leq 1$ measures how many ground-truth statements were conveyed in the predicted answer. We particularly focus on those statements marked as crucial by medical experts and do not penalize models for not covering supplemental statements, as these may be somewhat open-ended and arbitrary.

Hallucination rate. This metric measures how many of the ground-truth statements contradict the model's answer.

$$Hall(\hat{P}) = |\{x \in S | \hat{P} \text{ contradicts } x\}| \tag{2}$$

I.e., $Hall(\hat{P}) \in \{0, 1, ..., |S|\}$ penalizes answers that contradict any of the ground-truth statements and hence discourages models from making any sort of false medical statements. Similar to precision, a model can trivially achieve a perfect hallucination score by generating an empty answer $\hat{P} = \emptyset$ since, by definition, no hypothesis contradicts an empty premise.

Automatic evaluation. Following work on NLI-based evaluation, we approximate the metrics above via an automated NLI model. We used GPT-4 in conjunction with few-shot Chain-of-Thought (CoT; Wei et al., 2022) prompt that generates sequential intermediary text representations. The full prompt can be found in Appendix D.2. To assess the quality of the evaluation framework, we randomly selected 50 pairs of questions and their corresponding generated answers from the models described in Section 5.1. This process yielded 398 unique statements pertaining to the specified set of 50 questions.

Three physicians received instructions on how to classify the logical relationship for each triplet (*question*, *answer*, *statement*) into one of three NLI categories. The inter-agreement among annotators was assessed using Fleiss' kappa (κ; Fleiss, 1971) and pairwise agreement. For the three human annotators, the pairwise agreement was 83.2%, and the κ was calculated to be 0.70, signifying moderate to substantial agreement among raters. The agreement with the majority vote of the annotators and the automated model was 83.0%, indicating that the model can perform at a level comparable to human annotators for this complex task.

5 Evaluating State-of-the Art Models

Following the creation of K-QA and the formulation of evaluation metrics, we turn to evaluate the current state of the art in this challenging task.

5.1 Experimental Setup

Models. We use K-QA to evaluate the medical capabilities of 7 recent LLM-based models from diverse families and model sizes. Specifically, we evaluate two 7B instruction-tuned open access models: Mistral (Jiang et al., 2023), and MedAlpaca (Han et al., 2023) which was built upon LLaMA (Touvron et al., 2023) and trained specifically for biomedical tasks, three recent closed instruction-tuned LLMs: Open AI's GPT-3.5 and GPT-4 (Brown et al., 2020; OpenAI, 2023), and Google's PALM-2 (Anil et al., 2023), and finally two recent commercial closed generation search engines: BARD,[7] and Bing Chat.[8] We use zero temperature sampling for all models, except for BARD and Bing Chat, which do not allow setting temperature.

Retrieval augmented generation (RAG). We note that BARD and Bing Chat differ from the other 5 models in our evaluation in that they can reportedly augment their prompt with content retrieved from external sources, albeit in an undisclosed manner. To examine the effect that retrieved content may have on the performance of the other models, we implement Retrieval Augmented Generation approach (RAG; Lewis et al., 2020), which produces responses by conditioning the language model on both the input query and retrieved content. To achieve this, we index publicly available medical documents aimed at the lay audience (such as MayoClinic[9] and NHS[10]) aiming for medical-specific RAG. All the documents in this RAG are publicly available, which is distinct from the primary sources that the physician annotators used to create their answers (Figure 3).

Prompts. Most of our evaluations use the same vanilla zero-shot prompt without prompt engineering which only presents the question, without any additional instructions. In addition, for some models we also report results on another empirically engineered prompt which includes three in-context examples, to explore some of the effect that in context learning (ICL) may have on performance. We did not extensively explore ICL across all models due to various constraints, including MedAlpaca's limited context window, and determining optimal prompts for specific tasks and models remains an open research question (Sclar et al., 2023; Mizrahi et al., 2024). Consequently, we focused on the models that demonstrated the best performance, which are GPT-3.5 and GPT-4. Further exploration of prompt optimization in our case presents an interesting avenue for future work.

[7]https://bard.google.com/
[8]https://www.bing.com
[9]https://www.mayoclinic.org/
[10]https://www.nhs.uk/

Model	*Comp* ↑	*Hall* ↓	*%resp*
MedAlpaca 7B	31.4	56.7	100
Mistral 7B	47.6	28.4	100
PALM-2	50.8	31.3	100
BARD†	62.5	28.4	95.0
Bing Chat†	57.3	25.9	99.5
GPT-3.5	56.2	27.9	100
GPT-3.5+ICL	59.5	23.4	99.5
GPT-3.5+RAG†	50.5	17.9	89.0
GPT-3.5+ICL+RAG†	62.9	**15.4**	96.0
GPT-4	57.5	23.9	100
GPT-4+ICL	**67.7**	25.4	100
GPT-4+RAG†	52.2	22.9	91.5
GPT-4+ICL+RAG†	65.2	24.4	100

Table 4: Model performance on K-QA according to the comprehensiveness and hallucination rate metrics. ICL represents the addition of three in-context examples, and RAG is a medical retrieval augmented setup, as detailed in Section 5.1. The performance of the highest scoring model appears in **bold** for each metric. *%resp* indicates the percentage of questions answered by each model. †Marks models which have a retrieval component.

5.2 Results

The results for all models are shown in Table 4, in terms of the comprehensiveness and hallucination rate metrics defined in Section 4. Below, we highlight key observations based on these results.

Attaining high comprehensiveness is challenging even for state-of-the-art models. Across all models and prompts, the comprehensiveness metric (*Comp*) consistently remains below 68%. This is evident even in cases where models generated longer texts, as seen in the BARD model, which emitted nearly three times as many words per answer (242.1) compared to the physician's response (88.36 words). This underscores the models' difficulty in capturing what physicians consider critically important. Additionally, ICL improves comprehensiveness by instructing the model to include elements beyond a direct response to the question, such as assuming underlying medical concerns in patient inquiries.

While hallucinations seem rare, they could potentially lead to unintended and unsafe medical recommendations. The minimal hallucination rate, achieved by GPT-3.5+ICL+RAG, represents a contradiction of roughly 30 statements out of 1500 annotated examples. Some of the hallucinations may lead to subtle yet dangerous advice. For example, in Figure 5, the physician's statement asserts that *"Combining Buspar and Zoloft may increase the risk of serotonin syndrome"*, in contrast, the model claims that *"there are no known major interactions between Buspar and Zoloft"*. Finding cause of error in such cases is hard, and physicians are also prone to making dangerous errors. This particular error can be attributed to a combination of missing information within publicly available medical sources and by the LLM assuming that their omission implies the drug combination is safe.

For the GPT models in our evaluation, it seems that larger models lead to improved comprehensiveness, yet the larger the GPT model, the more it seems to introduce new hallucinations. In Table 4, we observe that GPT-4 outperforms GPT-3.5 under every comparable setting. However, the improved comprehensiveness comes at the cost of an increase in the hallucination rate.

Domain-specific RAG reduces hallucinations. Among all configurations, GPT-3.5+ICL+RAG demonstrates the fewest hallucinations while maintaining a comparatively good comprehensiveness score. We found that the tradeoff with comprehensiveness is partly thanks to its tendency to abstain from answering in certain questions, e.g., responding *"I'm unable to help, and don't have the ability to process and understand that."* (see *%resp* column in Table 4), which may be desired over misinformation in a patient-facing application. Computing the metrics only over the answered questions, this model receives a *Comp* score of 65.5% and a *Hall* score of 16.1, which is still lower than all other models, with the second-highest comprehensiveness score. However, Bing Chat and BARD, which also abstain, appear to underperform compared to their base models. This discrepancy might stem from our prompts lacking task optimization and their generic web retrieval, especially failing to focus on consumer health inquiries in the medical domain from reliable sources.

MedAlpaca performs poorly on K-QA. Even though MedAlpaca was fine-tuned specifically for the biomedical domain and intended for use as medical conversational AI, it exhibites the poorest results on both metrics, with an especially high hallucination rate. These findings indicate a mismatch between closed-QA (e.g., medical exams and short answers) and real-world patient questions which require the generation of long medical answers.

6 Conclusion

We introduce K-QA, a question-answering benchmark with real-world patients' questions and carefully curated physician answers. We formulate metrics that quantify how well a predicted answer covers important information and to what extent it contradicts gold answers. LLMs improve with size and augmented generation, but there is still a lot of room for improvement in both comprehensiveness and hallucination rate.

Limitations

One of the major limitations of our evaluation approach is its reliance on LLMs for approximating the entailment relation between ground-truth and predicted answers. In addition, the model which is used for the evaluation (GPT-4) is also then tested on the QA task, which may further confound our findings. While this was done in various recent works, it may propagate noise into the evaluation process, and yield a costly evaluation protocol. To mitigate this concern, we measure the agreement between human annotators and predicted labels, finding overall good agreement, while reducing evaluation costs an important avenue for future work (Perlitz et al., 2023). For the closed models, GPT-3.5, GPT-4, PALM-2, BARD and Bing, the responses are based on API calls, which are subject to changes in model versions, making reproducibility difficult.

Ethics Statement

The data in K-QA originates from deidentified real-world patient conversations that have been manually reviewed to ensure there it contains no personal information and revolves around general medical questions. The answers in K-QA were manually written by medical doctors, who did not use any automated writing assistance and wrote their answers with a general audience in mind. Our legal team has reviewed and approved the methodology used.

7 Acknowledgments

We would like to express our gratitude to Carmel Vider, Lee Herzog, Tamar Brufman, Nadia Mordenfeld, Lior Hayat, Ohad Simons, Jessica Thien, Shira Ben Porat, Itay Omer, Assaf Frajman, and Noa Bleistei n for curating and reviewing the medical answers. In addition, we would like to thank Nadav Rabani and Ya'ara Arkin for insightful discussion that influenced the direction of this work.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. 2024. Large language model validity via enhanced conformal prediction methods. *arXiv preprint arXiv:2406.09714*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

OpenAI. 2023. Gpt-4 technical report.

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

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2023. Efficient benchmarking (of language models). *ArXiv*, abs/2308.11696.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *ArXiv*, abs/2311.12022.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):322.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

WHO. 1993. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, volume 2. World Health Organization.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

A General Scoring Framework

In order to expand upon the definitions provided in 4, we define $w \in \mathcal{W}$ as the weight assigned to a specific statement $s \in \mathcal{S}$ for a given question. Additionally, we consider an NLI model $\mathcal{N}(premise, hypothesis)$ designed to classify each pair of *(premise, hypothesis)* into one of three labels: $l \in \{entail, contradict, neutral\}$. Within this framework, we view the hypothesis as the statement, and the generated response ($\hat{P}$) as the premise, denoted as $\mathcal{N}(\hat{P}, s)$. We define $f(\mathcal{N}(\hat{P}, s), l)$ as a function that takes the output of the NLI model and a predefined logical relation, such as "does not contradict," and returns a boolean value. The formula representing this process is as follows:

$$\sum_{s \in \mathcal{S}} w(s) \cdot f(\mathcal{N}(\hat{P}, s), l)$$

This mathematical expression quantifies how well the generated response aligns with a predefined logical condition, while taking into account the weights assigned to individual statement. The formulation of this equation is aligned with the metrics presented at section 4. For *Hall* computation, $w(s)$ is set to 1, whereas for *Comp*, $w(s)$ takes the value of 1 if s is in *Must Have* and 0 if s falls within *Nice to Have*.

B Annotators' Guidelines

B.1 Decomposition to Self-Contained Statements

We aim to evaluate the medical accuracy of responses generated by language models, specifically concerning stand-alone questions within medical conversations initiated by users. Given the potential for these answers to be open-ended, the evaluation task presents inherent challenges. To address this, we've devised a set of guidelines that break down answers into two components: "Must-have" statements, deemed essential for inclusion, and "Nice-to-have" statements, which, while beneficial, are not obligatory. Our objective is to formulate statements that are concise, accurate, definitive, and self-contained. It is imperative to ensure that the curated statements are medically correct and logically well-structured. The evaluation process is conducted independently for each statement, emphasizing the importance of avoiding overlap between statements to maintain clarity.

Guidelines for Various Scenarios

Below, a specific explanation is provided for different scenarios, accompanied by examples of good and bad options for decomposing an answer.

- **List of Unrelated Crucial Entities:** If the answer comprises a list of entities (e.g., red flags, vaccines, symptoms), and each entity is independently significant, consider treating each entity as a separate and distinct statement. However, if the list of entities is not critical to include in its entirety (e.g., suggestions for weekly menu options), these entities should be combined into a single statement and categorized as *Nice to Have*. The validation for these statements should ensure non-contradiction with the physician's input.

Question	*Must Have* - Good	*Must Have* - Bad
I am a young healthy adult, flying to Brazil next month. What vaccination should I take?	-Vaccination for Yellow Fever is recommended before traveling to Brazil. -Vaccination for Typhoid is recommended before traveling to Brazil.	Since you are traveling to Brazil next month, it is recommended to get fully vaccinated for several vaccines, including Typhoid and yellow fever.

Why is this example considered suboptimal?

- Single Answer Instead of Two: Instead of presenting two separate statements, the response combines both vaccines into a single answer. To enhance clarity and evaluation, it is recommended to break down such responses into distinct and independent statements, especially when the mentioned vaccines are not interdependent.

- Excessive Length: The response is too long and contains unnecessary information, particularly with the inclusion of prefixes like "Since you...". The focus should be on keeping the information concise and relevant.

- Overly Specific: The mention of the timeframe "next month" is overly specific and potentially misleading. It is crucial to provide information that is essential and directly related to the question.

- **AND/OR statements:** When entities have a logical relationship (e.g., treatment options), express them in statements following their logical connection rather than separating them. Entity 1 AND/OR Entity 2.

Question	*Must Have* - Good	*Must Have* - Bad
What is the best hypertension treatment for patients who are also diabetic?	Either angiotensin-converting enzyme (ACE) inhibitors OR angiotensin receptor blockers (ARBs), but not both.	-Angiotensin-converting enzyme (ACE) inhibitors -Angiotensin receptor blockers (ARBs)
	***Nice to Have* - Good**	***Nice to Have* - Bad**
	-angiotensin receptor blockers (ARBs) can be taken alone or with the following medications (thiazide and/or ccb). -angiotensin-converting enzyme (ACE) can be taken alone or with the following medications (thiazide and/or ccb)	

Why is this example considered suboptimal?

- Misleading statements: The entities ACE and the ARB are dependent on each other. If the language model (LLM) provides an answer recommending the patient take both ACE and ARB, it would be medically incorrect but might receive a high score in our evaluation method. In such cases, as ACE and ARB are distinct treatment options, we need to combine them in a statement to emphasize that only one of them can be prescribed, not both.

- Lack of inclusivity: Besides ACEs and ARBs, the treatment plan may involve other medications. We want to ensure that the response does not contradict the LLM's answer.

Moreover, if there are additional medications that can be prescribed alongside ACE or ARB, they should be listed in parentheses next to thiazide and/or CCB.

- **IF Statements:** Similar to AND/OR statements, maintain the coherence of "IF statements" if the cause is not a stand-alone factor.

	***Nice to Have* - Good**	***Nice to Have* - Bad**
What is the best hypertension treatment for patients who are also diabetic?	If lifestyle modifications alone are not effective in reducing blood pressure, medications may be necessary;	-Medications are needed to reduce the blood pressure. -Sometimes life modifications are not enough to reduce the blood pressure.

Why is this example considered suboptimal?

- Misleading statements: Splitting the IF statement leads to misleading statements. Even if the "bad examples" are medically correct, we might deviate from the intended verification of the IF statement.

- **Drugs Inclusion:** Include both the family drug name (generic) and trade name when dealing with drugs. For instance, "Low-risk drugs during pregnancy include aminosalicylates, such as sulfasalazine (Azulfidine) and mesalamine (Asacol, Pentasa).

C Annotators' Interface

Figures 6 display the annotation interface used for human evaluation during dataset creation. In this interface, annotators executed the second and third steps (described in 3.2 & 3.2). They were presented with the patient question, the free-form answer written by the physician, and the suggested decomposition of statements provided by GPT-4. Annotators were tasked with confirming, modifying, or removing the suggested decomposition to ensure relevance and self-containment. Additionally, they categorized the statement into one of the categories: *Must Have*, *Nice to Have*, or deemed it irrelevant.

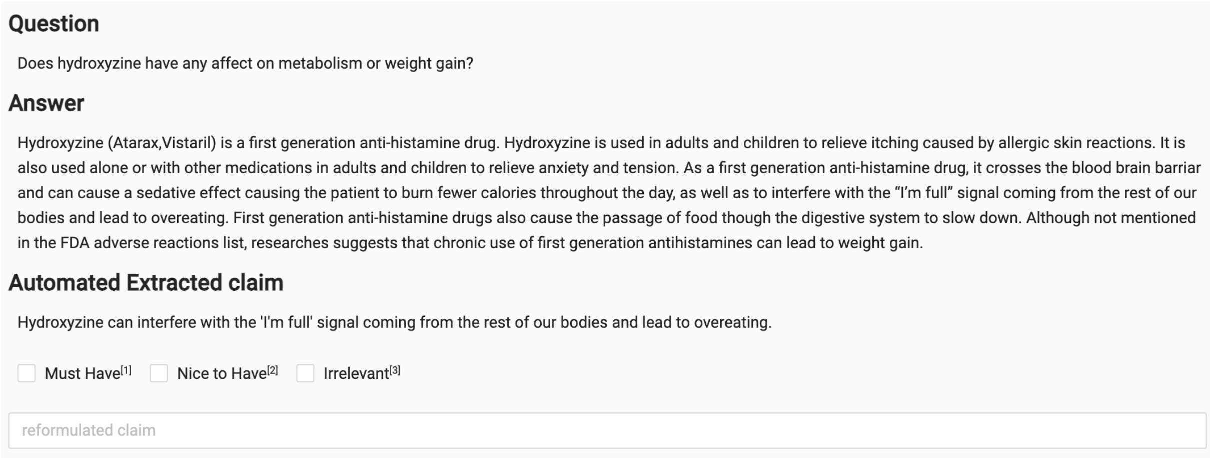

Figure 6: An example of the annotator's interface.

D Prompt Templates

D.1 Decomposition Free-Form to Statements

```
# OVERALL INSTRUCTIONS
You are an expert in understanding logical relationships. This is a Semantic Content Unit (SCU) extraction task. Given a pair of Question and Answer, your goal is to create a list of self-contained and concise claims. Each claim should be able to stand alone and be independent of other claims. Your claims should encompass all the information present in the answer.

# TASK INSTRUCTIONS
- List of Possible Causes: For scenarios involving multiple entities like red flags, vaccines, symptoms, etc., generate separate claims for each entity. This increases the number of claims.
- OR Claims: When medical entities are presented in an "OR" context, treat them as distinct claims.
- IF Claims: When an "if statement" is present, preserve the "if statement" context while creating the claim.
- XOR Claims: When entities have an XOR logical relationship (e.g., treatment options), create separate claims for each option.

# EXAMPLE CLAIM FORMAT - List Format: "Possible cause for [CONDITION] in [DEMOGRAPHIC] can be [ENTITY]."
- OR Format: "Possible causes include: [ENTITY X], [ENTITY Y], and [ENTITY Z]."
- OR Format: "The [CONTEXT] of treatments such as [TREATMENT X], [TREATMENT Y], and [TREATMENT Z], is not well established." - IF Format: "[CONTEXT], please seek medical attention if [CONDITIONS]."
- XOR Format: "Either take [TREATMENT X] or [TREATMENT Y], but not both."
——
{format_instructions}
——

# TASK EXAMPLE
Question: I am a 33-year-old female with right lower abdominal pain, what could it be? Answer: Possible causes for right lower abdominal pain in a young female are Appendicitis, Inflammatory bowel disease, Diverticulitis, Kidney stone, urinary tract infection, Ovarian cyst or torsion, Ectopic pregnancy, Pelvic inflammatory disease, endometriosis. Please seek medical attention if the pain is sudden and severe, does not go away, or gets worse, is accompanied by fever, nausea and vomiting, or if you have noticed blood in urine or in stool.
Claims: [ Possible cause for right lower abdominal pain in a young female: Appendicitis, Possible cause for right lower abdominal pain in a young female: Ovarian cyst or torsion, Possible cause for right lower abdominal pain in a young female: Ectopic pregnancy, Possible cause for right lower abdominal pain in a young female: Pelvic inflammatory disease, Possible cause for right lower abdominal pain in a young female: Kidney stone, Possible cause for right lower abdominal pain in a young female: Urinary tract infection, Possible cause for right lower abdominal pain in a young female: Diverticulitis, Possible cause for right lower abdominal pain in a young female: Inflammatory bowel disease, Possible cause for right lower abdominal pain in a young female: Endometriosis, Please seek medical attention if the pain is sudden and severe, Please seek medical attention if the pain is accompanied by fever, Please seek medical attention if the pain is accompanied by nausea and vomiting, Please seek medical attention if the pain is accompanied by blood in urine, Please seek medical attention if the pain is accompanied by blood in stool, Possible cause for right lower abdominal pain in a young female: Emotional stress ]

# TASK EXAMPLE
Question: So what does the non reactive mean for the hep a igm Answer: Hep A IgM refers to a specific type of antibody called Immunoglobulin M (IgM) against the virus hepatitis A. When infected with hepatitis A, these antibodies are detectable at symptom onset and remain detectable for approximately three to six months. These antibodies might also be detectable in the first month after hepatitis A vaccination. A negative or non-reactive result means no IgM antibodies against hepatitis A found in your serum, meaning the absence of an acute or recent hepatitis A virus infection.
Claims: [ A negative or non-reactive result means that there were no IgM antibodies against hepatitis A found in your serum, The absence of IgM antibodies against hepatitis A in your serum indicates the absence of an acute or recent hepatitis A virus infection, Hep A IgM refers to a specific type of antibodies called Immunoglobulin M (IgM) against the virus hepatitis A, These antibodies might also be detectable in the first month after hepatitis A vaccination, These antibodies remain detectable for approximately three to six months after infection, When infected with hepatitis A, these antibodies are detectable at the time of symptom onset ]

# TASK EXAMPLE
Question: What medications are contraindicated for a pregnant woman with ulcerative colitis? Answer: methotrexate (Otrexup, Rasuvo, RediTrex) and thalidomide (Contergan, Thalomid) are both considered contraindicated for treatment of UC in pregnancy. possible treatment for UC during pregnancy include low-risk drugs such as aminosalicylates (sulfasalazine and mesalamine), immunomodulators (azathioprine, cyclosporine A ,6-mercaptopurine) and corticosteroids. Biological agents such as Infliximabl, Adalimumab, Vedolizumab and Ustekinumab is generally avoided during pregnancy as their safety in pregnancy is not well established yet.
Claims: [ Methotrexate (Otrexup, Rasuvo, RediTrex) is contraindicated for treatment of ulcerative colitis in pregnancy, Thalidomide (Contergan, Thalomid) is contraindicated for treatment of ulcerative colitis in pregnancy, Aminosalicylates (sulfasalazine and mesalamine) are considered low-risk drugs for treatment of ulcerative colitis during pregnancy, Immunomodulators (azathioprine, cyclosporine A, 6-mercaptopurine) are considered low-risk drugs for treatment of ulcerative colitis during pregnancy, Corticosteroids are considered low-risk drugs for treatment of ulcerative colitis during pregnancy, Treatment for ulcerative colitis during pregnancy with biological agents such as Adalimumab is generally avoided during pregnancy as their safety in pregnancy is not well established yet, Treatment for ulcerative colitis during pregnancy with biological agents such as Vedolizumab is generally avoided during pregnancy as their safety in pregnancy is not well established yet, Treatment for ulcerative colitis during pregnancy with biological agents such as Infliximab is generally avoided during pregnancy as their safety in pregnancy is not well established yet, Treatment for ulcerative colitis during pregnancy with biological agents such as Ustekinumab is generally avoided during pregnancy as their safety in pregnancy is not well established yet, ]

# YOUR TASK
Question: {question}
Answer: {answer}
Claims:
```

D.2 Automatic Evaluation

```
# OVERALL INSTRUCTIONS
- You have a deep understanding of logical relationships, such as entailment and contradiction, to evaluate given triplets of (question, premise, hypothesis).

# TASK INSTRUCTIONS
Your goal is to determine whether the Premise effectively contradicts the corresponding Hypothesis. Carefully analyze each triplet, focusing on details. - If the premise and the hypothesis are unrelated or lack sufficient evidence to ascertain their truthfulness, label your answer as False. - be vigilant in identifying cases where the premise doesn't rule out the possibility of an entity (e.g., vaccine, symptom) appearing in the hypothesis. In such cases, classify the answer as False. - Approach each question methodically, considering the step-by-step process outlined below.

# INPUT DATA
Question: What does trich test for? Let's think step by step.
Premise: The term "trich test" can refer to two different medical tests, depending on the context. Here are the two possibilities: Trichomoniasis Test: Trichomoniasis is a sexually transmitted infection (STI) caused by the parasite Trichomonas vaginalis. The trichomoniasis test, also known as a trich test or trichomonas test, is used to detect the presence of this parasite in the body. The test is typically performed on a sample of vaginal discharge in women or urine in men. Trichogram: A trichogram is a diagnostic test used to evaluate hair loss and assess the health and condition of hair follicles. It involves plucking a small number of hairs from the scalp and examining them under a microscope. It's important to note that without additional context, it's difficult to determine which specific test you are referring to. Hypothesis: Trichamoniasis- a parasitic infection that can cause your symptoms.
Answer: According to the premise "trich test" refer to two different medical tests. A Trichamoniasis test is one of them, which is used to detect this parasite's presence. As stated in the hypothesis, the trich test is used to diagnose parasitic infections. Ths premise entails the hypothesis. The answer is False.
###
Question: Can diabetics eat sweets? Let's think step by step.
Premise: Individuals with diabetes are recommended to limit their consumption of sweets to one or two times per week. It is also suggested being selective with desserts and to focus on foods with a low glycemic index, such as high fiber foods like whole grains and legumes, as well as certain lower sugar fruits like berries, melons, and apples.
Hypothesis: It is recommended that diabetics avoid sweets.
Answer: The premise suggests that diabetics can eat sweets but limit their consumption. According to the hypothesis diabetics should avoid sweets. Diabetics are allowed to consume sweets according to the premise, but they are prohibited according to the hypothesis. There is a contradiction between the premise and the hypothesis. The answer is True.
###
Question: 25 yo female with right lower abdominal pain, what might be causing it? Let's think step by step.
Premise: Right lower abdominal pain in a 25-year-old female could be caused by a variety of medical conditions.
Some potential causes include: Ovarian cyst: a fluid-filled sac on the ovary - Ectopic pregnancy: a pregnancy that occurs outside the uterus.
Hypothesis: possible cause for right lower abdominal pain in a young female can be Appendicitis.
Answer: The premise lists several potential causes of right lower abdominal pain in a 25-year-old female, not including appendicitis. The hypothesis states that Appendicitis could be a cause of right lower abdominal pain in a young female. There is no direct contradiction between the premise and the hypothesis, as the premise does not exclude the possibility of appendicitis as the cause of the pain. The answer is False.
###
Question: Can a headache last longer than a few days? Let's think step by step.
Premise: Yes, it is possible. If you are experiencing a headache that lasts longer than a few days, it is important to see a doctor to get the appropriate treatment.
This will help to relieve the pain and prevent any further complications.
Hypothesis: It is not a cause for concern if a headache lasts longer than a few days.
Answer: This premise acknowledges that a headache can last for several days, but emphasizes that seeing a doctor to prevent further complications is important. According to this hypothesis, headaches lasting longer than a few days are not cause of concern. There is a contradiction between the premise and hypothesis due to the premise emphasizing the importance of seeking medical consultation, while the hypothesis posits that there is no cause for concern. The answer is True.
###

Question: {question} Let's think step by step.
Premise: {llm_answer}
Hypothesis: {physician_answer}
Answer:
```

D.3 Question-Answering

vanilla prompt

```
Question: {question}
Answer:
```

vanilla prompt for RAG

```
# TASK:
Answering question.

# INSTRUCTIONS:
- Use ONLY our verified sources to answer the question.
- If the answer is not in our sources, answer ONLY "I don't know".

# VERIFIED SOURCES
Our verified sources: {context}

# INPUT
Question: {question}

# OUTPUT
Answer:
```

engineered prompt

```
# TASK:
You are a medical expert and your task is to write a helpful human-like text and answer to the point.

# INSTRUCTIONS:
- Identify and address the patient's focus of interest, but include relevant information even if the patient didn't directly ask it.
- If relevant, provide information about specific cases, including children, pregnant women, allergies, or chronic conditions.
- If you don't know the answer, answer ONLY 'I don't know'.
- If the answer contains enumeration of factors/symptoms/conditions use bullet points to organize the information.

# EXAMPLES:
Question: Is it safe to take Macrobid?
Answer: Macrobid is generally considered safe to take when prescribed by a doctor and used according to their instructions.
Rationale: The answer is good, but it is missing important information regarding special cases, such as pregnant women, kids and comorbidities.
A better answer would include that Macrobid is not recommended for use in the last 2 to 4 weeks of pregnancy
###
Question: Ok Do I need any antibiotics or something for yeast infection?
Answer: Over-the-counter antifungal creams or suppositories applied internally to the vagina can effectively treat vaginal yeast infections, while a single oral dose of prescription fluconazole (Diflucan) may also be an option.
Rationale: the answer is good however it does not include relevant explanation about the condition such as Yeast infections are commonly caused by fungi, particularly Candida albicans, which is a natural inhabitant of the body but can cause infections in different areas such as the skin, mouth, and vagina. In addition, it assumes the patient is asking about vaginal yeast infection and excludes information about other possible infections caused by yeast.
###
Question: the vomiting, tight abdomen pain and burning pain may just be result of pcos?
Vomiting and tight abdominal pain are not common clinical presentations of PCOS. Vomiting and abdominal pain are nonspecific symptoms which could be an indication for a variety of illnesses. In case of this presentation, you should be evaluated by a physician, and treated accordingly.
Rationale: the answer is good, but it does not emphasize enough that the patient's symptoms are most likely caused by something else and what is the differential diagnosis in this case.

# INPUT
Question: {question}
Answer:
```

engineered prompt for RAG

```
# TASK:
You are a medical expert whose task is to provide a helpful, concise, human-like response to a patient's question based on verified medical sources.

# INSTRUCTIONS:
- Use our verified sources to answer the question at the end.
- If the answer is not in our sources, answer ONLY 'I don't know'.
- It is important to assume that there may be a medical concern underlying the patient's questions.
- If relevant, provide information about specific cases, including children, pregnant women, allergies, or chronic conditions.
- If the answer contains enumeration of factors/symptoms/conditions use bullet points to organize the information.
- Avoid repetition of details.

# EXAMPLES
Question: Is it safe to take Macrobid?
Answer: Macrobid is generally considered safe to take when prescribed by a doctor and used according to their instructions.
Rationale: The answer is good, but it is missing important information regarding special cases, such as pregnant women, kids and comorbidities.
A better answer would include that Macrobid is not recommended for use in the last 2 to 4 weeks of pregnancy
###
Question: I had my gallbladder removed in 2015, so that rules out gallstones, right?
Answer: Yes, as gallstones are stones that form inside the gallbladder, if the gallbladder is removed, gallstones cannot be the cause of future symptoms.
Rationale: The answer is correct however it does not provide the patient with an important clinical tie as even without a gallbladder, stones can develop anywhere in the biliary system and cause similar symptoms to gallstone disease.
###
Question: the vomiting, tight abdomen pain and burning pain may just be result of pcos?
Answer: Polycystic ovary syndrome (PCOS) is an endocrine (hormonal) insufficiency which is characterized by polycystic ovaries and has a variety of effects including anovulation and irregular menstrual cycles which in turn could causes fertility issues, hyperandrogenism (high levels of androgens causing coarse body hair growth in a male pattern and acne) and insulin resistance causing type 2 diabetes mellitus, obesity, and hypertension.
Vomiting and tight abdominal pain are not common clinical presentations of PCOS. Vomiting and abdominal pain are nonspecific symptoms which could be an indication for a variety of illnesses. In case of this presentation, you should be evaluated by a physician, and treated accordingly.
Rationale: The answer is good, but it does not emphasize enough that the patient's symptoms are most likely caused by something else and what is the differential diagnosis in this case.

# CONTEXT
Our verified sources: {context}

# INPUT
Question: {question}
Answer:
```