MedRedQA for Medical Consumer Question Answering: Dataset, Tasks, and Neural Baselines

Vincent Nguyen[1,2] **Sarvnaz Karimi**[1] **Maciej Rybinski**[1] **Zhenchang Xing**[1,2]
[1]CSIRO Data61, Sydney, Australia
[2]The Australian National University, Canberra, Australia
`{firstname.lastname}@csiro.au`

Abstract

Medical question answering for consumers aims to assist consumers in finding trustworthy and relevant information for their concerns. Although some datasets exist for consumer question answering, they use synthetic questions or present difficult-to-understand answers. We introduce `MedRedQA`, a large non-factoid English consumer Question Answering (QA) dataset containing 51,000 pairs of *consumer questions* and their corresponding *expert answers*. `MedRedQA` facilitates research that aims to provide consumer-friendly responses to real-world consumer questions. We propose and benchmark three tasks for consumer medical question answering for our dataset, including (1) candidate answer ranking, (2) open-ended answer generation, and (3) answer generation with scientific evidence. Our benchmarking experiments reveal that, for the ranking task, it is feasible to retrieve expert answers within five responses in an oracle retrieval. Though, in an answer generation task, it remains challenging to align the generation toward expert answers. However, our experiments show that including scientific evidence in the prompt may reduce hallucinations in an answer generation setup.[1]

1 Introduction

Those without expertise in the medical domain, hereafter referred to as *consumers*, often seek answers to their medical questions online (Van Riel et al., 2017). A startling number of these consumers do not follow up their online search with a medical professional and believe in their self-diagnosis (Kuehn, 2013), which may be corroborated by less reputable sources (Nelson et al., 2020). This is concerning, given the amount of health misinformation circulating the general web (Hussain et al., 2018), and that non-official high-traffic health websites likely have high amounts of misinformation (NewsGuard, 2021). As a result, a misinformed consumer might be less likely to adhere to health guidelines (Jon et al., 2020), such as vaccination advice (Garett and Young, 2021). On the other hand, there is a positive correlation between patients researching their conditions and improved patient-physician rapport. This occurs as dialogue improves between the physician and their patient when the patient has an improved understanding of their potential diagnosis and prognosis (Cocco et al., 2018). Importantly, when the consumer is wrong about their potential diagnosis, adherence to the physician's recommendations does not decrease (Van Riel et al., 2017).

Given the aforementioned benefits of patients' being proactive in searching their symptoms, some existing systems allow consumers to search for potential diagnoses. Examples of these systems include health cards, which provide disease information alongside search results (Jimmy et al., 2019), a curated multi-document answer synthesis that takes answers from an expert annotated database (Demner-Fushman et al., 2020), and a search system that allows users to get expert-level recall using consumer queries over peer-reviewed COVID-19 literature (Nguyen et al., 2022). These systems are important and require continual research as consumers require high-quality medical advice for a broad range of topics.

However, consumer QA systems often use datasets of limited scope and scale. Many of these datasets often make assumptions on consumer input (e.g., a singular short closed question), retrieve answer excerpts from difficult-to-understand medical articles (consumer unfriendly), and are too small to train contemporary models. We aim to address these problems by introducing `MedRedQA`, the largest English consumer question answering (QA) dataset. `MedRedQA` consists of 51,000 consumer questions and verified expert answer pairs collected from a health subreddit called AskDocs

---

[1] To preserve user anonymity, the dataset will be made available through a script to download the data.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 629–648
November 1–4, 2023. ©2023 Association for Computational Linguistics

which is dedicated to improving access to health information for consumers. We present benchmarking experiments for retrieval-based QA and generative QA for these pairs. We also introduce a novel task for generative QA with the availability of relevant scientific sources. To this end, we introduce a collection of *question–expert answer–source article* triplets containing PubMED[2] scientific articles that are referenced by a medical expert. Our paper provides the following research contributions:

1. Construction of a dataset for research on real-world consumer questions providing expert answers similar in language to the consumer.
2. Proposal of three tasks including evaluation and analysis of baselines for Consumer QA: (a) expert answers as a ranking problem; (b) expert answer generation from a question; and (c) novel expert answer generation grounded by question-source article pairs.
3. Analysis of the difference in language used by the consumer and the medical expert and their response behaviors on the forum.

2 Related Work

Literature in biomedical QA primarily focuses on providing answers to professionals.

Professional QA The annually run BioASQ (Krithara et al., 2016; Nentidis et al., 2017, 2018) provides tasks and datasets for QA posed by medical experts as a summarization and information retrieval task. Similarly, tasks run by TREC including TREC-COVID (Voorhees et al., 2020) and TREC-CDS (Voorhees and Hersh, 2012; Roberts et al., 2017, 2018, 2019, 2021) are focused on solving search tasks for medical experts which include clinical decision support (Roberts et al., 2015, 2016), clinical trial retrieval and precision medicine (Roberts et al., 2017, 2018).

Consumer QA While there has been considerable effort to help experts, less emphasis has been put forward on consumers. MEDIQA (Ben Abacha et al., 2019), which is an extended collection of MedNLI (Shivade, 2019), MedQUAD (Ben Abacha and Demner-Fushman, 2019) and RQE (Abacha and Demner-Fushman, 2016), is proposed to facilitate consumer biomedical QA. However, these datasets are limited in magnitude, have different latent distributions between the training – testing distributions (Nguyen et al., 2019), and have assumptions that reduce the difficulty of the consumer biomedical QA task (such as how users ask questions) or do not provide consumer-understandable answers.

Furthermore, MEDIQA derives some of its data from MIMIC-III (Johnson et al., 2016), a dataset of critical-care patient reports that does not reflect the medical needs of an everyday consumer. We summarize datasets for the consumer QA domain in Table 1.

MashQA (Zhu et al., 2020) presents long-form medical documents or websites for span-based question answering from typically short-form factoid questions. However, MashQA is suitable for question answering with context rather than open-ended question answering. Furthermore, the questions are typically short and address a single information need: ‘How many available treatments for X disease?’, without incorporating contextual information of the consumer, such as their medical history, age or gender. Moreover, there is an implicit assumption that question context contains the answer (SQuAD-style question answering), which may not be comprehensible to consumers.

RedHot (Wadhwa et al., 2023) presents a task of retrieving evidence for Reddit questions from peer-reviewed sources to verify a medical claim made in the post using patient/population, intervention, comparison and outcomes (PICO) elements for retrieval. The data is sourced from subreddits (or forums) where the diagnosis is known: /r/ADHD, /r/Psychosis, and is used to provide interventions or treatments. Although well annotated, it does not present answers in a consumer-friendly manner that the consumer will likely understand and appreciate, as the answers are given as PubMed abstracts (Graham and Brookey, 2008). Furthermore, the dataset does not leverage the answers from the users to the questions. This may be because there is no verified distinction between an expert response and a layperson response. These distinction labels are present in /r/askdocs, used in this work.

cMedQA (Zhang et al., 2018) and ChiMED (Tian et al., 2019), are large-scale corpora that use crowdsourced data from existing online forums with doctor-patient response pairs. They are well suited for consumer question answering, as they are sourced from real-world data. These datasets, however, are not in English.

---

[2] pubmed.ncbi.nlm.nih.gov

Dataset	**Size**	**Language**	**Description**
Mash-QA (Zhu et al., 2020)	34.8k	English	Span-based QA for consumer health, where contexts are given as long documents and questions are of the short-form.
ChiMed (Tian et al., 2019)	24.9k	Chinese	A corpus where doctor-patient answer pairs are taken from an existing online forum (XunYiWenYao).
cMedQA v2 (Zhang et al., 2018)	108k	Chinese	A corpus where doctor-patient answer pairs are taken from an existing online forum (wywy.com)
MedQuAD (Ben Abacha et al., 2019)	48k	English	A collection of curated English question-answer pairs from *cancer.gov* with consumer language in the answers.
MeQSum (Ben Abacha and Demner-Fushman, 2019)	1k	English	Summarization of long-form consumer context questions to short-form questions.
BioReddit (Basaldella and Collier, 2019)	N/A	English	Collection of Biology StackExchange and Reddit Health Subreddits corpus to produce static word embeddings.
RedHot (Wadhwa et al., 2023)	22k	English	Collection of annotated Reddit posts from health subreddits of a particular disease e.g. /r/ADHD /r/Psychosis
`MedRedQA` (Ours)	51k	English	An English Corpus from `/r/askdocs` where long-form consumer questions are mapped to short-form doctor comments

Table 1: Comparison of different biomedical consumer QA datasets.

There is a gap in the literature when it comes to real-world datasets that: (1) accurately reflect how consumers ask questions; (2) take into account the consumers' medical context; (3) are of a practical size to meet contemporary neural models data requirements; (4) provide trustworthy answers in a consumer-friendly way. Our dataset addresses this gap in the literature by sourcing real-world consumer questions that include medical context and providing the answers in a language that the consumer understands.

3 MedRedQA Dataset Creation

Raw Data Collection We collected data from posts and comments to subreddit `/r/askdocs`, published between July 10, 2013, and April 2, 2022, totalling $600,000$ submissions (original posts) and $1,700,000$ comments (replies). Askdocs (`/r/askdocs`) is a forum for consumers, referred to as laypeople in the forum, to ask medical questions and receive responses from physicians. There were 71 unique occupations for medical experts and three labels for the consumer. Each submission had an average of 3.03 (± 10.75) comments, a median of 3.00, with an overall maximum comment count of 5,888, indicating a long-tail distribution. The average submission contained a short patient profile: demographics, medication, symptoms and a question, with the format agreed on as part of the forum's rules. Only a third of the submissions had a comment, and 20% of the responses were from verified medical professionals. From n-gram frequency analysis, we found that consumers tend to ask about acute illnesses (symptoms within the last two weeks) with the leading advice given on the forum to *go see a doctor* from both consumers and medical experts.

Upon manual inspection, we find that submissions with the highest comment count often contained emotionally charged content (e.g., suicide) or were general information threads (e.g., a compilation of COVID questions/answers). Longer submissions, likely adhering to community guidelines, had more comments. The average number of tokens was 231.9 (± 217.2) per submission, 21.59 (± 10.46) per comment, and 11.21 (± 8.435) for titles. Physicians' comments (50.79 ± 68.61) were more concise than those from other consumers (53.89 ± 70.39). The Jaccard Index (Jaccard, 1912) between the vocabularies of consumer's comments and medical experts' comments was 0.89 after stopword removal, with the main difference in tokens being unique medical terminology.

Data Cleaning We use a set of heuristics to clean and organize the data. Firstly we remove all submissions that have been removed by moderators, submissions from banned users, bot content, and submissions with fewer than five words. To protect privacy, posts from any users with deleted accounts or those who have removed their posts, as the title and comments remain, have also been removed. We further remove posts with images and remove any URLs.

We then filter comments using the same criteria but with two additional heuristics: (1) the comment must be from a verified physician, and (2) the comment must have a score higher than 1.0 (the default score), which indicates that *at least* one other user agrees with the comment. We then use the highest-scoring comment as the answer for the submission. After filtering and cleaning heuristics, a total of $51,000$ question-answer pairs remained. We then

Title: Currently have traveler's diarrhea for 7 days in ME, no access to a doctor
Body Age: 32 Gender: F Height: 170cm Weight: 66kg Race: European Location: [omitted] Current medications: Dexilant 30mg for stomach acid Other medical conditions: None, non-smoker, otherwise healthy overall aside the PPI prescribed due to stomach acid (h. pylori tested negative). Present complaint: I have travelerś diarrhea that...
Answer Sounds like you are in the "mild" category. You could try Loperimide or Pepto Bismol. I recommend medicine if youŕe having pain, bloody stools, continued fever. Here are some guidelines you could read and decide: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6610510/
PMCIDS(s): PMC6610510, 31320958

Table 2: An example containing a PubMed Reference.

randomly divide the dataset into an 80-10-10 training/validation/testing split.

Scientific evidence Some comments contain references to medical articles as evidence or are presented as further reading for the consumer. There were $3,300$ PubMed articles mentioned in the comments. However, we retained only 1000 references (30.3%) which were from verified medical professionals in the final dataset. Table 2 shows an example from the dataset. We check the provided abstracts by annotating a random 10% sampling of the abstracts for quality verification (Appendix C).

4 Tasks

We explore three tasks, where the goal is to produce a consumer-friendly expert answer, given a consumer question: (1) Question Answering as a **ranking task** over candidate answers (Nguyen et al., 2016); (2) an **open-ended generation task** where questions are used to generate answers; and, (3) a **generation with scientific evidence task** where questions and the corresponding evidence (PubMed articles' abstract) are used to generate an expert response. These tasks are formulated as below:

Ranking Consumer question answering can be tackled as an ad-hoc ranking task:

$$P(relevance = 1, \mathbb{D}, q) = score(q, \mathbb{D}), \quad (1)$$

where the relevance of the ad-hoc consumer question (query), q to a candidate expert answer, d, from the corpus of expert answers, $d \in \mathbb{D}$, is computed with a function that estimates relevance $score$ for ranking. Retrieval of expert responses is desirable as they are written by experts, whose quality has been assessed via crowdsourcing (i.e., the forum score).

For dense retrieval (Zamani et al., 2018), typically a bi-encoder is used for initial stage ranking:

$$\boldsymbol{h}_d = \text{BERT}(d)[CLS] \quad (2)$$
$$\boldsymbol{h}_q = \text{BERT}(q)[CLS] \quad (3)$$
$$score(q, d \in \mathbb{D}) = \cos(\boldsymbol{h}_q, \boldsymbol{h}_d), \quad (4)$$

which estimates relevance score as the cosine similarity between the hidden representations of the consumer question, $\boldsymbol{h}_q$ and expert answer, $\boldsymbol{h}_d$.

The initial retrieval can be re-ranked in a secondary stage ranker with a filtered set of expert answers D_k, which is typically set to the top 1000 candidate answers, $D_{k=1000}$ from the initial retrieval:

$$score(q, d \in D_k) = Softmax(\boldsymbol{h}_{q;d}). \quad (5)$$

Sparsity However, for our dataset, we also need to account for the sparsity of the labels:

$$P(label = 1 | \mathbb{D}, q) \in \{0, 1\}, \quad (6)$$

where there is only a single expert answer, $d \in \mathbb{D}$, which has a label of 1 a query, q. We therefore validate if the baselines can reasonably model the sparse data distribution.

Open-ended Generation The importance of the ranking task is that it allows answers to be restricted to a set of verified expert answers. However, there is no guarantee that future answers are contained in the set of verified answers. We therefore can also model the task of providing consumer-friendly expert answers with auto-regressive generative Language Modeling (LM):

$$P(w_1, ..., w_t) \approx P(w_t | w_1, ..., w_{t-1}), \quad (7)$$
$$w_{t+1} \sim P(\cdot | w_{1:t}), \quad (8)$$

where a word, w_{t+1} is sampled from the probability distribution over the vocabulary $P(\cdot, w_{1:t})$ given the sequence generated, $w_{1:t}$ up until time step, t.

When finetuned with instructions, we can auto-regressively generate an output sequence, $O_t = \{w_p, ..., w_t\}$, given a fixed prompt $P = \{w_1, ..., w_{p-1}\}$ that is more aligned with human preference.

Generation With Scientific Evidence However, the output of the open-ended generation may be too broad, prone to hallucination or irrelevant. We, therefore, introduce the novel task of using Large Language Models (LLMs) fine-tuned with the context provided by PubMed documents to generate expert consumer-oriented answers. The context is provided to produce a better auto-regressive response using the evidence from the PubMed document. Moreover, providing source documents in prompts has reduced hallucination in LLMs (Madaan et al., 2022; Guu et al., 2020; Lazaridou et al., 2022), a vital property for sensitive tasks such as consumer medical QA. This generation task uses a subset of submissions that contain a PubMed document (provenance) which we use as scientific evidence in a fixed prompt P.

5 Methods

We investigate the following methods and their implementations to validate the feasibility of the tasks we propose for the `MedRedQA` dataset taking into account the sparsity of the labels.

5.1 Ranking

We use BM25 (Robertson et al., 1994) as a ranking baseline. BM25 is a robust statistical model that uses lexical overlaps or hard-matching as a relevance signal. Although not as competitive as a well-tuned neural model, it is a stable baseline and works well out-of-the-box (Armstrong et al., 2009). Furthermore, it captures different relevance signals when compared to dense retrievers (Wang et al., 2021).

The ranking pipeline includes a zero-shot Bi-Encoder (BiEnc) and Cross-Encoder (CEnc). We use the PubMEDBERT model by Gu et al. (2021) as our BiEnc, pretrained on the PubMed corpus (NCBI, 2023), without additional finetuning. Contradictory to (Reimers and Gurevych, 2019), we found that finetuning the sentence encoder hurts ranking in our preliminary experiments (Appendix B).

We select the top 1000 answers from BiEnc for second-stage ranking with a cross-encoder, MonoBERT (Nogueira et al., 2019). This CEnc was trained with a learning rate of 2×10^{-5} for 4500 steps (stopping when the model over-fit) using the AdamW optimizer (Loshchilov and Hutter, 2017) with a warm-up ratio of 0.1. We select the best checkpoint based on the validation set performance for downstream task prediction. We used random sampling to select negatives.

5.2 Open-Ended Generation

For the expert answer generation, we use T5 (Raffel et al., 2020), an encoder-decoder conditional LM, which has shown to be as effective as a GPT-3 model (Liu et al., 2022b; Brown et al., 2020).

We test various T5-large models for answer generation: (1) T5 version 1.0, (2) T5 version 1.1, which is pretrained without task prompts, as testing showed task confusion was present in the v1.0 model; (3) T5 Flan, a T5 model (Chung et al., 2022) pretrained on various instruction-based tasks (Wei et al., 2022) to improve alignment with instruction prompting, and (4) Flan-T5 pretrained on the in-domain PubMedQA (Jin et al., 2019), which may improve alignment with expert answer generation.

For training, we truncated questions to 510 tokens and trained the models using maximum likelihood estimation (teacher forcing) as

$$L_{CE}(\hat{y}_t, y_t) = -\log \hat{y}_t[w_{t+1}], \tag{9}$$

where the cross-entropy loss is the negative log probability assigned to the next word, w_{t+1}. We reuse the same hyperparameters and optimizers for finetuning, namely Adafactor (Shazeer and Stern, 2018) and a constant learning rate, as the original authors did (Raffel et al., 2020). We adopt a similar method to Flan for prompt training, where we construct four variations of prompt templates (Appendix D.1). We train for three epochs, regenerating the prompts for each epoch.

For generation decoding (Equation 8), we select generated words based on maximum likelihood[3] for reproducibility

$$w_{t+1} = \arg\max P(\cdot|w_{1:t}) \tag{10}$$

and set the number of generated tokens to 150.

5.3 Generation with Scientific Evidence

We employ the use of open-source LLMs, namely Llama 7B (Touvron et al., 2023) and Alpaca 7B (Taori et al., 2023) (a Llama checkpoint finetuned to align with human preference), as they can be finetuned. This removes the need to provide few-shot examples for in-context learning (Liu et al., 2022a). LLMs are also viable for consumer-oriented answer generation and have garnered

---

[3] Preliminary tests show that sampling or beam search do not have much of an effect on the MoverScore

Ranking	Facet	Acc@1000 ↑	HR ↓
BiEnc	Title	0.1389	785.8
	Body	0.1734	600.1
	Title+Body	0.1560	628.6
BM25	Title	0.0890	1332.7
	Body	0.0675	1548.2
	Title+Body	0.0588	1597.1

Table 3: Results for first-stage retrieval experiments.

Ranking	Acc@1 ↑	Acc@10 ↑	HR ↓
BM25 (Title)	0.0000	0.0002	1333
BiEnc (Body)	0.0002	0.0014	600.1
BiEnc + CE	0.0633	0.1018	*27.13
BiEnc + CE (Oracle)	**0.3654**	**0.5791**	**4.174**

Table 4: Second stage re-ranking of candidate answers. Oracle refers to evaluating the subset of query–answer retrieved that is known to have a relevant answer. *The Harmonic Mean Rank of BiEnc + CE is undefined as the BiEnc had topics that did not retrieve a relevant answer; we estimate the cross-encoder will perform at least as well as the bi-encoder.

	MoverScore ↑	Rouge-1
Pretrained		
Flan	0.501 ± 0.053	0.075 ± 0.081
T5-Large	0.479 ± 0.030	0.039 ± 0.060
Finetuned		
Flan	**0.507** ± **0.064**	**0.096** ± **0.01**
T5-Large	0.499 ± 0.032	0.094 ± 0.10
Flan+PubmedQA	0.504 ± 0.053	0.096 ± 0.01
T5-Large-1.1	0.502 ± 0.035	**0.096** ± **0.10**

Table 5: Generation of experts answers task using the T5-based models.

much attention for their fluent well-formed responses (Ouyang et al., 2022). Moreover, as the PubMed documents are fewer than 1000, we can use LLMs to compensate for this data scarcity, given that LLMs are a viable option with only a few training data points (Karimi Mahabadi et al., 2022). Another benefit is increased model input size over T5 (512 to 1024), which would not otherwise fit the question and PubMed abstract.

We use a checkpoint (Park, 2023) that is pretrained on the Alpaca dataset (Taori et al., 2023) and further finetune on our scientific evidence subset using Low-Rank Adaptation (LoRA) (Hu et al., 2021) learning rate of 3e-4 and LoRA settings (r=8, alpha=16, dropout=0.05 with an input size of 1024 tokens).

Our experiments are on zero-shot and finetuned models. For generation, we use the same decoding hyperparameters as T5. However, we increased the input size to 1024, and the generation limit to 250 as preliminary testing showed the Llama-based models were verbose. We experimented with longer generation limits (up to 500), but this led to repetitive degenerate text (Holtzman et al., 2020).

6 Results and Discussion

Metrics We report accuracy@k (where k is the retrieval size of the result set) as the primary metric to compensate for the sparse nature of the dataset. We also include the Harmonic mean of the Ranks (HR), the reciprocal of the mean reciprocal rank, which indicates the expected rank of the relevant answer. We report Rouge-1 (Lin, 2004) to capture vocabulary overlap as a supplementary metric for the generation task. To also capture semantics, we use MoverScore (Zhao et al., 2019) as our primary evaluation metric for the generation tasks.

Ranking Task Retrieval is expected to be difficult for `MedRedQA` due to the sparsity of labels and the nature of the questions. The questions are non-factoid, noisy (typographic errors and grammatical errors), and share little similarity with the answers (based on the results of the BM25 baseline, a keyword-matching probabilistic model). First-stage retrieval underperformed for both neural and probabilistic models (Table 3). Interestingly, we found that the bi-encoder model performed best with the body of the post as the query and that the title facet introduced noise. This was the opposite of the BM25 baseline, which performed best with the title as the query. This is a notable result, as it reinforces our assumption that the two ranking models capture different signals and that BM25 performed better with fewer terms in the query for a recall-based task.

Using a cross-encoder, a more expensive operation, showed substantial improvement over the initial retrieval. Given a perfect first-stage retrieval, where the retrieved set contains the relevant answer for the query (oracle in Table 4), reaching an expected ranking (HR) of 4.174 in the second-stage retrieval for a relevant answer is possible. Assuming the cross-encoder performs no better than the bi-encoder, we achieve an HR of 27.13. Overall, the results indicate that the task is feasible and that our setup is reliable for benchmarking much-needed first-stage improvements.

Open-ended Generation Task We found that the best-performing model, on average across the metrics, was finetuned Flan-T5 (Table 5). This highlights that the Flan dataset has better alignment with our dataset, given that we are using instruction prompting. However, we found that Flan-T5 pretrained with PubMedQA was less effective. This indicates that there is a difference in language between our dataset and scientific literature. Although Rouge-1 scores are low, this is generally expected in the biomedical domain and open-ended text generation, given the difficulty of mapping expert biomedical answers to questions with high lexical gaps (Weißenborn et al., 2013). We expect scores to be lower than those in BioASQ (Nentidis et al., 2022), as the task involves mapping consumer questions to expert answers, rather than expert questions to expert answers and can be modified to be extractive question answering task for higher Rouge-1 scores.

Overall, the T5 models produced much shorter responses than the LLM counterparts. Interestingly, the Flan-T5 generated 10.18±12.77 words after finetuning and 16.53±27.11 words before. This is reflected in the lower Rouge-1 scores, as the chance of word overlap between the generated text and the reference is reduced with fewer words; meaning medical experts tend to be concise and direct, which the T5 models learned during training.

Qualitatively, the T5 responses often instructed the consumer to see a specialist and prefaced with "I'm not a doctor." Although appearing in the training set, as medical experts are not all doctors, it is interesting that the model reinforced this type of response. Before finetuning, the model would respond with "I'm not a doctor, I'm a hairstylist". However, after finetuning, the model would still preface with "I'm not a doctor," but would then provide relevant health advice to the user. This suggests that this phrase frequently appears alongside consumer biomedical questions in the C4 dataset (Raffel et al., 2020) used for T5 pretraining.

Furthermore, T5-generated responses often advised the consumer to see a specific health professional based on their medications (e.g., "see a psychiatrist", "see an audiologist"). This is a useful generation, as maximizing recall—sensitivity is desirable when providing medical advice. That being said, we did find cases where Flan-T5 produced responses that contradicted the expert response; for instance, a user would ask if a particular aspect of their health is within normal range. The model would suggest it is not normal, but the expert response suggests it is within the normal range.

	MoverScore ↑	**Rouge-1**
Zero-shot		
Llama	0.509 ± 0.017	0.141 ± 0.075
Alpaca	0.515 ± 0.018	0.164 ± 0.087
Finetuned		
Llama	0.505 ± 0.019	0.133 ± 0.073
Alpaca	**0.516 ± 0.016**	**0.164** ± **0.088**

Table 6: Generation of experts answers task with the PubMed document in the prompt.

	MoverScore ↑	**Rouge-1**
Zero-shot		
Llama	0.510 ± 0.019	0.143 ± 0.087
Alpaca	0.517 ± 0.016	0.168 ± 0.095
Finetuned		
Llama	0.506 ± 0.014	0.117 ± 0.075
Alpaca	**0.518 ± 0.015**	**0.172** ± **0.093**

Table 7: Generation of experts answers task. There is no PubMed document in the prompt, however, the finetuned models are still finetuned with documents in the prompt.

One major differentiating factor between the T5 and expert responses was that the T5 responses were concise, factual, and instructive. While many expert responses were similar, expert responses were likely to be longer and provided reassurance rather than advice: "This is normal, nothing to be worried about" (further examples of generation and further analysis in Appendix D.3.1.

Generation with Evidence Task The results (Table 6), show that the Alpaca model performs best across the metrics, aligning better with the expert responses. We found that the Llama models fell behind in performance, which is expected as these models are not aligned to instructions, but are traditional LMs. However, it is surprising that tuning the Llama model yielded worse performance after finetuning, which may result from the small dataset size. Quantitatively, adding the document to the prompt produced slightly worse scores (Table 7) which may result from irrelevant abstracts being included in the prompt (Appendix C). This would have a significant effect, given the test sample size.

For the LLMs, the responses were more empathetic than their T5 counterparts. This is unsurprising, given that they were trained to align with human preference. This was shown as they often repeated the same beginning phrases—"Thank you for your question, you are concerned about [disease]"—which did appear in the pretrained model, but became more refined after finetuning.

We observed that the Alpaca model generation typically thanks the user and responds sincerely, which is also reflected in the average generation length being 136 tokens (10× more than T5). In contrast, Llama generated 98 tokens on average. It also is more likely to speak in the first person than as a third party, similar to Llama.

Inspection of the finetuned Alpaca-generated responses showed mostly factual responses, where the model would provide textbook-like definitions and excerpts describing a particular disease or condition: "Thank you for your question... NRBC count is a lab test measuring the amount of nucleated blood cells...[100 tokens omitted]" which appears off-topic and irrelevant to the user's question about lab results. This contrasts with expert responses which tended to be concise and pertinent: "The labs are acceptable. No reason to be concerned here." This finding is in line with recent research, showing that LLMs tended to produce more empathetic, verbose responses (Ayers et al., 2023). In contrast, some physician responses were dismissive of the user's concerns.

Qualitatively, adding the document to the prompt is important as LLMs produce textbook-like responses and, therefore must contextualize their responses in relevant literature to reduce hallucination. Without adding the document to the prompt, we find there are three primary sources of error: (1) the LLM hallucinates (replies as customer support), (2) strays off-topic (provides textbook definitions of a particular word in the prompt), or (3) provides incorrect advice (saying a lab result is normal when it is abnormal). This may indicate that the evaluation metrics used are not aligned with the response consumers want, but are aligned with the expert response that the consumer can understand. Examples and further analysis are in Appendix D.3.2.

Overall, the generation models provide coherent, well-formed sentences. Nevertheless, they are still unlikely to be useful without contextual evidence. We found that the T5 model's generations aligned more with expert responses (in length and tone). However, this is likely not aligned with the consumer expectations. LLMs are likely more aligned with consumer expectations, but have a more prominent tendency to hallucinate. We find that adding PubMed evidence to the prompt allows for improved responses from the model. Even without evidence in the prompt, training the model with these prompts produces responses better aligned with the expert. An application of this type of generation could be to draft the answers for experts interacting with consumers, where the expert could also provide the PubMed document.

Vocabulary differences between consumer and medical expert Analyzing the distinctions between the consumer and medical expert language could verify whether the experts' answers are consumer-friendly. We first analyze diction distinctions between a consumer and a medical expert by training a Random Forest model and analyzing SHapley Additive exPlanation (SHAP) values (Lundberg and Lee, 2017). A highly relevant signal was the grammatical person (we, your, I) and the mention of *virus* or *COVID*. The mention of COVID or using the first person was highly indicative of a layperson. In contrast, using the second or third person alongside a more general term such as virus or coronavirus was indicative of an expert.

We also measured the percentage of medical terminology used between an expert and a consumer. To do this, we used the Medical Subject Headings (MeSH) vocabulary as a proxy for the medical lexicon. We found that 36% of tokens from consumers were medically related, while 35% of tokens from the expert were medically related. This indicates that the lexicon of the consumer and experts are similar regarding medical terminology usage, which suggests that experts may use language that the consumer understands instead of medical jargon. We additionally found that the most common phrases were *feels like* and *see doctor*, and references to symptoms (area, duration). Most common words by doctors and consumers' comments were similar and mainly focused on instructions (see a doctor) and prepositions (looks like a benign tumor).

7 Conclusions

Consumer question answering in medicine requires real-world questions with expert-verified consumer-friendly answers. We presented `MedRedQA`, a large-scale dataset to facilitate research in this field. We explored *ranking* and *generation* tasks for question answering that contemporary neural models struggle with. Empirically, we showed that: (1)

for ranking, it is possible to find relevant expert responses within the first five responses with oracle retrieval; and (2) for the generation models, there is potential to generate human-aligned responses given PubMed evidence in the prompt to prevent model hallucination. For future studies, consumer-aligned metrics for generation should be explored.

Limitations

One limitation of our dataset is that it is only in English. However, other datasets of this size exist for other languages, such as in Chinese (Table 1). Other limitations include limited negative sampling and the use of weak supervision over expert gold labels and very sparse training labels (one per question), making the tasks more difficult. This may be remedied by crowdsourcing more data points by labelling the auto-regressive generation output or the ranking candidates produced in second-stage re-ranking.

Furthermore, the experimented Transformer models are not designed for arbitrarily long sequences of text, which may fail to capture the context of the longer questions.

A shortcoming of our current study is that, although we experimented with fine-tuning using medical abstracts for expert generation, it is more suitable for a ranking task to retrieve expert evidence from PubMed. However, as an open-ended ranking task, it is difficult to encapsulate the provenance task, as it becomes unclear what the set to be sampled is. Since the set contains 1000 documents, sparsity becomes a larger issue when retrieving from the entire PubMed corpus. Otherwise, ranking over the set of 1000 documents becomes a trivial task. We, therefore, leave these labels in the dataset as exploration for future work.

Furthermore, we could only use open-access transformer (Vaswani et al., 2017) models, which may not be state-of-the-art compared to private larger-scale ones such as GPT-3. This can potentially be explored in future work.

For our metrics, we could not report the harmonic mean of the bi-encoder with the cross-encoder as it would require running the cross-encoder across the entire corpus for every permutation of data points, which would take many months to complete on our limited hardware.

Another problem would be that we did not have direct annotation over the dataset. This could leave undesirable answers in the dataset. However, given the voting system, we assume that curt responses are unlikely to be voted highly and by using only expert answers, this may be mitigated enough so as to not be a major concern.

Finally, measuring user-oriented metrics such as brevity, coherence, and factual correctness for the auto-regressive generated responses is desirable as our metrics do not consider the user's preference. This, however, would require ample resources and cost for annotation and may be suitable for future exploration.

Ethics Statement

This work has received ethics approval from the authors' institutes' ethics committee. We also have approval from /r/askdocs for releasing the dataset and complying with Reddit's terms of service.

Consumer medical data may contain sensitive information about users. We partially mitigate this problem by removing users who have deleted their post or their accounts and removing posts containing links to images. Furthermore, we paraphrase any examples drawn from the dataset, including in this paper to avoid singling out any individual. We also do not include pseudonyms in the dataset, and require users of the dataset to sign an agreement acknowledging they have received both: (1) ethics approval from their internal review board and (2) have received the Reddit license. In order to allow users to opt out of the dataset, we provide a script to generate the dataset rather than releasing it as a static dataset.

Our experiments indicate that contemporary neural models cannot replace medical experts regarding advice. Instead, the aim should be to provide recommendations for specialists and provide further reading (through provenance) for the consumer or for drafting responses to consumers from the experts.

Furthermore, as with data from Reddit, a website where many users can contribute, there is a potential for bias or misinformation when using data from this website. However, the `/r/askdocs` forum has an effective verification strategy for medical experts and strong moderation, allowing for a higher-than-average quality of questions and answers. However, in times of mass posts such as the pandemic period, there could be potential for questions and answers to be under-moderated.

In our dataset, there may be triggering question-answer pairs for some people and we ask users of the dataset to execute caution in this regard. For instance, the most popular post on the `/r/askdocs` forum is suicide-related.

We make it clear that the dataset is to be used only for research purposes, as we find current models ill-suited for automatic responses or real-world usage.

References

Ben Abacha and Demner-Fushman. 2016. Recognizing Question Entailment for Medical Question Answering. *American Medical Informatics Association Annual Symposium Proceedings*, 2016:310–318.

T. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *CIKM*, pages 601–610, Hong Kong, China.

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*.

Marco Basaldella and Nigel Collier. 2019. BioReddit: Word embeddings for user-generated biomedical NLP. In *LOUHI*, pages 34–38, Hong Kong.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *ACL*, pages 2228–2234, Florence, Italy.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *CoRR*, abs/1901.08079.

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

Edward Loper Bird, Steven and Ewan Klein. 2009. Natural language processing with python. `http://www.nltk.org/`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Anthony Cocco, Rachel Zordan, David Taylor, Tracey Weiland, Stuart Dilley, Joyce Kant, Mahesha Dombagolla, Andreas Hendarto, Fiona Lai, and Jennie Hutton. 2018. Dr Google in the ED: searching for online health information by adult emergency department patients. *The Medical Journal of Australia*, 209:342–347.

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. *JAMIA*, 27(2):194–201. 31592532[pmid].

Renee Garett and Sean D Young. 2021. Online misinformation and vaccine hesitancy. *Transl. Behav. Med.*, 11(12):2194–2199.

Suzanne Graham and John Brookey. 2008. Do patients understand? *The Permanente journal*, 12(3):67–69.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*, Online.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Azhar Hussain, Syed Ali, Madiha Ahmed, and Sheharyar Hussain. 2018. The anti-vaccination movement: A regression in modern medicine. *Cureus*, 10(7):e2919–e2919. PMC6122668[pmcid].

Paul Jaccard. 1912. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11:37–50.

Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health card retrieval for consumer health search: An empirical investigation of methods. In *CIKM*, page 2405–2408, New York, NY, USA.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, pages 2567–2577, Hong Kong, China.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Anthony Celi, and Roger Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Roozenbeek Jon, Schneider Claudia, Dryhurst Sarah, Kerr John, Freeman Alexandra, Recchia Gabriel, van der Bles Anne Marthe, and van der Linden Sander. 2020. Susceptibility to misinformation about covid-19 around the world. *R. Soc. Open. Sci.*, 7.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and efficient few-shot learning with language models. In *ACL*, pages 3638–3652, Dublin, Ireland.

Anastasia Krithara, Anastasios Nentidis, Georgios Paliouras, and Ioannis Kakadiaris. 2016. Results of the 4th edition of BioASQ challenge. In *Proceedings of the Fourth BioASQ workshop*, pages 1–7, Berlin, Germany.

Bridget M. Kuehn. 2013. More Than One-Third of US Individuals Use the Internet to Self-diagnose. *Journal of the American Medical Association*, 309(8):756–757.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *CoRR*, abs/2203.05115.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*, pages 74–81, Barcelona, Spain.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *NIPS*, volume 35, pages 1950–1965.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *ICLR*, Toulon, France.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. In *EMNLP*, pages 2833–2861, Abu Dhabi, United Arab Emirates.

NCBI. 2023. Pubmed - ncbi. `https://pubmed.ncbi.nlm.nih.gov/`. [Online; accessed 20-Jan-2023].

Taylor Nelson, Nicole Kagan, Claire Critchlow, Alan Hillard, and Albert Hsu. 2020. The danger of misinformation in the COVID-19 crisis. *Mo. Med.*, 117(6):510–512.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. 2017. Results of the fifth edition of the bioasq challenge. In *Biomedical Natural Language Processing*, pages 48–57, Vancouver, Canada.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2022. Overview of BioASQ 2022: The tenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 337–361, Cham. Springer International Publishing.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Georgios Paliouras, and Ioannis Kakadiaris. 2018. Results of the sixth edition of the BioASQ challenge. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 1–10, Brussels, Belgium.

NewsGuard. 2021. Sizing the infodemic: Newsguard analysts have now found more than 500 'news' sites peddling covid-19 misinformation and identified 50 hoaxes relating to the covid-19 vaccines.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *NIPS Cognitive Computations Workshop*, volume 1773, Barcelona, Spain.

Vincent Nguyen, Sarvnaz Karimi, and Zhenchang Xing. 2019. ANU-CSIRO at MEDIQA 2019: Question answering using deep contextual knowledge. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 478–487, Florence, Italy.

Vincent Nguyen, Maciej Rybinski, Sarvnaz Karimi, and Zhenchang Xing. 2022. Search like an expert: Reducing expertise disparity using a hybrid neural index for covid-19 queries. *Journal of Biomedical Informatics*, 127:104005.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Chansung Park. 2023. chansung/gpt4-alpaca-lora-7b. `https://huggingface.co/chansung/gpt4-alpaca-lora-7b`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, pages 3982–3992, Hong Kong, China.

Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, William R. Hersh, Steven Bedrick, Alexander Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine track. In *TREC*, Gaithersburg, MD.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. 2021. Overview of the TREC 2020 Precision Medicine Track. In *TREC*, Gaithersburg, MD.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the TREC 2018 Precision Medicine Track. In *TREC*, Gaithersburg, MD.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the TREC 2019 Precision Medicine Track. In *TREC*, Gaithersburg, MD.

Kirk Roberts, M. Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. 2016. State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS track. *Inf Retr.*, 19(1-2):113–148.

Kirk Roberts, Matthew Simpson, Ellen Voorhees, and William Hersh. 2015. Overview of the TREC 2015 clinical decision support track. In *TREC*, Gaithersburg, MD.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *TREC*, Gaithersburg, MD.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

Chaitanya Shivade. 2019. Mednli - a natural language inference dataset for the clinical domain (version 1.0.0. *PhysioNet*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Noor Van Riel, Koen Auwerx, Pieterjan Debbaut, Sanne Van Hees, and Birgitte Schoenmakers. 2017. The effect of dr google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. *BJGP Open*, 1(2):bjgpopen17X100833.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 6000–6010.

Ellen Voorhees, Alam Tasmeer, Demner-Fushman Dina, Hersh William, and Kyle Lo. 2020. TREC-COVID: Constructing a pandemic information retrieval test collection. *ACM SIGIR Forum*, 54(1):1–12.

Ellen M. Voorhees and William R. Hersh. 2012. Overview of the TREC 2012 medical records track. In *TREC*, Gaithersburg, MD.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. Redhot: A corpus of annotated medical questions, experiences, and claims on social media. In *EACL: Findings*.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. BERT-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *ICTIR*, page 317–324, New York, NY, USA.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*, Online.

Dirk Weißenborn, George Tsatsaronis, and Michael Schroeder. 2013. Answering factoid questions in the biomedical domain. In *CEUR Workshop Proceedings*, volume 1094, Valencia, Spain.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, page arXiv:1910.03771.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *CIKM*, page 497–506, Torino, Italy.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *EMNLP*, Hong Kong, China.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *EMNLP (Findings)*, pages 3840–3849, Online.

In the appendix, we produce further experimental details and results.

A Packages and Experimental Setup

For dataset analysis, we used the NLTK (Bird and Klein, 2009) library for tokenization and stopword removal. For the experiments, we used the Huggingface Transformers (Wolf et al., 2019) library for training, evaluating and loading the ranking and generation models. We also use this library for tokenizing the data. Furthermore, all experiments were reported using a fixed seed of 42 for all random initialization or random sampling. We report our experiments for a single run as the primary purpose of the work is to illustrate the feasibility of the task and the usability of the dataset. All experiments (training, evaluation, testing) were conducted on a single Jetson NX Orin 32 GB developer kit.

B Negative Results

Universal sentence embedding training and warm-starting of the embedding space As the dataset has sparse training data, we assumed that warm-starting the embedding space with in-domain data before finetuning it with a task similar to the downstream task would produce a better ranking. To do this, we warm-start our embedding space with our curated medical subset of the MS-Marco (Nguyen et al., 2016) dataset. We then filtered MS-Marco using MetaMap, to ensure in-domain training and the model learns how to rank expert answers by relevance rather than semantics.

After warm-starting, it is beneficial to ensure that the bi-encoder sentence embeddings are universal (Reimers and Gurevych, 2019). To do this, we finetuned the PubMEDBERT model on the NLI and STS-B datasets using the settings indicated by the original authors (Reimers and Gurevych, 2019). We then use this universal sentence encoder for further finetuning. We train the model with a constant learning rate of 2×10^{-5} for 5 epochs and select the best checkpoint based on the validation set. Rather than directly sampling negatives, we opt for a loss function designed for datasets with only positive samples. We use the Multiple Negatives Ranking Loss (Henderson et al., 2017) which uses in-batch negative sampling. These results are shown in Table 8 and indicate that the baseline model works best.

	Facet	**Acc@1000** ↑	**HR** ↓
BM25 (Baseline)	Title	0.0890	1333
Bi-Encoder (Baseline)	Body	0.1734	600.1
Bi-Encoder (NLI)	Body	0.1080	999.4
Bi-Encoder (NLI + Warm-Start)	Body	0.0975	1144

Table 8: Negative results indicating that NLI (universal sentence embeddings) and warm-starting the embedding space lead to worse results for ranking.

We found similar shortcomings with MonoBERT (Nogueira et al., 2019) when warm-starting, including strong overfitting immediately at the start of training. This is notable given that MonoBERT was pretrained using MS-Marco, indicating that the medical subset of MS-Marco is likely noisy or has reinforced inductive biases in the model that were not useful for the downstream task.

However, experiments with provenance retrieval reveal that warm-started models were better at retrieving the relevant PubMed abstracts. This reveals there is still merit to warm-starting the embedding space as this aligns better with the data distribution of the layperson questions to PubMed articles rather than layperson to the expert response.

C PubMed annotation

C.1 Annotation instructions

We randomly sampled 10% of examples and annotated 100 examples between 3 annotators of academic backgrounds in biomedical natural language processing. We labeled each example with one of three labels: irrelevant, relevant, or unknown. To assess for relevance, we compared the user's question, and expert response and see if the provided PubMed abstract is relevant. Specifically, the abstract is relevant if it either: (1) addresses the user's information need (contains the user's mentioned disease); (2) explains the expert's diagnosis (contains the mentioned diagnosis); or (3) is used for differential diagnosis (for example, it was provided to rule out the disease the author may have proposed). We did not, however, determine if the PubMed abstract contained the correct diagnosis or prognosis for the individual and only judged the article's relevance from the criteria, as this was outside the area of expertise for the annotators.

C.2 Annotation results

We found that 69 out of 100 were deemed relevant, and 3 were considered unknown. Many articles deemed irrelevant were in response to posts about a general topic, for example, a weekly post asking physicians to post articles they found interesting or were a result of a parsing error. For instance, PMIDXXXXXX was unavailable so PMCXXXXXXX was retrieved instead.

D Additional Experimental Details

D.1 T5 Prompt Templates

The following four prompts were used during the training of the T5 models, and were inspired by the Flan prompts[4]:

1. Provide a response for the question: Question: *context* Answer:
2. title: *user_title* question: *context* answer:
3. Q: *context* A:
4. Write a title for the following question: Question: *context*

The prompt template was randomly picked from a random uniform distribution for each training example. However, we used a fixed seed of 42 for the random generation for reproducible results. During the generation of validation or test examples, we exclude the fourth prompt template as the response generated would otherwise be a title rather than an expert response. After each prompt, the model is trained to conditionally generate a response to the full prompt, which is the template prompt with the fields filled with a sample from the training set.

D.2 Llama and Alpaca Prompt Templates

We use the following instruction-based prompt for generation based on the following template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction: Given the following user question and document, provide useful, medical advice, citing evidence from the document where possible.

### Question: {question}

### Document: {document}

### Response:

The prompt used is derived from the original Alpaca prompt template but adjusted to suit the current task of producing an expert response given a PubMed abstract and user question.

D.3 Generation Output Examples

D.3.1 T5-based generation analysis

We sample our generated examples based on their mover score. To avoid cherry-picking, we look at the best (Table 9) and worst (Table 10) generations based on their MoverScore, and randomly sample a generation from the final set (Table 11). We consider the Flan T5 model that has been finetuned as it performed the best and find the corresponding generation in the untuned model for comparison.

In the worst case, we find both the Flan-FT and untuned models output the end-of-sentence token too early for the same input. In the best-performing case, the generated Flan-T5 finetuned model produces an answer more aligned with human experts, whereas the untuned model copied the prescription from the input text. We see this strategy occurring several times in the generation for the untuned model as it falls back to squad-style question answering, where it copies spans from the context as an answer to the question.

On average, we find that the Flan-T5 model, finetuned or otherwise, recommends the consumer to see a medical specialist. That being said, the untuned model tended to be more verbose and indirect, which was refined after finetuning. This suggests that the model learns to be more direct with its responses during the training process.

Interestingly, we find it difficult to get T5 to generate longer responses, where it would often generate singular-word responses. However, the LLMs have the opposite problem where they are almost always hitting the generation limit.

D.3.2 LLM-based generation analysis

Similar to the previous section, we adopt the same sampling strategy to select generation examples. We found that in the worst case, the Alpaca hallucinates immediately when the source document is removed – acting as healthcare customer support, rather than providing answers to the user's question. Llama models tended to act textbook-like in the worst-case scenario, but produce more coherent responses to less difficult questions (Table 12).

With the source document, we find that Alpaca is more direct in its responses, and can address the user's information need sooner in the generation. Interestingly, we find that alignment between the reference response and the LLM-generated responses is worse (Table 13) when the physician dismisses the user's concerns or questions. We find

---

[4] `https://github.com/google-research/FLAN/blob/main/flan/templates.py`

that LLMs will always address the user's concerns, and in the event that the physician is dismissive, it attempts to ask clarifying questions to better understand the user's intent.

On average (Table 14), we found that Llama will respond as a third party, whereas Alpaca models tended to speak in the first person. With the document in the prompt, the Alpaca models were more direct and less likely to hallucinate.

However, in the paper, we find that the metrics do not perfectly align with consumer expectations. We find that MoverScore, is a useful metric for relative importance and ranking. As shown in the Tables, it still describes the best, average and worst-case scenarios.

Flan FT	Prediction	It’s a good antibiotic.
Flan	Prediction	Prescribed Bactrim DS for bladder infection.
	Reference	Bactrim is a very safe and well-tolerated antibiotic.

Table 9: Best performing generation examples based on MoverScore. Reference here is the expert response to the user’s question.

Flan FT	Prediction	Consider
Flan	Prediction	summer
	Reference	It’s just a boil. Sweat, dirt, and dead skin cells block the pores and infect them. I’d know better if I could examine you, but is the affected area around the boil turning red and getting hard to the touch? It looks like there is cellulitis present as well (or if it’s not there it’s coming). \n\nYou need antibiotics pronto. You should also use antibacterial soap and let your breasts hang fee as much as you can. Go get antibiotics for MRSA (methicillin resistant staph aureus) as soon as possible. This can get really bad quickly.

Table 10: Worst performing generations based on MoverScore.

Flan FT	Prediction	You need to see a doctor.
Flan	Prediction	I think you’re right. Stress and anxiety can cause a variety of symptoms. It can also cause other health problems. You should seek medical attention if you have any of these symptoms."
	Reference	Were you educated about HSV after your diagnosis? Stress manifests in all different ways for different people. You could make an appointment to speak about your concerns with your new diagnosis. Im sure there’s a community on Reddit who have all gone through what you are now. Just breathe. You’ll get through it and learn how to live with it.

Table 11: Randomly sampled from the typical distribution i.e. average case generations.

Alpaca Finetuned (FT)	Thank you for your question. It is important to note that there is currently no conclusive evidence that the Gardasil vaccination can prevent the development of cervical cancer. The vaccine protects against four high-risk human papillomavirus (HPV) types that are responsible for approximately 70% of cervical cancers. However, it does not protect against all types of HPV, including those that cause genital warts. Therefore, even after receiving the Gardasil vaccine, it is still possible to develop genital warts. It is important to note that the strains of HPV that cause genital warts are different from the ones that cause cervical cancer. In fact, most people who get genital warts do not go on to develop cervical cancer.
Alpaca FT w/o Document	Dear [user], Thank you for reaching out to us for assistance. We are sorry to hear about your concerns regarding your health. It is important to note that there is no direct correlation between the Gardasil vaccination and the development of plantar warts. The Gardasil vaccine protects against four high risk human papillomavirus (HPV) types which can lead to cervical cancer. These include HPV types 6, 11, 16, and 18. Plantar warts are caused by different types of HPV, namely HPV type 17. Therefore, it is unlikely that the Gardasil vaccine would have prevented the development of plantar warts. It is important to note that the Gardasil vaccine does not protect against all types of HPV.
Alpaca	Thank you for reaching out to us! It sounds like you have a lot going on and we want to help you navigate through your concerns. Firstly, let's start with the basics. The Gardasil vaccine protects against certain strains of Human Papillomavirus (HPV). These strains can cause cervical cancer, genital warts, and other health issues. The vaccine does not prevent all strains of HPV, however, it does reduce the risk of developing these conditions. It is important to note that even if you were exposed to HPV before receiving the vaccine, it does not mean that the vaccine did not work. In fact, the vaccine works best when given before exposure to HPV.
Llama	The answer is yes, you probably have HPV. You can get tested for HPV through your primary care physician. If you test positive for HPV, then you need to get a Pap smear every six months until you are done having children. If you test negative for HPV, then you don't need to get a Pap smear anymore. ## External links * Official website
Reference	The Gardasil vaccine is formulated to protect first and foremost against the strains of HPV that cause cervical cancer. More recent version of the vaccine have included more of the strains that just cause warts, but it doesn't protect against every strain of HPV. Having genital warts doesn't mean that your Gardasil vaccine didn't work. There is no need for you to be getting Pap smears any more frequently than is normally suggested.

Table 12: Best performing generation example based on MoverScore (Alpaca FT) and the baseline counterpart.

Alpaca FT	Thank you for reaching out to us. We appreciate your interest in our platform and we are happy to help you with your query. The provided document does not contain sufficient information to provide a proper medical advice. It would be helpful if you could share more details about your symptoms, such as the type of pain, its location, duration, severity, associated symptoms, and any other relevant information. Also, it would be beneficial to know your age, gender, and any past medical history. We recommend that you consult a doctor for a thorough evaluation and diagnosis of your condition. They will be able to determine the cause of your pain and suggest appropriate treatment options based on your individual needs. Please note that we are not a medical facility and cannot provide medical advice or diagnose conditions over the internet. If you need immediate assistance, please call 911 or visit your nearest emergency room. Best wishes for a speedy recovery.
Alpaca FT w/o Document	Hello there, thank you for reaching out to us. We appreciate your interest in learning more about medicine and healthcare. As a community-based platform, we encourage open dialogue between members and medical professionals. However, due to HIPAA regulations, we cannot provide personalized medical advice without having access to your full medical history. If you would like to receive personalized medical advice, please consider scheduling an appointment with a doctor through our partner telemedicine service, [Teladoc](https://www.teladoc.com/). Teladoc offers 24/7 access to board-certified physicians who can diagnose, treat, and prescribe medications when medically necessary. You can find more information about Teladoc by visiting their website or calling 1-800-831-6599. We hope this helps! We look forward to continuing the conversation with you.
Alpaca	Thank you for reaching out to us. We appreciate your interest in our platform and we are happy to help you with your query. The provided document does not contain sufficient information to provide a reliable medical advice. It is important to note that the results presented in the document are based on a small sample size and therefore cannot be generalized to the entire population. Furthermore, it is unclear whether the authors controlled for confounding factors such as age, lifestyle habits, and other environmental factors which could affect the outcome measures.It is recommended that you consult a qualified physician who can conduct a thorough physical examination and review your medical history before providing any medical advice. We hope this helps. Please feel free to reach out to us again if you need more assistance. Best regards, AskDocs Team
Llama	The first thing we need to consider is whether the patient is experiencing symptoms related to their heart. If they are, it would be important to rule out other causes such as anxiety, depression, or stress. If these are ruled out, then we should consider the possibility of a cardiac cause.\em{If the patient is experiencing no symptoms, then we could consider the possibility of a cardiac cause. However, it is more likely that the patient is experiencing symptoms related to another condition. For example, the patient may be experiencing symptoms related to anxiety, depression, or stress. It is important to rule out these conditions before considering a cardiac cause.}\em{If the patient is experiencing symptoms related to anxiety, depression, or stress, then we should consider treating those conditions.
Reference	It's pseudoscience.

Table 13: Worst performing generation example (Alpaca FT) based on MoverScore and the baseline counterpart.

Alpaca FT	Thank you for reaching out to us for assistance. Based on the information provided, it appears that your niece may have contracted a type of Human Papillomavirus (HPV) which causes warts. There are many types of HPV, some of which can lead to more serious conditions such as genital warts and cervical cancer. The good news is that most cases of HPV clear up on their own within two years. In terms of treating the wart on the top of her hand, it would be best to consult with a dermatologist. They will be able to assess the wart and determine the best course of action based on its location, size, and appearance. Some dermatologists may choose to use cryotherapy (freezing) to remove the wart, while others may opt for other methods such as laser treatment, electrosurgery, or curettage
Alpaca FT w/o Document	Dear Aunt, Thank you for reaching out to us for help. Based on your description, it sounds like your niece may be suffering from HPV infection. Human papillomavirus (HPV) is a common virus that can cause warts in various parts of the body including the hands, feet, face, genital area, and mouth. The most common type of HPV that causes warts is called HPV-6 and HPV-11. These types of HPV usually do not cause cancer but can lead to warts. The best way to prevent HPV infection is by getting vaccinated against HPV before becoming sexually active. The HPV vaccine protects against several strains of HPV, including those that cause genital warts.
Alpaca	Dear Customer, Thank you for reaching out to us for assistance. Based on the information provided, it seems like your niece may have multiple warts on her hands and face. The most effective way to treat these warts would be to see a dermatologist as soon as possible. They can assess the condition and determine the best course of action based on the location, size, and type of warts. The dermatologist may use one of several methods to remove the warts including cryotherapy, laser surgery, curettage, electrosurgery, or chemical peels. These procedures are usually done in the office under local anesthesia. Some warts may require more than one procedure to completely eliminate them. It is important to note that warts can recur even after successful removal, especially if the patient does not follow proper hygiene practices. Therefore, it is essential to maintain good personal hygiene
Llama	The patient needs to see a dermatologist. The wart is likely a sebaceous cyst. Sebaceous cysts can be treated with surgery, cryotherapy, electrocautery, laser ablation, or liquid nitrogen. The best option would be to remove the cyst. If the cyst is removed, then the patient may need to take antibiotics to prevent infection. ## External links * Official website
Reference	There are many nonfreezing treatments for warts.

Table 14: Average randomly sampled generations and the baseline counterpart.