Enhancing Consumer Health Question Reformulation: Chain-of-Thought Prompting Integrating Focus, Type, and User Knowledge Level

Jooyeon Lee, Luan Huy Pham, Özlem Uzuner
Department of Information Sciences and Technology
George Mason University
{jlee252, lpham6, ouzuner}@gmu.edu

Abstract

In this paper, we explore consumer health question (CHQ) reformulation, focusing on enhancing the quality of reformation of questions without considering interest shifts. Our study introduces the use of the website of the Genetic and Rare Diseases Information Center (GARD) at the National Institutes of Health (NIH) as a gold standard dataset for this specific task, emphasizing its relevance and applicability. Additionally, we developed other datasets consisting of related questions scraped from Google, Bing, and Yahoo. We augmented, evaluated and analyzed the various datasets, demonstrating that the reformulation task closely resembles the question entailment generation task. Our approach, which integrates the Focus and Type of consumer inquiries, represents a significant advancement in the field of question reformulation (QR). We provide a comprehensive analysis of different methodologies, offering insights into the development of more effective and user-centric AI systems for consumer health support.

Keywords: Consumer Health Question Answering, Chain-of-Thought Prompting, Entailment Question Generation, Question Reformulation

1. Introduction

Our research explores the field of consumer health question answering, a specialized area within question-answering systems that aims to provide medical knowledge to the general public. This area presents unique challenges, as it requires AI systems to communicate complex medical information in a clear and accessible manner to users who may have little to no medical training. Despite improvements in AI systems, consumers often face the burden of formulating effective queries to obtain the information they need. This process can involve a manual series of trial-and-error attempts where consumers refine their questions to learn more about their health concerns. Our study focuses on understanding how consumers reformulate their questions in the consumer health domain, drawing on insights from a study by Chen et al. that examined user behavior in Question Reformulation (QR). This study provides valuable understanding of the reasons and methods consumers use to modify their queries in search of health information. According to Chen et al. there are four primary reasons for QR: satisfaction with results, dissatisfaction leading to modification, user-initiated improvements for better alignment with search intent, and shifts to different Foci. Our research specifically addresses the first three reasons, excluding interest shifts, as a shift generally indicates a change in intent. We aim to enhance consumer satisfaction by providing suggestions that match their intended query Focus, such as maintaining the core focus while avoiding shifts to new Foci. In this paper, the definition focus of a question indicates disease names as defined by Roberts et al. for the purpose of decomposing Consumer Health Questions (CHQs) and type is nondisease information of the question (e.g., symptom and treatment) (Demner-Fushman et al., 2019). Our experiments reveal a significant similarity between entailment and QR tasks, addressing a key challenge in CHQ reformulation - the lack of a standardized dataset for system evaluation. By aligning our task with entailment, we utilize multiple data sources to improve the robustness and relevance of our findings. Our main contributions are as follows: **[1.]** We propose datasets that are particularly suited for the task of CHQ reformulation. **[2.]** Our study compares three question generation methods: QR definition (ref-def) prompting, entailment definition (ent-def) prompting, and Chain-of-Thought (CoT) Prompting (Reynolds et al., 2021), to find the best way to reformulate CHQs. **[3.]** We introduce a CoT prompting technique that focuses on focus and type specifics for CHQ reformulation. This approach aims to make the reformulated questions (RQs) more relevant, accurate, and helpful.

2. Datasets

For the purposes of this work, we employed multiple datasets. This section provides a description of each dataset, including origin and characteristics.

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 220–228
20 May, 2024.

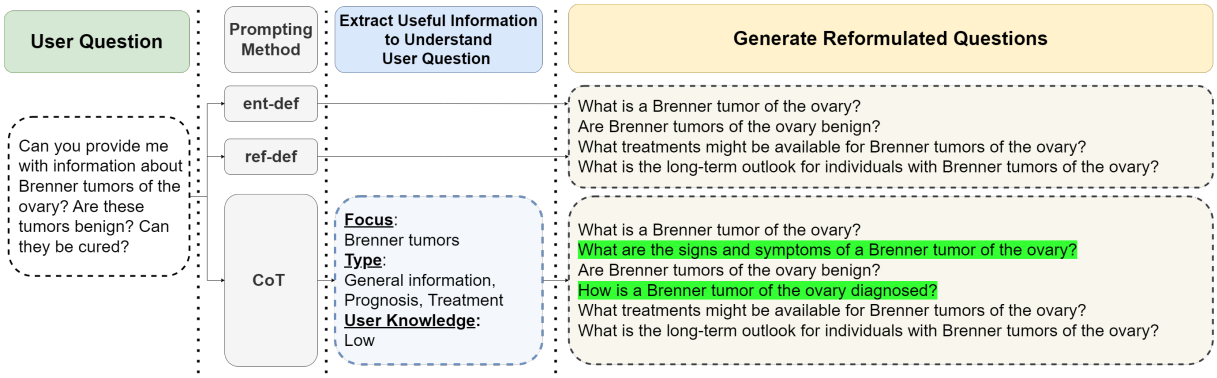

Figure 1: Prompt flow for standard prompting (ent-def and ref-def prompt) vs CoT prompting with example.

2.1. Entailment Question

In our study, we employed the Recognizing Question Entailment (RQE) dataset to develop algorithms for identifying similarities between CHQs and expert-answered queries (Ben Abacha and Demner-Fushman, 2016). This dataset includes real-world CHQs from the National Library of Medicine and FAQs from various NIH websites, encompassing a wide range of medical Foci. The goal of the RQE task is to determine if an existing FAQ answer can also respond to a new CHQ. We select the 129 true entailment pairs to test the QR models.

2.2. GARD Reformulated Questions

We created a dataset using content scraped from the website of the Genetic and Rare Diseases Information Center (GARD), a part of the US National Institutes of Health (NIH). GARD provides public information regarding rare diseases, including causes, treatments, and submitted CHQs paired with expert-provided responses (GARD). For instance, when we go to GARD website, section cysticercosis, the page has a "see answer" link under "GARD Answer", which redirects to a different page, that lists the CHQs, their answers with references, and the expert suggested questions are introduced with the phrase "The following information may help to address your question:". It is important to note that direct access to these expert suggestions is not available in the current version of the GARD website. However, an earlier version of the site, preserving these valuable expert insights, is accessible through Internet-in-a-Box, an innovative solution designed to provide offline access to various educational content, including medical resources (Internet-in-a-Box Team, 2016). This digital resource, available at `https://iiab.me/modules/en-nih_rarediseases/diseases/categories/index.html`, offers an archived version of the GARD site.

We scraped these questions upto 250 CHQ-RQ pairs. Of the pairs CHQs in the dataset, 48 were modified to explicitly state the topic, which could be inferred from the section title on the GARD website but was not directly mentioned in the CHQ. For example, "*this* condition..." is changed to: " *Myostatin-related muscle hypertrophy* condition...".

2.3. Search Engine API

We developed QR datasets using the RQE and GARD CHQs using SerpApi. SerpApi is a commercially-available web engine scraping service which has People Also Ask/Related Questions APIs for Google, Bing, and Yahoo (jvmvik et al., 2024). According to the web traffic analysis website StatCounter, these three search engines comprise over 96% of the global search market in February 2024 (Chen et al., 2021). SerpApi furnishes code libraries in various programming languages. We used version 2.4.1, which allowed text queries to be submitted (specifying the search engine as a parameter) and the RQs be returned in JSON format. Our inputs were the GARD and RQE datasets.

3. Experiment Setup

Our experimental framework is structured around GPT-4 (OpenAI et al., 2024), a model recognized for its advanced language processing and generation proficiency. The methodology employed in this study was based on a one-shot prompting technique, which was consistently applied across all experiments.

3.1. Prompts

Our study involved a comparative analysis of three distinct prompts to assess their effectiveness in the reformulation of CHQs: 1) ref-def prompt, 2) ent-def prompt, and 3) Chain-of-Thought (CoT) Prompting. Actual prompts are shown in Table 1.

The ref-def prompt was designed for QR task and is aimed to rephrase a given CHQ into several

Method	Actual Prompt
ref-def Prompt	The goal is to reformulate the given consumer health question into several, clearer single-sentence questions that could potentially answer the original, given consumer health question to increase satisfaction of the consumer."
ent-def Prompt	The goal is to reformulate the given consumer health question into several clearer single-sentence questions which are in an entailment relationship to the original given consumer health question. The definition of an entailment relationship is 'when question A can answer question B partially or fully, then question B entails question A.'"
CoT Prompt	The goal is to reformulate the given consumer health question into several, clearer single-sentence questions that could potentially answer the original, given consumer health question to increase satisfaction of the consumer. **(1)** Identify the Main Health Topic (Focus): Determine the primary health condition in the user's question to establish the central subject of inquiry. **(2)** Assess User Knowledge Level (Knowledge): Evaluate the consumer's familiarity with the health topic based on the language and concepts used in their question. This assessment categorizes knowledge as LOW, MEDIUM, or HIGH. LOW Knowledge Level: Indicates a basic or minimal understanding of the health condition or topic. The user may be unfamiliar with the condition or its implications. This level typically includes general inquiries or seeks foundational information. Example questions might be: "What is [health condition]?" "What causes [health condition]?" "Are there common symptoms associated with [health condition]?" MEDIUM Knowledge Level: Suggests a moderate understanding of the condition. The user might know what the condition is or some of its symptoms, but seeks more detailed or specific information. This level often involves questions about management, treatment options, or lifestyle impacts. Example questions might be: "What are the treatment options for [health condition]?" "How does [health condition] typically progress over time?" "Can lifestyle changes impact the course of [health condition]?" HIGH Knowledge Level: Indicates an advanced understanding or familiarity with the health condition. Users at this level often have detailed knowledge about the condition and seek highly specific, nuanced, or recent information. This might include questions about the latest research, complex treatment options, or specific subtypes of the condition. Example questions might be: "What are the latest research findings on [health condition]?" "Are there new or experimental treatments for [health condition]?" "How does [health condition] interact with other coexisting conditions?" **(3)** Determine Information Needs (Type): Identify what specific aspects of the condition the user is interested in, such as symptoms, causes, treatments, prognosis, or lifestyle impacts. **(4)** Question Segmentation: Segment the consumer question into individual, focused questions using the determined Focus, Type and User Knowledge Level. Each question should address a single aspect of the Focus.

Table 1: This table shows actual prompts used for the experiments.

clearer, single-sentence questions that could potentially provide answers to the original CHQ, thereby increasing consumer satisfaction.

The ent-def prompt was based on the question entailment definition as defined by Ben Abacha and Demner-Fushman, particularly suited for question-answering tasks: "When question A can answer question B partially or fully, then question B entails question A."

For the CoT Prompting, we expanded on our prior research that underscored the significance of identifying the Focus and Type in the question entailment recognition task for CHQs (Lee and Pham, 2022). This approach integrates the CoT method, introduced by Reynolds et al., which enhances performance by incorporating reasoning steps. Our prompt has three chains: 1) extract the Focus; 2) determine the Type; and 3) evaluating the user's knowledge level about the Focus. 4) considering 1, 2, 3, reformulate the CHQ.

3.2. Evaluation Method

Our approach for evaluating the accuracy of generated questions involves text similarity measurement analysis. We aligned the output generated by our models with a predefined gold standard of questions. We consider the output of models as a single output by concatenating the list of questions and doing the same for the gold standard (GARD expert suggestion). This enables a direct comparison to obtain a similarity score. For the similarity metric, we utilized UniEval, a multifaceted tool designed for evaluating text generation tasks. It measures aspects like consistency, coherence, and relevance (Zhong et al., 2022). We also applied the ROUGE (R) metric, with an emphasis on R-1 for unigram overlap and R-L for the longest common subsequence (Lin, 2004), to analyze the lexical similarity and coherence of the generated research questions. This methodological blend offered a robust framework for assessing the effectiveness of our models.

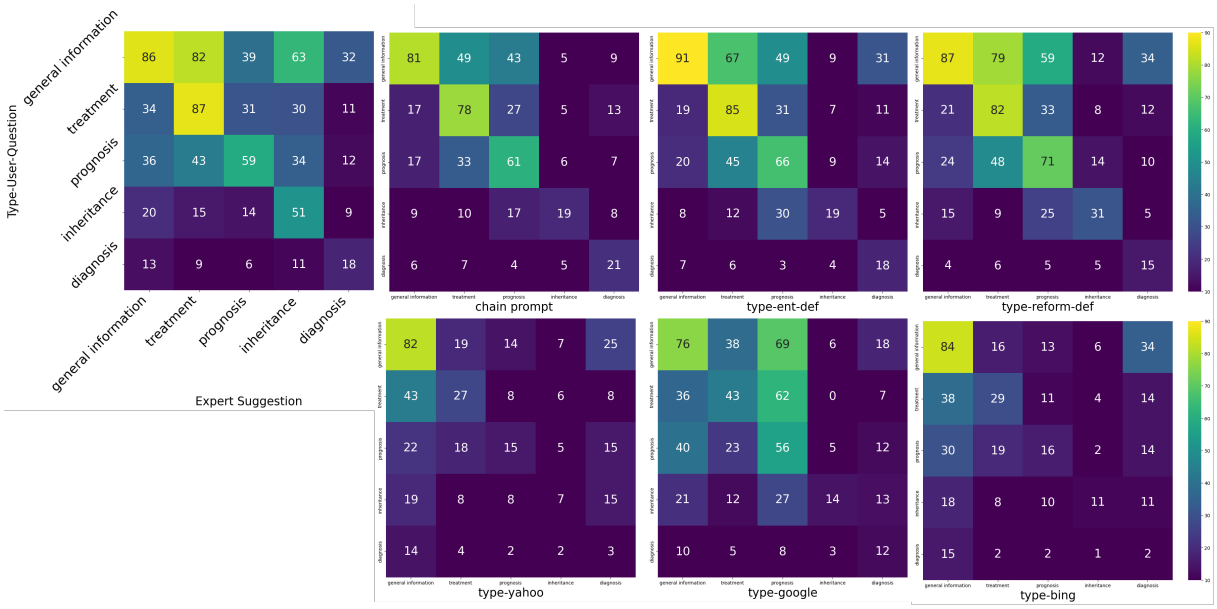

Figure 2: Heatmap of GARD dataset vs generated output (5 most frequent types only).

Data	Method	Consistency	Coherence	Relevance	R-1	R-L
GARD	CoT prompt	**0.9338**	**0.9604**	0.8945	**0.3115**	**0.2630**
	ref-def prompt	0.9003	0.9472	**0.8975**	0.2964	0.2344
	ent-def prompt	0.9087	0.9456	0.8833	0.2892	0.2300
	Bing	0.7978	0.8078	0.7475	0.1487	0.1231
	Google	0.7704	0.7262	0.6383	0.2101	0.1711
	Yahoo	0.7979	0.8037	0.7442	0.1572	0.1318
GARD**	CoT prompt	**0.9339**	**0.9620**	**0.9414**	**0.4269**	**0.4241**
	ref-def prompt	0.9068	0.9530	0.9345	0.4128	0.4061
	ent-def prompt	0.9151	0.9525	0.6512	0.1946	0.1918
	Bing	0.8193	0.8728	0.8690	0.3924	0.3910
	Google	0.7927	0.8325	0.7941	0.3523	0.3510
	Yahoo	0.8151	0.8784	0.8676	0.3634	0.3614
RQE	CoT prompt	0.7436	0.8058	0.7690	**0.1208**	**0.1152**
	ref-def prompt	0.7411	**0.8798**	**0.8506**	0.1097	0.1030
	ent-def prompt	0.7424	0.8532	0.8226	0.1118	0.1055
	Bing	0.7690	0.7937	0.7620	0.0743	0.0662
	Google	0.7060	0.7290	0.6934	0.0868	0.0819
	Yahoo	**0.7868**	0.8170	0.7864	0.0742	0.0698

Table 2: Evaluation using with UniEval (consistency, coherence and relevance), R-1 and R-L. GARD** designates results on the GARD dataset where output and user questions possess identical Focus (to have a "fairer" comparison with search engine suggestions by excluding "intention shift").

Furthermore, we incorporated quantitative metrics for a comprehensive assessment. We calculate the frequency of Focus and Type elements in the responses, thereby measuring their alignment with the gold standard, which is shown in Table 3. We use the Euclidean Distance to measure the similarity between gold standard vs. prompt outputs and gold standard vs. search engine. The formula is $\text{Distance}_i = \sqrt{(F_0 - F_i)^2 + (T_0 - T_i)^2}$ where, F_0 and T_0 are representing an average number of Focus and an average number of Type values, respectively, for the gold standard. F_i and T_i are the corresponding values for the row being compared.

4. Analysis

In this section, we provide an analysis of results.

4.1. Question Reformulation Analysis

The performance of prompts on the GARD dataset using ref-def, ent-def, and CoT prompting (which incorporates Focus and Type along with reformulation definition) was remarkably similar when tested against an expert-suggested dataset. This similarity indicates a strong alignment between the task and the dataset. Notably, CoT prompting demonstrated superior performance over other methods

	Focus	Type	Distance
GARD User Question	1.6447	2.0220	2.897
Gold standard	1.5311	4.8462	-
CoT Prompting	2.2308	4.3187	**0.707**
Ref Def Prompting	2.2418	**5.2601**	0.715
Ent Def Prompting	**1.7289**	3.2015	1.648
Bing API	2.3736	2.8425	2.034
Google API	2.7509	3.3663	1.596
Yahoo API	2.3040	2.8352	2.028

Table 3: Comparative analysis of prompting methods against the gold standard, using an average occurrence of Focus and Type. Method alignment with the gold standard calculated using the Euclidean distance (lower values = greater similarity).

in terms of consistency, coherence, and R-1 and R-L scores. This suggests that understanding the Focus and Type of a question before generating a reformulated version is crucial for this specific task.

4.2. Entailment Task Analysis

In our analysis of entailment prompts, we noted that the ref-def prompt and the ent-def prompt demonstrated remarkably similar patterns. As shown in Table 2, the performance scores of both prompts were closely aligned when applied to the GARD dataset and RQE dataset, across multiple evaluation methods. This parallel trend is also evident in Figure 2, which further corroborates our observation. These results align with the patterns of human behavior observed in previous surveys, as cited in (Chen et al., 2021). For future research, especially in scenarios where there is no “interest shift" in user intent, augmenting the reformulation task with question entailment datasets emerges as a promising strategy to overcome data limitations.

Regarding the entailment dataset, we conducted the same experiment using the RQE dataset to assess its similarity to the QR task. Considering that the RQE dataset consists of only 129 entailment question pairs and was originally designed for entailment recognition tasks, the reformulation pairs are limited to user questions with a single corresponding entailment question. Therefore, the evaluation might not fully represent real-world scenarios. However, similar to the findings with the GARD dataset, the CoT Prompt method outperformed others on the ROUGE metrics.

4.3. Search Engine Behavior Analysis

Our study also delved into the behavior of search engines in offering QR suggestions. Utilizing the GARD dataset, Google provided QR suggestions for 99.6% of queries, significantly outperforming Yahoo and Bing, with both providing only 81.2% each.

Interestingly, the Google’s heatmap in Figure 2, illustrates a more evenly distributed range of Foci compared to prompt results and expert suggestions. Also indicated in Table 3, the average number of Foci in Google’s suggestions exceeds those in other methods. This suggests that Google’s QR approach provides more varied results.

Given this indication that a large portion of questions suggested by search engine consist of interest shift suggestions, we explored excluding the ‘intent shift’ questions from the search engines, by selecting questions from the GARD dataset where both input and output were determined to have the same Focus. After filtering, only 11%, 9%, and 8% of questions remained in the Google, Bing, and Yahoo datasets, respectively - confirming our suspicions. In comparison, the prompting methods had higher Foci alignment as the GARD expert suggested RQs: 33% (ent-def prompt), 35% (ref-def prompt), and 34% (CoT prompt). We then ran our evaluation method upon this filtered dataset with results shown in Table 2, row with GARD**. Compared to the full GARD dataset, we see that overall search engine accuracy significantly increased, yet still lower than our three prompting results. We conclude that filtering on Focus would not be a desirable dataset augmentation and that further investigation is required on how to separate interest shift task vs. QR with search engine dataset.

Also, while this diversity is advantageous for users seeking to shift the Focus of their inquiries, it may be undesirable for those who simply wish to rephrase their existing questions. In contrast, a mechanism to filter or categorize these options may be preferred. Such a system would enhance the user experience by streamlining the process of navigating through the multitude of suggestions, thereby catering more effectively to the specific needs of users, whether they seek focus diversity or question refinement.

5. Conclusion

In our study, we have made significant strides in the field of CHQ reformulation by conducting a comprehensive comparative analysis of three distinct question generation methodologies: ref-def, ent-def and CoT prompt. Our CoT prompting approach, which integrates Focus and Type specificity, represents a novel method tailored for CHQ reformulation. Furthermore, we have identified and recommended specific datasets that are instrumental for ongoing research in this domain. These datasets are poised to aid other researchers in conducting similar studies, thereby driving continuous innovation and exploration in the field. Our contributions lay the groundwork for future advancements in CHQ reformulation, setting a new benchmark for research.

6. Bibliographical References

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*.

Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the Web Conference 2021*, WWW '21, page 743–755, New York, NY, USA. Association for Computing Machinery.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2019. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association : JAMIA*.

GARD. About GARD | genetic and rare diseases information center (GARD) – an NCATS program. .

Internet-in-a-Box Team. 2016. Internet-in-a-box (iiab). https://github.com/iiab/iiab.

jvmvik, hartator, ilyazub, Dmitiry Zub, Lőrinc Pap, Kenneth Reitz, Lenny Fishler, Alexej, gbcfxs, Justin O'Hara, Manoj Nathwani, ajsierra117, and elizost. 2024. *serpapi/google-search-results-python*.

Jooyeon Lee and Luan Huy Pham. 2022. Recognizing question entailment in consumer health using a query formulation approach. In *Proceedings of The First Workshop on Context-aware NLP in eHealth (WNLPe-Health 2022) co-located with The nineteenth International Conference on Natural Language Processing (ICON-2022), Delhi, India, December 15-18, 2022*, volume 3416 of *CEUR Workshop Proceedings*, pages 56–69. CEUR-WS.org.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly

Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Laria Reynolds, Jason Wei, Daphne Ippolito, Noah Fiedel, Emily Reif, Andy Coenen, Ann Yuan, Adam Roberts, and Colin Raffel. 2021. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014. Decomposing consumer health questions. In *Proceedings of BioNLP 2014*, pages 29–37, Baltimore, Maryland. Association for Computational Linguistics.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023a. SESCORE2: Learning text generation evaluation via synthesizing realistic mistakes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5166–5183, Toronto, Canada. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023b. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A. Appendix

A.1. Data Source Links

This section specifies the precise location of the data source.

A.1.1. GARD Questions

The dataset was obtained from the following website: https://rarediseases.info.nih.gov/search?keyword=the%20following%20information%20may%20help%20address%20your%20question&page=1&filters=contentType%3Dgardcase

However, access to the expert suggestions is no longer available on the current version of the site. Nevertheless, an older version of the website, which includes these expert suggestions, can be accessed via this link: https://iiab.me/modules/en-nih_rarediseases/diseases/categories/index.html. In this version, the expert responses can be found in the "GARD Answer" section.

For instance, to find information on cysticercosis, one can visit https://iiab.me/modules/en-nih_rarediseases/diseases/8194/cysticercosis/index.html. In this section, clicking on the "see answer" link under "GARD Answer" redirects to a different page (https://iiab.me/modules/en-nih_rarediseases/diseases/8194/cysticercosis/cases/26056/index.html). This page lists the CHQ, suggested questions, their answers with references, and the expert suggested questions are introduced with the phrase "The following information may help to address your question:".

A.1.2. RQE Dataset

The RQE datasets, as referenced in Ben Abacha and Demner-Fushman (2016), are made publicly accessible at the following URL: https://github.com/abachaa/RQE_Data_AMIA2016

A.1.3. Search Engine API

The dataset for the search engine was not obtained from a repository; instead, we directly scraped search results using SerpAPI, as detailed in

(jvmvik et al., 2024). The relevant information and resources for this API can be found at the following URL: https://github.com/serpapi/google-search-results-python?tab=readme-ov-file.

A.2. Examples

In this section, we provide samples of each dataset and prompt results, which are listed in Table 4, Table 5, Table 6.

CHQ	How is fibrosing mediastinitis treated? How can I learn more about research?
Focus	['Fibrosing Mediastinitis']
Type	['Treatment', 'Research']
User Knowledge Level	MEDIUM
CoT Prompt Result	['What is the treatment for Fibrosing Mediastinitis?', 'Where can I find recent research on Fibrosing Mediastinitis?']
GARD Expert Suggestion	How might fibrosing mediastinitis be treated? Are there any research studies or clinical trials enrolling people with fibrosing mediastinitis?

Table 4: CoT Prompt Good Result.

CHQ	My son presented with a sudden onset of partial complex seizures and a friend provided us with information about neurocysticercosis. My son has not yet been tested, but I would like to know if this disease has an affect on a person's appetite or weight?
Focus	['Neurocysticercosis']
Type	['Symptoms', 'Effects']
User Knowledge Level	MEDIUM
CoT Prompt Result	['What is neurocysticercosis?', 'What are the symptoms of neurocysticercosis?', "Does neurocysticercosis affect a person's appetite?", "Does neurocysticercosis affect a person's weight?"]
GARD Expert Suggestion	What is cysticercosis? What is neurocysticercosis? Can neurocystocercosis affect a person's appetite or weight? With whom should I speak if I suspect neurocysticercosis may be affecting someone I know?

Table 5: CoT Prompt Error Result. Expert suggested a related question about cysticercosis, but this Focus is not shown in CHQ nor prompt output. Cysticercosis and neurocysticercosis are related but distinct conditions, both caused by the same parasitic infection but affecting different parts of the body. Thus, it is useful information to users who do not have a HIGH knowledge level about the disease. This is a question the user may ask without an "interest shift" intention.

CHQ	I have a Robertsonian translocation 13;14. My husband and I just had our trisomy 13 baby boy pass at 24 weeks gestation. My question is, does the type of trisomy 13 a baby has correlate with how long it will live in utero? For example, does full trisomy 13 have a better chance of survival than translocation trisomy 13? Or do they all have an equal prognosis?
Focus	['Robertsonian translocation 13;14', 'trisomy 13', 'survival in utero', 'prognosis']
Type	['cause', 'prognosis']
User Knowledge Level	MEDIUM
CoT Prompt Result	['Does the type of trisomy 13 a baby has correlate with how long it will live in utero?', 'Does full trisomy 13 have a better chance of survival than translocation trisomy 13?', 'Do all types of trisomy 13 have an equal prognosis?']
GARD Expert Suggestion	Does the prognosis differ depending on the type of trisomy 13 a fetus has? How can I find a genetics professional in my area?

Table 6: CoT Prompt Bad Result. Prognosis belongs to Type, not Focus. Survival in utero is also not a disease name. This is a case of multiple and incorrect Foci, which led to incorrect final results.