

# MedREQAL: Examining Medical Knowledge Recall of Large Language Models via Question Answering

Juraj Vladika, Phillip Schneider, Florian Matthes

Technical University of Munich

TUM School of CIT

Garching, Germany

{juraj.vladika, phillip.schneider, matthes}@tum.de

## Abstract

In recent years, Large Language Models (LLMs) have demonstrated an impressive ability to encode knowledge during pre-training on large text corpora. They can leverage this knowledge for downstream tasks like question answering (QA), even in complex areas involving health topics. Considering their high potential for facilitating clinical work in the future, understanding the quality of encoded medical knowledge and its recall in LLMs is an important step forward. In this study, we examine the capability of LLMs to exhibit medical knowledge recall by constructing a novel dataset derived from systematic reviews – studies synthesizing evidence-based answers for specific medical questions. Through experiments on the new MedREQAL dataset, comprising question-answer pairs extracted from rigorous systematic reviews, we assess six LLMs, such as GPT and Mixtral, analyzing their classification and generation performance. Our experimental insights into LLM performance on the novel biomedical QA dataset reveal the still challenging nature of this task.

## 1 Introduction

The field of natural language processing (NLP) has been transformed with the advent of pre-trained large language models (LLMs). During their process of pre-training to predict the next token on massive amounts of text data, these models learn and construct an internalized representation of world knowledge (Zhang et al., 2023). A popular domain of application of LLMs is healthcare, where they have the potential to democratize medical knowledge and facilitate access to healthcare, but also introduce risks of misinformation and lack of transparency (Clusmann et al., 2023). Recent work has hinted at the fact that LLMs encode clinical knowledge rather well (Singhal et al., 2023).

In the medical world, *systematic reviews* are overview studies that synthesize the best available

studies on a clearly defined medical research question. In them, the studies on a topic are analyzed, critically appraised, and their interpretations summarized by experts into a refined evidence-based conclusion (Pollock and Berge, 2018). Considering they synthesize the best available knowledge for a medical question into a concise answer, we see systematic reviews as a very well-suited proxy for testing the level of *knowledge recall* in LLMs.

When prompted in a zero-shot setting, instruction-tuned LLMs also tend to generate a conclusion addressing the question based on their recalled knowledge from diverse sources. To examine how well the LLMs perform medical knowledge recall, we constructed a question-answering (QA) dataset originating from systematic reviews and evaluated the performance of six LLMs, probing their medical knowledge recall. Based on experiments conducted on the newly constructed MedREQAL dataset, we analyze both the classification and generation performance, discussing insights about individual capabilities and limitations.

Our contributions include: (1) a novel dataset of biomedical question-answer pairs originating from rigorous systematic reviews, (2) experiments testing the zero-shot medical knowledge recall of six LLMs, and (3) deeper analysis of the dataset and experimental results. To ensure reproducibility, we provide the dataset and code in a public repository.<sup>1</sup>

## 2 Related Work

Medicine is a common domain of application for NLP tasks (Thirunavukarasu et al., 2023). It is marked by the highly complex biomedical language and terminology. Previous studies have highlighted the potential of LLMs in medical knowledge recall and exposed key gaps indicating the importance of further method development for creating safe

<sup>1</sup><https://github.com/jvladika/MedREQAL>

and effective LLMs for health-related applications (Liévin et al., 2022; Singhal et al., 2023).

Biomedical question answering (BQA) can be split into four main categories: scientific, clinical, consumer, and examination (Jin et al., 2022). Our dataset best fits into the scientific category, where questions and answers originate directly from biomedical research publications, and the most well-known datasets are PubMedQA (Jin et al., 2019) and BioASQ (Krithara et al., 2023). While the usual QA setting works with provided documents, recent efforts have started advancing the open-domain QA setting, where evidence first has to be discovered in order to answer the question (Jin et al., 2021; Vladika and Matthes, 2024a,b; Tian et al., 2024).

Also related is the task of scientific fact-checking, focusing on assessing the veracity of claims based on relevant scientific evidence (Wadden et al., 2020; Vladika and Matthes, 2023). The most similar dataset to MedREQAL is HealthFC (Vladika et al., 2024), which also uses systematic reviews as its evidence source for answers but focuses on popular health inquiries and retells the reviews in lay language. Systematic reviews are most commonly associated in NLP with the task of their automatic construction (Marshall and Wallace, 2019; van Dijk et al., 2023), which is a highly complex task eliciting advanced biomedical knowledge and refined reasoning skills. To the best of our knowledge, we introduce the first BQA dataset directly generated from systematic reviews, all adhering rigorously to a predefined research protocol.

### 3 Dataset

**Data Source** Our data originates from systematic reviews conducted by the Cochrane Collaboration. Systematic reviews synthesize the best available evidence on a clearly defined medical research question and then provide a concise evidence-based conclusion (Pollock and Berge, 2018). The Cochrane Collaboration is a global charitable organization formed with the aim of improving evidence-based healthcare decision-making through systematic reviews of the effects of healthcare interventions (Henderson et al., 2010). Cochrane Reviews are done by 30,000 volunteer medical experts, who follow a strict methodology and a highly structured format (with same seven sections in every abstract) for easier publishing in the Cochrane database (Cumpston and Chandler, 2022). We deem that

---

**Objective:** To assess the effects of listening to music on sleep in adults with insomnia and to assess the influence of specific variables that may moderate the effect.

**Generated question:** Can listening to music improve sleep in adults with insomnia?

**Conclusion:** The findings of this review provide evidence that music may be effective for improving subjective sleep quality in adults with symptoms of insomnia. (...)

**Verdict:** **Supported** (Jespersen et al., 2022)

---

**Objective:** To assess the effects of alpha-lipoic acid as a disease-modifying agent in people with diabetic peripheral neuropathy.

**Generated question:** Does alpha-lipoic acid have a disease-modifying effect on diabetic peripheral neuropathy?

**Conclusion:** Our analysis suggests that ALA probably has little or no effect on neuropathy symptoms or adverse events at six months, and may have little or no effect on impairment at six months. (...)

**Verdict:** **Refuted** (Baicus et al., 2024)

---

Table 1: Example of two instances from the MEDREQAL dataset. The question is generated from the original objective. The original conclusion is used for the answer generation task and the verdict is used for the classification task.

the focus on important healthcare interventions, wide domain coverage, standardized structure, and rigorous process followed by medical experts to construct these reviews are all factors that made them a highly suitable source for the construction of a novel biomedical question-answering dataset.

**Dataset Construction** To construct the dataset, we first scraped the abstracts of all the available Cochrane systematic reviews in the database PubMed, published from 2018 to 2023,<sup>2</sup> using the Python library Beautiful Soup. The final QA dataset consists of (1) questions, (2) labels, and (3) long answers. Questions were generated by leveraging the *Objectives* section present in every review. The objective text (usually one declarative sentence) was given to an LLM and instructed to form a question from it using GPT-3.5 (Turbo-0125) (see Table 6 in Appendix for all prompt details). Long answers are original words taken directly from the *Authors' conclusions* section present in each review. On top of these answers, we also generated a discreet label for each of these conclusions, which was one of the three: SUPPORTED, REFUTED, NOT ENOUGH INFORMATION. The motivation for this

---

<sup>2</sup><https://pubmed.ncbi.nlm.nih.gov/?term=%22Cochrane+Database+sys+rev%22%5BJournal%5D&filter=years.2018-2023>

is two-fold: (1) better alignment with existing comparable biomedical fact-checking and QA datasets that have these three labels, and (2) this allows for an easier evaluation of experiments with classification metrics like F1. The generation of these labels was also done with GPT-3.5 (see prompt in Table 6 in Appendix). Ultimately, the generated questions were sorted into various health areas using a classification prompt featuring 13 options, with a fourteenth category created afterward to consolidate questions concerning the renal and urinary systems. The classification process was performed using the same GPT-3.5 model with the temperature parameter set to 0. Two authors evaluated the generated questions and labels by randomly selecting 100 examples ( $\sim 4\%$  of all questions) and found 94% of questions and 92% of labels to be correct, which we deem satisfying. The most common error in question generation was generating a question about the study itself (e.g., "*Did the study analyze interventions to (...)?*"), while the most common label error was mislabeling a refuting verdict with a low level of evidence as "not enough information".

**Dataset Description** The constructed dataset comprises a total of 2,786 generated questions, each averaging 16.6 words in length with a std. dev. of 5.4. The distribution of labels shows that a significant portion of the medical questions (2057) lack sufficient information for a conclusive verdict, while 543 questions are supported by evidence, and 186 are refuted based on available data. While the proportion of NEI answers seems high (74%) as opposed to supported & refuted (26%), this is the result of strict guidelines followed by Cochrane reviewers for giving a conclusive verdict. A similar distribution was found by Howick et al. (2022), where 2,428 Cochrane reviews were manually analyzed, and the authors note "that only 26% of interventions had effects that were supported by moderate quality evidence according to GRADE<sup>3</sup>".

As visualized in Figure 1, the distribution of questions across the health categories in our dataset exhibits a relatively balanced distribution, with most topics having a comparable percentage, indicating an even spread and thus contributing to the dataset’s quality. Being the majority class, cognitive and mental health comprises the largest proportion at 23%. Respiratory, cardiovascular, and cancer-related inquiries each represent approx-

<sup>3</sup><https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/>

imately every tenth question, highlighting significant attention to these vital areas. Other important areas include musculoskeletal health, sexual and reproductive health, and immune system-related questions. The dataset also covers questions concerning digestive health, nutrition, sensory organs, and other health topics, reflecting the diversity of topics encountered in clinical studies.

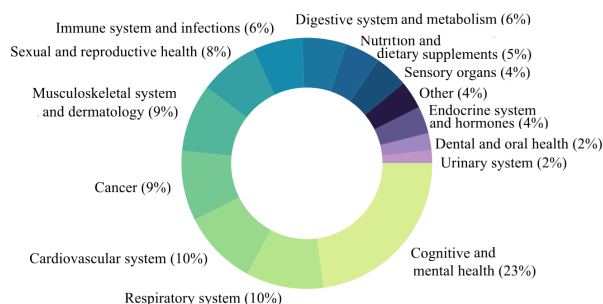


Figure 1: Percentage distribution of classified health areas in the MEDREQAL dataset.

## 4 Experimental Setup

We conduct our experiments using a variety of large language models. The models were instructed to answer the question (thus producing a long answer) and also predict a final label (one of the S/R/NEI). For evaluation, predicted labels were collected and evaluated using classification metrics accuracy and macro-averaged F1 score, while the long answers were compared to the ground truth authors’ conclusions with NLG evaluation metric ROUGE (Lin, 2004). Since this metric focuses on lexical overlaps, we also use the BERTScore metric, which captures semantic similarity (Zhang et al., 2020).

We chose three general-purpose LLMs and three biomedical LLMs. As a commercial state-of-the-art LLM, we include **GPT-4** (Turbo-0125) (OpenAI, 2023) in our comparison since it has demonstrated remarkable zero-shot performance on various NLP tasks. Consequently, it is often used as a benchmark for comparing LLMs. We also include two recent open-source models that achieved impressive performance, namely the **Mistral-7B** (Instruct-v2) (Jiang et al., 2023); and **Mixtral** (Jiang et al., 2024), based on a sparse mixture-of-experts architecture (Fedus et al., 2022).

We found three domain-specific biomedical models that fit our use case: (1) **PMC-LLaMa 13B** (Wu et al., 2023), which is an extension of LLaMa (Touvron et al., 2023) additionally pre-trained on biomedical research articles from the database

PubMed (Canese and Weis, 2013) and later fine-tuned on various datasets; (2) **MedAlpaca 7B** (Han et al., 2023), a biomedical extension of Alpaca (instruction-tuned version of LLaMa, Taori et al., 2023) fine-tuned on medical texts, encompassing resources such as medical flashcards, wikis, and dialogue datasets; and (3) **ChatDoctor (7B)** (Li et al., 2023), another extension of LLaMa fine-tuned on patient-physician conversation datasets.

Since the models were instruction-tuned with different templates, we used for each of them the base prompt reported by authors in technical reports. The prompt additionally instructed the models to predict one of the three classes and write an explanation for its answer (which we deem the "long answer"). All prompts can be found in Table 7. GPT-4 was prompted through the OpenAI API, and Mixtral through the FastChat API<sup>4</sup> platform, which replicates OpenAI’s chat completion API endpoint. The other four models were small enough to be loaded onto a single A100 GPU w/ 80 GB VRAM, and took around two computation hours each to generate all 2786 answers. We set the token limit to 256 and the temperature parameter to 0 for GPT-4 and Mixtral, maximizing deterministic generation by favoring high-probability words.

All the models were tested only in a zero-shot setting with just the question provided. Providing the full text of the abstract would make the experiments focus on *machine reading comprehension*, which is not the goal of our study. We are interested in medical knowledge recall, which is best evaluated with a zero-shot setup. Initially, we also experimented with few-shot learning, but it only slightly helped in learning the formatting of answers and did not affect F1 performance since answers are still dependent on internal knowledge.

## 5 Results and Discussion

The performance metrics and results are summarized in Table 2. The three general-purpose models performed better on average, both in the classification and the generation task. To our surprise, GPT-4 performed worse than expected and was beaten in classification by the smaller Mistral and MedAlpaca models, as well as Mixtral. In total, Mixtral exhibited the best performance.

The difference in final classification results is apparent in Table 3 – our dataset contains mostly NEI labels and the best-performing models tended

	Acc	F1	R-L	BSc
<b>Mistral</b>	50.8	32.5	20.3	84.5
<b>Mixtral</b>	<b>62.0</b>	<b>34.8</b>	<b>21.1</b>	85.6
<b>GPT-4</b>	41.3	28.3	20.5	85.1
<b>ChatDoctor</b>	22.1	16.0	16.4	<b>86.0</b>
<b>MedAlpaca</b>	55.2	33.4	19.3	85.4
<b>PMC-LLaMa</b>	32.4	27.8	15.4	84.3

Table 2: Performance results of six models, measured by accuracy (Acc) and F1-macro score for classification, Rouge-L (R-L), and BERTScore (BSc) for generation.

to predict this label frequently. On the other hand, GPT-4 was very prone to answering questions affirmatively and saying they are supported by the latest medical research. Conversely, Mixtral was more reluctant to give definitive affirmative or negative verdicts, but correctly predicted in most cases that there is not enough reliable evidence to answer the question. This is demonstrated in an example in Table 4. When predicting NEI, the models did not refuse to answer but rather gave some background to the question and hints of positive or negative results in the past while finishing by saying the conclusive evidence is missing.

	SUP.	REF.	NEI
<b>Mistral</b>	1277	30	1404
<b>Mixtral</b>	753	3	2030
<b>GPT-4</b>	1835	5	946
<b>ChatDoctor</b>	2046	200	103
<b>MedAlpaca</b>	1010	43	1733
<b>PMC-LLaMa</b>	1417	682	687
<b>Gold</b>	<b>543</b>	<b>186</b>	<b>2057</b>

Table 3: Sum of predicted final labels (supported, refuted, not enough information) by each model.

Among the three biomedical LLMs, ChatDoctor had by far the worst performance and resorted to answering almost every question affirmatively (we observed this even when accounting for different values of hyperparameters *temperature* or *top p*). Interestingly enough, it was also the model with the highest BERTScore value of its generated text, showing that its fine-tuning on patient-doctor conversations most likely led to more authentic phrasing and word choices when delivering medical information. On the opposite side of the spectrum, PMC-LLaMa was the model that had by far the most REFUTED predictions (which were basically nonexistent for GPT-4 and Mixtral).

<sup>4</sup>FastChat: <https://github.com/lm-sys/FastChat>



<b>Question</b>	Do interventions to improve appropriate polypharmacy result in clinically significant improvement?
<b>Conclusion</b>	It is unclear whether interventions to improve appropriate polypharmacy resulted in clinically significant improvement. (Cole et al., 2023) [label: <b>NOT ENOUGH INFO</b> ]
<b>GPT-4</b>	Interventions aimed at improving appropriate polypharmacy, which refers to the use of multiple medications in a manner that maximizes benefit and minimizes harm, have been shown to result in clinically significant improvements in various studies. These improvements include enhanced medication adherence, reduced adverse drug events, and better overall patient outcomes. [predicted: <b>SUPPORTED</b> ] ✕
<b>Mixtral</b>	While there is evidence suggesting that interventions can improve appropriate polypharmacy (the use of multiple medications for a patient), the impact on clinically significant improvement can vary depending on the specific patient population, interventions, and outcomes measured. Some studies have shown positive results, while others have not found significant improvements. Therefore, more research is needed to draw definitive conclusions. [predicted: <b>NOT ENOUGH INFO</b> ] ✓

Table 4: Example of a question and conclusion from our dataset with generated answers by Mixtral and GPT-4 in a zero-shot setting. On average, GPT-4 produced more agreeable answers supporting the claims from the questions, while Mixtral on average had more critical answers and reported on uncertainty.

The disparity between REFUTED and NEI predictions is a consequence of another common pattern we observed – the models struggle to differentiate between these two classes. Questions labeled with NEI will usually have a conclusion saying there is not enough high-quality evidence to definitively answer the hypothesis. On the other hand, the REFUTED questions will conclude that reliable studies show no effect of the healthcare intervention on the outcome, i.e., there is no difference between the tested intervention and a placebo drug or treatment. Since both of these classes contain "negative" phrasing, they were commonly mislabeled by models even when they were able to recall correct clinical evidence.

# mentions	Mis.	Mix.	GPT-4	CD	MA	PL
Cochrane review*	321	105	21	2	3	52
meta-analys*	759	663	390	8	614	477
	810	306	214	587	75	35

Table 5: No. of model responses mentioning the given term, showcasing its recall of relevant meta-reviews.

Another interesting finding is shown in Table 5. The models were prompted in a zero-shot setting with no reference to reviews, and yet in hundreds of answers, the models referred to reviews and meta-analyses related to the posed question, sometimes even referring to Cochrane itself. This was most evident in Mistral-7B, followed by Mixtral and MedAlpaca. This clearly demonstrates the encoded internal medical knowledge and the ability of models to refer to systematic reviews as the highest type of clinical evidence to answer given questions.

Example response of all six LLMs for the same question is shown in Table 9. Both Mistral and Mixtral referred to (different) systematic reviews

on the given topic, while PMC-LLaMa referred to a specific randomized control trial. GPT-4 and MedAlpaca mentioned that the answer is based on clinical research, and ChatDoctor gave a plain affirmative answer. Still, it is evident that models tend to cite studies that are sometimes rather old. This can also lead to incorrect predictions and quoting outdated knowledge. An example is in Table 8, where the systematic review from 2016 did not have enough information but the updated version of the same review from 2023 in our dataset did have enough new studies to give a positive verdict. How to update the outdated knowledge contained within LLMs is an ongoing challenge of *knowledge editing* (Yao et al., 2023; Cohen et al., 2024).

In the future, we see the dataset MedREQAL as a challenging testbed for medical knowledge recall in LLMs, but also envision its potential for testing tasks like multi-document summarization, evidence retrieval, or retrieval-augmented generation.

## 6 Conclusion

We constructed a new biomedical QA dataset for testing the medical knowledge recall of LLMs. The dataset originates from systematic reviews, synthesized evidence-based studies on research questions, a well-suited proxy for knowledge recall probing. We tested the performance of three general LLMs and three biomedical LLMs, showing that the scale of the model or its domain-tuning is not always tied to better performance. Deeper insights show that models have a moderately high level of recall and awareness of systematic reviews as quality evidence but still struggle with decisively concluding when there is not enough evidence to answer the question, leaving space for future improvements.

## Limitations

Our dataset was built semi-automatically and relied a lot on using automated generative methods for its construction, which could have led to certain incorrect labeling. While the questions are usually just a simple interrogative rewording of the original objective and the long answers are original authors' conclusions, labels were completely generated from the given question and conclusion text. Our manual analysis of a 4% subset of data showed that the performance is above 92% correct. Keeping in mind that even human annotation is not perfect and always has inconsistencies in labeling, we deemed this a satisfying performance.

Our comparative analysis has certain limitations. We focus solely on a zero-shot setting and direct medical knowledge recall probing, and we acknowledge that settings of machine reading comprehension and in-context learning are also worthy of exploring in the future. We do not benchmark all relevant biomedical LLMs; some, like Med-PaLM, were computationally too expensive for us to run. Our study also lacks human evaluation of generated model responses, which could have shed more light on the qualitative performance and user-friendliness of the answers.

## Ethical Considerations

Our dataset and experiments deal with the highly sensitive domain of healthcare and medical NLP. While we probe the models in a zero-shot setting to elicit their internal medical knowledge for our research purposes, this is not a recommended way of their usage by end users or patients. Some responses still contained hallucinations and misleading medical advice that should be taken with a grain of salt and always manually checked within reliable sources or consulted with medical professionals.

## Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions. This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 Software Campus 2.0 (TU München).

## References

Cristian Baicus, Adrian Purcarea, Erik von Elm, Caterina Delcea, and Florentina L Furtunescu. 2024.

Alpha-lipoic acid for diabetic peripheral neuropathy. *Cochrane Database Syst. Rev.*, 1(1):CD012967.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Judith A Cole, Daniela C Gonçalves-Bradley, Mubarak Alqahtani, Heather E Barry, Cathal Cadogan, Audrey Rankin, Susan M Patterson, Ngaire Kerse, Chris R Cardwell, Cristin Ryan, and Carmel Hughes. 2023. Interventions to improve the appropriate use of polypharmacy for older people. *Cochrane Database Syst. Rev.*, 10(10):CD008165.

Miranda Cumpston and J Chandler. 2022. Chapter ii: planning a cochrane review. *Cochrane handbook for systematic reviews of interventions*. *Cochrane*.

William Fedus, Jeff Dean, and Barret Zoph. 2022. [A review of sparse expert models in deep learning](#).

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

Lorna K Henderson, Jonathan C Craig, Narelle S Willis, David Tovey, and Angela C Webster. 2010. How to write a cochrane systematic review. *Nephrology*, 15(6):617–624.

Jeremy Howick, Despina Koletsi, John PA Ioannidis, Claire Madigan, Nikolaos Pandis, Martin Loeff, Harald Walach, Sebastian Sauer, Jos Kleijnen, Jadbinder Seehra, et al. 2022. Most healthcare interventions tested in cochrane reviews are not effective according to high quality evidence: a systematic review and meta-analysis. *Journal of clinical epidemiology*, 148:160–169.

Kira V Jespersen, Victor Pando-Naude, Julian Koenig, Poul Jennum, and Peter Vuust. 2022. Listening to music for insomnia in adults. *Cochrane Database Syst. Rev.*, 8(8):CD010459.

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

- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. **PubMedQA: A dataset for biomedical research question answering**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.
- Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. **Bioasqqa: A manually curated corpus for biomedical question answering**. *Scientific Data*, 10:170.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).
- Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.
- 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.
- Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8:1–10.
- OpenAI. 2023. **Gpt-4 technical report**.
- Alex Pollock and Eivind Berge. 2018. How to do a systematic review. *International Journal of Stroke*, 13(2):138–156.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. **Large language models encode clinical knowledge**. *Nature*, 620(7972):172–180.
- 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](https://github.com/tatsu-lab/stanford_alpaca).
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. **Llama: Open and efficient foundation language models**. *arXiv:2302.13971*.
- Sanne HB van Dijk, Marjolein GJ Brusse-Keizer, Charlotte C Bucsán, Job van der Palen, Carine JM Doggen, and Anke Lenferink. 2023. Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ open*, 13(7):e072254.
- Giovanni Vita, Beatrice Compri, Faith Matcham, Corrado Barbui, and Giovanni Ostuzzi. 2023. Antidepressants for the treatment of depression in people with cancer. *Cochrane Database Syst. Rev.*, 3(3):CD011006.
- Juraj Vladika and Florian Matthes. 2023. **Scientific fact-checking: A survey of resources and approaches**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.
- Juraj Vladika and Florian Matthes. 2024a. **Comparing knowledge sources for open-domain scientific claim verification**. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2103–2114, St. Julian’s, Malta. Association for Computational Linguistics.
- Juraj Vladika and Florian Matthes. 2024b. **Improving health question answering with reliable and time-aware evidence retrieval**.
- Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. **HealthFC: Verifying health claims with evidence-based medical fact-checking**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. **Fact or fiction: Verifying**

- scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. [Pmc-llama: Towards building open-source language models for medicine](#).
- Janet Yamada, Mariana Bueno, Lucia Santos, Sarah Haliburton, Marsha Campbell-Yeo, and Bonnie Stevens. 2023. [Sucrose analgesia for heel-lance procedures in neonates](#). *Cochrane Database of Systematic Reviews*, 2023(8).
- Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. [Editing large language models: Problems, methods, and opportunities](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. [How do large language models capture the ever-changing world knowledge? a review of recent advances](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.



## A Appendix

The Appendix provides supplementary material about this study, including the model prompts in full length (Tables 6 and 7) and example questions and model answers (Tables 8 and 9).

<b>Use Case</b>	<b>Prompt Content</b>
Classification	<p>SYSTEM: Your task is to classify an input with a medical question into one of several medical classes. If none of the classes fits, output miscellaneous. Output only one class from the following options:</p> <ul style="list-style-type: none"><li>cancer</li><li>cardiovascular system</li><li>dental and oral health</li><li>digestive system and metabolism</li><li>endocrine system and hormones</li><li>immune system and infections</li><li>cognitive and mental health</li><li>musculoskeletal system and dermatology</li><li>nutrition and dietary supplements</li><li>respiratory system</li><li>sensory organs</li><li>sexual and reproductive health</li></ul>
Question & Label generation	<p>SYSTEM: You're a helpful assistant. Your task is to help with generating questions and labels in the medical and clinical domain.</p> <p>AGENT You will be given an excerpt of an abstract of a clinical systematic review. Based on the given background, objectives, and author's conclusions, generate only ONE SINGLE question, answerable with yes/no/uncertain, that sums up the main medical objective that was investigated. Please keep the question short and general and use the "Objectives" section to construct the question. The question should be about a general medical hypothesis, not about this specific review. Afterwards, please also give a label for the author's conclusions. Label tries to answer the objective by looking at the conclusion. The label may be ONLY from one of the following three: (1) SUPPORTED; (2) REFUTED; (3) NOT ENOUGH INFORMATION. Do not try to make up a new label. Please only select the third label if not enough evidence was found to reach a verdict, not if certainty of the conclusion is low! Please aim to predict "SUPPORTED" or "REFUTED" even if certainty of these conclusions by authors is low or moderate. Please structure the output in two lines, as:</p> <p>QUESTION: (question) LABEL: (label)</p> <p>The documents begins now.</p>

Table 6: Overview of applied prompts for data generation and annotation.

<b>Model</b>	<b>Prompt Content</b>
<b>Mistral</b>	<p>&lt;s&gt;[INST] Based on your knowledge, please answer this clinical question only with SUPPORTED (if the question is supported by the clinical research) or REFUTED (if the hypothesis is refuted by the current clinical research) or NOT ENOUGH INFORMATION (if there is insufficient evidence for the question in current research). Please give your output in form of LABEL: (label) . Briefly explain your answer.</p> <p>QUESTION: {question}</p> <p>[/INST]</p>
<b>Mixtral, GPT-4</b>	<p>SYSTEM You are an AI assistant helping answer clinical and medical questions based on your best knowledge.</p> <p>AGENT Please answer this clinical question only with SUPPORTED (if the question is supported by the clinical research) or REFUTED (if the hypothesis is refuted by the current clinical research) or NOT ENOUGH INFORMATION (if there is insufficient evidence for the question in current research). Please give your output in form of LABEL: (label) . Briefly explain your answer.</p> <p>QUESTION: {question}</p>
<b>ChatDoctor, MedAlpaca, PMC-LLaMa</b>	<p>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.</p> <p>### Instruction: Based on your best current knowledge, please answer the following medical question. If you think there is not enough evidence to answer, then say so. Please answer the question with "SUPPORTED" or "REFUTED" or "NOT ENOUGH INFORMATION". Briefly explain your answer.</p> <p>### Input: {question}</p> <p>### Response:</p>

Table 7: Prompts used for each of the LLMs in experiments.

<b>Question</b>	Does sucrose effectively relieve pain from heel lance in neonates?
<b>Mixtral</b>	While some studies suggest that sucrose may help to reduce procedural pain in neonates, including heel lance procedures, the evidence is not conclusive. The <b>Cochrane review on this topic from 2016</b> states that "there is considerable uncertainty about the effects of sucrose on procedural pain in newborn infants." Therefore, more research is needed to definitively answer this question. [predicted label: <b>NOT ENOUGH INFORMATION</b> ]
<b>Latest Review (in the dataset)</b>	We reviewed 55 studies that included a total of 6273 babies. Twenty-nine studies had full-term babies only, 22 had preterm babies only, and four had both full-term and preterm babies. (...) Main results – Sucrose compared to the control treatment <b>probably reduces pain from single heel lances</b> in babies. (...) (Yamada et al., 2023) [original label: <b>SUPPORTED</b> ]

Table 8: Example of a question from our dataset with the original study conclusion and the generated answer by Mixtral. In this case, the model recalled a systematic review on the correct topic, but which is outdated and deprecated. This shows the challenging nature of updating and editing the knowledge encoded within LLMs.

<b>Question</b>	Do antidepressants show efficacy in treating depressive symptoms in adults with cancer?
<b>Conclusion</b>	Despite the impact of depression on people with cancer, the available studies were few and of low quality. This review found a potential beneficial effect of antidepressants against placebo in depressed participants with cancer. (Vita et al., 2023) [label: <b>Supported</b> ]
<b>GPT-4</b>	Antidepressants have been studied for their efficacy in treating depressive symptoms in adults with cancer. <b>Clinical research, including randomized controlled trials and systematic reviews</b> , has generally supported the use of antidepressants for this purpose. These studies have found that antidepressants can be effective in reducing depressive symptoms in cancer patients, improving their quality of life. However, the effectiveness can vary depending on the type of antidepressant, the type of cancer, and the individual patient's response to the medication.
<b>Mixtral</b>	There is evidence to support the use of antidepressants in treating depressive symptoms in adults with cancer. <b>A 2015 meta-analysis</b> of 52 randomized controlled trials found that antidepressants were more effective than placebo in reducing depressive symptoms in adults with cancer (Kissane et al., 2015). However, it is important to note that the efficacy may vary depending on the type of antidepressant and individual patient factors. Therefore, a careful evaluation and monitoring by healthcare professionals is necessary when prescribing antidepressants for adults with cancer.
<b>Mistral</b>	Numerous studies have demonstrated that antidepressants are effective in reducing depressive symptoms in adults with cancer. <b>A meta-analysis of 28 randomized controlled trials</b> involving over 3,000 patients found that antidepressants were significantly more effective than placebo in treating depressive symptoms in individuals with cancer (Kanji et al., 2011). Additionally, <b>a systematic review and meta-analysis of 35 studies</b> found that antidepressants were effective in reducing depressive symptoms in individuals with cancer, and that selective serotonin reuptake inhibitors (SSRIs) were more effective than tricyclic antidepressants (TCAs) (Cassidy et al., 2015).
<b>MedAlpaca</b>	Based on the <b>available research</b> , antidepressants appear to be effective in treating depressive symptoms in adults with cancer. Multiple studies have found that antidepressants can reduce depressive symptoms and improve quality of life in this population.
<b>PMC-LLaMa</b>	The use of antidepressants to treat depressive symptoms in adults with cancer is supported by <b>a single randomized controlled trial (RCT) and a number of observational studies</b> . <b>The RCT, which was conducted in the United Kingdom</b> , included 152 patients with advanced cancer who were receiving palliative care. Patients were randomized to receive up to 3 months of treatment with either amitriptyline, a tricyclic antidepressant (TCA), or a nonsteroidal anti-inflammatory drug (NSAID). The study found that patients who received amitriptyline reported fewer depressive symptoms and better quality of life at the end of the study period than those who received the NSAID. (...)
<b>ChatDoctor</b>	Yes, antidepressants have been shown to be effective in treating depressive symptoms in adults with cancer.

Table 9: Example of a question and conclusion from our dataset with generated answers by all six benchmarked LLMs. **Bolded** text includes references to literature and studies that were recalled by the models. Mixtral and Mistral named concrete systematic reviews (high-level evidence), PMC-LLaMa referred to a single randomized controlled trial (moderate-level evidence), while GPT-4 and MedAlpaca mentioned the research work in general. All six LLMs correctly predicted the **Supported** label.