Large Language Models for Binary Health-Related Question Answering: A Zero- and Few-Shot Evaluation

Marcos Fernández-Pichel, David E. Losada, and Juan C. Pichel

Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela, Santiago de Compostela, Spain
{marcosfernandez.pichel,david.losada,juancarlos.pichel}@usc.es

Abstract. In this research, we investigate the effectiveness of Large Language Models (LLMs) in answering health-related questions. The rapid growth and adoption of LLMs, such as ChatGPT, have raised concerns about their accuracy and robustness in critical domains such as Health Care and Medicine. We conduct a comprehensive study comparing multiple LLMs, including recent models like GPT-4 or Llama2, on a range of binary health-related questions. Our evaluation considers various context and prompt conditions, with the objective of determining the impact of these factors on the quality of the responses. Additionally, we explore the effect of in-context examples in the performance of top models. To further validate the obtained results, we also conduct contamination experiments that estimate the possibility that the models have ingested the benchmarks during their massive training process. Finally, we also analyse the main classes of errors made by these models when prompted with health questions. Our findings contribute to understanding the capabilities and limitations of LLMs for health information seeking.

Keywords: Binary Question Answering, Health, Large Language Models

1 Introduction

The emergence of Large Language Models (LLMs) has induced significant improvements in performance on various Natural Language Processing (NLP) downstream tasks [28,29]. The appearance of BERT [7], GPT-2 [30], and GPT-3 [3], among others, has accelerated the development of LLMs. With the increasing reliance of users on online medical information [10], the reliability of these models to provide correct responses to health-related information needs must be put under scrutiny. The potential consequences of incorrect health-related information can result in personal harm [36,26]. Hence, the evaluation of the robustness of these models in this critical domain is of utmost importance. In addition, it must be taken into account that the performance of LLMs is highly dependent on the prompt and context provided by the questioner [3,13,20].

In this paper, we present a systematic evaluation of LLMs, exploring their potential to correctly answer health-related questions. To that end, we compare multiple LLMs and examine their performance on a range of binary health questions extracted from standardised Information Retrieval (IR) collections. Our evaluation considers a wide range of context and prompt conditions and we discuss the potential challenges and implications of using LLMs for health information needs. Our ultimate goal here is not to attain the highest possible performance but, rather, to gain insights into these AIs' responses given an assorted set of input conditions. Through the conducted experiments, we try to answer the following research questions:

- To what extent do LLMs provide correct answers to binary health-related questions? How different models perform for this task?
- To what extent does the provided context and demonstrations influence the models' answers?
- Are these models really responding to *unseen* questions? Is their effectiveness conditioned by some form of data contamination?
- What kinds of mistakes do these LLMs tend to make?

2 Related work

Current LLMs have great potential for addressing health-related and medical information needs. However, their reliability for such critical task remains largely unknown, as most efforts have focused on general domain tasks. For instance, Jiang et al. [13] tried to optimise knowledge discovery in LLMs by generating high quality prompts (manual or automatic) and by exploiting ensemble methods. Liu et al. [19] focused their efforts on another critical aspect, the optimal configuration of in-context examples to enhance GPT-3's few shot capabilities. They found that this is specially crucial in Text Generation tasks. Other recent studies [17,20] performed systematic reviews of different models, prompts, metrics and tasks. The appearance of ChatGPT has also stimulated targeted studies to gauge the model's knowledge and utility for a number of tasks [1,2,34].

Several fine-tuned models have been specifically built for the medical domain [37,16]. However, existing evaluations of these models have been restricted to a single specialised topic, like genetics or radiology, and there is a lack of comparisons across multiple models [8,12,14,31,35,4]. Evaluating the accuracy of general-purpose LLMs for multiple types of medical queries has received little attention. A recent study [38] analysed the impact of prompts in health information seeking. However, the study was confined to a single LLM (ChatGPT) and the main goal was to evaluate prompts that incorporate supporting and contrary evidence obtained from a search engine.

Our contribution consists of a systematic evaluation of LLMs' capabilities to correctly answer health questions. We will only focus on these models' internal knowledge and we will assess their performance when prompted with different inputs and in-context examples. In contrast to previous studies, we will systematically compare several models and we will focus on general health questions,

without restricting the analysis to specialised topics. Moreover, we include the recently released Llama2 model in our comparison[1]. We also report our endeavours to estimate if the models really generalise well or, by the contrary, they have seen these benchmarks during its pre-training process (i.e, we study the so-called data contamination [23,11,22,33]). Finally, we also provide an initial exploration of the most common mistakes (e.g., about a medical treatment).

3 Experimental design

3.1 Models

We considered language models of different nature (close and open source) and architecture. restrict the study to general-purpose models that are freely available to end-users. Thus, fine-tuned models like ChatDoctor [37] or BioBERT [16] are out of the scope of this research (as standard web users do not have the knowledge to install and invoke these tools). For a rigorous experimentation, we considered recent LLMs of different nature (including both proprietary and open source models):

- **GPT-3** is a series of models with a decoder-only structure with 175 billion parameters. Its training corpus is extensive, encompassing a variety of web sources and the entire Wikipedia, with information up to June 2021. These models were built on top of InstructGPT [25] and were fine-tuned with human feedback using reinforcement learning (RLHF). For these experiments we considered two different versions: text-davinci-002 (d002) and text-davinci-003 (d003).
- **ChatGPT** is similar to InstructGPT, but it meant a paradigm shift towards more conversational interaction [9]. Its training data goes up to September 2021. For these experiments we used gpt-3.5-turbo version (a snapshot from June 2023).
- **GPT-4**. It is a bot also designed for conversational purposes. It serves as a cutting-edge advancement in this field and surpasses ChatGPT's performance in various tasks that require human-like intelligence, such as passing an exam [24]. Its training data also goes up to September 2021. For these experiments we used gpt-4-8k version (a snapshot from June 2023).
- **Flan T5** is a sequence-to-sequence model developed by Google. It was fine-tuned on instruction-based datasets, which include a wide range of information collected up until 2022 [21]. For these experiments we used the flan-t5-xl version.
- **Llama2** is the most recent model developed by Meta AI. It was trained with over 1 million human annotations on conversational data. Its training data goes up to September 2022, but its fine-tuning also includes data up to July 2023. We used the llama-13b-chat version for these experiments.

[1] https://ai.meta.com/llama/

```
<topic>
  <number>1234</number>
  <query>dexamethasone croup</query>
  <description>Is dexamethasone a good treatment for croup?</description>
  <narrative>Croup is an infection of the upper airway and causes swelling,
    which obstructs breathing and leads to a barking cough. As one kind of
    corticosteroids, dexamethasone can weaken the immune response and
    therefore mitigate symptoms such as swelling. A very useful document
    would discuss the effectiveness of dexamethasone for croup, i.e. a very
    useful document specifically addresses or answers the search topic's
    question. A useful document would provide information that would help
    a user make a decision about treating croup with dexamethasone, and
    may discuss either separately or jointly: croup, recommended treatments
    for croup, the pros and cons of dexamethasone, etc.</narrative>
  <disclaimer>We do not claim to be providing medical advice, and medical
    decisions should never be made based on the stance we have chosen.
    Consult a medical doctor for professional advice.</disclaimer>
  <stance>helpful</stance>
  <evidence>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5804741/</evidence>
</topic>
```

Fig. 1: A topic from the TREC 2021 Health Misinformation Track (Topic 101).

The first three models were tested through OpenAI's official Python API[2], while the two latter ones were tested through their Hugging Face implementation.

We are aware that there is a growing concern in the scientific community about evaluations performed on proprietary models. Sometimes, it is difficult to guarantee reproducibility of a model that suffers constant updates and where the technical intricacies (e.g. architectural design or training data) are unknown. However, we believe that the adoption of these conversational AI systems by the general population makes it necessary to put them under scrutiny. Regardless of the open or close nature of each platform, the reality is that systems such as ChatGPT are currently used by millons of users worldwide. Therefore, in this paper we made an effort for reproducibility by providing the code[3], the outputs of each round, and all the dates of the execution of the experiments[4].

3.2 Datasets

To conduct the evaluation, we used three different collections from the TREC Health Misinformation (HM) Track [5,6]. The collections consist of health-related topics, in the form of questions (e.g., "*Can wearing masks prevent COVID-19?*"), and web documents. For our experimentation, we only used the questions and their binary ground truth answers (yes/no), which represent the best understanding of current medical practice (gathered by the task organisers when creating the collection). Figure 1 shows an example of a topic. The 2020 questions are all related to COVID-19, while the 2021 and 2022 questions encompass general health information needs. The 2020 questions were released in mid 2020 and, thus, we cannot discard that the LLMs have seen this benchmark within

[2] https://openai.com/blog/openai-api
[3] https://anonymous.4open.science/r/llm-binary-health-qa-8743
[4] These experiments were run between September and December 2023.

their training data. The 2021 questions were released in mid July and, thus, they might have been available for all models, except for GPT-3, whose training ended earlier. The 2022 questions, instead, could only have been seen by Flan T5 or Llama2. This therefore conforms an assorted set of health questions, with varying levels of difficulty for the models (depending on their exposure to this type of data and the level of specificity of the information needs). In any case, section 6 further analyses the possibility of data contamination.

3.3 Contexts

As a core part of this research, we want to determine the effectiveness of these models for health information seeking. First of all, we try to see how well they would respond to non-expert end-users who give little or no context at all. This leads to the following input prompts:

- **no-context**: a prompt composed only of the medical question, i.e. "*Can Vitamin D cure COVID-19?*".
- **non-expert**: The text "*I am a non-expert user searching for medical advice online*" plus the corresponding question. This prompt might be representative of a regular user searching for medical advice.

In a second series of experiments, we also test more sophisticated prompts and, additionally, evaluate the effect of in-context examples. These artifacts are unlikely employed by normal users but, still, they can help to further understand and exploit the models' internal knowledge. We tested the following prompt:

- **expert**: The text "*We are a committee of leading scientific experts and medical doctors reviewing the latest and highest quality of research from PubMED. For each question, we have chosen an answer, either 'yes' or 'no', based on our best understanding of current medical practice and literature.*" plus the corresponding medical question. This prompt was designed by Waterloo's team in their participation in the TREC 2022 HM track [27]. The rationale is to bias the LLM towards reputed contents associated to high quality sources.

More elaborate prompt engineering techniques, like Chain-of-Thought (CoT) could further enhance performance, but this was left as future work. The models' temperature was set to 0, with the intention of minimising randomness in their responses. To perform an automatic evaluation, we restricted the response of the models to a single "yes" or "no" token via the model's APIs.

4 Zero-shot evaluation

As can be seen in Table 1, text-davinci-003 and Llama2 are the best perfomers for TREC HM 2020 and 2021 collections. With the TREC HM 2022 collection, GPT-4 and ChatGPT outstand. There are also some differences in performance among the selected prompts. As expected, the most robust context seems to be

	TREC HM 2020						TREC HM 2021					
prompt	d-002	d-003	Chat GPT	GPT 4	Llama2	FT5	d-002	d-003	Chat GPT	GPT 4	Llama2	FT5
no-context	0.84	0.91	0.84	0.79	**0.92**	0.24	**0.72**	0.76	0.68	**0.68**	**0.76**	0.44
non-expert	0.78	**0.92**	0.84	**0.90**	0.86	0.31	0.40	0.62	0.56	0.66	0.70	0.54
expert	**0.86**	0.9	**0.86**	0.86	**0.92**	**0.79**	0.36	**0.80**	**0.70**	0.66	0.72	**0.64**
avg.	0.83	0.91	0.85	0.85	0.90	0.39	0.47	0.72	0.64	0.67	0.72	0.54
std. dev.	0.04	0.01	0.01	0.06	0.03	0.30	0.20	0.09	0.08	0.01	0.03	0.10

	TREC HM 2022					
prompt	d-002	d-003	Chat GPT	GPT 4	Llama2	FT5
no-context	**0.76**	**0.76**	0.76	0.86	0.74	0.56
non-expert	0.48	0.72	0.80	0.86	0.68	0.54
expert	0.68	0.72	**0.90**	**0.88**	**0.84**	**0.74**
avg.	0.63	0.73	0.82	0.87	0.75	0.61
std. dev.	0.14	0.02	0.07	0.01	0.08	0.11

Table 1: Zero-shot experiments, proportion of correct answers of each model-prompt combination for the three TREC datasets.

expert one. We hypothesise that this is due to the inclusion of keyphrases such as "*research from PubMed*" or "*medical practice and literature*", which bias the model towards reputable sources of knowledge.

Although models are relatively stable, they still exhibit some variations depending on the input. This is concerning, as a model's effectiveness can range from 90% of correct answers to $\approx$ 75% of correct responses. The overall levels of effectiveness are remarkable but, still, these inconsistencies are a cause of discomfort. Even adopting the most consistent prompt (*expert*) we observe concerning outcomes. For example, GPT-4 suffers from poor performance (66%) in the 2021 dataset.

The three datasets vary in their level of difficulty. The 2020 health questions (related to COVID-19) appear to be easier for the LLMs. A plausible explanation for this phenomenon could be that the models might have already been exposed to these health questions during their massive training. We will further explore this possibility in Section 6. Another explanation could be that the highly relevant and significant nature of COVID-19 as a topic might have motivated a specialised curation process for the relevant data.

We also employed McNemar's test to assess the significance of the differences between the top-performing models [15]. Between ChatGPT and GPT-4, we found no significant difference in 7 out of 9 comparisons (3 collections $\times$ 3 prompts). The pair ChatGPT vs Llama2 revealed no difference in 7 out of 9 comparisons and GPT4 vs Llama2 revealed no significant difference at all. The pairwise comparisons d-003 vs ChatGPT, d-003 vs Llama2 and d-003 vs GPT-4

revealed more cases of statistical significance but, still, more than a half of the compared instances yielded a no significance result.

5 Few-shot evaluation

To perform an analysis of the effect of demonstrations, we focused on the test questions from TREC HM 2022. Each question was prompted to the models prepended by one-to-three demonstrations extracted from TREC HM 2021. We randomly chose three pairs of *(medical question, correct answer)* from the 2021 dataset as in-context examples and explored the effect of including them[5]. Past research [17] has shown that evaluating a narrow range of in-context examples is a solid choice.

As can be seen in Table 2, the effect of the demonstrations strongly depends on the model. For instance, both versions of GPT-3 (davinci-002 and davinci-003) and FlanT5 are the models that benefit the most from the inclusion of these in-context examples. For these models, some in-context variants led to statistical significant benefits. On the other hand, the best performing models under the zero shot-setting do not seem to benefit from the inclusion of demonstrations. Regarding types of prompts, the *expert* variant is the one that benefits most from the inclusion of the few-shot examples. In terms of the number of examples, the results suggest that prompting with more than one does not boost performance.

In Figure 2, we plot the proportion of correct answers for Llama2 and text-davinci-002 models (expert prompting) with varying number of demonstrations. This evolution shows that the weakest model benefits the most, and more number of in-context examples does not always translate into better performance.

6 Data Contamination

LLMs have shown excellent performance in multiple NLPs but, in some cases, this might be attributed to the presence of golden truth data from the evaluated benchmarks within the LLM's training corpora. This is particularly concerning with proprietary LLMs that do not disclose information about their training data, as there is no direct way to consult the sources of the training data. A fair evaluation of these models needs to test their generalisation abilities beyond the training data. A system that directly *copies the answer* from an existing ground truth file should not be considered as intelligent. A really intelligent system is the one that learns about the world from the training corpora and, next, makes proper inferences to answer new questions. Making an analogy with education, a student who had access to the responses of the exam should fail while a student who studied all the relevant material and submitted correct answers should pass.

[5] We used these question-answer pairs (in this same order): (*Will wearing an ankle brace help heal achilles tendonitis?*, *No*), (*Does yoga improve the management of asthma?*, *Yes*), (*Is starving a fever effective?*, *No*)

prompt	d002				d003			
	0-shot	1-shot	2-shot	3-shot	0-shot	1-shot	2-shot	3-shot
no-context	0.76	0.7	**0.78**	**0.78**	0.76	**0.86**	**0.86**	**0.86**
non-expert	0.48	**0.64**	**0.74**	**0.76**	0.72	**0.82**	**0.82**	**0.82**
expert	0.68	**0.74***	**0.76***	**0.78***	0.72	**0.82***	**0.84***	**0.84***

prompt	FT5				ChatGPT			
	0-shot	1-shot	2-shot	3-shot	0-shot	1-shot	2-shot	3-shot
no-context	0.56	**0.66**	**0.64**	**0.7**	0.76	**0.82**	**0.88**	**0.84**
non-expert	0.54	**0.68***	**0.66***	**0.64**	0.8	0.8	**0.88**	**0.86**
expert	0.74	0.68*	0.72	0.72	0.9	0.84	0.88	0.88

prompt	Llama2				GPT-4			
	0-shot	1-shot	2-shot	3-shot	0-shot	1-shot	2-shot	3-shot
no-context	0.74	0.7*	**0.84**	**0.76**	0.86	0.84	0.86	0.86
non-expert	0.68	**0.72**	**0.74**	0.62*	0.86	0.86	**0.88**	**0.88**
expert	0.84	0.64*	**0.76**	0.6*	0.88	0.88	**0.92**	**0.9**

Table 2: Few-shot experiments, proportion of correct answers of each model-prompt combination with three shot samples. For each row, if few shot surpasses the 0-shot is marked in bold and the symbol * marks those cases where McNemar's test ($\alpha = .05$) finds a significant difference between both variants.

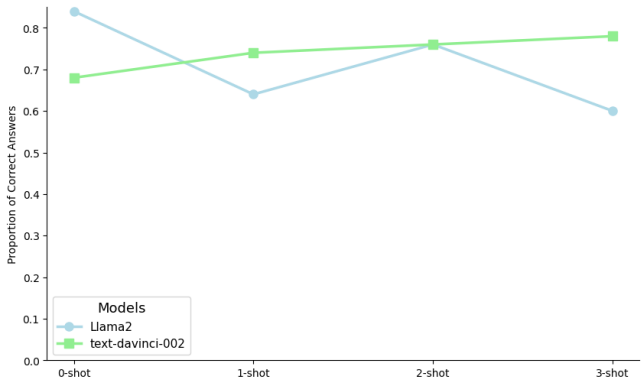


Fig. 2: Proportion of correct answers for Llama2 and text-davinci-002 models with expert prompting and different number of in-context examples.

Data contamination is an active area of research [23,11,22,33] that tries to estimate whether or not a NLP benchmark was ingested during the pre-training process. As part of our study, we have conducted data contamination estimation experiments to further validate the capacity of the LLMs to correctly answer medical and health questions.

Model	Version	TREC HM 2020			TREC HM 2021		
		Levenshtein	BLEURT	ROUGE	Levenshtein	BLEURT	ROUGE
ChatGPT	General	0.45	0.41	0.27	0.46	0.48	0.29
	Guided	0.30	0.39	0.10	0.38	0.47	0.23
GPT-4	General	0.45	0.42	0.26	0.45	0.46	0.29
	Guided	0.44	0.50*	0.24	0.42	0.51*	0.25
Llama2	General	0.42	0.37	0.23	0.44	0.43	0.28
	Guided	0.43*	0.35	0.24	0.45	0.40	0.27

Model	Version	TREC HM 2022		
		Levenshtein	BLEURT	ROUGE
ChatGPT	General	0.48	0.41	0.26
	Guided	0.49	0.42	0.27
GPT-4	General	0.47	0.41	0.24
	Guided	0.47	0.43	0.25
Llama2	General	0.41	0.32	0.14
	Guided	0.40	0.25	0.15

Table 3: Results for the data contamination experiments across different models and datasets. For each guided vs general comparison, the symbol * marks those cases where the guided completion surpassed the general completion and Wilcoxon test ($\alpha = .05$) found a significant difference between both variants.

We implemented Golchin et al. [11] heuristics, which were validated under controlled contamination experiments[6]. Essentially, the approach consists of i) prompting the model with a general instruction (with no information about the benchmark or specific identifiers of the task or split), ii) prompting the model with a guided instruction (which identifies the benchmark and task), and iii) comparing the responses against the golden truth text. If the model's output obtained from the guided prompt has higher similarity to the ground truth text compared to the model's output from the general prompt then this is evidence that the model might have seen the benchmark during its training. A statistical significance test that compares the two means of similarity can thus mark those cases where we can suspect that the model ingested the benchmark during its pre-training[7].

[6] The code of our data contamination experiments is available at: https://anonymous.4open.science/r/data-contamination-45F2

[7] Note that this is a high precision but low recall method. Meaning that if the method estimates that the model ingested the dataset, we can state with some certainty that it did. However, if the heuristic does not provide evidence, we cannot state with certainty that the model did not ingest it.

In our case, the general prompt was: "*Complete the narrative field based on the query, question and answer fields such that the narrative provides an explanation for the answer to the given question. Query:* {*query*}*, Question:* {*question*}*, Answer:* {*answer*}*, Narrative:*". Meanwhile, the guided prompt was: "*You are provided with the query, question and answer fields of a topic from the TREC* {*year*} *Health Misinformation topic set. Complete the narrative field of the topic as exactly appeared in the dataset. Only rely in the original form of the topic in the dataset to complete the narrative field. Query:* {*query*}*, Question:* {*question*}*, Answer:* {*answer*}*, Narrative:*". As can be observed, the latter prompt specifies the name of the dataset and forces the model to produce the completion in its original form.

These two completions produce a pair of texts that are compared against the narrative text from the actual benchmark. For computing the overlappings (general completion vs golden truth and guided completion vs golden truth), we adopted the same metrics as in the original data contamination study [11]: BLEURT to estimate lexical similarity [32] and ROUGE-L [18] to estimate semantic relevance. Additionally, we also computed the Levenshtein distance, which measures the number of character permutations needed to transform the completion into the original data. This analysis was conducted for each available topic (i.e., each TREC question produced two completions and we report the average similarity across all topics).

We have performed this data contamination study for the most recent models, as can be seen in Table 3. Our results show some evidence that GPT-4 might have ingested TREC HM 2020 and TREC HM 2021 datasets, since we found statistically significant improvements for the guided completion with respect to the general one (in terms of semantic similarity). Levenshtein metric also shows some evidence that Llama2 might have been trained with the TREC HM 2020 collection.

For ChatGPT we found no evidence that it has ingested any benchmark. Still, it performs similarly to GPT-4 and Llama2 in the TREC HM 2020 and 2021 collections (under the zero-shot setting) and there is no statistical difference between ChatGPT and these two models. This also seems to indicate that the good results by ChatGPT in the TREC HM 2020 collection are not due to contamination effects but, rather, to the peculiar characteristics of the COVID-19 topics.

Furthermore, no model seems to have seen the TREC HM 2022 collection and many models performed effectively for this dataset (see Table 1). These results speak well of the capabilities of the LLMs models to correctly transfer the knowledge acquired during training and produce accurate answers to health and medical questions.

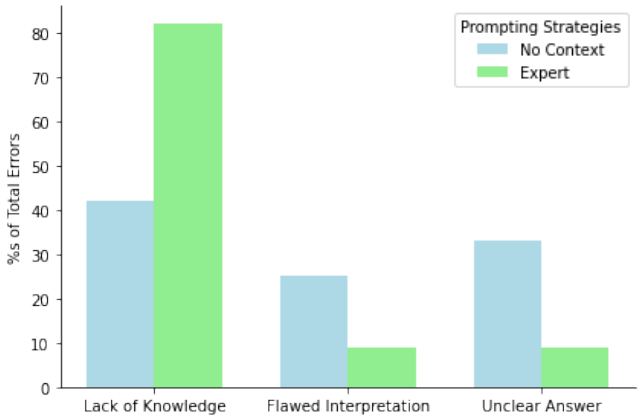


Fig. 3: Percentages of the different type of errors for the analysed prompting strategies.

7 Error Analysis

To further understand the LLMs' behaviour for this task, we inspected the questions where none of the models provided a correct answer[8]. This analysis was done for the best performing prompt (*expert*) and for the *no-context* prompt, which arguably reflects the type of input submitted by a regular user.

We found that for the TREC HM 2020 collection, models answered incorrectly 8% and 6% of the questions with the *no-context* and *expert* prompts, respectively. For the TREC HM 2021, they failed in providing the correct answer for 12% of the cases for both prompting strategies. Finally, in the TREC HM 2022, they answered incorrectly 4% of the queries for both strategies. These results confirmed that TREC HM 2021 is the most difficult collection, with a larger percentage of errors. It also seems that providing no context derives in a greater or similar percentage of failed queries than using the expert prompt. After manually inspecting the models' outputs, we could organise the errors found into a taxonomy that represents the most common health advice mistakes:

- **Lack of knowledge about current medical consensus**: Sometimes, models provide answers that go against the medical consensus. For instance, to the question "Can Hydroxychloroquine worsen COVID-19?", ChatGPT answered "no, there is no evidence that hydroxychloroquine worsens covid-19..." while medical evidence says otherwise[9].
- **Flawed interpretation of the question**: Here, LLMs misinterpret the question. For example, "Can bleach prevent COVID-19? No, bleach should

[8] We prompted again the models with these queries without restricting the form of token output. For the sake of simplicity, we used only the most recent models (as in the previous section).

[9] FDA cautions against use of hydroxychloroquine for COVID-19

not be ingested...". But the ground truth has the most obvious interpretation of this question (the use of bleach for surface disinfection can actually prevent COVID-19). A human would hardly interpret the question in this way.
- **Unclear answer**: We include in this category the responses in which models did not provide a blunt answer. These cannot be counted as correct responses but the LLM's output is arguably useful, for example, "sit-ups can be both beneficial and harmful, depending on your individual circumstances and the way you perform the exercise...".

Figure 3 plots the percentage of each type of errors for the different prompting conditions. In summary, lack of knowledge about medical consensus is the most common error for both prompting strategies. This is a concerning outcome as it is the most dangerous type of error. It also seems that providing expert context mitigates possible flawed interpretations of the questions and it prevents unclear answers, but it also comes with the cost of more mistakes about current medical consensus.

8 Concluding Remarks

We have conducted an exhaustive evaluation on the ability of a set of LLMs in providing the correct answer to health and medical questions. We have evaluated the models with three different collections of medical question-answer pairs and prompted them with different contexts, ranging from intricate prompts to simpler prompts (close to those possibly submitted by non-expert users).

Under the zero-shot setting, the most sophisticated and modern models performed similarly. However, there are still some causes of discomfort, e.g. in some cases the models provide less than 70% of correct answers. This is a low figure for a critical task such as health information seeking. We also found out that intricate prompting strategies enhanced the performance compared with simpler contexts. From our point of view, this is an obstacle for the adoption of these models for health question answering. Note that end users are unlikely to produce very sophisticated prompts. We also discovered that including few-shot examples enhanced the performance even with the most complex prompt. However, the effect of the demonstrations is tied to the model, being the simpler ones the most benefited from the provided examples.

On the other hand, our data contamination experiments have shed light on the generalisation abilities of the models. We found no evidence that ChatGPT has ingested any of the collections of health questions and, additionally, our results indicate that no model has seen the TREC HM 2022 collection. This breaks a lance in favour of the models, as many of the health questions were new for them but, still, their performance was remarkable.

Finally, we conducted an error analysis in which we inspected the models' answers. We organised the errors into a taxonomy and identified that, in some cases, models provided advice that goes against the well-known medical consensus. This behaviour, which also happens with the most sophisticated prompts, is a barrier for the wide adoption of these models in their current form.

Limitations

We are aware that these conversational AI systems are highly sensible to the input prompt. Our study represents an initial exploration with some manually defined prompts and further prompt optimisation was left for future work. Our ultimate goal here was not to achieve the highest possible performance but, rather, to test these AIs with an assorted set of input conditions. Other strategies like chain-of-thought (CoT) prompting or prompt tuning were also left for future work. With this research we do not intend to pursue the replacement of human professionals that provide health advice. In fact, we firmly believe that human validation is crucial to learn more about these systems and to leverage AIs systems. For example, exploiting LLMs for automating certain documentation tasks (e.g., collecting and curating recommendations generated by AI systems). Finally, we are also aware that the evaluation of proprietary systems causes some concern in the scientific community (because of the lack of transparency about crucial aspects, such as their design and training data). However, as scientists, we cannot ignore the fact that these tools are used by millions of individuals worldwide and, thus, we need to evaluate the risks involved. In our study, we have made a special effort for transparency providing the code, outputs and dates of all experiments.

Acknowledgements

The authors thank: i) the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System, and ii) the financial support supplied by projects PLEC2021-007662 and PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund). Finally, David E. Losada thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego).

References

1. Ahn, C.: Exploring chatgpt for information of cardiopulmonary resuscitation. Resuscitation **185** (2023)
2. Biswas, S.S.: Potential use of chat gpt in global warming. Annals of Biomedical Engineering pp. 1–2 (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

4. Chervenak, J., Lieman, H., Blanco-Breindel, M., Jindal, S.: The promise and peril of using a large language model to obtain clinical information: Chatgpt performs strongly as a fertility counseling tool with limitations. Fertility and Sterility (2023)
5. Clarke, C., Maistro, M., Smucker, M.: Overview of the trec 2021 health misinformation track. In: Proceedings of the Thirtieth Text REtrieval Conference, TREC (2021)
6. Clarke, C., Maistro, M., Smucker, M., Zuccon, G.: Overview of the trec 2020 health misinformation track. In: Proceedings of the Twenty-Nine Text REtrieval Conference, TREC. pp. 16–19 (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Duong, D., Solomon, B.D.: Analysis of large-language model versus human performance for genetics questions. European Journal of Human Genetics pp. 1–3 (2023)
9. Forbes: Introducing chatgpt (November 2022), https://openai.com/blog/chatgpt, [accessed April 4, 2023]
10. Fox, S.: Health topics: 80% of internet users look for health information online. Pew Internet & American Life Project (2011)
11. Golchin, S., Surdeanu, M.: Time travel in llms: Tracing data contamination in large language models. arXiv preprint arXiv:2308.08493 (2023)
12. Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T.T., McGee, L.A., Ashman, J.B., Li, X., Liu, T., Shen, J., et al.: Evaluating large language models on a highly-specialized topic, radiation oncology physics. arXiv preprint arXiv:2304.01938 (2023)
13. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423–438 (2020)
14. Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., et al.: Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model (2023)
15. Lachenbruch, P.A.: Mcnemar test. Wiley StatsRef: Statistics Reference Online (2014)
16. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
17. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
18. Lin, C.Y., Och, F.: Looking for a few good metrics: Rouge and its evaluation. In: Ntcir workshop (2004)
19. Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., Chen, W.: What makes good in-context examples for gpt-3? arXiv preprint arXiv:2101.06804 (2021)
20. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
21. Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H.W., Tay, Y., Zhou, D., Le, Q.V., Zoph, B., Wei, J., et al.: The flan collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:2301.13688 (2023)

22. Magar, I., Schwartz, R.: Data contamination: From memorization to exploitation. arXiv preprint arXiv:2203.08242 (2022)
23. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023)
24. OpenAI: Gpt-4 technical report. arXiv:submit/4812508 (2023)
25. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
26. Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.: The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In: Proceedings of the ACM SIGIR Int. Conf. on Theory of Information Retrieval. pp. 209–216 (2017)
27. Pradeep, R., Lin, J.: Towards automated end-to-end health misinformation free search with a large language model. In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval. pp. 78–86. Springer Nature Switzerland, Cham (2024)
28. Radfar, M., Mouchtaris, A., Kunzmann, S.: End-to-end neural transformer based spoken language understanding. arXiv preprint arXiv:2008.10984 (2020)
29. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
31. Samaan, J.S., Yeo, Y.H., Rajeev, N., Hawley, L., Abel, S., Ng, W.H., Srinivasan, N., Park, J., Burch, M., Watson, R., et al.: Assessing the accuracy of responses by the language model chatgpt to questions regarding bariatric surgery. Obesity surgery pp. 1–7 (2023)
32. Sellam, T., Das, D., Parikh, A.P.: Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696 (2020)
33. Sianz, O., Campos, J.A., García-Ferrero, I., Etxaniz, J., Agirre, E.: Did chatgpt cheat on your test? (2023), https://hitz-zentroa.github.io/lm-contamination/blog/, [accessed January 19, 2024]
34. Surameery, N.M.S., Shakor, M.Y.: Use chat gpt to solve programming bugs. International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290 **3**(01), 17–22 (2023)
35. Thirunavukarasu, A.J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M., Shah, S.: Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Medical Education **9**(1), e46599 (2023)
36. Vigdor, N.: Man fatally poisons himself while self-medicating for coronavirus, doctor says (March 2020), https://www.nytimes.com/2020/03/24/us/chloroquine-poisoning-coronavirus.html, [accessed June 9, 2022]
37. Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., You, Z.: Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070 (2023)
38. Zuccon, G., Koopman, B.: Dr chatgpt, tell me what I want to hear: How prompt knowledge impacts health answer correctness. arXiv preprint arXiv:2302.13793 (2023)