npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01546-w

Evaluating search engines and large language models for answering health questions

Check for updates

Marcos Fernández-Pichel ✉, Juan C. Pichel & David E. Losada

Search engines (SEs) have traditionally been primary tools for information seeking, but the new large language models (LLMs) are emerging as powerful alternatives, particularly for question-answering tasks. This study compares the performance of four popular SEs, seven LLMs, and retrieval-augmented (RAG) variants in answering 150 health-related questions from the TREC Health Misinformation (HM) Track. Results reveal SEs correctly answer 50–70% of questions, often hindered by many retrieval results not responding to the health question. LLMs deliver higher accuracy, correctly answering about 80% of questions, though their performance is sensitive to input prompts. RAG methods significantly enhance smaller LLMs' effectiveness, improving accuracy by up to 30% by integrating retrieval evidence.

Recent progress in natural language processing (NLP) has positioned Large Language Models as major players in numerous Information Access tasks [1–3]. The release of ChatGPT in November 2022 has been a game-changer globally, marking a significant milestone and revolutionizing many sectors. One of the outstanding features of current LLMs is their ability to generate coherent and human-like text, which has garnered attention and excitement among practitioners, researchers, and the general public. This breakthrough has precipitated a transformative shift in the orientation of information access research towards LLMs, their potential applications, and the interconnection between LLMs and other computer-based tools. The conversational paradigm has gained traction, enabling more interactive and user-friendly search experiences[4–7]; and many citizens currently turn to conversational AIs based on LLMs for consulting multiple types of information needs. However, traditional search still plays a crucial role in the generative AI era[8]. While LLMs may support advanced information access and reasoning capabilities, there is still a need to advance retrieval technologies. The role of traditional web search engines in answering user-submitted queries is far from being relegated. Web search is widely used to obtain health advice[9] and effective medical information retrieval has attracted the attention of the scientific community over the years[10]. For example, Wang et al. evaluated several search engines in terms of their usability and effectiveness for searching for breast cancer information[11]. They found that the results highly overlapped among the four search engines tested, all of them providing rich information about breast cancer. Zuccon and colleagues[12] studied the effectiveness of search engines for the so-called "diagnostic medical circumlocutory queries", which are searches issued by individuals seeking information about their health using casual descriptions of symptoms rather than medical words.

The emergence and global adoption of advanced LLMs have sparked the urgent need to explore and understand their capacities and knowledge acquisition attributes. Some research studies have focused on the capabilities of these models under specific language understanding and reasoning benchmarks [13–15] and, specifically, interest in assessing the correctness of health-related AI-based completions has escalated. For example, Chervenak et al. demonstrated ChatGPT's abilities to answer fertility questions[16], and Duong and Solomon[17] evaluated the effectiveness of LLMs in comparison to humans when tasked with answering multiple-choice questions on human genetics. Similarly, Holmes et al.[18] conducted a comparative study of LLMs' knowledge on the highly specialized subject of radiation oncology physics. Recently, Elgedawy and colleagues tested the capacity of LLMs to query an extensive volume of clinical records [19] and Kim et al. assessed ChatGPT's accuracy in answering 57 epilepsy-related questions[20]. Kim and others [21] examined the diagnostic accuracy of GPT-4 compared to mental health professionals and other clinicians employing clinical vignettes of obsessive-compulsive disorder (OCD). In other papers, the authors reported studies on the role of LLMs for biomedical tasks[22], patient-specific EHR questions[23], and bariatric surgery topics[24]. Tang et al.[25] evaluated LLMs' ability to perform zero-shot medical summarization across six clinical domains. With a broader perspective, other authors involved physicians in a thorough evaluation of the accuracy of ChatGPT in answering health queries[26] or evaluated ChatGPT using the applied knowledge test (AKT) from the Royal College of General Practitioners[27]. All of the aforementioned studies are restricted to a single model, usually ChatGPT, and/or to a specific medical

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain.
✉e-mail: marcosfernandez.pichel@usc.es

area. Kusa et al.[28] analyzed the impact of users' beliefs and prompt formulations on completions related to health diagnoses. The study explored the sensitivity of two GPT models to variations in the user's context. Caramancion[29] explored users' preferences between search engines and LLMs. This analysis revealed interesting trends in user preferences, indicating a distinct tendency for participants to choose search engines for straightforward, fact-based questions. In contrast, LLMs were more frequently favored for tasks needing detailed comprehension and language processing. The results clearly showed that users prefer to search for medical information using traditional search engines. Oeding and colleagues[30] compared GPT-4's and Google's results for searches concerning the Latarjet procedure (anterior shoulder instability). These authors discovered that GPT-4 provided more information based on academic sources than Google in response to the patient's queries.

In the area of retrieval-augmented models for the health domain, Li et al.[31] fine-tuned Llama with medical conversations and injected medical evidence from Wikipedia and other medical sources, while Koopman and Zuccon[32] evaluated ChatGPT's capacity to answer health questions (based solely on its internal knowledge or, alternatively, fed with offline retrieval evidence). Xiong et al.[33] proposed MedRAG, a system that indexes multiple medical corpora and retrieves relevant evidence to ground different LLMs. This study found that simpler models can reach the performance of GPT-4 when grounded with relevant medical information.

In line with these developments, it is crucial to acknowledge the significance of accurate health information and there is a pressing need to evaluate the abilities of classic and new information access tools in answering medical questions. There is a lack of comprehensive studies that compare the effectiveness of LLMs with that of traditional SEs in the context of health information seeking. Furthermore, given that the input prompt significantly influences the efficacy of LLMs[13,34,35], understanding models' effectiveness with different types of prompts is of utmost importance. It is also crucial to explore the effect of combining both classes of tools and, for example, explore the behavior of LLMs when prompted with medical questions together with related search results. This paper aims to contribute towards filling these gaps by conducting a thorough study aimed at responding to the following research questions:

- RQ1. To what extent do search engines retrieve results that help to answer medical questions? Does the correctness of the information provided drop as we go down the search engine result page?
- RQ2. Are LLMs reliable in providing accurate medical answers? How do different models compare regarding their effectiveness in responding to medical questions?
- RQ3. Does the given context influence the capabilities of LLMs in providing correct answers? Is there an observable improvement in these models when they are exposed to a few in-context examples?
- RQ4. Do LLMs improve their performance when fed with web retrieval results?

Our comparison of SEs included experiments with Google, Bing, Yahoo!, and DuckDuckGo. We found that Bing seems to be the most solid choice. However, it is not significantly better than the others. Our evaluation also suggests that extracting the answers from the top-ranked webpage often produces good results. This is good news, as web users are known to be reluctant to inspect many items from the search engine result pages (SERPs). Moreover, LLMs show an overall good performance, but they are still very sensitive to the input prompt, and, in some instances, they provide highly inaccurate responses. Finally, augmenting LLMs with retrieval results from SEs looks promising and, according to our experiments, even the smallest LLMs can achieve top-tier performance when provided with suitable retrieval evidence. These findings significantly contribute to our understanding of health information seeking using the new LLMs and traditional search engines.

Results

In this section, we report the effectiveness of different search engines in providing correct responses to binary health questions and analyze web user experience following two user models (see "Methods"). We also evaluate the performance of LLMs when prompted with the same health questions (under zero- and few-shot settings), propose a taxonomy of errors made by LLMs, and present results for retrieval-augmented strategies.

Search engines

Figure 1 shows the results for three different collections of questions. From these plots, we observe no significant drop in performance as we move down the rankings, which speaks well of the SEs' retrieval capabilities. With respect to different engines, Bing seems a solid choice in all datasets. However, we ran pairwise comparisons between SEs (Mann–Whitney U-test, $\alpha = 0.05$), for the cutoff positions 1 and 20 (i.e., statistical differences between their P@1 or their P@20 values), and found no significant differences between any pair of engines. This suggests that the SEs have similar performance and, according to our experiments, we cannot declare a clear winner. By comparing the plots for the three collections, we can observe that the 2022 topics produced the best results. A possible explanation for this outcome is that the 2022 collection contains fewer specialized health questions, and it might be easier to retrieve correct answers from the web.

We also wanted to analyze the extent to which retrieval results provide an actual answer to the health questions. Ranked lists contain off-topic results and, additionally, many on-topic pages do not provide a clear response to the user's request. For each of the 50 health questions available in each collection, we processed its top retrieved results to identify whether they answered the health question. This identification was automatic, supported by passage extraction and reading comprehension technology (see "Methods"). For example, for Google's rankings and $n = 1$, we computed the number of entries that actually provided a response to the health query. The same count was obtained for each position (up to rank #20) and, finally, a global answering score was obtained by averaging the number of questions that were actually answered at each position. These scores, plotted in Fig. 2, are much lower than the number of questions in these collections (50), reflecting that many retrieval results do not provide a clear yes/no recommendation to the health question. According to this experiment, Bing is the engine whose rankings provided more answers for 2020 questions (around 30 questions answered) and, from Fig. 1a we infer that Bing's responses tend to be slightly better than those provided by the three other SEs. For 2021, Google and DuckDuckgo supplied responses to more questions but at the cost of providing more incorrect responses throughout the ranking (Fig. 1b). Finally, for 2022 questions, Google provided many responses and, according to Fig. 1c, it was not inferior to the other SEs. Note also that, on average, between 15 and 20 questions did not receive a response at any given ranked position. This suggests that restricting the analysis to a few top-ranked web pages, which is the common behavior of end users, often results in a failure to meet these health information needs.

The precision of the SEs is low (60–70%), but the good news is that this is because many ranked results did not provide an answer. In fact, if we do not count the entries that do not provide a response as failures, then precision rises to 80–90% for all search engines. Although high, these values still reflect a worrying percentage of 10–15% of incorrect responses in the search engine results pages.

Let us now evaluate the search results using two user behavior models: (i) lazy, modeling a user who stops inspecting results when presented with the first entry that gives a yes/no response to his question, and (ii) diligent, modeling a user who only stops after contrasting three responses (taking a majority vote decision, see "Methods"). In Fig. 3, we report the percentage of correct decisions, incorrect decisions, and no decisions for the three collections. Note that in a few instances, the user exhausted the ranked list and found no response (gray area in Fig. 3). These cases are more frequent in the diligent user scenario because this model needs three answers to make an informed decision. We also report in Table 1 the effort required to make a decision, measured as the mean number of results a user needs to inspect before reaching the stopping criterion.

Observe that the lazy behavior produces similar or even better results, requiring less effort, compared with those achieved by diligent behavior.

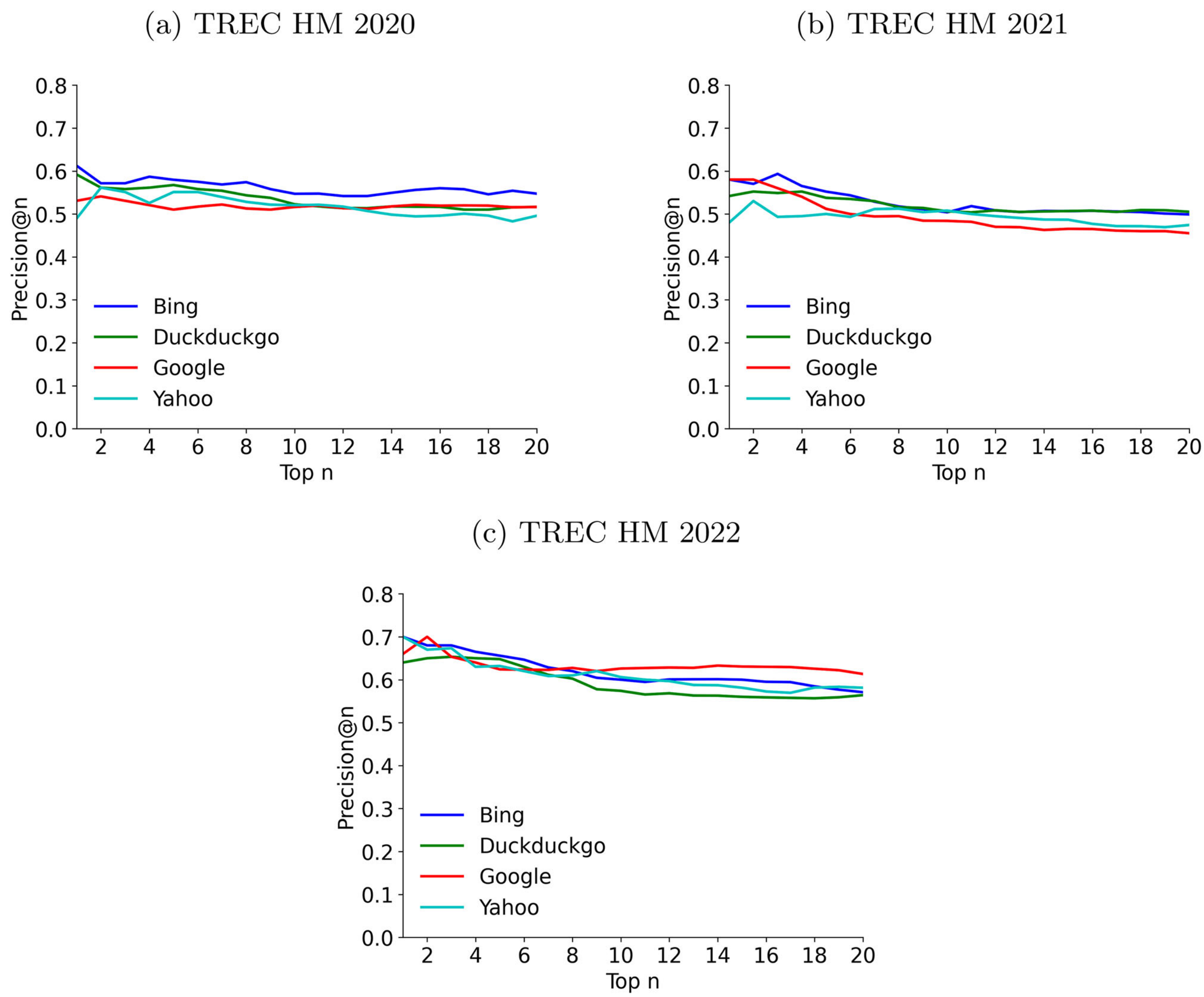

Fig. 1 | Precision@n obtained from the different search engines. Each data point on the graphs represents the precision at the *n*-th position of the ranking (Precision@n) averaged over all queries in each collection. For example, the fourth point in each line (position = 4) represents the average ratio of correct responses obtained after inspecting the top 4 results (P@4). Each line represents one search engine (Bing is the blue line; Duckduckgo, is the green line; Google, is the red line; and Yahoo, is the light blue line). **a** Plots results for the TREC HM 2020 collection. **b** Plots results for the TREC HM 2021 collection. **c** Plots results for the TREC HM 2022 collection.

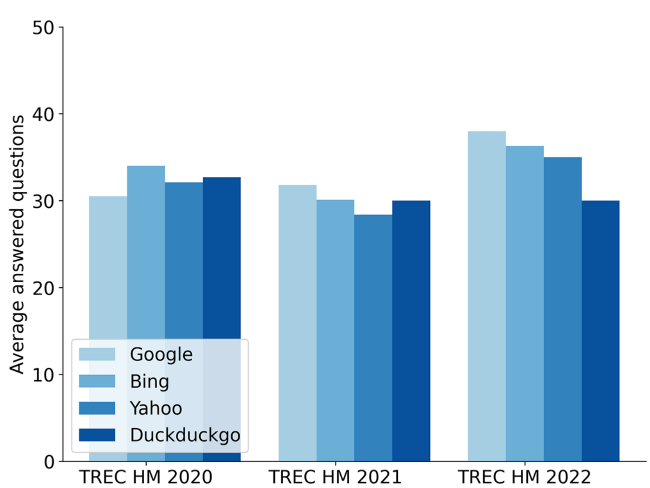

Fig. 2 | Search Engine Answering Scores. An average number of health questions (out of 50) were answered at the top 20. Each engine is represented by an increasingly darker blue bar in the following order: Google, Bing, Yahoo, and Duckduckgo.

Indeed, the additional effort spent by the diligent user does not translate into a lower percentage of incorrect decisions. This may sound surprising, but it somehow reinforces the confidence in the search engine's top result (the top answer suffices). In fact, our results suggest that the diligent user, who goes deeper in the ranking to find additional responses, would make poorer health-related decisions.

Large language models

Next, we evaluate the performance of LLMs for our binary question-answering task under zero- and few-shot settings. We also propose a taxonomy for the errors that these models tend to produce and perform an error analysis based on this taxonomy. Finally, the results for retrieval-augmented strategies are presented.

Figure 4 plots the 0-shot accuracy for seven LLMs and three prompting strategies ("no context", "non-expert" and "expert", see "Methods"). All models show remarkable performance for the TREC HM 2020 collection, with the exception of FlanT5. In this collection, Llama3, MedLlama3, and text-davinci-003 stand out from the rest of the LLMs. For the 2021 questions, the accuracy generally drops, but the same models dominate, with special mention to MedLlama3. Regarding the TREC HM 2022 data, ChatGPT and GPT-4 stand out. Additionally, there are variances in

Fig. 3 | Lazy and Diligent user models. Percentages of correct, incorrect, and no decision for the different datasets, search engines, and user models. Each panel is divided into two parts: the left side for lazy user behavior and the right side for diligent user behavior. Each bar presents the percentage of correct (green stacked bar), incorrect (red stacked bar), and no decision (gray stacked bar) for the corresponding user behavior. **a** Plots results for the TREC HM 2020 collection. **b** Plots results for the TREC HM 2021 collection. **c** Plots results for the TREC HM 2022 collection.

performance linked to the use of different prompts. In general, the "expert" context is the most effective, and it seems to be guiding the model toward more reputed knowledge sources.

The models generally show relative stability, but their performance still varies based on the inputs provided. For example, FlanT5 and text-davinci-002 are highly sensitive to the type of prompt. This raises concerns, as a model's accuracy can drop from 90% to 75% or even lower numbers. While the overall performance levels are high, these inconsistencies are troubling. Even when using the most reliable prompt ("expert"), there are still concerning situations. For instance, GPT-4's accuracy drops to 68% on the 2021 dataset.

The difficulty levels of the three datasets vary. The 2020 health questions (centered on COVID-19) seem easier for the large language models. A possible reason could be the significant relevance of COVID-19 as a topic, potentially encouraging specialized data curation processes.

Table 1 | Effort required to make a decision by each user behavior model

	TREC HM 2020		TREC HM 2021		TREC HM 2022	
	Lazy	Diligent	Lazy	Diligent	Lazy	Diligent
Google	1.7	5.2	1.7	5.2	2	4.9
Bing	1.8	5.4	1.5	4.3	1.6	4.6
Yahoo	1.6	5.5	1.6	5.2	1.7	4.4
Duckduckgo	1.7	5.1	1.4	4.9	1.7	4.6

The effort is measured as the mean number of results inspected.

We ran McNemar's test to ascertain the statistical significance of the pairwise performance disparities between the leading models. The comparison between GPT-4 and ChatGPT showed no significant difference in 7 out of 9 cases (3 datasets × 3 prompts). The comparison between Llama3 or its fine-tuned version, MedLlama3, and ChatGPT revealed no significant difference in almost all cases. However, when compared against GPT-4, both Llama3 and MedLlama3 were significantly better for the 2021 collection. Additionally, MedLlama3 was statistically better than Llama3 when no context was provided. On the other hand, the statistical tests confirmed that GPT-4 dominated Llama3/MedLlama3 in the 2022 collection.

The pairwise comparisons of ChatGPT vs d-003, Llama3 vs d-003, MedLlama3 vs d-003, and GPT-4 vs d-003 produced a higher frequency of statistically significant outcomes, but still, a majority of compared cases resulted in no significant differences.

Comparing the 0-shot results achieved by LLMs (Fig. 4) with those achieved by SEs (Fig. 1), we can observe that, in general, LLMs produce higher performance. Even the user behavior evaluation (Fig. 3), which simulates users skipping non-answers from SE results, yields effectiveness statistics that are inferior to those obtained with LLMs. This suggests that the huge amounts of training data of the LLMs and their advanced reasoning capabilities are key advantages compared to the extraction of responses from a few top-ranked search results.

To examine the impact of in-context examples on LLMs, we made additional tests with the health questions of TREC HM 2022. We sent each health question to the models preceded by one to three examples taken from TREC HM 2021. To this end, we selected three random pairs from the TREC HM 2021 dataset to serve as in-context examples and investigated their influence. Previous studies have indicated that a small set of in-context examples suffices for instructing the LLMs[35].

Table 2 reports the accuracy when providing a varying number of in-context examples (from 1 to 3). The impact of in-context examples varies significantly across different models. Specifically, FlanT5 and the two versions of GPT-3 (d-002 and d-003) showed a clearly positive effect with in-context examples. For these three models, the inclusion of in-context cases resulted in statistically significant improvements. Conversely, the models that performed best in the zero-shot experiments did not experience any noticeable benefit from adding in-context examples. Regarding the type of prompts, the weakest prompts ("non expert" and "no context") gained the most from in-context learning. Furthermore, these experiments suggest that using one example is good enough (adding more than one does not consistently improve performance).

In order to gain a deeper understanding of the performance of the LLMs, we scrutinized the instances in which all models were unsuccessful in delivering a correct response. This analysis aimed to clarify the reasons for such low effectiveness. Specifically, we sent these "low-performance questions" again to the models, but this time, we did not set a limit to the token output. In this way, we could manually analyze the explanations provided and try to understand the reasoning behind the models. This examination was conducted with the most effective prompt ("expert" variant) and with

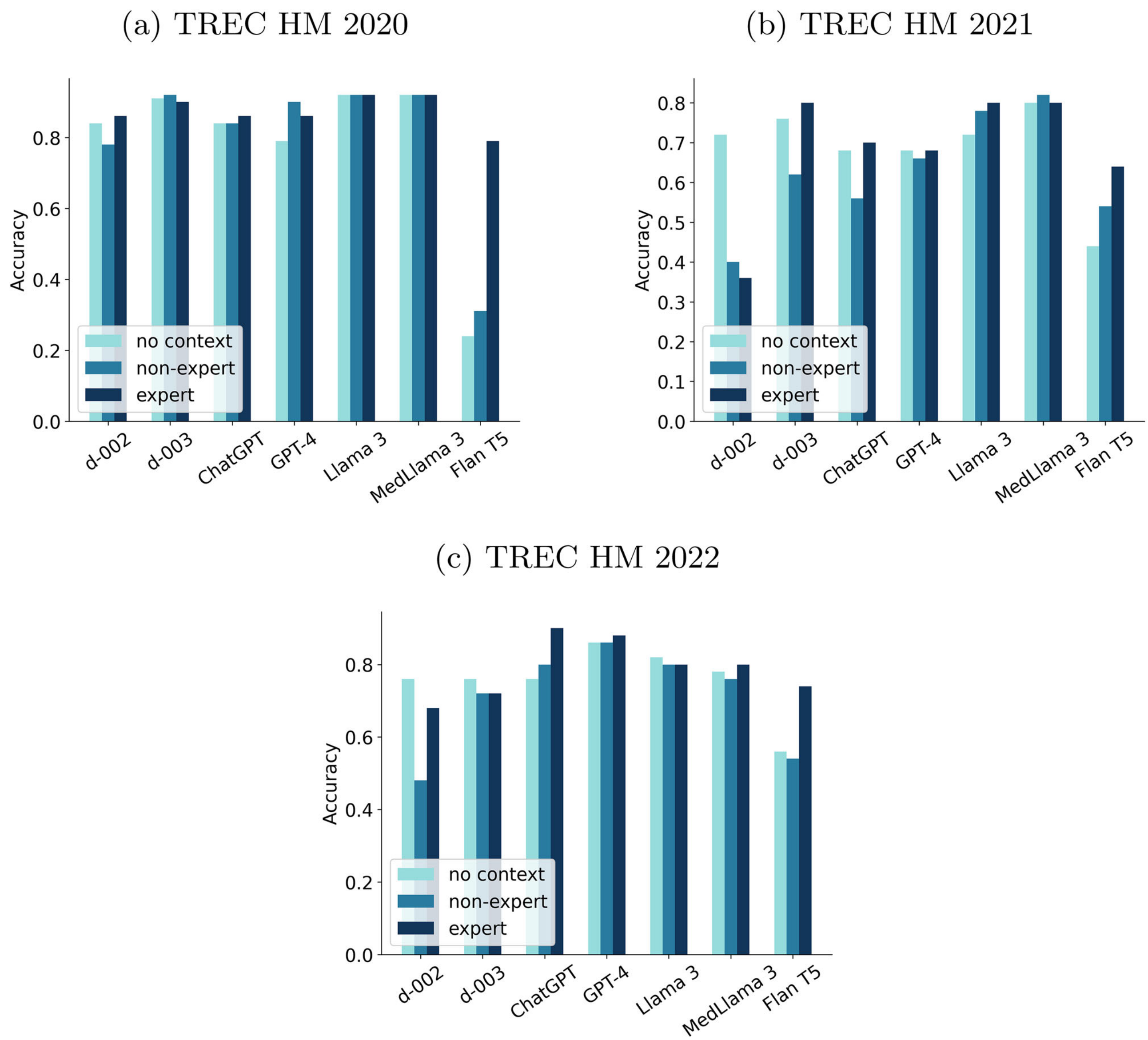

Fig. 4 | Zero-shot experiments. For each LLM, the bars represent the accuracy for the three prompting strategies: "no context" (light blue bar), "non-expert" (blue bar), and "expert" (navy blue bar). **a** Plots results for the TREC HM 2020 collection. **b** Plots results for the TREC HM 2021 collection. **c** Plots results for the TREC HM 2022 collection.

the "no context" variant, which arguably mirrors the type of interaction done by a lay user.

In the TREC HM 2020 collection, 4% of the questions had an incorrect answer produced by all models using both prompts. The TREC HM 2021 collection had 2% of the questions assigned with wrong answers by all models for the "no context" prompt and 6% for the "expert" prompt. Meanwhile, 2% of the questions from the TREC HM 2022 collection made that all LLMs failed for the "no context" prompt (and 4% for the "expert" prompt). These figures confirm that, regardless of the LLM, some questions pose serious difficulties to the models.

After manually reviewing the models' completions, we were able to categorize the errors into a taxonomy that embodies the most frequent mistakes in health guidance, namely:

- Incorrect understanding of prevailing medical consensus: sometimes, the models provided responses that directly contradict the medical consensus. For example, GPT-4 answered "Some research suggests that it may help in reducing symptoms associated with eczema..." in response to "Can evening primrose oil help treat eczema?". This answer clearly contradicts clinical evidence (see e.g., NIH does not support the use of evening pimrose oil to treat eczema).
- Misinterpretation of the question: in this class of errors, the LLMs misunderstood the question. For instance, in response to "Can bleach prevent COVID-19?" the models produced completions such as "No, bleach should not be ingested...". However, the correct interpretation of this inquiry is that the application of bleach for surface sterilization can indeed inhibit COVID-19. It is hardly conceivable that a human would interpret the question in the same fashion as the LLM.
- Ambiguous answer: this category includes responses where models did not generate a clear answer. These responses cannot be considered correct, but the LLM's output could arguably be of value. For example, in reference to the potential effects of ankle braces to help heal an ankle fracture the model says "No, ankle braces alone do not heal an ankle fracture. They can, however, provide support, stability, and aid in managing pain during the healing process....".

Figure 5 reports the percentage of each class of error for the different input conditions. The predominant category of error is the lack of knowledge about medical consensus. This is worrying because this is the most harmful type of mistake. Note also that the "expert" prompt had a few questions misinterpreted and a few questions whose answers contradicted

Table 2 | Few-shot experiments (with up to three in-context examples)

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	Llama3				GPT-4			
	0-shot	1-shot	2-shot	3-shot	0-shot	1-shot	2-shot	3-shot
No-context	0.82	**0.86**	**0.84**	0.8	0.86	0.84	0.86	0.86
Non-expert	0.8	0.76	0.8	0.8	0.86	0.86	**0.88**	**0.88**
Expert	0.8	0.76	**0.82**	0.78	0.88	0.88	**0.92**	**0.9**
prompt	MedLlama3							
	0-shot	1-shot	2-shot	3-shot				
No-context	0.78	**0.8**	0.76	0.76				
Non-expert	0.76	**0.78**	**0.78**	**0.82**				
Expert	0.8	**0.82**	0.8	0.8				

Accuracy of each model and prompt. For each row, if a few-shot instance outperforms the 0-shot case then the few-shot case is bolded, and the symbol "*" marks those instances where the improvement was deemed as statistically significant (McNemar's test, $\alpha = 0.05$).

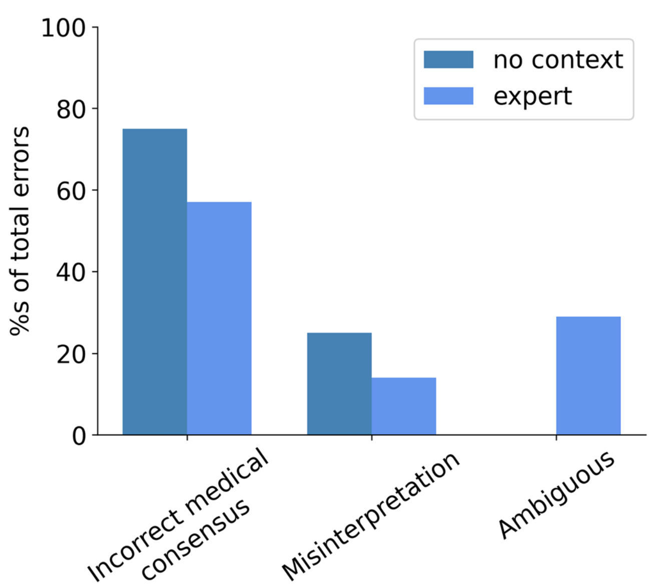

Fig. 5 | Most common errors made by the LLMs. Percentage of different types of errors over the total number of errors made by the "no context" (navy blue bar) and "expert" prompting (blue bar) strategies.

the medical consensus, but at the cost of providing more ambiguous answers. In future work, we would like to involve clinicians in reviewing the models' outputs and evaluating different dimensions, such as understandability or degree of correctness.

To further understand the relative merits of LLMs and SEs, we now analyze the behavior of the SEs with these difficult health questions. To that end, we calculated the average P@10 across all search engines for all questions and the average P@10 across all search engines for the difficult questions. The comparison of these two averages helps to understand whether or not the SEs also struggle with these topics: TREC HM 2020 topics (P@10 all: 0.53 vs P@10 difficult: 0.06), TREC HM 2021 topics (P@10 all: 0.50 vs P@10 difficult: 0.23), TREC HM 2022 topics (P@10 all: 0.60 vs P@10 difficult: 0.04). These figures suggest that the problem is not so much with LLMs as with the questions themselves, as search engines also return incorrect information in the top results for these difficult questions.

Retrieval-augmented large language models

A stimulating line of research consists of combining the potential of LLMs and SEs. Exploring the effect of including extracts from SE results in the LLMs' instructions is interesting. To reduce the monetary cost and computational load of these retrieval-augmented experiments, we decided to extract the passages from a single engine (Google). The same passages from Google were fed to five LLMs: text-davinci-002, ChatGPT, GPT-4, Llama3, and MedLlama3 under two prompting strategies ("no context" and "expert").

Figure 6 depicts the results of augmenting LLMs with each of the passages from Google's top 5 for the "no context" prompting strategy. For the 2020 questions, the LLMs do not seem to require these additional passages and, actually, four of the LLMs got worse results when presented with additional evidence. For the two other datasets, the LLMs seem to improve their performance with textual evidence from the search engine. It must be noticed that in some cases (e.g., 2022), the provided passages make that the least sophisticated model, text-davinci-002, becomes comparable or even superior to more recent models, such as GPT-4 and reaches the performance of fine-tuned models in medical data. We perceive this as an important outcome of our study since it demonstrates that lighter models can achieve state-of-the-art performance when provided with additional evidence. In some cases, we can see a trend where the most sophisticated models only improve when provided with the first result. For example, this happened for MedLlama3 augmented with the top passage in the 2021 collection. We estimated statistical significance using McNemar's tests to compare the augmented versions of each LLM and its non-augmented counterpart. For the "no context" prompting strategy, two comparisons yielded significant differences (MedLlama3 no RA vs MedLlama3 + top 2

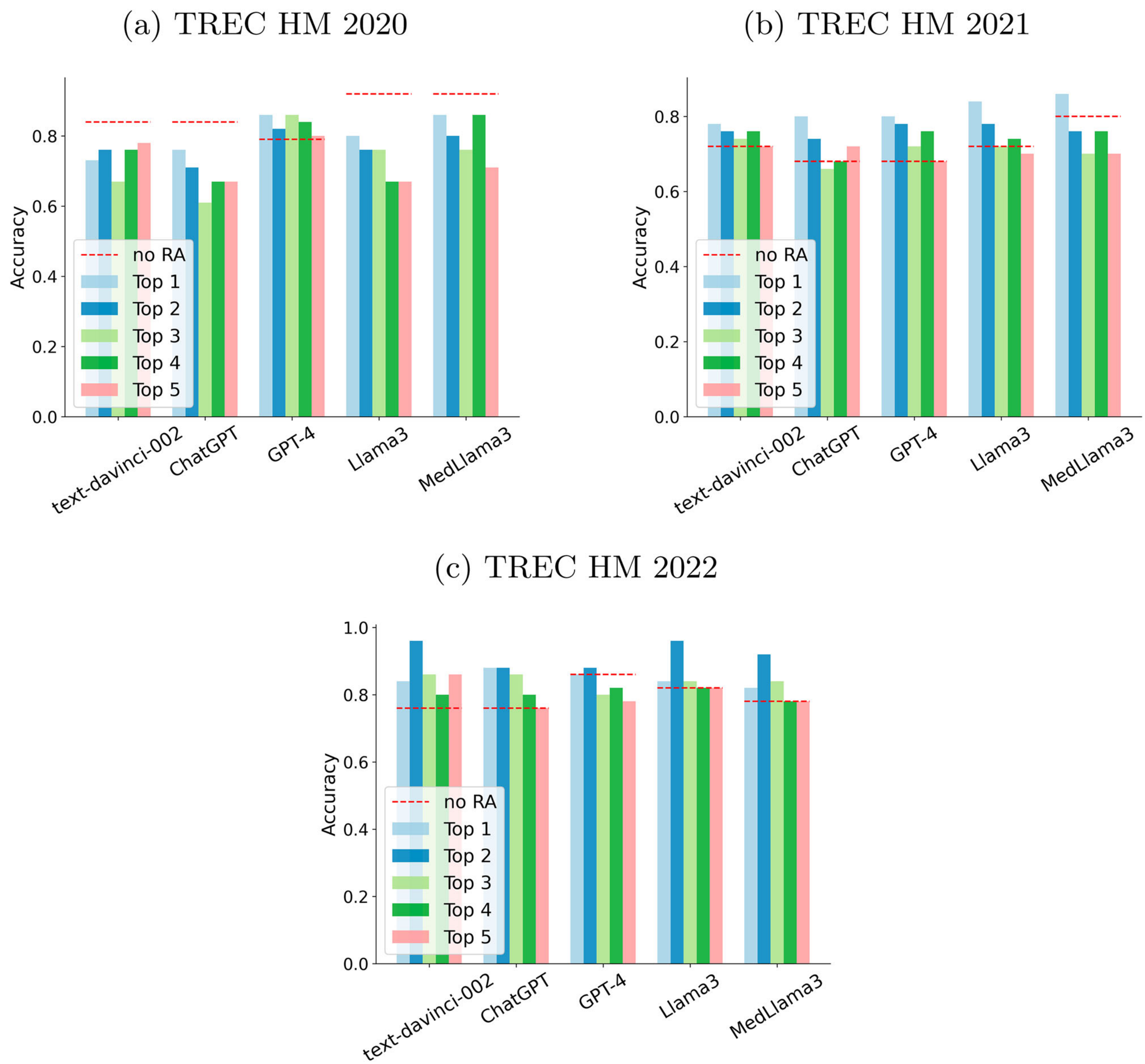

Fig. 6 | Retrieval-augmented LLMs, accuracy for "no context" prompt. Each "Top n" bar depicts the performance obtained by feeding the n-th result from Google until the top 5 (top 1 results are represented as a light blue bar, top 2 as a navy blue bar, top 3 as a light green bar, top 4 as a green bar and top 5 as a salmon bar). The baseline (no retrieval augmentation) is represented by a red dashed line. **a** Plots results for the TREC HM 2020 collection. **b** Plots results for the TREC HM 2021 collection. **c** Plots results for the TREC HM 2022 collection.

result -TREC HM 2022-, and ChatGPT no RA vs ChatGPT + top 1 result -TREC HM 2021-). The rest of pairwise comparisons did not yield a significant outcome. This might be attributed to the relatively small size of the sample (number of queries in each collection).

On the other hand, Fig. 7 shows the behavior of the retrieval augmented language models with the "expert" prompt. Tendencies remain similar, but we can highlight a larger increase in performance for text-davinci-002 in the 2021 collection. Its performance grows from 0.36 up to values above 0.70 with retrieval augmentation. Again, we ran statistical tests and some comparisons declared statistical significance (ChatGPT no RA vs ChatGPT + top 5 result -TREC HM 2021-, GPT-4 no RA vs GPT-4 + top 1 result -TREC HM 2021-, Llama3 no RA vs Llama3 + top 2 result -TREC HM 2022-, Llama3 no RA vs Llama3 + top 3 result -TREC HM 2022-, and MedLlama3 no RA vs MedLlama3 + top 5 result -TREC HM 2022-). We can conclude that there seems to be some positive signal from prompting the LLMs with SEs results, but the results are still inconclusive.

We also evaluated the influence of augmenting with multiple passages. To that end, we ran experiments where the top 3 passages (extracted from Google's top 3 results) were concatenated and fed to the LLM as a single input. The results of this experiment were not very conclusive about the effect of injecting multiple passages. To further understand how the correctness of the injected passages influences the accuracy of the LLMs' responses, we also analyzed the quality of the responses with varying numbers of correct passages in the input. More specifically, for each passage obtained from Google, we extracted the answer (if any) that the passage gives to the question (using, again, the GPT3-based reading comprehension module, see "Methods"). By contrasting these answers with the ground truth response of the health question, we can compute how many passages actually provide the LLM with a correct response. Since we injected three passages, we have four possible situations (0 correct passages out of 3, 1 correct passage out of 3, and so forth). Figure 8 plots the results of this analysis (for the sake of simplicity, we only ran this analysis for the expert

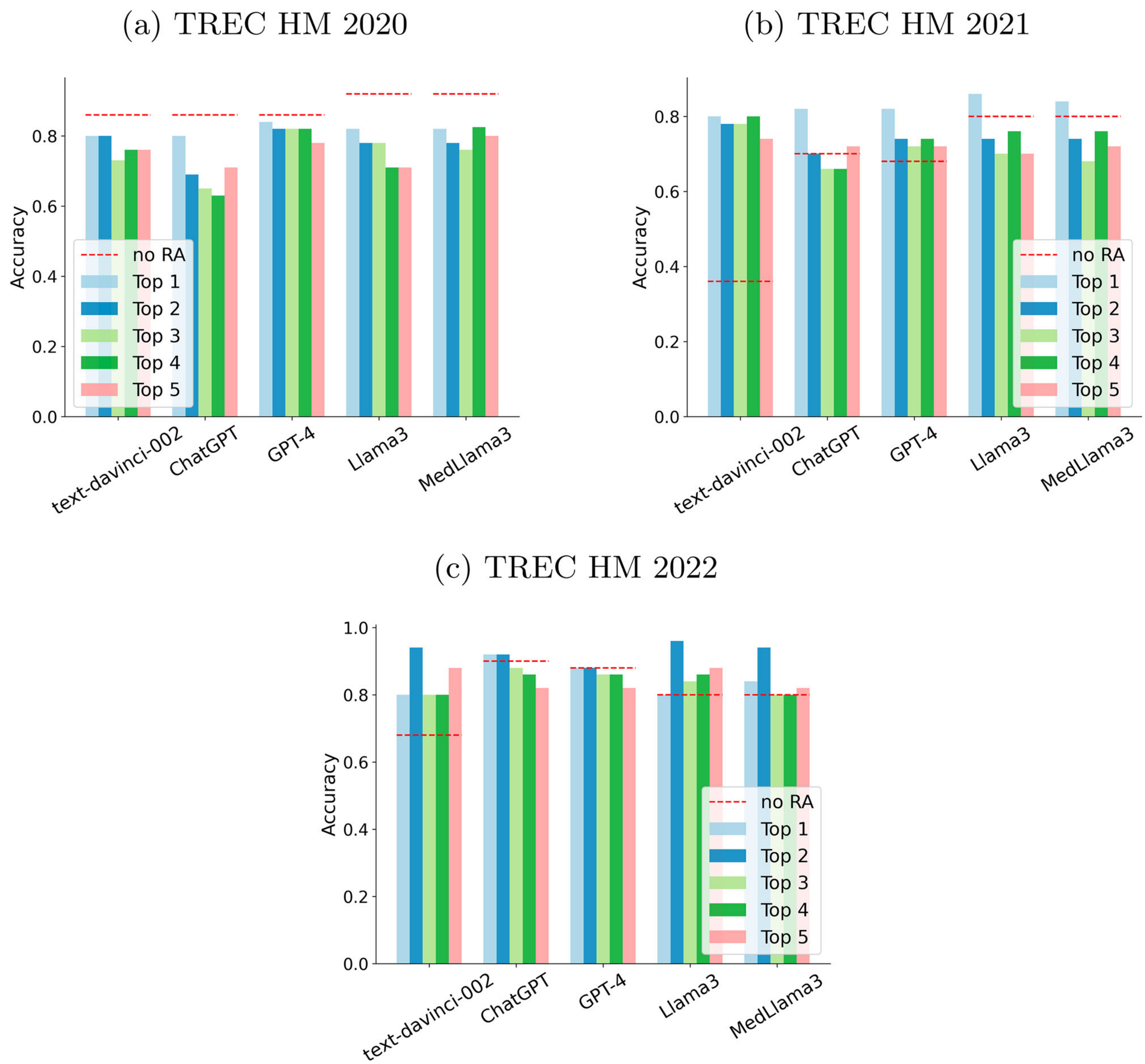

Fig. 7 | Retrieval-augmented LLMs, accuracy for "expert" prompt. Each "Top *n*" bar depicts the performance obtained by feeding the *n*-th result from Google until the top 5 (top 1 results are represented as a light blue bar, top 2 as a navy blue one, top 3 as a light green one, top 4 as a green one and top 5 as a salmon one). The baseline (no retrieval augmentation) is represented by a red dashed line. **a** Plots results for the TREC HM 2020 collection. **b** Plots results for the TREC HM 2021 collection. **c** Plots results for the TREC HM 2022 collection.

prompt). As expected, accuracy improves as we feed the LLMs with higher proportions of correct passages. In fact, if all input passages are correct, then the LLMs hardly fail. In contrast, when all passages are incorrect, accuracy drops significantly, particularly in the 2020 and 2021 collections. All models benefit from higher counts of correct passages. This experiment confirms that the correctness of the injected passages is crucial to the performance of a health RAG system.

In the future, it would be interesting to explore other RAG variants and the interactions among LLM complexity, types of prompts, size, variety of retrieval results, and types of health questions.

Discussion

Search engines are the classic information access tools to retrieve content from online sources. The first goal of our study was to estimate their capacity to provide correct answers to binary health questions (RQ1). We have evaluated four popular SEs and observed that the percentage of correct answers found in the SERPs is in the range of 50–70%. These figures are low and might sound highly concerning. However, this outcome is partially explained by the presence of many results that do not provide an answer. By focusing on the retrieved pages that provide an unequivocal answer to the reference health questions, we obtained much higher precision. Still, SE companies have room for improvement, as many top-ranked web pages do actually contain harmful health recommendations (10–15% of wrong answers). This is a natural consequence of the open and unmoderated nature of the web, and we encourage developers of search technologies to further advance in the removal of low-quality content from their indexes and SERPs.

Regarding the presence of correct and incorrect answers over the ranked positions, the quality of the responses does not seem to decrease as we go deeper in the rankings (at least for the top 20 results). Our results also show that Bing seems to be the most solid choice among the four SEs. As part of the analysis, we modeled two search user behaviors: lazy and diligent. We found out that lazy behavior, based on making decisions from the first observed response, is not inferior to a more

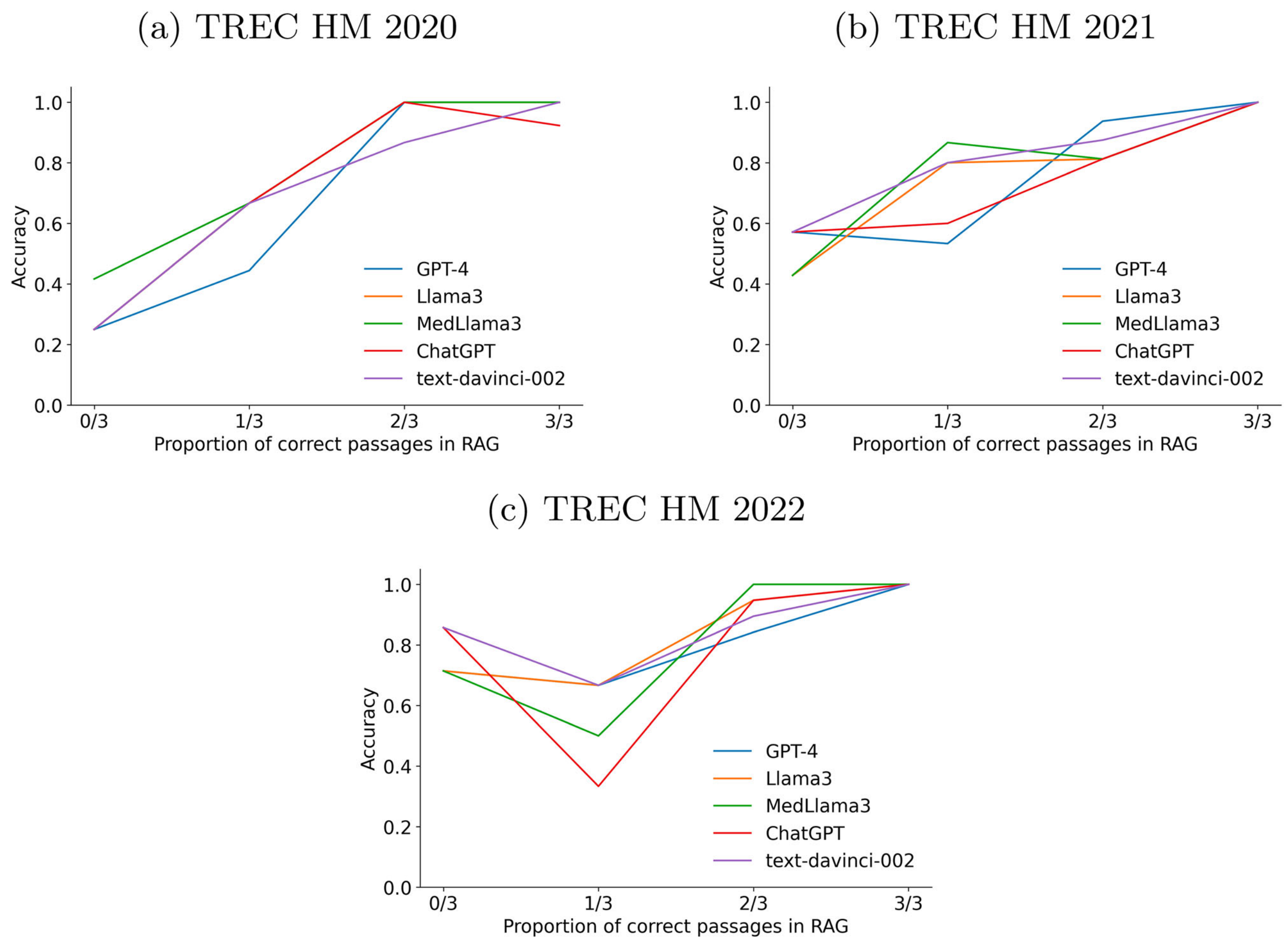

Fig. 8 | RAG experiments with 3 passages injected. Accuracy of the LLMs with varying numbers of correct passages (0/3, 1/3, 2/3, or 3/3). On each panel, GPT-4 is represented by a blue line, LLama3 by an orange line, MedLlama3 by a green line, ChatGPT by a red line, and text-davinci-002 by a purple line. **a** Plots results for the TREC HM 2020 collection. **b** Plots results for the TREC HM 2021 collection. **c** Plots results for the TREC HM 2022 collection.

diligent method based on acquiring and comparing three responses to the health questions.

Our next goal was determining whether LLMs are reliable for providing accurate medical answers (RQ2). Our results suggest that, in general, the most capable LLMs generate better answers compared with those extracted from top web pages ranked by SEs. It seems that the extensive training data of the LLMs, coupled with their superior reasoning abilities, offer significant advantages over extracting responses from a handful of top-ranked search results. Among the largest models (GPT-4, ChatGPT, LLama3, and MedLLama3), we found no clear winner; and we observed poor performance from models such as FlanT5. Despite showing an overall good performance, there are still some barriers to adopting LLMs. For instance, under some circumstances, LLMs provide more than 30% of incorrect results.

Another concerning outcome is that the quality of the LLMs' completions in response to health questions was highly sensitive to the input prompt (RQ3). We found that some input prompts, which guide the models towards reputed sources, are much more effective than basic prompts (or prompts that give no context at all). However, lay users would hardly resort to sophisticated prompts or complex interactions with the LLMs. This suggests that the future adoption of LLMs to support QA in the health domain would need to wrap the user's questions into automatically extended contexts (or, alternatively, design health-oriented assistants that guide the AIs toward high-quality knowledge).

We also conducted a thorough error analysis and demonstrated that LLMs made errors due to a lack of medical knowledge, even with the most sophisticated prompts. However, we also discovered that search engines also fail more in these "low-performance questions", suggesting that the problem is not so much with the language models as with the questions themselves.

We also discovered that retrieval-augmented generation is promising for answering health questions (RQ4). We demonstrated that smaller LLMs reach the level of superior models when grounded in evidence provided by a search engine. This opens up the debate on whether it is worthwhile to persist in generating massive and computationally demanding models, or alternatively, we can direct our efforts towards leveraging lighter models enriched with search evidence. Moreover, we also demonstrated that the correctness of the passages used for augmenting the language model plays a critical role in its final performance.

In the literature, some teams focused their efforts on designing solutions that estimate the correctness of medical information. For instance, Vera[36] is a transformer-based ranker that achieved state-of-the-art performance in health misinformation detection tasks. This model was fine-tuned with assessments from TREC 2019's Decision Track and it achieved the best results in TREC 2022's Health Misinformation Track. This shared-data task fosters the development of systems prioritizing accurate and trustworthy documents over misinformation. However, systems like Vera take a health question and its correct response as input and then search for harmful or helpful documents. These technological tools could support, for example, moderation services for online platforms. However, the need for pairs of questions and correct responses represents a limitation. Thus, the creators of Vera also conducted research on how to automatically infer the correct response for a health question[37]. Specifically, the authors evaluated two approaches: (i) using LLMs under different settings (zero-, few-shot, and chain-of-thought), and (ii) using the ranking produced by Vera and averaging the decisions extracted from the top 50 retrieved results. Their results suggest that the LLM-based approach (powered by GPT-4) outperformed the rest of the strategies. Our comparison of SEs and LLMs is related to this study, but we compared multiple commercial search engines (while Vera is not a tool that is widely available to the public). Furthermore, we do not only evaluate GPT models but also consider seven LLMs of different families. Our study, therefore, is reflective of the type of answers that the general public might get when interacting with popular information access tools. Additionally, we compared the answering capabilities of the LLMs with and without search results provided by the web search engines. This is an aspect that has not received enough attention in health information seeking. We also evaluated the effectiveness of the answers provided as we went down in the ranked lists of the SERPs.

We are aware that our study presents some limitations. For instance, the automatic extraction of answers from specific passages for evaluating search engines can be prone to error. However, the passage retrieval and reading comprehension stages were based on state-of-the-art technologies that were effectively tested elsewhere (see "Methods"). We do not claim that the process is error-free but we are confident about the robustness of the trends found. In any case, in the near future, we intend to further validate these findings by involving human evaluators to manually annotate the correctness of the search results. This is, however, a costly process that is not exempt from problems. Moreover, our current experiments did not consider the effect of personalization. Retrieval results are known to be dependent on multiple user factors (e.g., geolocalization or user preferences) and it will be important to study the relative quality of the medical responses taking into account geographical factors or other user-dependent variables. Related to this, it will be interesting to study the relative adoption of LLMs and SEs by different groups of the population (e.g., educated/uneducated or wealthy/low-income). Access to free vs premium services provided by these platforms might represent a major factor in access to high-quality health-related information. We think that this adoption should be scrutinized to ensure that the health gap between vulnerable people and other segments of the population does not enlarge.

Some research teams focused on other crucial aspects, such as readability or understandability. For example, Yan et al.[38] argued that search results should be re-ranked by descending readability and proposed several readability formulas for health and medical information retrieval. Related to this, Zuccon and Koopman proposed an innovative method that integrates the notion of understandability into the evaluation of retrieval systems for consumer health search[39]. In our study, we did not specifically analyze the understandability or readability of search results (or LLM's completions). These factors are known to affect user experience and we plan to extend our research to study the interaction between these two dimensions and health-related decisions.

About LLM evaluation, we are aware that using proprietary models poses the difficulty of reproducing these experiments. Their architectural design is unknown, the training data is not disclosed and, often, these are black-box models in constant evolution. However, we must acknowledge that these models are currently being used by millions of people worldwide and it is of paramount importance to assess their answering capabilities in the health domain. Note that SEs are also closed systems whose core elements (e.g., ranking algorithms) are largely unknown, and additionally, their indexes undergo continuous updates. Our experiments and comparison, therefore, give a specific picture of LLMs and SEs at a certain point in time. The forthcoming evolution of these systems will call for new trials and experiments to supplement the present study. In any case, we facilitate the reproducibility of our experiments by providing the code (see "Code availability"), the generated outputs, and all the dates of execution of the experiments.

Another limitation affects the type of information needed. In these experiments, we restricted ourselves to binary question answering. This was a practical decision because we then were dealing with yes/no responses whose correctness could be automatically assessed. This represents a valuable first step towards understanding the relative effectiveness of SEs and LLMs in the health domain. In the near future, we would like to explore other classes of information needs that require more elaborate answers.

Methods

Binary question answering

In this research, we focus on estimating the ability of web search engines and LLMs to retrieve correct health responses. To that end, we selected a binary health question answering (QA) task as our reference for evaluation. This type of information needs is prevalent, as many users usually search for specific advice, e.g., "Can X (treatment) cure Y (disease)?". This binary setting facilitates the quantitative assessment of the system's responses. Automatic systems must provide the correct answer to these health questions and, as shown below, the target or reference response comes from the established medical consensus for each search topic.

To obtain a solid set of health-related search topics we leveraged the data created under the TREC Health Misinformation Track. This is a three-year shared-data task that fostered the creation of systems adept at detecting false health information, thereby empowering individuals to make health-related choices grounded on reliable and factual information[40,41]. These datasets contain health-related queries posed as questions (for instance, "Does wearing masks prevent COVID-19?") and their correct responses (yes/no). Each topic represents a searcher who is looking for information that is useful for making a "yes" or "no" decision regarding a health-related information need. The binary ground truth field represents the current understanding of contemporary medical practice (compiled by the organizers of the tasks in the process of constructing the dataset). Figure 9 shows an example of a topic (the question is stored in the description field while the binary response is stored in the stance field, encoded as "helpful" or "unhelpful").

The TREC HM 2020 dataset was restricted to questions related to COVID-19, while topics from TREC HM 2021 and 2022 collections covered a wide range of health topics. The 2020 questions were disclosed in the middle of 2020, which raises the possibility that they could have been included in the pre-training phase of some LLMs. Such inclusion could give them an advantage over generative models trained earlier. The 2021 dataset was made available in the middle of 2021 and, therefore, it could have been used for training models such as GPT-4 or ChatGPT (but GPT-3, instead, ended its training earlier). The topics of 2022, on the other hand, were released after all the LLMs had been created. Therefore, this configuration of health topics shapes an assorted evaluation with questions created and released at different dates.

Summing up, the selection of health topics integrates a diverse set of binary health-related questions, which pose different degrees of difficulty related to the exposure of the models to such data and to the specificity of the information needs.

Search engines

To evaluate the responses provided by SEs, each health question was submitted to four well-known search engines: Google, Bing, Yahoo, and Duckduckgo. We used a scraping tool (see "Code availability") on the organic search results to collect the top retrieved web pages. We gathered the top 20 ranked entries since users rarely go beyond the second page of results, and thus, the extraction focuses on web pages that have some chance of being inspected by a standard user. Then, we obtained the raw content of the webpage, converted it to workable plain text, and split it into passages. To obtain on-topic passages, the health question and each passage were embedded into a vectorial contextual representation using MonoT5[42], a highly effective model fine-tuned for passage retrieval. This constitutes a well-established approach for the automatic extraction of excerpts answering user queries[43]. Rosa and colleagues[44] demonstrated the effectiveness of this model for several retrieval tasks.

After the passages were chosen, it was necessary to ascertain whether they provided an affirmative, negative, or non-responsive answer to the health question. To that end, we exploited the GPT-3 model's capabilities to perform reading comprehension. Brown et al.[34] evaluated the ability of this

Fig. 9 | Example of a health question. Topic from the TREC health misinformation track. The data fields used in the experiments are presented in red boxes.

```
<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>
```

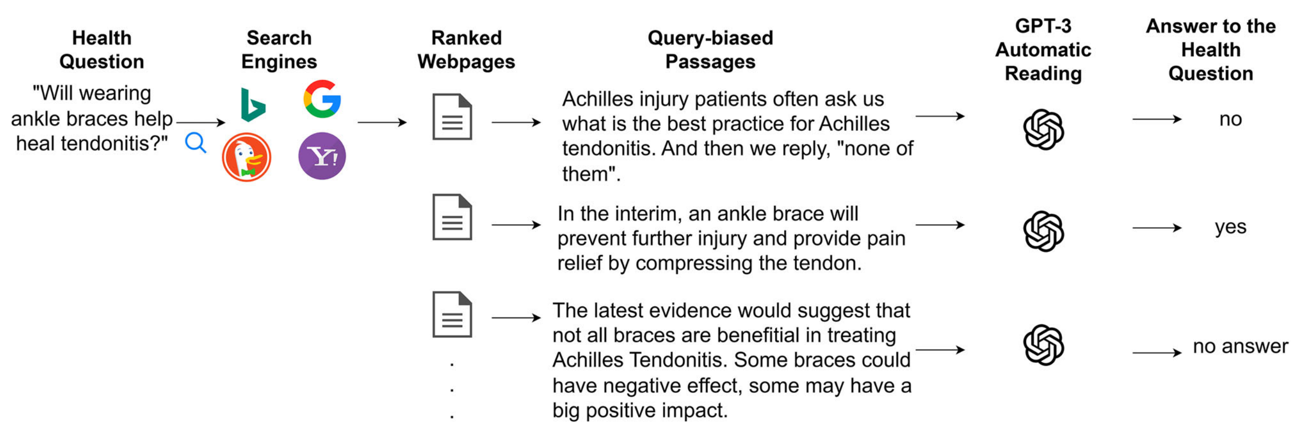

Fig. 10 | Extracting responses for health questions from the search engine results. The passages shown are real passages obtained from the top search results for the question "Will wearing ankle braces help heal tendonitis?". The process consists of: first, retrieving the top 20 results for the query (provided by each search engine); second, selecting the most relevant passages from the retrieved pages (query-biased passages, using MonoT5); third, asking GPT-3 to read each passage and determine if it answers 'yes', 'no', or does not provide an answer to the health question.

model in different reading comprehension settings. The model achieved remarkable results, for example, F1 of 85% in the CoQA dataset. Dijkstra and others[45] also demonstrated this model's ability to perform reading comprehension tasks in the educational domain. In our case, we defined a prompt that included the passage and health question, and specifically tasked the model with answering the question based on the provided passage (without resorting to its internal knowledge). The following prompt template was used: " <Passage>. Based on the previous text, answer 'yes', 'no' or 'no answer provided' to the following question: <health question> ". This estimation process arguably produces the sequence of answers that a search engine user would obtain from inspecting the SE results. Figure 10 illustrates the whole process.

Note that a retrieval result might not provide an answer to the question. In our accounting of correct answers across ranking positions, these cases are recorded as failures, as they do not respond with the correct answer. It is important, though, to bear in mind that there is an important distinction between a non-answer and an incorrect answer. We will further delve into this issue in the future, but in the current study, we focused on analyzing the relative trends of correct responses.

Our interest here is not only to analyze the overall effectiveness of web search systems but also to study the quality of the responses as we go deeper into the ranked lists. To that end, we considered two models of user behavior that simulate alternative forms of inspection of the retrieved results, see Fig. 11. The lazy user model represents a user who stops inspecting results when presented with the first entry that gives a yes/no response to the user's question. Note that on-topic passages might not respond to the user's question, and thus, the lazy user model does not necessarily stop at the top 1 entry. This user, therefore, sticks to the first answer found and does not spend time searching for contrasting evidence. We therefore assume that this user believes the answer provided by the passage. This class of user behavior (blindly believing the first response found) is clearly suboptimal, given the high presence of misinformation on the web.

The second model, referred to as the diligent user model, represents a user who traverses the ranking from the top position and stops after finding three responses. We assume that this user makes a decision about his health question based on majority voting from the three provided responses.

Under these models, the user's decision is the binary resolution the user takes after seeing the first answer (lazy model) or after seeing three answers (diligent model). It might be helpful or harmful, depending on the correctness/incorrectness of the documents inspected.

Web users are known to be reluctant to explore many search results, and these two models represent two natural forms of exploration in the quest for the response to the health question. For future work, we leave the study of additional user models, including click models for web search[46].

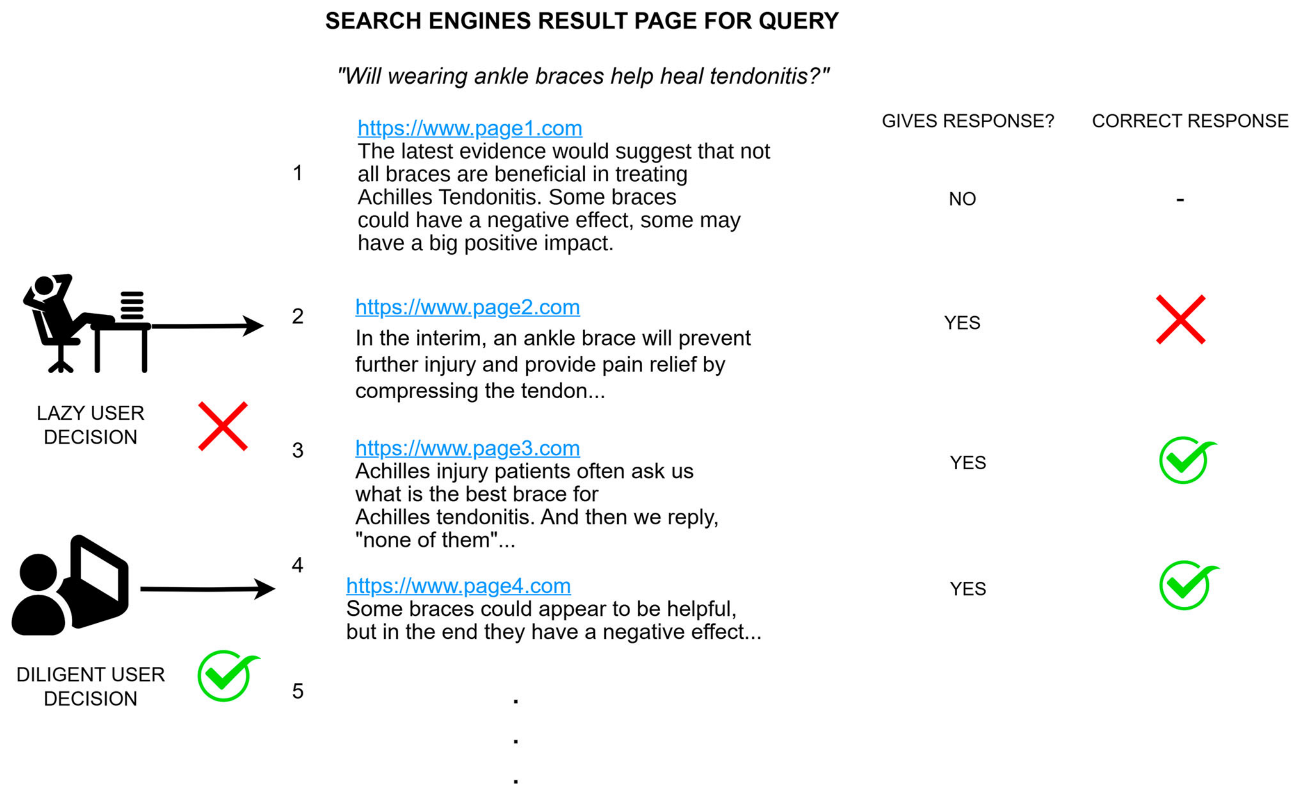

Fig. 11 | Example of the user behavior models. The lazy user makes a wrong decision because he stops right after seeing the first response (second document). In contrast, the diligent user makes the right decision after reading three responses (four documents). Note that an entry can give no response to the question and that the provided responses can also contain incorrect information.

Large language models

To assess the LLMs, we submitted the same health question to several generative models and obtained their completions. We worked under different settings, including zero- and few-shot strategies (providing examples of correct responses to other health questions), and different types of prompts.

Specifically, we evaluated seven models of different nature and architecture (closed and open source), including a model fine-tuned to the clinical domain. The fine-tuned model helps to test the influence of in-domain data in accurately answering health questions. The models included in our comparative study are:

- GPT-3, also known as text-davinci-002 (d-002). This model has 175 billion parameters and a decoder-only structure. It was extensively trained with corpora from diverse sources (including a full Wikipedia corpus). The training information goes up to June 2021.
- text-davinci-003 (d-003), the subsequent iteration of GPT models, built upon its predecessor by incorporating InstructGPT methodologies[47]. This model has been built through reinforcement learning with human feedback (RLHF) and has the same training data timeline as d-002.
- ChatGPT represents an evolution of OpenAI models towards a more dialog-oriented and user-friendly behavior. Its knowledge cutoff goes until September 2021. We experimented with the GPT-3.5-turbo version (from June 2023).
- GPT-4, another conversational agent by OpenAI represents a significant leap forward, outperforming ChatGPT in complex tasks that demand human-like reasoning, such as solving academic exercises[48]. The training examples of GPT-4 were collected until September 2021.
- Flan T5 (FT5), a sequence-to-sequence model from Google. This model underwent fine-tuning with a diverse array of instruction-based datasets[1]. The data for this model was sourced from the "Flan 2022" open-source repository, which includes online content collected until the end of 2022. For our experimentation, we worked with flan-t5-xl.
- Llama3, the most recent language model developed by Meta AI. Llama3's pretraining data covers until the end of 2023. It also underwent several post-training rounds, aligning the model with human feedback. Each post-training round included supervised fine-tuning and preference optimization through reinforcement learning. For these experiments, we used the version named llama3.1-8b-instruct.
- MedLlama3, a fine-tuned Llama3 with medical instruction data[49]. It was specially tuned for user alignment in healthcare settings. We used the Med42-Llama3-8B version, released in August 2024.

The four OpenAI models were tested through its official Python API, while Llama3, MedLlama3, and Flan T5 were executed through the Ollama and HuggingFace APIs, respectively (see "Code availability" for further details). We fixed the temperature of the models to 0, with the intention of reducing the randomness or creativity of the completions.

A pivotal part of this research consists of determining the effectiveness of these models for answering health questions under different input conditions or contexts. Having in mind that online users are known to be reluctant to have complex interactions with automated systems, we first tested the LLMs' effectiveness when responding to non-expert individuals who send the question and give little or no context at all:

- No-context prompt: a context formed only with the health question, i.e., "Can Vitamin D cure COVID-19?".
- Non-expert prompt: the text "I am a non-expert user searching for medical advice online" is added before the health question. This prompt intends to mirror a lay user seeking medical counsel.

As a next step, we tested more elaborate prompts and we also assessed the influence of including some in-context examples. While it may be uncommon for a typical user to adopt these sophisticated strategies, they nonetheless offer valuable insights into understanding and leveraging the inherent knowledge of the models:

- Expert prompt: 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." followed by the corresponding medical question. These contextual instructions were produced by the research team from Waterloo University for their TREC 2022 HM experiments[37]. The core idea is to guide the LLM towards renowned sources.

To automatically record responses, we forced the models to answer only with "yes" or "no" tokens. In contrast to the scenario with search engines, there can, therefore, be no unanswered questions here. More complex prompt engineering alternatives, like chain-of-thought (CoT)[50], were left for future work.

For the few-shot experiments, we selected three random pairs from the TREC HM 2021 dataset. The selected question-answer pairs were: ("Does yoga help manage asthma?", "Yes"), ("Will wearing an ankle brace help heal Achilles tendonitis?", "No"), and ("Is starving a fever effective?", "No").

In this study, we also performed an "online" retrieval augmented generation experiment[51]. There are several approaches for augmenting generative language models with retrieved evidence. Following standard practice, we injected textual chunks in the input layer of the LLM[52]. In essence, we prompted different large language models with passages extracted from the top entries of the SERPs. In this way, our experimental setup allows the comparison of three classes of outputs (SE alone, LLM alone, and an RAG variant that injects real-time SE results into the input of the LLM). For these RAG experiments, we fed some LLMs with on-topic passages obtained from the search results produced by Google. We injected passages from Google's top five results and we clearly instructed the LLM to compare the provided evidence with its internal knowledge prior to providing a definitive answer. The following prompt was used for RAG: "Provide an answer to the question using the provided evidence and contrasting it with your internal knowledge. Evidence: <evidence from search engine>. Question: <query>. Your answer:". The extraction of on-topic passages from Google's top five results was also supported by MonoT5.

Data availability

All data used in this study is public and it can be found in the Text Retrieval Conference (TREC) repository (https://trec.nist.gov/data.html).

Code availability

All code used in this experimentation is openly published under a GPLv3 License (https://github.com/MarcosFP97/information-seeking-health). Special mention to the open-source scraping tool used to evaluate the search engines (https://github.com/tasos-py/Search-Engines-Scraper). The four OpenAI models were tested through its official Python API (https://openai.com/blog/openai-api), while Llama3 (https://ollama.com/library/llama3.1:8b-instruct-q8_0, a quantized version of eight bits of the model to run on a CPU), MedLlama3 (https://ollama.com/thewindmom/llama3-med42-8b, quantized version, eight bits) and Flan T5 (https://huggingface.co/google/flan-t5-xl) were executed through the Ollama and HuggingFace APIs, respectively.

Received: 1 October 2024; Accepted: 27 February 2025;
Published online: 10 March 2025

References

1. Longpre, S. et al. The flan collection: designing data and methods for effective instruction tuning. In *International Conference on Machine Learning* 22631–22648 (PMLR, 2023).
2. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://doi.org/10.48550/arXiv:2303.12712 (2023).
3. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://doi.org/10.48550/arXiv:2307.09288 (2023).
4. Mao, K., Dou, Z., Chen, H., Mo, F. & Qian, H. Large language models know your contextual search intent: a prompting framework for conversational search. Preprint at https://doi.org/10.48550/arXiv:2303.06573 (2023).
5. Friedman, L. et al. Leveraging large language models in conversational recommender systems. Preprint at https://doi.org/10.48550/arXiv:2305.07961 (2023).
6. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering—example of ChatGPT. *Nat Commun.* **15**, 1569 (2024).
7. O'Leary, D. E. Massive data language models and conversational artificial intelligence: emerging issues. *Intell. Syst. Account. Financ. Manag.* **29**, 182–198 (2022).
8. Hersh, W. Search still matters: information retrieval in the era of generative AI. *J. Am. Med Inform. Assoc.* **31**, 2159–2161 (2024).
9. Wu, B., Liu, Q. B., Guo, X. & Yang, C. Investigating patients' adoption of online medical advice. *Decis. Support Syst.* **176**, 114050 (2024).
10. Sivarajkumar, S. et al. Clinical information retrieval: a literature review. *J. Healthc. Inform. Res.* **8**, 313–352 (2024).
11. Wang, L. et al. Using internet search engines to obtain medical information: a comparative study. *J. Med. Internet Res.* **14**, e74 (2012).
12. Zuccon, G., Koopman, B. & Palotti, J. Diagnose this if you can: on the effectiveness of search engines in finding medical self-diagnosis information. *Adv. Inf. Retr.* **9022**, 562–567 (2015).
13. Jiang, Z., Xu, F. F., Araki, J. & Neubig, G. How can we know what language models know? *Trans. Assoc. Comput. Linguist.* **8**, 423–438 (2020).
14. Liang, P. et al. Holistic evaluation of language models. Preprint at https://doi.org/10.48550/arXiv:2211.09110 (2022).
15. Chang, Y. et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **15**, 1–45 (2024).
16. 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. *Fertil. Steril.* **120**, 575–583 (2023).
17. Duong, D. & Solomon, B. D. Analysis of large-language model versus human performance for genetics questions. *Eur. J. Hum. Genet.* **32**, 466–468 (2023).
18. Holmes, J. et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front. Oncol.* **13**, 1219326 (2023).
19. Elgedawy, R., Srinivasan, S. & Danciu, I. Dynamic Q&A of clinical documents with large language models. Preprint at https://doi.org/10.48550/arXiv:2401.10733 (2024).
20. Kim, H.-W., Shin, D.-H., Kim, J., Lee, G.-H. & Cho, J. W. Assessing the performance of chatgpt's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. *Seizure. Eur. J. Epilepsy* **114**, 1–8 (2024).
21. Kim, J. et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Med.* **7**, 193 (2024).
22. Jahan, I., Laskar, M. T. R., Peng, C. & Huang, J. Evaluation of ChatGPT on biomedical tasks: a zero-shot comparison with fine-tuned generative transformers. Preprint at https://doi.org/10.48550/arXiv:2306.04504 (2023).
23. Hamidi, A. & Roberts, K. Evaluation of AI chatbots for patient-specific EHR questions. Preprint at https://doi.org/10.48550/arXiv:2306.02549 (2023).
24. Samaan, J. S. et al. Assessing the accuracy of responses by the language model chatgpt to questions regarding bariatric surgery. *Obes. Surg.* **33**, 1790–1796 (2023).

25. Tang, L. et al. Evaluating large language models on medical evidence summarization. *NPJ Digital Med.* **6**, 158 (2023).
26. Johnson, D. et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. Preprint at https://doi.org/10.21203/rs.3.rs-2566942/v1 (2023).
27. Thirunavukarasu, A. J. et al. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med. Educ.* **9**, e46599 (2023).
28. Kusa, W., Mosca, E. & Lipani, A. "Dr LLM, what do I have?": the impact of user beliefs and prompt formulation on health diagnoses. In *Proc Third Workshop on NLP for Medical Conversations* 13–19 (MDPI, 2023).
29. Caramancion, K. M. Large language models vs. search engines: evaluating user preferences across varied information retrieval scenarios. Preprint at https://doi.org/10.48550/arXiv:2401.05761 (2024).
30. Oeding, J. F. et al. ChatGPT-4 performs clinical information retrieval tasks utilizing consistently more trustworthy resources than does Google Search for queries concerning the Latarjet procedure. *Arthroscopy* **41**, 588–597 (2024).
31. Li, Y. et al. ChatDoctor: a medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus* **15**, e40895 (2023).
32. Koopman, B. & Zuccon, G. Dr ChatGPT tell me what I want to hear: how different prompts impact health answer correctness. In *Proc 2023 Conference on Empirical Methods in Natural Language Processing* 15012–15022 (Association for Computational Linguistics, 2023).
33. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics* ACL 6233–6251 (2024).
34. Brown, T. et al. Language models are few-shot learners. *Adv. neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
35. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).
36. Pradeep, R., Ma, X., Nogueira, R. & Lin, J. Vera: prediction techniques for reducing harmful misinformation in consumer health search. In *Proc 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2066–2070 (Virtual Event, Canada, 2021).
37. Pradeep, R. & Lin, J. Towards automated end-to-end health misinformation free search with a large language model. In: *Advances in Information Retrieval* 78–86 (Springer Nature, Switzerland, 2024).
38. Yan, X., Song, D. & Li, X. Concept-based document readability in domain specific information retrieval. In *Proc 15th ACM International Conference on Information and Knowledge Management* 540–549 https://doi.org/10.1145/1183614.1183692 (CIKM, 2006).
39. Zuccon, G. & Koopman, B. Integrating understandability in the evaluation of consumer health search engines. *CEUR Workshop Proc.* **1276**, 32–35 (2014).
40. Clarke, C., Maistro, M., Smucker, M. & Zuccon, G. Overview of the TREC 2020 health misinformation track. In *Proc Twenty-Nine Text REtrieval Conference, TREC,* 1266 (NIST Special Publication, 2020).
41. Clarke, C., Maistro, M. & Smucker, M. Overview of the TREC 2021 health misinformation track. In *Proc Thirtieth Text REtrieval Conference, TREC,* 500–335 (NIST Special Publication, 2021).
42. Pradeep, R., Chen, H., Gu, L., Tamber, M. S. & Lin, J. Pygaggle: a gaggle of resources for open-domain question answering. In: *Advances in Information Retrieval 45th European Conference*, Vol. 13982, 148–162 (Springer, 2023).
43. Nogueira, R. & Cho, K. Passage re-ranking with BERT. Preprint at https://doi.org/10.48550/arXiv:1901.04085 (2019).
44. Rosa, G. M. et al. No parameter left behind: how distillation and model size affect zero-shot retrieval. Preprint at https://doi.org/10.48550/arXiv:2206.02873 (2022).
45. Dijkstra, R., Genç, Z., Kayal, S. & Kamps, J. et al. Reading comprehension quiz generation using generative pre-trained transformers. *CEUR WS Proc.* **3192**, 4–17 (2022).
46. Chuklin, A., Markov, I. & de Rijke, M. *Click Models for Web Search* (Springer, 2015).
47. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
48. OpenAI. Gpt-4 technical report. Preprint at https://doi.org/10.48550/arXiv:submit/4812508 (2023).
49. Christophe, C., Kanithi, P. K., Raha, T., Khan, S. & Pimentel, M. A. Med42-v2: a suite of clinical LLMS. Preprint at https://doi.org/10.48550/arXiv:2408.06142 (2024).
50. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
51. Asai, A., Min, S., Zhong, Z. & Chen, D. Retrieval-based language models and applications. In *Proc 61st Annual Meeting of the Association for Computational Linguistics* 41–46 (Association for Computational Linguistics, 2023).
52. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M. Retrieval augmented language model pre-training. *International Conference on Machine Learning* 3929–3938 (2020).

Acknowledgements

The authors thank the financial support supplied by the Xunta de Galicia—Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19) and the ERDF, which acknowledges the CiTIUS Research Center in Intelligent Technologies of the USC as a Research Center of the Galician University System. They also thank the financial support obtained from (i) project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund) and (ii) project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU).

Author contributions

All authors contributed equally to this work as co-senior authors. Conceptualization: M.F., J.C., and D.L. Experimentation: M.F. Visualization: M.F., J.C., and D.L. Writing—original draft preparation: M.F., J.C., and D.L. Writing—review and editing: M.F., J.C., and D.L.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Marcos Fernández-Pichel.