Can Large Language Models Match the Conclusions of Systematic Reviews?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Systematic reviews (SR), in which experts summarize and analyze evidence across individual studies to provide insights on a specialized topic, are a cornerstone for evidence-based clinical decision-making, research, and policy. Given the exponential growth of scientific articles, there is growing interest in using large language models (LLMs) to automate this process. However, the ability of LLMs to critically assess evidence and reason across multiple documents to provide expert-quality observations remains poorly characterized. We therefore ask: **Can LLMs match the conclusions of systematic reviews written by clinical experts when given access to the same studies?** To explore this question, we present MedEvidence, a benchmark pairing findings from SRs with the studies they are based on. We benchmark 24 LLMs on our MedEvidence dataset, including reasoning, medical specialist, and models of varying sizes. We find that reasoning does not necessarily improve performance, larger models do not consistently yield greater gains, and knowledge-based fine-tuning tends to degrade accuracy on MedEvidence. Instead, most models exhibit similar behavior: performance tends to degrade as token length increases, their responses show overconfidence, and all models show a lack of scientific skepticism toward low-quality findings. These results suggest that more work is still required before LLMs can reliably match the observations from expert-conducted SRs, even though these systems are already deployed and being used by clinicians.

1 Introduction

As the number of published articles grows exponentially [1], manually synthesizing findings from multiple sources has become highly time-consuming. Thus, there is growing interest in developing automatic tools to process, synthesize, and extract insights from scientific literature [2, 3]. In particular, large language model (LLM)-based systems could offer a promising solution for supporting and automating tasks such as conducting systematic reviews (SRs). For example, several LLM-assisted tools such as Deep Research [4, 5], Elicit [6], and Open Evidence [7], have already been deployed. The momentum behind these technologies is further exemplified by the U.S. Food and Drug Administration's launch of an LLM-assisted scientific review pilot on May 2025 [8].

However, despite multiple deployments and efforts assessing scientific synthesis generation, the behavior of LLMs across key variables that influence generation remains poorly understood. In particular, their ability to synthesize findings from multiple studies—each varying in study type, population size, and risk of bias—and to navigate conflicting evidence (as medical findings can often contradict one another) is not well-characterized. Understanding these behaviors is essential, as medical knowledge is continually reshaped by new clinical trials, cohort studies, and expert opinions. Thus, like medical professionals do, LLMs must be capable of integrating the latest findings (e.g.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

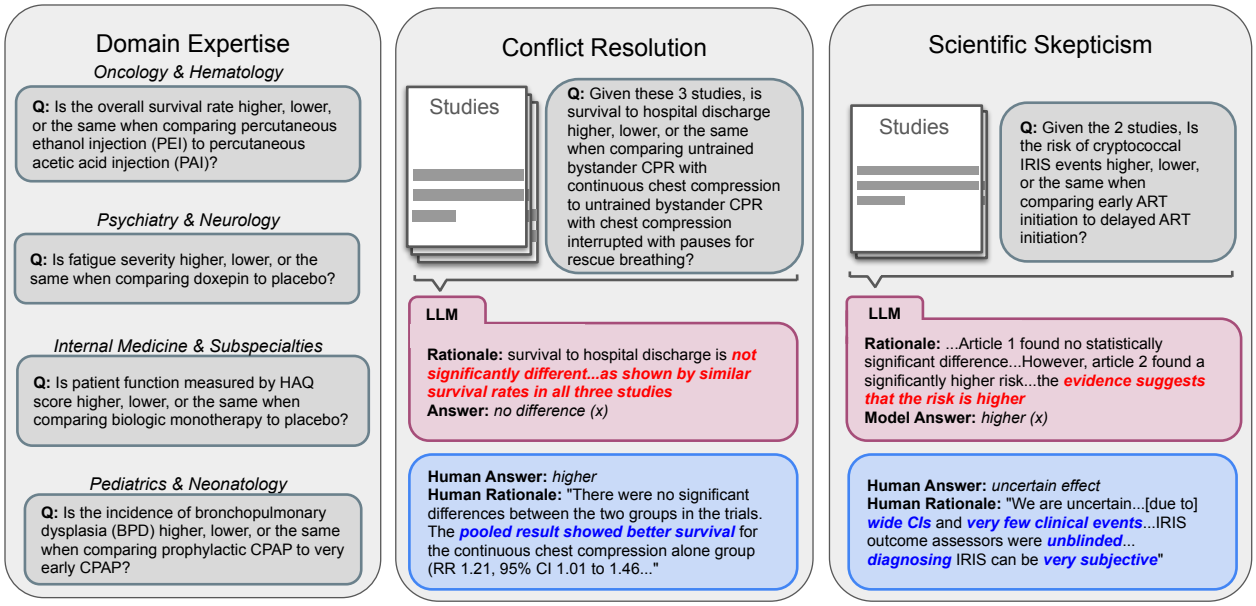

Figure 1: Core skills evaluated by MedEvidence including: medical domain expertise across 10 different specialties, synthesizing conflicting evidence, and applying scientific skepticism when studies exhibit a high risk of bias (e.g. due to small sample sizes or insufficient supporting evidence).

via retrieval augmentation) [9], weighing the strength of varying evidence, and applying appropriate skepticism when needed to produce reliable, up-to-date recommendations (as shown in Figure 1).

While prior work has successfully evaluated LLMs on their internal "static" medical knowledge [10, 11], assessing LLMs' capability to reason across multiple sources and draw expert-level conclusions remains a significant challenge. Specifically, previous efforts have often evaluated LLMs' ability to generate summaries on a given topic. This approach requires a thorough review of every detail in the generated content and lacks easily verifiable ground truth; therefore, medical experts are typically needed to assess output accuracy [12, 13, 14, 15, 16], making evaluation time-consuming and hard to scale. To address this, we remove the complexity of evaluating long-format summaries and retrieving relevant papers to pose an even simpler, but fundamental question: **Can LLMs replicate the individual conclusions of expert-written SRs when provided with the same source studies?** We explore this question in a controlled setting by collecting open-access SRs along with their associated reference articles. We then extract individual findings and reformat them into a closed question-answering (QA) task to enable straightforward evaluation. This allows us to test whether LLMs, when provided with the same evidence selected by experts, can reproduce each conclusion. To this end, we introduce the following contributions:

- **MedEvidence Benchmark** We introduce MedEvidence, a human-curated benchmark of 284 questions curated from the conclusions of 100 open-access SRs across 10 medical specialties. Each question evaluates comparative treatment effectiveness on clinical outcomes. All questions are manually transformed into closed-form question answering to enable large-scale evaluation. In addition, human annotators extract evidence quality (based on the SR's analysis), determine whether full-text access is necessary, and collect the relevant sources needed to replicate the SR findings.

- **Large-scale evaluation on MedEvidence** We leverage MedEvidence to perform an in-depth analysis of 24 LLMs spanning general-domain, medical-finetuned, and reasoning models. By utilizing MedEvidence's metadata, we dissect and examine success and failure modes, helping to identify targeted directions for future work.

Table 1: Comparison of factuality and evidence reasoning benchmarks with medical focus. We compare MedEvidence to prior datasets across attributes relevant to systematic review-style reasoning. MedEvidence is the only dataset to satisfy all criteria.

Dataset	Size	Topic	Curation	Expert-Grounded Answer	Automated Evaluation	Multiple Sources	Evidence Quality	Source-Level Concordance
Reason et al.	4	Medicine	Human	✓	✗	✓	✗	✗
Schopow et al.	1	Medicine	Human	✓	✗	✓	✗	✗
MedREQAL	2786	Medicine	LLM	✓	✓	✗	✓	✗
HealthFC	750	Consumer Health	Human	✓	✓	✗	✓	✗
ConflictingQA	238	Multi-Domain	LLM	✗	✗	✓	✗	✓
MedEvidence	284	Medicine	Human	✓	✓	✓	✓	✓

2 Related work

An overview of related works and the key distinct contributions of our current work are summarized in Table 1.

LLM-based medical systematic review Numerous studies have explored the potential of LLMs to automate various aspects of scientific literature review, including literature search, query augmentation, screening, data extraction, bias assessment, narrative synthesis, and answering simple clinical inquiries [17, 18]. However, larger-scale evaluations of LLM-based SR or meta-analyses generation remain relatively underexplored. Reason et al. [12] examined the ability of LLMs to extract numerical data from abstracts and generate executable code to perform meta-analyses. While their results are promising, the study is limited to just four individual case studies. Schopow et al. [13] and Qureshi et al. [14] investigate LLM usage across a range of systematic review stages, including meta-review and narrative evidence synthesis, but also present findings on a very small-case study scale ($N < 10$) and rely on comparison to humans. Overall, these investigations have been limited in scope and require substantial amounts of review from medical experts, highlighting the need for automated benchmarks to help evaluate LLMs' progress.

Verification of medical facts derived from systematic reviews Several studies have leveraged SRs to benchmarked LLMs' ability to perform medical fact verification, where a model must decide whether to support or refute a given claim. For instance, MedREQAL [19] is an LLM-curated closed QA dataset designed to investigate how reliably models can verify claims derived from Cochrane SRs. However, it does not provide the sources used by the SRs. Instead, the dataset evaluates models on their internal knowledge, making the task a form of fact recall. HealthFC [20], on the other hand, tasks models with verifying claims analyzed by the medical fact-checking site Medizin Transparent, but it only provides pre-synthesized analysis from the web portal as evidence. In contrast to real SR generation, this task primarily involves retrieving information from a pre-synthesized source, removing the real complexity of reasoning across unsynthesized evidence. Unlike prior work, MedEvidence requires extracting, reasoning over, and synthesizing relevant information across single or multiple sources (each with different levels of evidence) to match the expert-derived conclusion of a SR (without access to the original SR itself). It resembles the intricacies of SR analysis, as the raw sources (articles/abstracts) are directly provided to the model.

LLM Behavior in the Presence of Conflicting Sources ConflictingQA [21] examines how models respond to conflicting arguments supporting or refuting a claim. However, it focuses on inherently contentious questions without definitive answers, spans domains beyond medicine, and uses diverse online sources rather than peer-reviewed literature. ClashEval [22] investigates conflicts between a model's internal knowledge and external evidence, including a drug-related (medical) subset, but limits evaluation to single-source conflicts with artificially perturbed values. ConflictBank [23] and KNOT [24] assess model performance on specific conflict types—such as temporal inconsistencies, misinformation, and logic-based contradictions—but rely on factoid-style questions sourced from Wikipedia.These benchmarks only leverage relatively small and synthesized inputs.

To the best of our knowledge, no existing studies or datasets provide richly annotated data to systematically benchmark models' ability to align with the conclusions of medical systematic reviews while using the same underlying research documents as the original medical experts.

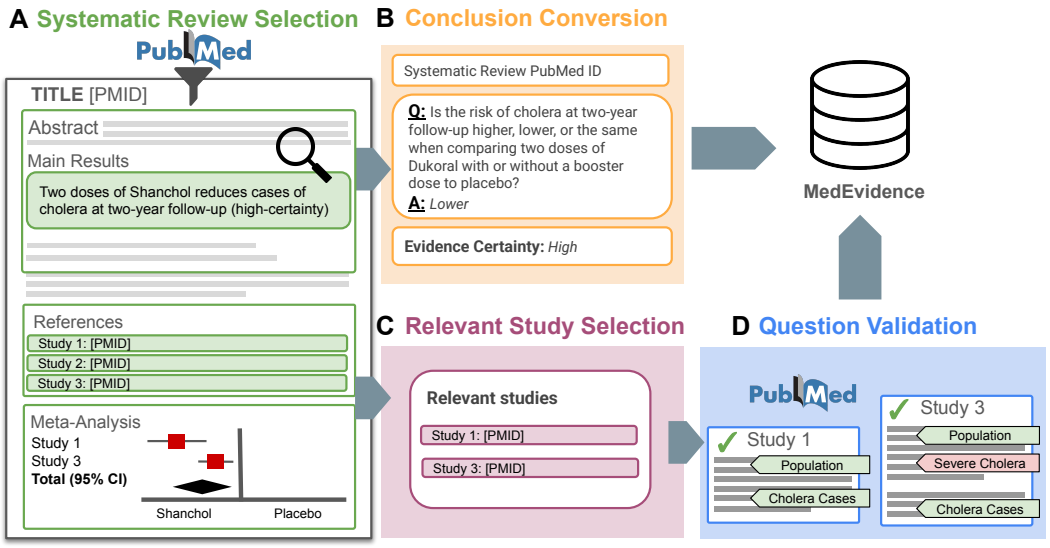

Figure 2: Overview of the dataset curation process for MedEvidence.

3 Dataset Curation Process

Data provenance We collect open-source systematic reviews, available via PubMed, conducted by Cochrane, an international non-profit organization dedicated to synthesizing evidence on healthcare interventions through contributions from over 30,000 volunteer clinician authors [25]. Cochrane is a long-standing and widely respected source of clinical evidence [26, 27], offering open-access content and analyses presented in a standardized format. Additionally, for each SR, we collect all the cited studies that are relevant for a given conclusion (we refer to these studies as 'sources'). When the source article's full text is available (i.e. the article is open-source), we obtain it using the existing BIOMEDICA dataset [28]; otherwise, abstracts are retrieved directly via PubMed's Entrez API [29]. All retrieved full-text articles use a CC-BY 4.0 license, which allows for re-distribution.

Dataset curation pipeline The core challenge in creating our dataset is ensuring that an LLM is provided with sufficient information to reproduce a given conclusion. To ensure a high-quality dataset, we developed a four-stage pipeline consisting of: (1) systematic review selection, (2) conclusion to questions conversion, (3) relevant study selection, and (4) question feasibility validation (as shown in Figure 2).

1. **Systematic review selection** We use Entrez to retrieve all Cochrane SRs published between January 1, 2014 to April 4, 2024 [30]. We only include systematic reviews for which all sourced studies are indexed in PubMed (with at least an abstract available). We additionally retrieve all data and metadata for the sourced studies, including: full-text via BIOMEDICA (when it is available), abstract, mesh terms, title, and publish date.

2. **Conclusion to question conversion.** Cochrane reviews follow a standardized format, allowing for a systematic conversion process. To identify potential questions, we followed the protocol below: Human annotators were instructed to review the SR abstract and examine the "Main Results" subsection (see Appendix Figure 9 for an example) to identify individual conclusive statements that statistically compare an intervention with a control group. These individual statements were then converted into question–answer pairs by the annotators, with answers belonging to a fixed set of classes. To be clear, `insufficient data` was used for statements by the SR authors explicitly indicating that no study investigated—or included sufficient data to analyze—the combination of treatment, control, and outcome; `uncertain effect` referred to cases where analysis was performed but definitive conclusions could not be made (see Appendix Section B.3 for more conversion details). Evidence certainty was extracted only when it was explicitly provided by the original SR authors, who use the standardized GRADE framework [31] to assess the quality of evidence in the included

studies. This certainty is often stated in the abstract, indicating the strength or quality of each observation.

3. **Relevant study selection** To identify relevant studies for a given SR, annotators used the analysis section provided in the appendix, which "weighs" the contributions of sources supporting each conclusion. For questions with insufficient data (where it is not possible to determine weights), reviewers were instructed to include studies cited in the SR that either (1) discuss the specified treatment and control but not the outcome, or (2) evaluate the treatment and outcome but compare against a different control.
4. **Question feasibility validation** Finally, given the question–answer pair and the source studies, annotators were tasked with determining whether the question was answerable based on the provided information. A question was considered answerable if at least 75% of the total weight in the analysis came from "valid" studies included in the meta-analysis. We define a study as "valid" if it (1) provides numerical data on both the intervention and control groups specified in the question, and (2) includes statistical or numerical details about the difference between the groups on the specified outcome—such as raw counts, p-values, confidence intervals, or risk ratios. The most common reason for discarding conclusions was when review authors pooled outcome data across studies, but the outcome was omitted or discussed without clear statistical detail in the abstracts of relevant studies.

In addition to these human-curated metadata, we use an LLMs to assess the percentage of individual source studies whose answer to the question aligns with the final answer provided in the systematic review. Thus, to calculate source-level agreement (which we call 'source concordance') we prompt DeepSeekV3 (the strongest model in our benchmark) to answer the question using only one single relevant source; the source is deemed to 'agree' with the final answer if and only if the LLM's classification with the one source matches the ground truth classification.

Medical domain taxonomy assignment To identify the relevant medical specialties in our dataset, we extract the Medical Subject Headings (MeSH terms)—a controlled vocabulary used by PubMed to index papers—from the 100 systematic reviews included in our dataset. We then feed this list into DeepSeek to generate a simplified categorization of specialties, resulting in 10 categories. Finally, we prompt DeepSeek to assign each question to the most relevant category, or to an "Other" category if no specific specialization is applicable.

4 Dataset Description

Table 2: Sample question from the dataset. Fields marked with an asterisk (*) use LLMs to assist the generation. Relevant source details are omitted here for brevity.

Question	Is stroke prevention higher, lower, or the same when comparing Transcatheter Device Closure (TDC) to medical therapy?
Answer	no difference
Relevant Sources (PubMed IDs)	22417252, 23514285, 23514286
Systematic Review (PubMed ID)	26346232
Review Publication Year	2015
Evidence Certainty	n/a
Open-Access Full-Text Needed	no
***Source Concordance**	1.0
***Medical Specialty**	Surgery

MedEvidence contains a total of 284 questions derived from 100 systematic reviews with 329 referenced individual articles, of which 114 have full-text available (see Appendix Figure 8 for a cohort diagram of the dataset). Questions were systematically collected by three human annotators with between one and five years of graduate education. Figure 3 shows the dataset distribution stratified by specialty, outcome effect, and source concordance with the expert-assessed treatment outcome effect (i.e. the correct answer). The benchmark covers topics from 10 medical specialties (e.g. public health, surgery, family medicine, etc.), five different outcome effects (`higher`, `lower`, `no difference`, `uncertain effect`, `insufficient data`), and three broad levels of concordance between the source paper and the correct answer (full agreement, no agreement, mixed agreement). Additional characteristic distributions of the dataset can be found in Appendix Figure 17.

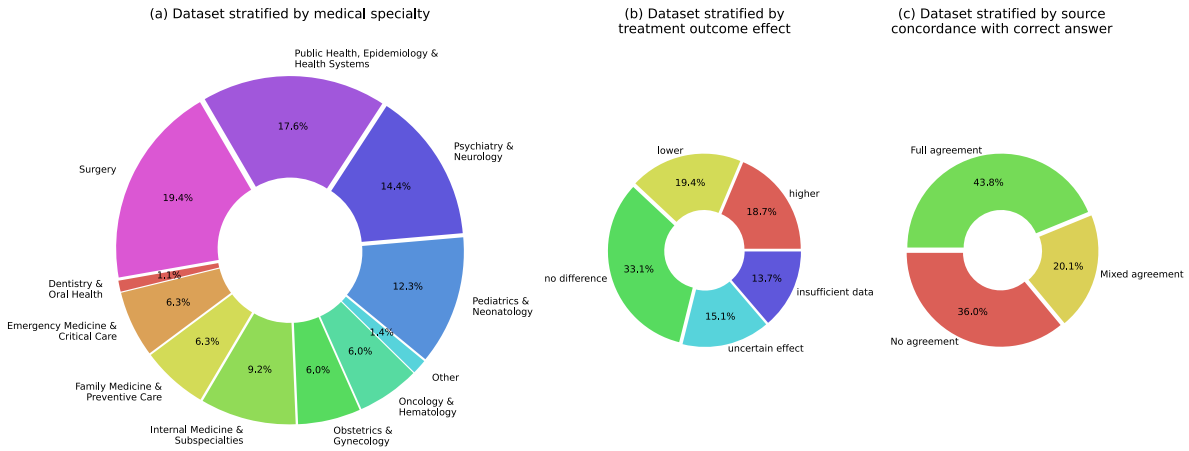

Figure 3: Key statistical characteristics of the questions in MedEvidence. (a) shows the dataset distribution stratified by medical specialty. (b) presents the distribution stratified by outcome effect. (c) shows the distribution stratified by source concordance with the expert-assessed treatment outcome effect (i.e. the correct answer).

Data format. MedEvidence is grouped by question; each question includes core data for evaluation, metadata, as well as the content details for the relevant sources. The core data consists of: a human-generated question of the form "Is [quantity of medical outcome] higher, lower, or the same when comparing [intervention] to [control]?"; the taxonomized answer to the question (`higher`, `lower`, `no difference`, `uncertain effect`, `insufficient data`); and the list of relevant studies (sources) used by the review authors to perform the analysis, identified by their unique PubMed IDs. We additionally provide the following metadata: the systematic review from which the question was extracted; the publication year of the systematic review; the authors' confidence in their analysis, also referred to as the 'evidence certainty' (`high`, `moderate`, `low`, `very low`, or `n/a` if not provided); a Boolean identification of whether full-text is available and needed to answer the question; the exact fractional source concordance; and the medical specialty associated with the question. Separately, for each source, we provide the unique PubMed ID, title, publication date if available, and content (full-text if available in PMC-OA, abstract otherwise). An individual data point example is shown in Table 2.

5 Benchmarking LLM performance

5.1 Experimental settings

LLM selection We selected 24 LLMs across different configurations, including a variety of sizes (from 7B to 671B), reasoning and non-reasoning capabilities, commercial and non-commercial licensing, and medical fine-tuning. This selection includes GPT-o1 [32], DeepSeek R1 [33], OpenThinker2 [34], GPT-4.1 [35], Qwen3 [36], Llama 4 [37], HuatuoGPT-o1 [38], OpenBioLLM [39], and more (please see Appendix Table 3 to see details of all selected models). This selection is non-exhaustive; rather, it is designed to investigate overarching trends across different model types.

Prompting setup

1. **Basic prompt** We evaluated all models in a zero-shot setting, prompting them to first provide a rationale for their answer, followed by an 'answer' field containing only one option from the list of five valid treatment outcome effects (higher, lower, no difference, uncertain effect, or insufficient data). To assess the models' "natural" behavior, we provided minimal guidance in the prompt beyond specifying the required response format, and supplied the abstracts or full text of the relevant studies as context (see Appendix Figure 11).

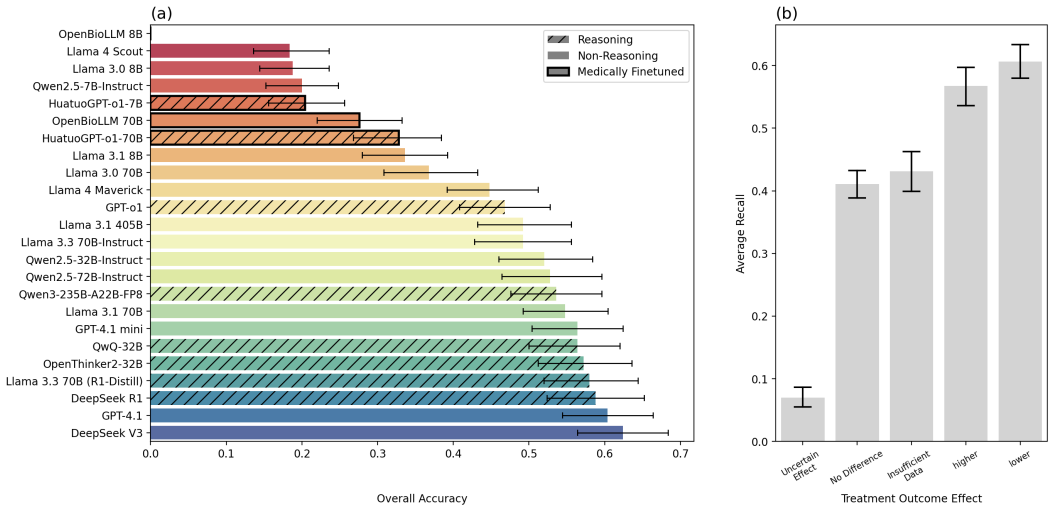

Figure 4: (a) Average model accuracy (and 95% interval) on MedEvidence. (b) Average recall by ground truth treatment outcome effect, aggregated across all models (with overall 95% interval). Per-model average recall by treatment outcome effect can be found in Appendix Figure 19.

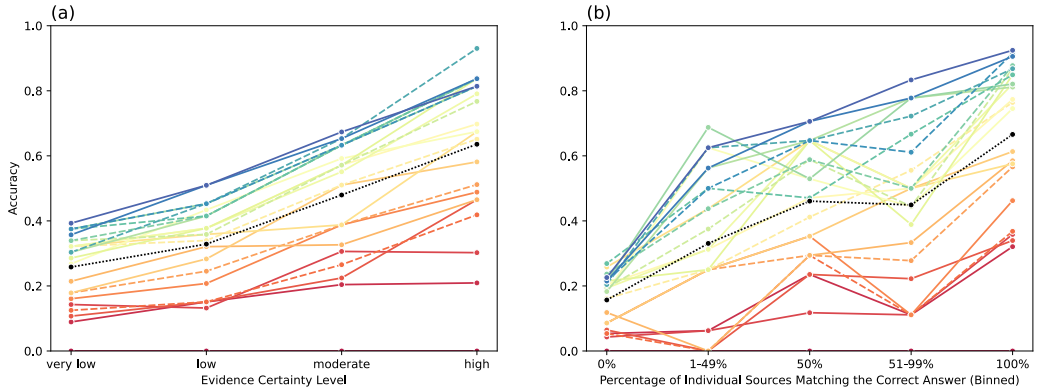

Figure 5: (a) Accuracy as a function of evidence certainty, shows a monotonically increasing trend. (b) Accuracy as a function of source concordance, defined as the percentage of relevant sources that agree with the final systematic review (SR) answer, also exhibits a monotonically increasing trend.

2. **Expert-guided prompt** LLMs may not natively understand how to handle multiple levels of evidence, which can lead to unfair evaluations. To address this, we explicitly design a prompt that instructs the LLM to summarize the study design and study population, and to assign a grade of evidence based on established definitions of grades of recommendation (see Appendix Figure 12 for the full prompt).

For both cases, if the input exceeded the LLM's context window, we used multi-step refinement (via LangChain's RefineDocumentsChain [40]) to iteratively refine the answer based on a sequence of article chunks. All models were evaluated with zero temperature to maximize reproducibility.

LLM evaluation Model performance was evaluated using accuracy based on an exact match between the answer field and the ground truth. Model outputs were lower-cased and stripped of whitespace before comparison. If no 'answer' field was provided, or if its content was not an exact rule-based match with the correct answer, the output was deemed incorrect. Confidence intervals (CIs) were calculated via bootstrap (95%, N=1000) [41].

Compute Environment Experiments were performed in a local on-prem university compute environment using 24 Intel Xeon 2.70GHz CPU cores, 8 Nvidia H200 GPUs, 16 Nvidia A6000 GPUs, and

40 TB of Storage. Large-scale models that could not be run locally in this environment were queried in the cloud using public APIs available from together.ai or OpenAI.

6 Discussion

As shown in Figure 4 (a), even frontier models such as DeepSeek V3 and GPT-4.1 demonstrate relatively low average accuracy of 62.40% (56.35, 68.45) and 60.40% (54.30, 66.50), respectively—far from saturating our benchmark. We identify four key factors that influence model performance on our benchmark: (1) token length, (2) dependency on treatment outcomes, (3) inability to assess the quality of evidence, and (4) lack of skepticism toward low-quality findings. Additionally, we found that (5) medical finetuning does not improve performance, and (6) model size shows diminishing returns beyond 70 billion parameters. We explore each of these factors in more detail below using the basic prompt setup.

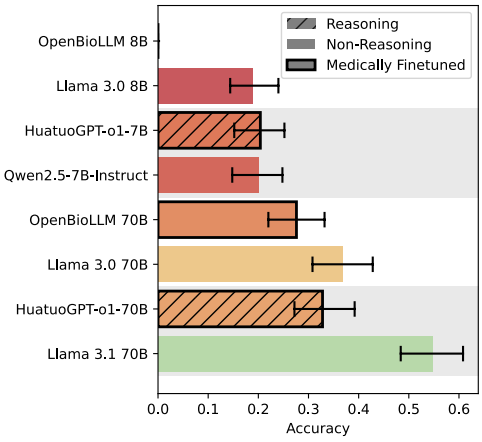

Figure 6: Medically-finetuned models vs their base generalist counterparts. Pairs of medical and base models are adjacent. 95% confidence intervals are calculated via bootstrapping with $N = 1000$.

Reasoning vs non-reasoning LLMs We highlight that, in general, reasoning models do not consistently outperform non-reasoning models of the same class or size on MedEvidence (Figure 4 (a)), as evidenced by DeepSeek V3 outperforming its reasoning counterpart (DeepSeek R1), while LLaMA 3.3 70B distilled from DeepSeek R1 outperforms the LLaMA 3.3 70B base model.

Model performance decreases as token length increases Generally, performance on MedEvidence drastically reduces as the number of tokens increases (Appendix Figure 16). Naturally, training LLMs on long contexts does not guarantee improved long-context understanding, as models may still struggle to utilize information from lengthy inputs [42, 43].

Model performance dependency on treatment outcome effect Figure 4 (b) shows the per-class recall stratified by treatment outcome effect. Overall, all models perform best on questions where the correct answer corresponds to `higher` or `lower` effects—cases where a strong stance can be taken. They are slightly less successful on `no difference` and `insufficient data` questions, where a definitive conclusion is available but there is no clear preference for either treatment. Performance is lowest on the most ambiguous class, `uncertain effect`. Notably, as shown in Appendix Figure 15, models are generally reluctant to express uncertainty, often committing to a more certain outcome that appears plausible. Notably, previous work has observed LLMs are verbally overconfident [44, 45] and shown that reinforcement learning via human feedback (RLHF) amplifies this effect [46].

Model performance improves with increasing levels of evidence We leverage the evidence certainty levels reported by experts in each systematic review (SR). As shown in Figure 5(a), the overall ability of models to match SR conclusions improves as the level of evidence increases. We therefore explore whether model performance is also associated with the level of source concordance. As shown in Figure 5(b), models' ability to match human conclusions increases as the proportion of sources agreeing with the correct answer increases (e.g., DeepSeek V3 achieves 92.45% accuracy at 100% source agreement vs. 41.21% at 0% source agreement). This suggests that, unlike human experts, current LLMs struggle to critically evaluate the quality of evidence and to remain skeptical of results. We observe that this behavior persists even when models are prompted (using the expert-guided prompt) to consider study design, population, and level of evidence (Appendix Figure 20).

Medical finetuning does not improve performance Figure 6 compares the average performance of medically finetuned models to their base model counterparts. Across all comparisons, medical finetuning fails to improve performance (even for medical-reasoning models) and, in most cases, actually degrades it. Indeed, fine-tuning without proper calibration can harm generalization, some-

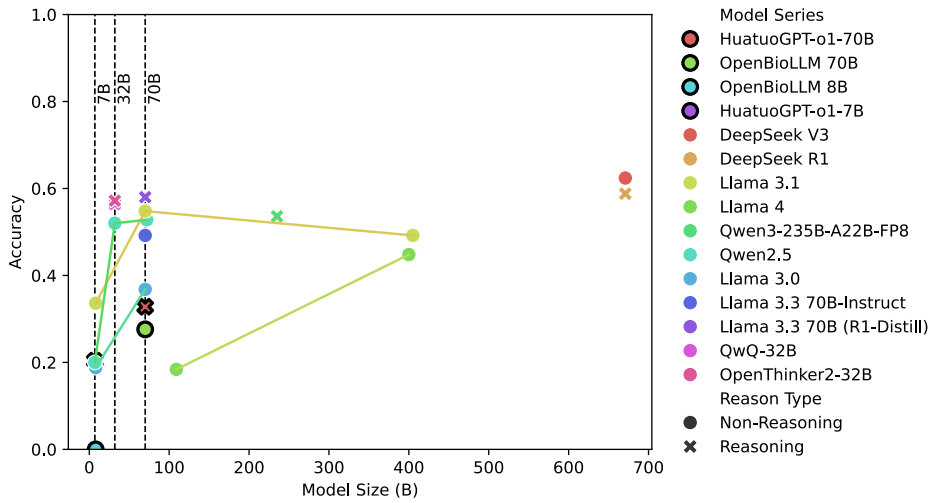

Figure 7: Average model accuracy on MedEvidence as a function of model size. We observe diminishing returns beyond 70 billion parameters.

times resulting in worse performance than the base model [47, 48, 49]. Similar behavior has been previously reported in long-context medical applications [11].

Model size shows diminishing returns beyond 70B parameters As shown in Figure 7, within the same model families, increasing size from 7B to 70B parameters yields substantial accuracy gains on MedEvidence. However, beyond this point, we observe rapidly diminishing returns, both within specific model families and across our suite of evaluated models more broadly.

Combined, our results suggest that synthesizing information across sources to match individual systematic reviews' conclusions eludes current scaling paradigms. Increasing test-time compute (i.e., reasoning) does not necessarily improve performance, larger models do not consistently yield greater gains, and knowledge-based fine-tuning tends to degrade performance. Instead, most models exhibit similar behavior: model performance tends to degrade as token length increases, their responses show overconfidence, and all models exhibit a lack of scientific skepticism toward low-quality findings. These results suggest that more work is still required before LLMs can reliably match the observations from expert-conducted SRs, even though LLM systems are already deployed and being used by clinicians.

Limitations Our study has several limitations. First, the dataset is subject to selection bias, as we only include a SR if all its sources are available (either full text/abstract). Second, while our benchmark is designed to isolate and provide a controlled environment to test LLMs' ability to reason over the same studies experts used to derive conclusions, it does not assess the full SR pipeline, including literature search, screening, or risk-of-bias assessment. Future work could incorporate multi-expert consensus or update findings based on newer studies to strengthen benchmark reliability.

7 Conclusion

Benchmarks drive advancements by providing a standard to measure progress and enabling researchers to identify weaknesses in current approaches. While LLMs are already deployed for scientific synthesis, our understanding of their failure modes still requires broader investigation. In this work, we present MedEvidence, a benchmark derived from gold-standard medical systematic reviews. We use MedEvidence to characterize the performance of 24 LLMs and find that, unlike humans, LLMs struggle with uncertain evidence and cannot exhibit skepticism when studies present design flaws. Consequently, given the same studies, frontier LLMs fail to match the conclusions of systematic reviews in at least 37% of evaluated cases. We release MedEvidence to enable researchers to track progress.

References

[1] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224, 2021.

[2] Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 8–23. World Scientific, 2023.

[3] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. *arXiv preprint arXiv:2409.04600*, 2024.

[4] OpenAI. Deep research system card, 2025. Accessed: 2025-05-15.

[5] Google. Gemini deep research – your personal research assistant, 2025. Accessed: 2025-05-15.

[6] Elicit. Elicit: The ai research assistant, 2025. Accessed: 2025-05-15.

[7] OpenEvidence. Open evidence: Ai-powered medical information platform, 2025. Accessed: 2025-05-15.

[8] U.S. Food and Drug Administration. Fda announces completion of first ai-assisted scientific review pilot and aggressive agency-wide ai rollout timeline, May 2025. FDA News Release.

[9] YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, and Daniel Shu Wei Ting. Development and testing of retrieval augmented generation in large language models–a case study report. *arXiv preprint arXiv:2402.01733*, 2024.

[10] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

[11] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22021–22030, 2024.

[12] Tim Reason, Emma Benbow, Julia Langham, Andy Gimblett, Sven L Klijn, and Bill Malcolm. Artificial intelligence to automate network meta-analyses: Four case studies to evaluate the potential application of large language models. *Pharmacoecon Open*, 8(2):205–220, Mar 2024.

[13] Nikolas Schopow, Georg Osterhoff, and David Baur. Applications of the natural language processing tool chatgpt in clinical practice: Comparative study and augmented systematic review. *JMIR Med Inform*, 11:e48933, Nov 2023.

[14] Riaz Qureshi, Daniel Shaughnessy, Kayden A. R. Gill, Karen A. Robinson, Tianjing Li, and Eitan Agai. Are chatgpt and large language models "the answer"to bringing us closer to systematic review automation? *Systematic Reviews*, 12(1):72, 2023.

[15] Honghao Lai, Long Ge, Mingyao Sun, Bei Pan, Jiajie Huang, Liangying Hou, Qiuyu Yang, Jiayi Liu, Jianing Liu, Ziying Ye, Danni Xia, Weilong Zhao, Xiaoman Wang, Ming Liu, Jhalok Ronjan Talukdar, Jinhui Tian, Kehu Yang, and Janne Estill. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Netw Open*, 7(5):e2412687, May 2024.

[16] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. *arXiv preprint arXiv:2501.07171*, 2025.

[17] Judith-Lisa Lieberum, Markus Töws, Maria-Inti Metzendorf, Felix Heilmeyer, Waldemar Siemens, Christian Haverkamp, Daniel Böhringer, Joerg J. Meerpohl, and Angelika Eisele-Metzger. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *Journal of Clinical Epidemiology*, 181:111746, 2025.

[18] Justin Clark, Belinda Barton, Loai Albarqouni, Oyungerel Byambasuren, Tanisha Jowsey, Justin Keogh, Tian Liang, Christian Moro, Hayley O'Neill, and Mark Jones. Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*, page 1–19, 2025.

[19] Juraj Vladika, Phillip Schneider, and Florian Matthes. MedREQAL: Examining medical knowledge recall of large language models via question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14459–14469, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[20] Juraj Vladika, Phillip Schneider, and Florian Matthes. HealthFC: Verifying health claims with evidence-based medical fact-checking. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia, May 2024. ELRA and ICCL.

[21] Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing?, 2024.

[22] Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence, 2025.

[23] Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm, 2024.

[24] Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. Untangle the knot: Interweaving conflicting knowledge and reasoning skills in large language models, 2024.

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

[26] Mark Petticrew, Paul Wilson, Kath Wright, and Fujian Song. Quality of cochrane reviews. quality of cochrane reviews is better than that of non-cochrane reviews. *BMJ*, 324(7336):545, Mar 2002.

[27] A Cipriani, T A Furukawa, and C Barbui. What is a cochrane review? *Epidemiol Psychiatr Sci*, 20(3):231–233, Sep 2011.

[28] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature, 2025.

[29] *Entrez Programming Utilities Help [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US), 2010-.

[30] Search strategy used to create the pubmed systematic reviews filter, 2019.

[31] Camila Torres Bezerra, Antonio José Grande, Vivianny Kelly Galvão, Douglas Henrique Marin dos Santos, Álvaro Nagib Atallah, and Valter Silva. Assessment of the strength of recommendation and quality of evidence: Grade checklist. a descriptive study. *Sao Paulo Medical Journal*, 140(6):829–836, 2022.

[32] OpenAI. Openai o1 system card, 2024.

[33] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[34] OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, January 2025.

[35] OpenAI. Gpt-4 technical report, 2024.

[36] Qwen Team. Qwen3, April 2025.

[37] AI@Meta. The llama 4 herd, 2025.

[38] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024.

[39] Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`, 2024.

[40] LangChain. Refinedocumentschain. Accessed: 2025-05-16.

[41] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[42] Longze Chen, Ziqiang Liu, Wanwei He, Yunshui Li, Run Luo, and Min Yang. Long context is not long at all: A prospector of long-dependency data for large language models. *arXiv preprint arXiv:2405.17915*, 2024.

[43] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning, 2024. *URL https://arxiv. org/abs/2404.02060*.

[44] Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. Large language models are overconfident and amplify human bias. *arXiv preprint arXiv:2505.02151*, 2025.

[45] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[46] Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024.

[47] Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, et al. Fine-tuning is fine, if calibrated. *Advances in Neural Information Processing Systems*, 37:136084–136119, 2024.

[48] Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in-and out-of-distribution data. *arXiv preprint arXiv:2010.11506*, 2020.

[49] Eric Wu, Kevin Wu, and James Zou. Finetunebench: How well do commercial fine-tuning apis infuse knowledge into llms? *arXiv preprint arXiv:2411.05059*, 2024.

[50] DeepSeek-AI. Deepseek-v3 technical report, 2025.

[51] AI@Meta. The llama 3 herd of models, 2024.

[52] Qwen Team. Qwen2.5 technical report, 2025.

[53] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract and Introduction 1 accurately reflect the main contributions and scope of this paper. We assess whether large language models can match the conclusions of medical systematic reviews. To do this, we collect a human-curated dataset of observations from expert-written and published systematic reviews. We observe that while performance can be saturated in certain agreement, we also observe that LLMs struggle to be skeptical of studies with major limitations and struggle to handle different levels of evidence, suggesting that LLMs struggle to match human experts when they create systematic reviews.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Potential limitations of this work and outlined future directions to enhance our benchmark are presented in the Limitations section 5.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical study. Thus, theoretical results are not derived.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Sections 4 and 5, we explicitly mention how the dataset was curated, its statistics, and the metrics we used for evaluation, as well as the methods for quantifying uncertainty.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to the code via a GitHub repository and release the dataset on HuggingFace Datasets. Along with the paper, the provided code and data are sufficient to reproduce the main results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: While we do not train a model, we do specify the provenance, size, and statistics of the dataset we used for evaluation in Section 4. Furthermore, we specify the parameters used to do inference with the large language models, which was consistent across all evaluations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we calculate 95% confidence intervals for all metrics via bootstrapping with N=1000, as mentioned in the main body of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5.1 details the computational resources used in our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No human subjects were used throughout the experiments. Throughout the curation of this dataset, we focused on the use of open-access reviews and studies that allow for the redistribution of their materials. We highlight the publication of datasets that have used systematic reviews from the same source in the Related Work section.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix Section A discusses the social impact of LLMs for systematic review generation and their implications in clinical practice. In particular, if the technology is used but produces incorrect results, there is a risk that clinicians or policymakers may rely on flawed evidence synthesis, potentially causing harm through inappropriate treatments or misinformed guidelines. Alarmingly, these systems are already deployed and used in the real world. Therefore, it is highly important to create benchmarks to systematically assess model performance.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The dataset is derived from de-identified data.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use open-source data that allows for redistribution; furthermore, for all data used, we provide the PubMed ID and metadata needed to retrieve the original work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We thoroughly characterize the dataset characterization process, its statistics, and its fields

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The protocol for data collection is shared both as a pipeline in Figure 2 as well as described in the Dataset Curation section. Dataset curation was performed by collaborators.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: An IRB was not required to conduct this research, as neither human subjects nor crowdsourcing were involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used to collect some metadata to stratify the dataset; details on all exact LLM use are described in the main body of the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.