CAN LARGE LANGUAGE MODELS MATCH THE CONCLUSIONS OF SYSTEMATIC REVIEWS?

Christopher Polzak*[1] **Alejandro Lozano***[1] **Min Woo Sun***[1]

James Burgess[1] **Yuhui Zhang**[1] **Kevin Wu**[1] **Chia-Chun Chiang**[2]

Jeffrey J. Nirschl[1] **Serena Yeung-Levy**[1]

[1]**Stanford University** [2]**Mayo Clinic**

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 SR generation. However, the ability of LLMs to critically assess evidence and reason across multiple documents to provide recommendations at the same proficiency as domain experts 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 100 medical SRs with the studies they are based on. We benchmark 25 LLMs on MedEvidence, including reasoning, non-reasoning, medical specialists, and models across varying sizes (from 7B-700B). Through our systematic evaluation, we find that reasoning does not necessarily improve performance, larger models do not consistently yield greater gains, and knowledge-based fine-tuning degrades accuracy on MedEvidence. Instead, most models exhibit similar behavior: performance tends to degrade as token length increases, their responses show overconfidence, and, contrary to human experts, 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 (Bornmann et al., 2021), 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 (Lozano et al., 2023; Scherbakov et al., 2024). In particular, large language model (LLM)-based systems could offer a promising solution for supporting and automating tasks in conducting systematic reviews (SRs), which typically take an average of 67 weeks of intensive human effort (Fabiano et al., 2024; Riaz et al., 2024; Wang et al., 2025). For example, several LLM-assisted tools such as Deep Research (OpenAI, 2025; Google, 2025), Elicit (Elicit, 2025), and Open Evidence (OpenEvidence, 2025) have already been deployed and can be incorporated into the SR process to improve efficiency (Fabiano et al., 2024). 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 (FDA, 2025).

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,

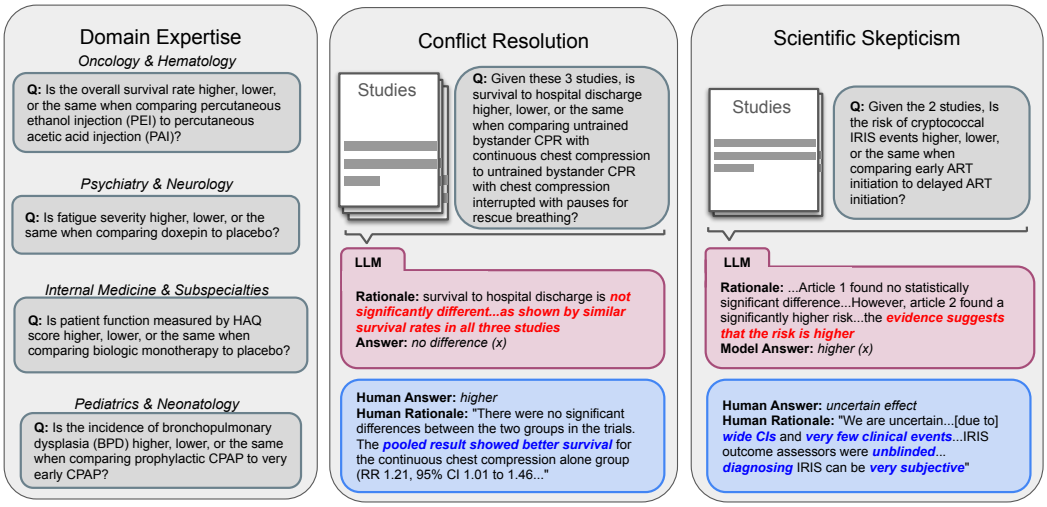

Figure 1: Core skills evaluated by MedEvidence including: medical domain expertise across 10 different specialties, synthesizing conflicting evidence, and applying scientific skepticism when studies exhibit a high risk of bias (e.g. due to small sample sizes or insufficient supporting evidence).

population size, and risk of bias—and to navigate conflicting evidence (as medical findings may 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. via retrieval augmentation) (Ke et al., 2024), 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 (Liévin et al., 2024; Fleming et al., 2024), 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 (Wallace et al., 2021; Schopow et al., 2023; Qureshi et al., 2023; Lai et al., 2024; O'Doherty et al., 2024), 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 simplify evaluation. 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 25 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.

2 RELATED 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	✓	✓	✓	✓	✓

Table 1 presents an overview of related works and their key distinctions with respect to our work.

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 (Lieberum et al., 2025; Clark et al., 2025). However, larger-scale and closed QA evaluations of LLM-based SR or meta-analyses generation remain relatively underexplored. Wallace et al. (2021) evaluates neural models on generating narrative summaries from RCTs referenced in Cochrane reviews, using the review as ground truth, ultimately requiring expert oversight to assess correctness. Building on this line of inquiry, O'Doherty et al. (2024) conducts a similar assessment across 45 systematic reviews. Reason et al. (2024) further explore whether LLMs can extract numerical data from abstracts and generate executable code for meta-analyses. While their findings are encouraging, the study is limited to four case studies, restricting generalizability. Schopow et al. (2023) and Qureshi et al. (2023) 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 require substantial amounts of review from medical experts, highlighting the need for automated benchmarks to facilitate evaluation.

Verification of medical facts derived from systematic reviews Several studies have leveraged SRs to benchmark LLMs' ability to perform medical fact verification, where a model must decide whether to support or refute a claim. For instance, MedREQAL (Vladika et al., 2024a) 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 (Vladika et al., 2024b), 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 SRs, this task primarily involves retrieving information from a pre-synthesized source, removing the 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 (Wan et al., 2024) 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 (Wu et al., 2025) 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 (Su et al., 2024) and KNOT (Liu et al., 2024) 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 succinct and synthesized inputs.

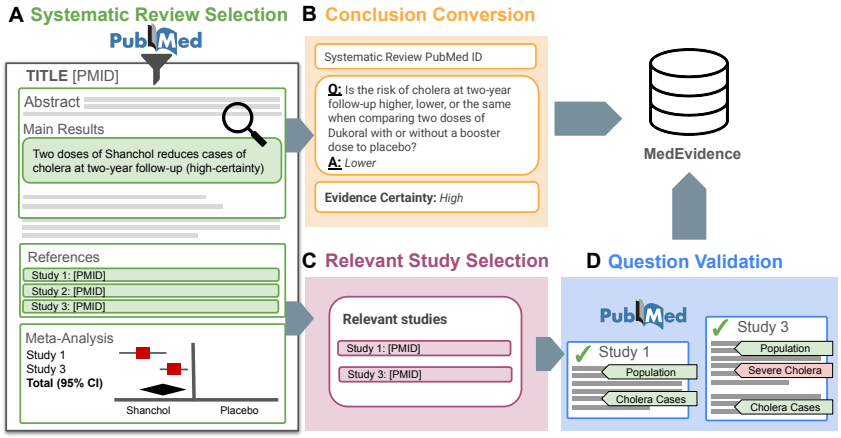

Figure 2: Overview of the dataset curation process for MedEvidence.

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.

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 (Henderson et al., 2010). Cochrane is a long-standing and widely respected source of clinical evidence (Petticrew et al., 2002; Cipriani et al., 2011), 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 (Lozano et al., 2025); otherwise, abstracts are retrieved directly via PubMed's Entrez API (PubMed, 2010-). 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 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 (PubMed, 2019). 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.2 for more conversion details). Evidence certainty was extracted only when it was explicitly provided by the

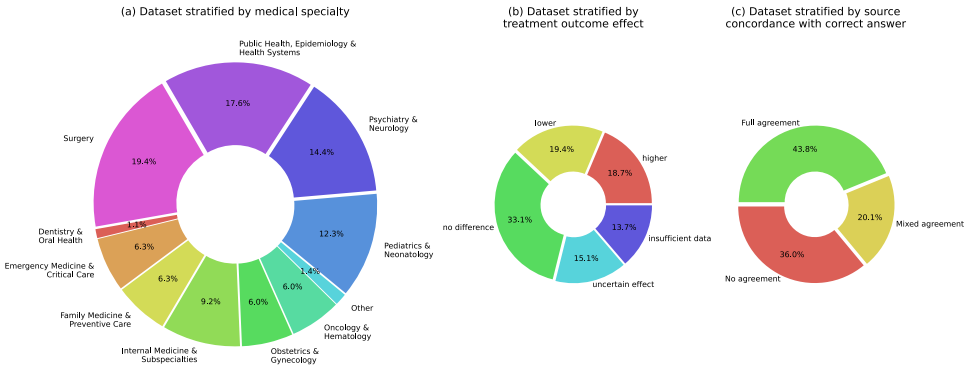

Figure 3: Key statistical characteristics of the questions in MedEvidence. (a) shows the dataset distribution stratified by medical specialty. (b) presents the distribution stratified by outcome effect. (c) shows the distribution stratified by source concordance with the expert-assessed treatment outcome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

original SR authors, who use the standardized GRADE framework (Bezerra et al., 2022) 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 LLM 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

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

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

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 11.

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 25 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 (OpenAI, 2024b), DeepSeek R1 (DeepSeek-AI, 2025a), OpenThinker2 (Team, 2025a), GPT-4.1 (OpenAI, 2024a), Qwen3 (Team, 2025c), Llama 4 (AI@Meta, 2025), HuatuoGPT-o1 (Chen et al., 2024a), OpenBioLLM (Ankit Pal, 2024), 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). 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 12 for the exact prompt).

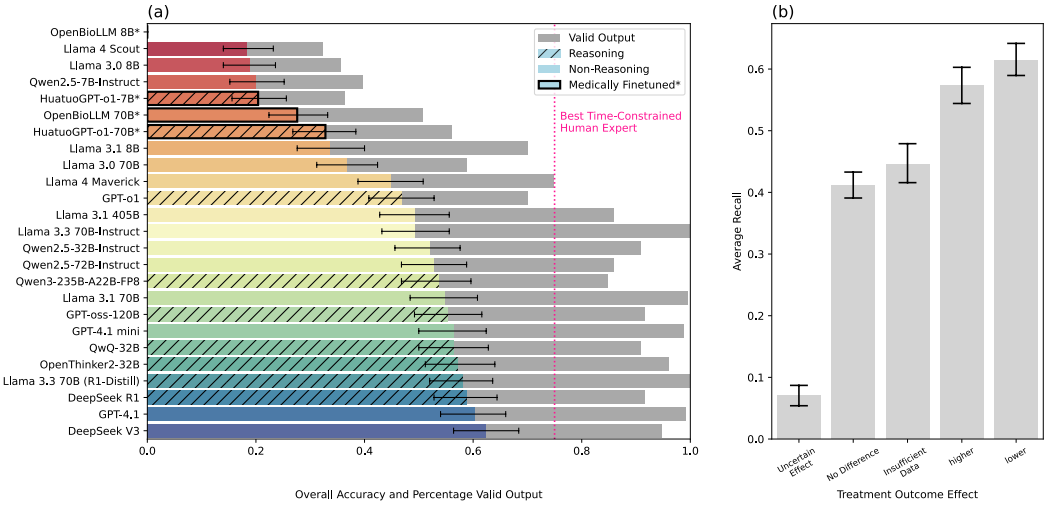

Figure 4: (a) Average model accuracy (and 95% CI) on MedEvidence, overlaid on the percentage of questions where the model provided valid output (details in Appendix E). Best expert performance is shown in a pink dashed line (more details in Appendix R). No model matches or surpasses the best expert performance, even though experts are time-constrained. (b) Average recall grouped by ground truth treatment outcome effect, aggregated across all models (with 95% CI). A per-model average recall by treatment outcome effect is shown in Appendix Figure 18.

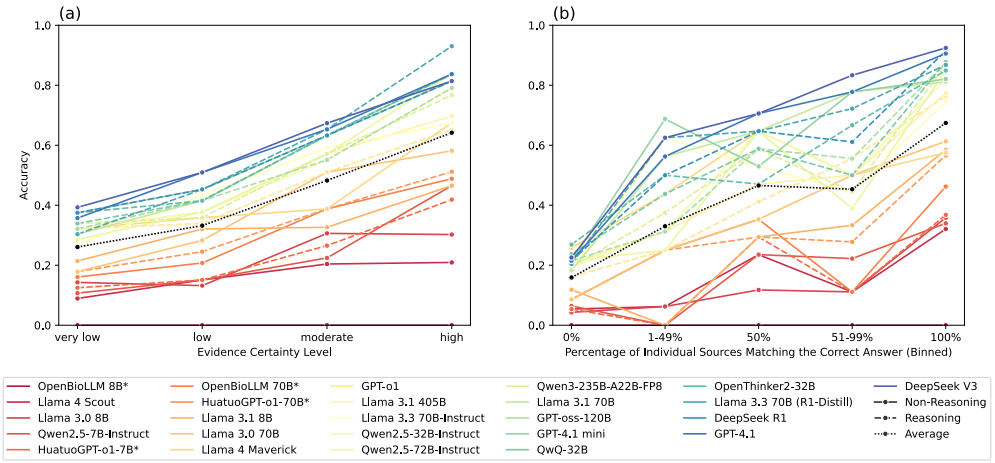

Figure 5: (a) Accuracy as a function of evidence certainty, shows a monotonically increasing trend. (b) Accuracy as a function of source concordance, defined as the percentage of relevant sources that agree with the final systematic review (SR) answer, also exhibits a monotonically increasing trend.

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 13 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) 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

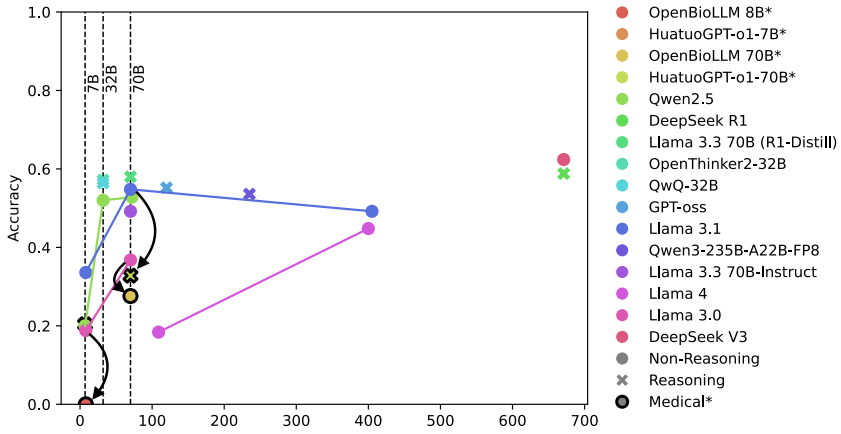

Figure 6: Average model accuracy as a function of model size. We observe diminishing returns beyond 70 billion parameters. Arrows point from base models to their medically-finetuned counterparts (arrow between HuatuoGPT-o1 7B and Qwen2.5 7B omitted due to very similar performance).

match with the correct answer, the output was deemed incorrect. Confidence intervals (CIs) were calculated via bootstrap (95%, N=1000) (Efron & Tibshirani, 1994). Models were evaluated both under zero-shot and few-shot settings.

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. More importantly, model performance still lags behind expert clinical accuracy (< 0.75), even when clinicians are limited by time and unable to conduct the in-depth analysis performed by the original SR authors. 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 fine-tuning 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.

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 15), even though all but two models can fit 80% of the dataset within one context window (see Appendix D). Naturally, training LLMs on long contexts does not guarantee improved long-context understanding, as models may still struggle to utilize information from lengthy inputs (Chen et al., 2024b; Li et al., 2024).

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 16, 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 (Sun et al., 2025; Xiong et al., 2023) and shown that reinforcement learning via human feedback (RLHF) amplifies this effect (Leng et al., 2024).

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). Prior work has noted that LLMs struggle to aggregate findings across multiple documents Shaib et al. (2023). Our results show that this limitation is amplified when sources present conflicting conclusions. 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 19).

Medical finetuning does not improve performance Figure 7 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, finetuning without proper calibration can harm generalization, sometimes resulting in worse performance than the base model (Mai et al., 2024; Kong et al., 2020; Wu et al., 2024). Similar behavior has been previously reported in long-context medical applications (Fleming et al., 2024).

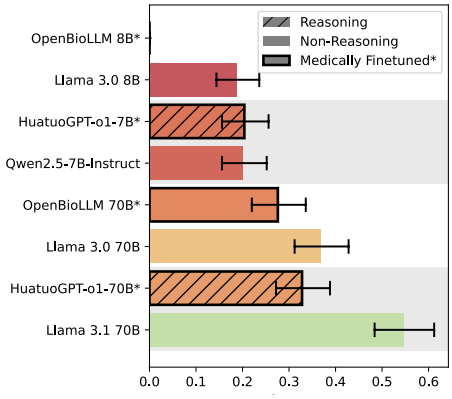

Figure 7: Medically-finetuned models vs their base generalist counterparts. Pairs of medical and base models are adjacent. 95% confidence intervals are calculated via bootstrapping with $N = 1000$.

Model size shows diminishing returns beyond 70B parameters As shown in Figure 6, 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.

Additional evaluations We further re-evaluate top performing models and find that randomizing the order of sources and omitting chain-of-thought prompting both do not significantly affect performance. Few-shot evaluation yields slight improvements, but high error rates persist and the performance gap to clinical experts remains. For detailed results, further experiments, and qualitative analysis, please see the Appendix.

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 required before LLMs can reliably match the observations from expert-conducted SRs, despite LLM systems already being deployed and 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 25 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 $\sim$37% of evaluated cases. We release MedEvidence to enable researchers to track progress.

REFERENCES

AI@Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

AI@Meta. The llama 4 herd, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

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

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.

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. doi: 10.1057/s41599-021-00903-w. URL https://doi.org/10.1057/s41599-021-00903-w.

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, 2024a. URL https://arxiv.org/abs/2412.18925.

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*, 2024b.

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

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*, pp. 1–19, 2025. doi: 10.1017/rsm.2025.16.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025a. URL https://arxiv.org/abs/2501.12948.

DeepSeek-AI. Deepseek-v3 technical report, 2025b. URL https://arxiv.org/abs/2412.19437.

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

Elicit. Elicit: The ai research assistant, 2025. URL https://elicit.com. Accessed: 2025-05-15.

Nicholas Fabiano, Arnav Gupta, Nishaant Bhambra, Brandon Luu, Stanley Wong, Muhammad Maaz, Jess G Fiedorowicz, Andrew L Smith, and Marco Solmi. How to optimize the systematic review process using ai tools. *JCPP advances*, 4(2):e12234, 2024.

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

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, pp. 22021–22030, 2024.

Google. Gemini deep research – your personal research assistant, 2025. URL https://gemini.google/overview/deep-research/?hl=en. Accessed: 2025-05-15.

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.

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.

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.

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.

LangChain. Refinedocumentschain. URL https://python.langchain.com/api_reference/langchain/chains/langchain.chains.combine_documents.refine.RefineDocumentsChain.html. Accessed: 2025-05-16.

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

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

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.

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

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. URL https://arxiv.org/abs/2404.03577.

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*, pp. 8–23. World Scientific, 2023.

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. URL https://arxiv.org/abs/2501.07171.

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.

OpenAI. Gpt-4 technical report, 2024a. URL https://arxiv.org/abs/2303.08774.

OpenAI. Openai o1 system card, 2024b. URL https://arxiv.org/abs/2412.16720.

OpenAI. Deep research system card, 2025. URL https://cdn.openai.com/deep-research-system-card.pdf. Accessed: 2025-05-15.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher,

Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b and gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

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

James O'Doherty, Cian Nolan, Yufang Hou, and Anja Belz. Beyond abstracts: A new dataset, prompt design strategy and method for biomedical synthesis generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 358–377, 2024.

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.

PubMed. *Entrez Programming Utilities Help [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US), 2010-. URL https://www.ncbi.nlm.nih.gov/books/NBK25501/.

PubMed. Search strategy used to create the pubmed systematic reviews filter, 2019. URL https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html.

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.

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.

Irbaz Bin Riaz, Syed Arsalan Ahmed Naqvi, Bashar Hasan, and Mohammad Hassan Murad. Future of evidence synthesis: Automated, living, and interactive systematic reviews and meta-analyses. *Mayo Clinic Proceedings: Digital Health*, 2(3):361–365, 2024.

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.

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.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron C Wallace. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1387–1407, 2023.

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. URL https://arxiv.org/abs/2408.12076.

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

OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, January 2025a.

Qwen Team. Qwen2.5 technical report, 2025b. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3, April 2025c. URL https://qwenlm.github.io/blog/qwen3/.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025d. URL https://qwenlm.github.io/blog/qwq-32b/.

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 (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14459–14469, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.860. URL https://aclanthology.org/2024.findings-acl.860/.

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 (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8095–8107, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.709/.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605, 2021.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing?, 2024. URL https://arxiv.org/abs/2402.11782.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509, 2025.

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.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence, 2025. URL https://arxiv.org/abs/2404.10198.

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.