Evaluating Large Language Models for Evidence-Based Clinical Question Answering

Can Wang [1] Yiqun T. Chen [1,2]

[1]Department of Biostatistics [2]Department of Computer Science

Summary: LLMs perform best on structured clinical guidelines (90% accuracy) compared to narrative guidelines and systematic reviews (60-70% accuracy), with performance strongly correlated to citation counts of source materials. Providing relevant abstracts through retrieval-augmented prompting significantly improves performance.

Introduction

Background. Large Language Models (LLMs) show promise in medical question answering, but existing benchmarks primarily use standardized exam questions that may not reflect real-world clinical practice, where evidence is often missing, contradictory, or continuously evolving. Current QA datasets often fail to examine how underlying evidence characteristics (such as study quality, citation frequency, and subject matter) affect model accuracy, and most evaluations test "zero-shot" ability without access to external literature search tools.

Our Contributions.

Construct a multi-source QA dataset from *Cochrane systematic reviews*, *AHA structured guidelines*, and *narrative guidelines*.

Evaluate GPT-4o-mini and GPT-5, analyzing how performance varies by evidence quality, domain, and citation prominence.

Assess the effect of *retrieval-augmented generation (RAG)* using PubMed as a proxy for external search.

Methods

Dataset Construction. We built a multi-source clinical QA dataset from:

- **Cochrane Systematic Reviews** (8,533 abstracts, 2010–2025) with metadata and derived QAs: *Answer (Yes/No/No Evidence)*, *Discrepancy (RCT vs. obs.)*, *Evidence Quality (5 levels)*.
- **AHA Guidelines** (2,581 structured recs, 2020–2025) mapped to *Class of Recommendation (COR)* and *Level of Evidence (LOE)*.
- **Narrative Guidelines** (289 docs) segmented into ∼2,000-char chunks; questions generated in PICO style with categorical answers.

Question Example.

(a) Cochrane Review	
Title	*Chemoradiotherapy for cervical cancer*
Question	Does chemoradiotherapy improve 5-yr survival vs. radiotherapy alone?
Answer	Yes (Evidence: High)
Discrepancy	No
Note	6% improvement in survival (HR=0.81, $p < 0.001$).

Evaluation. (1) Tasks: classification ({Yes, No, No Evidence}) with structured JSON output. (2) Settings: *No-context baseline* vs. *Context ablations* (gold abstract, PubMed top-3, random abstract).

Statistical Analysis. (1) Primary metric: exact-match accuracy. (2) Additional: valid output rate, confusion matrices, 95% CIs. (3) Stratified analyses: clinical field, year, citation count (logistic regression).

Results

Overview. GPT-5 consistently outperforms GPT-4o-mini by a small margin (2–6%). Performance is highest on structured data, varies with citation prominence and clinical domain, and improves substantially with contextual information.

(1) Performance on Systematic Reviews.

Task	GPT-4o-mini	GPT-5
Answer	60.3%	67.8%
Discrepancy	57.0%	59.1%
Evidence Quality	32.1%	38.8%

Table 1. Performance on systematic review questions.

Topic	N	4o-mini	GPT-5
Top 5			
Rheumatology	66	72.7	70.8
Pain & Anaesthesia	337	68.8	72.1
Tobacco/Drugs/Alcohol	169	68.6	74.6
Public Health	108	68.5	65.7
Urology	108	67.6	66.7
Bottom 5			
Health & Safety at Work	48	54.2	52.1
Complementary & Alt. Med.	92	53.3	57.6
Dentistry/Oral Health	216	53.2	61.1
Neonatal Care	400	50.5	68.8
Wounds	173	43.4	61.8

Table 2. Top/bottom 5 topics by answer accuracy (%).

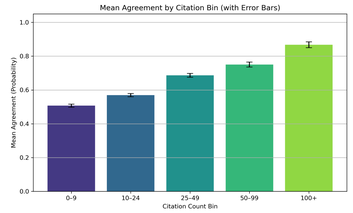

Figure 1. Accuracy by citation-count bin.

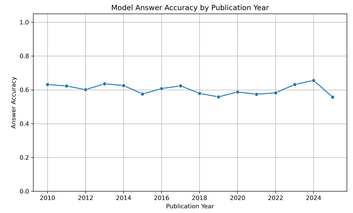

Figure 2. Accuracy by publication year.

(2) Structured Clinical Guidelines (AHA). GPT-4o-mini achieves 94.0% accuracy and shows moderate ability to reason about evidence quality.

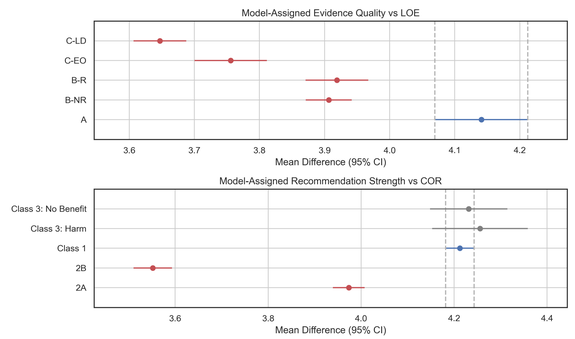

Figure 3. Tukey HSD: model-assigned scores align with LOE/COR hierarchy.

(3) Narrative Clinical Guidelines. GPT-4o-mini's accuracy fell to 56.3%, where it particularly failed to correctly interpret statements of negative findings.

Results

(4) Impact of Retrieval-Augmented Context. Providing the correct source abstract increased accuracy for both models, surpassing 90%. A more realistic setup using PubMed-retrieved abstracts also yielded substantial gains, boosting GPT-4o-mini from a baseline of 60.3% to 79.9% and GPT-5 from 67.8% to 75.2% on the tested subset. Irrelevant context (random abstracts) produced only a slight degradation in accuracy, suggesting both models are relatively robust to noisy inputs.

Context	4o-mini	GPT-5*
No Context	60.3%	67.8%
Gold Abstract	91.6%	93.2%
PubMed Top-3	79.9%	75.2%
Random (Noise)	58.1%	65.1%

Table 3. Model accuracy on abstract-based QA with different contexts.

Discussion & Conclusion

Main Findings.

- Base models achieve only **moderate accuracy** on systematic review questions, with common failures including domain knowledge gaps and poor handling of absent/ambiguous evidence.
- Performance is **markedly higher** on structured clinical guideline recommendations vs. free-text systematic reviews.
- Models show **systematic variation** by domain and citation impact, suggesting predictions are influenced by external visibility of underlying evidence.
- **In-context learning and retrieval-augmented prompting** substantially improve performance.

Future Directions. Extend dataset to additional evidence sources, incorporate richer answer structures, and integrate retrieval pipelines with high-quality biomedical databases for safer clinical decision support systems.

References

American Heart Association. Guidelines & Statements Search, 2025. URL `https://professional.heart.org/en/guidelines-statements-search`.

Cochrane Library. Search | Cochrane Library, 2025. URL `https://www.cochranelibrary.com/cdsr/reviews`.

Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, Alexander Löser, and Keno K. Bressem. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine*, 7(1):358, December 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01356-6. URL `https://www.nature.com/articles/s41746-024-01356-6`. Publisher: Nature Publishing Group.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering, September 2019. URL `http://arxiv.org/abs/1909.06146`. arXiv:1909.06146 [cs].

Xinyan Li, Zhihao Tang, Yuxing Zhang, et al. Benchmarking retrieval-augmented generation for clinical question answering. *arXiv preprint arXiv:2311.08417*, 2023.

cwang271@jh.edu