nature portfolio

Peer Review File

General-Purpose Large Language Models Outperform Specialized Clinical AI Tools on Medical Benchmarks

Corresponding Author: Mr Krithik Vishwanath

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
This manuscript addresses an important and timely question—how generalist frontier LLMs compare with specialized clinical AI tools—but the study design, evaluation framework, and inferential claims contain fundamental flaws that undermine the validity of its conclusions. While the core descriptive observation (that frontier LLMs score highly on standardized benchmarks) may be directionally correct, the paper does not meet the evidentiary or methodological standards required to support its strong claims about clinical AI performance or real-world utility.
The primary issues are evaluation validity, epistemic circularity, unfair comparability, and overgeneralization beyond experimental conditions. These are not cosmetic concerns and cannot be resolved through incremental revision alone.
Major Concerns
1. Epistemic Circularity and Evaluation Bias
A central limitation is the reliance on HealthBench (developed by OpenAI) combined with GPT-4.1 (also OpenAI) as the sole evaluator, while one of the top-performing systems evaluated is GPT-5, likewise developed by OpenAI. This creates an unavoidable epistemic circularity in which benchmark design, scoring norms, and model optimization share institutional and methodological lineage.
Even absent direct data leakage, this alignment systematically advantages models optimized under similar reward, instruction-following, and stylistic paradigms—particularly for subjective dimensions such as communication quality, uncertainty framing, and safety reasoning. HealthBench's own documentation emphasizes the need for meta-evaluation and validation against physician judgment, yet no such independent validation is presented here.
The absence of any blinded human (physician) evaluation, even on a subset of responses, severely weakens confidence in the reported performance differences.
2. Unfair and Non-Comparable Querying Conditions
The comparison framework is fundamentally asymmetric:
• Clinical tools (OpenEvidence, UpToDate Expert AI) were manually queried via browser interfaces
• Generalist models were queried via APIs
These modalities differ in ways that materially affect outputs, including hidden system prompts, retrieval behavior, response length defaults, latency constraints, safety guardrails, and UI-imposed truncation. As a result, observed differences cannot be cleanly attributed to "clinical vs. generalist" capability.
Moreover, the clinical tools are retrieval-augmented systems designed for reference use, not free-form benchmark answering. Evaluating them using a single-turn, free-text paradigm optimized for generalist LLMs introduces structural bias against their intended design goals.
3. Benchmark Contamination and Decontamination Gaps
Both MedQA and HealthBench are public benchmarks, and the paper does not include:
• Any decontamination analysis
• Any held-out or novel evaluation set
• Any justification for assuming newer frontier models have not seen benchmark content during training
This is especially problematic when comparing frontier models trained on large-scale web and academic corpora against tools that rely on RAG over curated medical knowledge bases. Without addressing contamination risk, claims of superiority lack epistemic grounding.
4. Statistical Fragility and Overinterpretation
The manuscript draws category-level conclusions from n=2 clinical tools vs. n=3 generalist models, using model-level statistical tests that are fragile at this scale. While item-level pairing exists (same questions across systems), the inferential

framing often shifts toward model-count-based comparisons that are underpowered and potentially misleading. Subgroup analyses further exacerbate this issue, with wide confidence intervals suggesting limited statistical power.

5. Limited Clinical Validity and Missing Safety Analysis

The evaluation focuses narrowly on benchmark correctness and alignment scores, which do not establish clinical safety or utility. There is:

- No analysis of harmful or contraindicated recommendations
- No structured error taxonomy
- No assessment of uncertainty handling or appropriate deferral
- No clinician-rated evaluation of usefulness in realistic workflows

The manuscript acknowledges that benchmarks are not equivalent to clinical trials, but nevertheless frames conclusions in ways that imply deployment-relevant superiority. This leap is not justified.

Notably, prior randomized and deployment-level studies (e.g., physician diagnostic reasoning trials, clinical operations RCTs, agent-based evaluation frameworks) already provide stronger translational evidence than benchmark-only comparisons, yet these are not meaningfully engaged.

6. Overgeneralized and Marketing-Adjacent Claims

Several claims—particularly around adoption rates and market penetration of clinical tools—appear sourced from company or press materials and are presented without adequate caveating. These should not be treated as epidemiologic facts.

More broadly, the conclusion that “specialization does not help” oversimplifies a complex literature and conflicts with peer-reviewed evidence showing benefits from medical fine-tuning, retrieval grounding, and clinician-in-the-loop evaluation.

7. Reproducibility and Transparency Gaps

The methods lack sufficient detail to enable independent replication. Missing or underspecified elements include:

- Exact prompting templates (including system messages)
- Sampling procedures and random seeds
- Generation parameters (temperature, top-p, token limits, retries)
- UI interaction details for clinical tools
- Response normalization or length constraints
- Inter-rater reliability for automated scoring

As written, it is not possible to disentangle model capability from evaluation setup.

Reviewer #3

(Remarks to the Author)

A. Summary of the key results

The short manuscript “Generalist Large Language Models Outperform Clinical Tools on Medical Benchmarks” addressed a very important and timely theme and is refreshingly clearly written and to the point.

The key result is that, at least on public benchmarks, the domain specific models tested, do not do better than mass market models. Even though public benchmarks are limited, it is important to publish this result. More discussion of the limitations and potential bias of the benchmarks is needed.

I have a number of presentational and methodological questions:

B. Originality and significance: if not novel, please include reference

It is certainly of interest to the readership of the journal.

There is an important recent overlapping work https://arxiv.org/pdf/2512.01241, but this is a preprint. The authors of the manuscript under review cite other preprints, and citing and discussing the results of (First, do NOHARM: towards clinically safe large language models
D Wu, FN Haredasht, SK Maharaj, P Jain, J Tran, M Gwiazdon, A Rustagi, J Jindal…
arXiv preprint arXiv:2512.01241, 2025•arxiv.org) would add to the manuscript.

On this NOHARM benchmark a domain specific model does outperform the best performing mass-market model, but only by a very narrow margin. Discussion of the context for the difference between these core results would be interesting for all readers of the manuscript under review.

C.E.F. - Data & methodology: validity of approach, quality of data, quality of presentation AND Conclusions: robustness, validity, reliability AND Suggested improvements: experiments, data for possible revision

Comment 1: [TITLE] “Generalist Large Language Models Outperform Clinical Tools on Medical Benchmarks”. Is the term “Generalist “appropriate here. …Or rather, it is correct and appropriate but also confusing. The term has two overlapping meanings and can confuse. All of the models tested in the manuscript study have “Generalist” properties. OpenEvidence and UpToDate Expert AI are” Medically “Generalist”, as they take broad questions across all medicine disciplines and all medical stages producing answers that can be graded on these multiple axes. Gemini/ChatGPT/Sonnet are generally-“Generalist”.

Both a “Generalist” in their own way.

As a second point, both OpenEvidence and UpToDate Expert AI are LLMs at their core and the original title could confuse the readers - who cold interpret it as "Generalist Large Language Models Outperform" [non-LLM] "Clinical Tools on Medical Benchmarks"

An accurate and more informative title would be:

“Mass-market Large Language Models Outperform Major Commercial Clinical Models on Medical Benchmarks” Tat is 99 characters and may be longer than allowed, but it is not for me to write a title - just to point out limitations of the existing one.

Comment 2: [MAIN]. “We assessed two widely deployed clinical AI systems (OpenEvidence and UpToDate Expert AI) against three state-of-the-art generalist LLMs (GPT-5, Gemini 3 Pro, and Claude Sonnet 4.5) using a 1,000-item mini-benchmark combining MedQA (medical knowledge) and HealthBench (clinician-alignment) tasks.”

It is important that the manuscript acknowledges that HealthBench is a system developed by OpenAI. All the authors of the preprint are from OpenAI. The preprint is relatively lacking in information on the development of the “5,000 multi-turn conversations between a model
and an individual user or healthcare professional”. The preprint states that “HealthBench comprises 5,000 realistic health conversations between a model and an individual user or healthcare professional. Conversations were selected for relevance, realism, and difficulty, and span a wide range of geographies, languages, and healthcare personas. Given a conversation, the task for an LLM is to respond to the last user message.” My exploration of the system is limited to the preprint alone, and some general community discussion of it. It may be very good, but it does rely on a relatively small number of physician opinions per rubric, and not highly transparent flexible criterial for the evaluation these rubrics. It is not really known how ‘good’ the physicians where and how “right” they were in the assessment of rubrics.

In the preprint, OpenAI then use the developed HealthBench to test a series of Mass-market LLMs, and report, that, surprisingly.... their model, GPT-4o is best of all models ... . It is important to acknowledge this linking of HealthBench to one of the tested models in the current manuscript under review - which, is the next generation of models tested in the preprint Arora et al. It is highly highly likely that OpenAI had privilege knowledge and access (including aspects not in the public domain)to HealtBench in the further development of GPT models.

In the limitations section of the current manuscript the risk of access of models to benchmark is discussed - this point about the HeathBench, and specifically that one of the test models is from the same company, needs to be addressed more clearly and directly. More emphasis is also needed on the point about models potentially having access to the benchmarks in their training. This is highly likely to be the case. To some extent, this is unavoidable and to an extent benchmarking mark test studies actually test the “newness” of models with respect to the benchmarks. This clearly benefits the mass market models with the current faster development turnaround than the medical domain fine-tuned models. More acknowledgement of this is needed.

Was “HealthBench Hard” from Arora used in the manuscript study? From the preprint “We also introduce HealthBench Hard, a subset of 1, 000 HealthBench examples chosen because they are difficult for current frontier models. HealthBench Hard is designed to be a difficult target for future models, with many current models achieving a score of zero on it. The procedure for selecting HealthBench Hard examples is described in Appendix C.” If not, why was HealthBench Hard” not used. It might provide important insights.

Comment 3 [MAIN] “GPT-5 achieving the highest scores, while OpenEvidence and UpToDate demonstrated deficits in completeness, communication quality” Can more information be provided on the assessment of communication quality. Should this necessarily be high for OpenEvidence and UpToDate Expert AI. These models have a specific target user, who are Health Care Professionals (HCPs). They are not designed to communicate with patients directly. Does the communication metric here only assess the communication with HCPs, or does it assess general or lay communication? If it does, then it is not a fair assessment for the task in hand. Much more context is needed on this metric and its applicability.

Comment 4 [MAIN]. “Frontier generalist models benefit from substantially larger training corpora and more advanced alignment, which may enable them to match or exceed specialist performance depending on the task.” Related to Comment 2. Frontier generalist models also potentially benefit from more frequent update iterations. Both GPT5 and Gemini3 are very recent models developed through enormous investment. The phenomena/results described in the manuscript may reflect our current short-term development curve / exponential stage of the development curve of the Frontier generalist models. The rate of increase of the performance of the very large generalist models may slow. Many experts have predicted that it will (or is already doing so). At this point the advantage of domain tunes model may become apparent. Do the authors agree? Some (brief) discussion of this point in the manuscript would make sense.

Comment 5 “Moreover, standardized benchmarks, while allowing quantitative comparison, have inherent shortcomings.10 It is possible that recently trained models or tools have been trained on the benchmarks themselves, which were incidentally swept up in their training data.”

- More detail on these “inherent shortcomings” is needed here, perhaps from reference 10.
- More consideration is needed of “It is possible that recently trained models or tools have been trained on the benchmarks themselves, which were incidentally swept up in their training data.”, as mentioned under my Comment 2.

D. Appropriate use of statistics and treatment of uncertainties

Appropriate as far as I can tell.

G. References: appropriate credit to previous work?
As discussed under B., There is an important recent overlapping work https://arxiv.org/pdf/2512.01241, but this is a preprint. The authors of the manuscript under review cite other preprints, and citing and discussing the results of (First, do NOHARM: towards clinically safe large language models
D Wu, FN Haredasht, SK Maharaj, P Jain, J Tran, M Gwiazdon, A Rustagi, J Jindal…
arXiv preprint arXiv:2512.01241, 2025•arxiv.org) woudl add to the manuscript.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions
The manuscript is very strong in these areas - More discusssion context on the following preprint would help the reader: https://arxiv.org/pdf/2512.01241, (First, do NOHARM: towards clinically safe large language models
D Wu, FN Haredasht, SK Maharaj, P Jain, J Tran, M Gwiazdon, A Rustagi, J Jindal…
arXiv preprint arXiv:2512.01241, 2025•arxiv.org) woudl add to the manuscript.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Below are a few more lingering/nagging issues.
1. The multi-model grading panel is a reasonable fix, but residual circularity remains. Replacing solo GPT-4.1 grading with consensus voting across GPT-5.2, Gemini 3.1 Pro, and Claude Opus 4.6 is better than what they had, but the authors correctly acknowledge that using frontier models as both evaluated systems and judges is still problematic. The RCQ physician evaluation serves as the real escape valve here. The human evaluation should be the primary evidence and HealthBench is supplementary. As Audre Lorde quipped, "The master's tools will never dismantle the master's house." We are being played by the tech companies.
2. The authors make a fair argument that the asymmetric querying concern is a structural constraint rather than a methodological choice: clinical tools don't offer APIs. The addition of the RCQ benchmark helps because it evaluates all systems on real clinical queries, but the fundamental issue that browser-queried systems face hidden UI constraints remains. How about matching response length or format across systems?
3. The revised manuscript reported that no model produced significantly more harmful content or hallucinations (Cochran's Q non-significant). It presented a scoring rubric with anchor points, but not a systematic categorization of error types, a structured error taxonomy. The OpenEvidence clarity finding (mean 2.84) is a nice granular insight, but this was not really a failure mode analysis. 4. There's still no cost, latency, or citation quality comparison. These are practical considerations for clinical deployment that benchmark studies typically ignore, and they are not even acknowledged as limitations.
In summary, please clarify inter-rater reliability distinctions, address model version changes between submission rounds, and add at minimum a paragraph engaging with alternative evaluation frameworks beyond industry-created benchmarks.

Reviewer #3

(Remarks to the Author)
I am satisfied that my comments have been reasonably addressed. For R3.3 the authors responded with a reasonable description in the manuscript of the limitations of HealthBench for testing OpenAi models. and stated "We are happy to strengthen this language further if the reviewer considers it necessary." I would feel more comfortable with strengthened language, but I do not ask for another review round just to address this point.

Response to reviewers

We thank the referees of our initial submission for their thoughtful and considerate comments. We addressed each point and feel the manuscript is substantially stronger thanks to their insights. Please find below a point-by-point response to the suggestions and criticisms raised by the referees of the first submission and how the changes are incorporated into the revised manuscript. For clarity, we label each reviewer comment as RX.Y , where X is the reviewer number and Y is the number of their comment. We then place our response immediately after each of their comments. Excerpts or quotes that were added to or changed in our manuscript are included in *blue italics*.

Reviewer Two:

Major Concerns

R2.1: *Epistemic Circularity and Evaluation Bias. A central limitation is the reliance on HealthBench (developed by OpenAI) combined with GPT-4.1 (also OpenAI) as the sole evaluator, while one of the top-performing systems evaluated is GPT-5, likewise developed by OpenAI. This creates an unavoidable epistemic circularity in which benchmark design, scoring norms, and model optimization share institutional and methodological lineage.*
Even absent direct data leakage, this alignment systematically advantages models optimized under similar reward, instruction-following, and stylistic paradigms—particularly for subjective dimensions such as communication quality, uncertainty framing, and safety reasoning. HealthBench's own documentation emphasizes the need for meta-evaluation and validation against physician judgment, yet no such independent validation is presented here.
The absence of any blinded human (physician) evaluation, even on a subset of responses, severely weakens confidence in the reported performance differences.

Response: We thank the reviewer for this important and well-articulated concern. We agree that reliance on a single OpenAI-developed benchmark scored by a single OpenAI model, while simultaneously evaluating an OpenAI model as a top performer, introduced a legitimate risk of epistemic circularity. We have addressed this concern substantively through three major changes in the revised manuscript.

1. Addition of blinded physician evaluation via the Real Clinical Queries (RCQ) Benchmark: In direct response to this concern, we developed a novel, independent benchmark comprising 100 de-identified queries drawn from actual clinician interactions with a HIPAA-compliant GPT instance at NYU Langone during routine clinical care. Responses from all six models were evaluated by a panel of 12 U.S. physicians in a randomized, blinded design (raters were unaware of model identity), producing 1,800 model–question annotations scored across four clinician-defined dimensions (clinical correctness, completeness, safety/harm avoidance, and clarity for clinicians) on a 1–4 Likert scale, with binary flags for harmful content and hallucination. This evaluation is entirely independent of any model developer's benchmark or grading infrastructure. Importantly, the RCQ results corroborated the automated benchmark findings: frontier LLMs formed a statistically distinct upper performance tier (Gemini 3.62, GPT 3.54, Claude 3.52), while clinical tools and Google AI Overview formed a lower tier (OpenEvidence 3.24,

UpToDate 3.17, Google AI Overview 3.27), with all significant pairwise differences occurring between tiers rather than within them. Inter-rater concordance was strong (Kendall's W = 0.651, P = 2.3×10^{-7}), and all 12 clinicians independently ranked frontier LLMs above clinical tools.

2. Multi-model grading panel for HealthBench: To mitigate single-model grading bias on the HealthBench component, we replaced the sole GPT-4.1 grader with a panel-majority voting scheme across three frontier LLM judges (GPT-5.2, Gemini 3.1 Pro, and Claude Opus 4.6). This cross-model consensus approach reduces the risk that any one developer's stylistic or optimization biases systematically influence scoring. We now describe this explicitly in the Online Methods.

3. Acknowledgment of residual limitations: We have expanded our limitations discussion to explicitly note that HealthBench is an OpenAI-developed benchmark, that scores should be interpreted as benchmark-specific, that evaluation of OpenAI models on HealthBench may be affected by benchmark–developer overlap, and that we cannot exclude partial benchmark exposure for any evaluated model. We further note that frontier models serving as both evaluated systems and grading panel members introduces a potential for same-model bias. In our revised manuscript, we state:

"Standardized benchmarks have known shortcomings: models may have been exposed to MedQA or HealthBench during training, though our RCQ benchmark is free from this contamination. HealthBench is an OpenAI-developed benchmark that relies on a small number of physicians for each rubric, and public documentation provides limited detail on its construction and evaluation. Evaluation of OpenAI models, including the highest-scoring model, GPT-5.2, on HealthBench may be influenced by potential benchmark–developer overlap. Grading bias is also possible, since frontier models served as both evaluated systems and judges, although we used a multi-model panel to mitigate this effect."

Taken together, these additions provide the independent, physician-grounded validation the reviewer rightly identified as missing.

R2.2: *Unfair and Non-Comparable Querying Conditions*
The comparison framework is fundamentally asymmetric:
• Clinical tools (OpenEvidence, UpToDate Expert AI) were manually queried via browser interfaces
• Generalist models were queried via APIs
These modalities differ in ways that materially affect outputs, including hidden system prompts, retrieval behavior, response length defaults, latency constraints, safety guardrails, and UI-imposed truncation. As a result, observed differences cannot be cleanly attributed to "clinical vs. generalist" capability. Moreover, the clinical tools are retrieval-augmented systems designed for reference use, not free-form benchmark answering. Evaluating them using a single-turn, free-text paradigm optimized for generalist LLMs introduces structural bias against their intended design goals.

Response: We thank the reviewer for highlighting this important concern regarding asymmetry in querying conditions. We agree that querying clinical systems via browser interfaces while querying generalist models via APIs introduces unavoidable differences, including hidden system prompts, retrieval configuration, response-length defaults, UI truncation, latency constraints, and safety guardrails.

These factors can materially influence outputs and limit the ability to attribute performance differences solely to "clinical vs. general-purpose" capability.

However, this asymmetry reflects a structural constraint rather than a discretionary methodological choice. Many clinical systems evaluated in this study do not provide API access, configurable decoding parameters, or transparent system prompts. In such cases, the most reproducible and externally valid approach is to evaluate systems through the method that minimizes possible interferences. While querying generalist models via browser interfaces could create superficial symmetry, doing so would introduce additional uncontrolled UI-specific variables without resolving the fundamental limitation that clinical tools remain non-transparent and non-configurable.

We also agree that MedQA and HealthBench may not fully capture real-world clinical usage, particularly for retrieval-augmented clinical reference systems. To address this, we introduced a new physician-derived benchmark based on real clinical queries, designed to better reflect practical clinical information-seeking behavior. This benchmark complements MedQA and HealthBench and allows evaluation under conditions more aligned with the intended use of clinical tools.

Finally, we note that some clinical-system developers themselves report performance on MedQA-style benchmarks (e.g., public claims of USMLE-style accuracy), suggesting that such evaluations are considered relevant within the field. Nonetheless, we have added explicit discussion of modality asymmetry and benchmark-design limitations to the manuscript and interpret cross-system comparisons conservatively.

We also note in our limitations:

"Clinical tools lacked public APIs, so we queried them through browser interfaces, which limited sample size and may have introduced differences in hidden prompts, retrieval behavior, and output formatting"

R2.3: *Benchmark Contamination and Decontamination Gaps. Both MedQA and HealthBench are public benchmarks, and the paper does not include:*
• Any decontamination analysis
• Any held-out or novel evaluation set
• Any justification for assuming newer frontier models have not seen benchmark content during training
This is especially problematic when comparing frontier models trained on large-scale web and academic corpora against tools that rely on RAG over curated medical knowledge bases. Without addressing contamination risk, claims of superiority lack epistemic grounding.

Response:

We appreciate this concern and agree that benchmark contamination is a serious threat to validity when evaluating models trained on large-scale web corpora against public benchmarks. We have addressed this directly in the revision through the introduction of a novel, held-out evaluation set and through expanded discussion of this limitation.

1. Novel held-out benchmark (RCQ). The Real Clinical Queries benchmark was constructed from 100 de-identified clinician queries submitted to a HIPAA-compliant institutional GPT instance during routine clinical care. These queries were never publicly available and could not have appeared in any model's training data. The RCQ therefore serves precisely as the held-out, contamination-proof evaluation the reviewer requests. Critically, the RCQ results were concordant with MedQA and HealthBench findings (frontier LLMs formed a distinct upper tier while clinical tools formed a lower tier) suggesting that contamination is unlikely to explain the observed performance gaps on the public benchmarks.

2. Explicit acknowledgment in limitations. We have expanded our revised manuscript to directly address this:

"Standardized benchmarks have known shortcomings: models may have been exposed to MedQA or HealthBench during training, though our RCQ benchmark is free from this contamination."

The convergence of results across two public benchmarks and one private, novel benchmark substantially mitigates the concern that contamination drives our conclusions.

R2.4: *Statistical Fragility and Overinterpretation*
The manuscript draws category-level conclusions from n=2 clinical tools vs. n=3 generalist models, using model-level statistical tests that are fragile at this scale. While item-level pairing exists (same questions across systems), the inferential framing often shifts toward model-count-based comparisons that are underpowered and potentially misleading.Subgroup analyses further exacerbate this issue, with wide confidence intervals suggesting limited statistical power.

Response: Thank you for this thoughtful comment. We agree that category-level comparisons based on a small number of systems (n=2 clinical, n=3 generalist) are inherently underpowered and should be interpreted cautiously. We have removed all grouped analysis from our manuscript and now rely only on inter-model comparisons.

R2.5: Limited Clinical Validity and Missing Safety Analysis
The evaluation focuses narrowly on benchmark correctness and alignment scores, which do not establish clinical safety or utility. There is:
• No analysis of harmful or contraindicated recommendations
• No structured error taxonomy
• No assessment of uncertainty handling or appropriate deferral
• No clinician-rated evaluation of usefulness in realistic workflows
The manuscript acknowledges that benchmarks are not equivalent to clinical trials, but nevertheless frames conclusions in ways that imply deployment-relevant superiority. This leap is not justified. Notably, prior randomized and deployment-level studies (e.g., physician diagnostic reasoning trials, clinical operations RCTs, agent-based evaluation frameworks) already provide stronger translational evidence than benchmark-only comparisons, yet these are not meaningfully engaged.

Response:

We appreciate this thorough critique and agree that benchmark performance alone does not establish clinical safety or deployment readiness. We have substantially addressed these concerns in the revision through the addition of the RCQ benchmark, which provides several of the elements the reviewer identifies as missing.

1. Clinician-rated evaluation in realistic workflows. The RCQ benchmark was built from 100 de-identified queries submitted by providers during live clinical care at NYU Langone, not synthetic or exam-style prompts. Twelve U.S. clinicians evaluated model responses in a randomized, blinded design.

"To develop the RCQ benchmark, we sampled 100 anonymous clinician queries to the NYU Langone HIPAA-compliant GPT instance. Twelve blinded clinicians scored six models' responses across four dimensions (clinical correctness, completeness, safety/harm avoidance, clarity) on a 1–4 scale (Extended Data Fig. 2). For each response, three raters were randomly assigned to evaluate."

2. Explicit safety and harm analysis. The RCQ evaluation included dedicated safety dimensions and binary harm/hallucination flags, directly addressing the reviewer's concern about harmful or contraindicated recommendations.

"Safety outcomes (Fig. 2f-g) did not differ across models: none produced more harmful content (Cochran's $Q = 4.00$, $P = 0.55$) or hallucinations ($Q = 5.00$, $P = 0.42$)."

3. Assessment of appropriate deferral: The revised manuscript now explicitly analyzes refusal rates as a proxy for uncertainty handling and deferral behavior.

"UpToDate AI refused 19% of queries (Fig. 2e), more than all other models (1–3%; $P < 0.01$) except Google AI Overview (6%; $P = 0.10$)."

4. Structured error characterization: The RCQ physician evaluation rubric (Extended Data Fig 2) provides a structured framework across four axes with defined anchor points (e.g., safety/harm avoidance ranges from "Unsafe/dangerous; misses critical red flags" to "Proactively safe: red flags/contraindications + escalation guidance"). The score distributions in Figure 2 further characterize where each model's failures concentrate; for example, OpenEvidence's weakness in clarity (mean 2.84) despite comparable correctness, as noted in our revised manuscript:

"OpenEvidence scored lowest on clarity (mean 2.84), suggesting its weakness was communication, not knowledge."

5. Tempered conclusions. We agree that benchmark results should not be framed as deployment-relevant superiority. Our revised conclusions explicitly call for further evaluation rather than implying readiness for deployment. In our revised manuscript:

"As generative LLMs become integrated into healthcare at the enterprise, individual clinician, and consumer levels, there is an increasing need for rigorous, independent evaluation on real world tasks."

We believe that the RCQ benchmark represents a meaningful step beyond benchmark-only comparison by grounding evaluation in real clinical workflows with blinded physician assessment of safety, correctness, completeness, and clarity.

R.2.6: *Overgeneralized and Marketing-Adjacent Claims*
Several claims—particularly around adoption rates and market penetration of clinical tools—appear sourced from company or press materials and are presented without adequate caveating. These should not be treated as epidemiologic facts.
More broadly, the conclusion that "specialization does not help" oversimplifies a complex literature and conflicts with peer-reviewed evidence showing benefits from medical fine-tuning, retrieval grounding, and clinician-in-the-loop evaluation.

Response: We thank the reviewer for this concern and have made targeted revisions to address both points.

1. Adoption and market claims: We agree that company-sourced adoption statistics should not be presented as epidemiologic facts. In the revised manuscript, we removed the specific market penetration claims that appeared in the original submission (e.g., "claimed to be used by 40% of U.S. physicians," "estimated 70% of major enterprise health systems"). The revised introduction introduces the clinical tools without unverified adoption statistics:

"These proprietary large language model (LLM) based tools promise superior clinical performance to general-purpose frontier LLMs through domain-specific post-training or retrieval-augmented generation (RAG)."

2. Tempered conclusions regarding specialization: we agree that the original framing risked oversimplifying a nuanced literature. The revised manuscript qualifies the conclusion with "for particular tasks" and avoids blanket claims that specialization is unhelpful. From our revised manuscript:

"The superior performance of frontier models in our study suggests that scale, alignment, and cross-domain reasoning may outweigh domain-specific tuning as determinants of medical competency for particular tasks which has implications for procurement, reimbursement, and regulatory oversight."

We removed the original claim that "the perceived distinction between 'clinical LLMs' and 'general-purpose LLMs' may become increasingly artificial." Additionally, we cite work that provides the opposing perspective – specifically, Jiang et al., which argues that general-purpose foundation models are not sufficiently clinical for hospital operations. The revised manuscript also hedges on mechanism:

"As the architecture of the proprietary systems is unknown, it is hard to definitively assess why they underperformed general-purpose models."

We wish to clarify that our findings do not claim specialization is inherently unhelpful across all settings. Rather, our results indicate that on the specific benchmarks and real-world queries evaluated here, two widely deployed specialist tools did not outperform frontier generalist models. We agree with the reviewer that medical fine-tuning, retrieval grounding, and clinician-in-the-loop evaluation have demonstrated benefits in other contexts, and our conclusions should not be read as contradicting that literature.

R2.7: *Reproducibility and Transparency Gaps*
The methods lack sufficient detail to enable independent replication. Missing or underspecified elements include:
• Exact prompting templates (including system messages)
• Sampling procedures and random seeds
• Generation parameters (temperature, top-p, token limits, retries)
• UI interaction details for clinical tools
• Response normalization or length constraints
• Inter-rater reliability for automated scoring
As written, it is not possible to disentangle model capability from evaluation setup.

Response: Thank you for highlighting the need for greater methodological transparency.

1. Generation parameters: the revised Online Methods now specifies deterministic generation settings:

"Frontier model generations were conducted using fixed, deterministic parameters. Search tools were enabled for all runs. The temperature was set to 0.0 to eliminate sampling variability, and a fixed generation seed of 62 was used."

"The 1,000 items were used to evaluate three frontier LLMs via API: GPT-5.2 (accessed 02/26, GPT-5.2-2025-12-11), Gemini 3.1 Pro Preview (accessed 02/26), and Claude Opus 4.6 (accessed 02/26). Two clinical tools, OpenEvidence (accessed 09/25, 02/26) and UpToDate Expert AI (accessed 11/25, 02/26), were queried manually through browser interfaces."

2. Sampling procedures and random seeds: the revised methods now documents the sampling approach:

"We randomly sampled (seed=62) 500 USMLE-style questions from MedQA and 500 single-turn prompts from HealthBench."

3. Scoring reliability and cross-validation: for MedQA, we now describe a dual-extraction approach:

"GPT-4.1 (gpt-4.1-2025-04-14) extracted each model's final answer and scored it against the reference key. Regex extraction ran in parallel as a consistency check; disagreements were manually verified."

For HealthBench, we replaced single-model grading with a consensus panel:

"Grading used panel-majority voting across three judges (Claude Opus 4.6, Gemini 3.1 Pro Preview, GPT-5.2)."

4. Inter-rater reliability for physician evaluation. The revised manuscript reports multiple reliability metrics for the RCQ. From the Online Methods:

"Inter-rater reliability was assessed with Krippendorff's alpha (ordinal). Item-level agreement was fair (α = 0.10-0.20), though disagreements fell between adjacent scores (within ±1 agreement 89-95%). When collapsed to acceptable (3-4) versus unacceptable (1-2), agreement was higher (PABAK = 0.55-0.83). Agreement on binary safety flags was high (PABAK = 0.86 for harm, 0.95 for hallucination)."

5. UI interaction details for clinical tools. We acknowledge that we cannot fully specify the hidden system prompts, retrieval behavior, or output constraints of browser-based clinical tools, as these are proprietary. This is explicitly noted as a limitation in our revised manuscript:

"Clinical tools lacked public APIs, so we queried them through browser interfaces, which limited sample size and may have introduced differences in hidden prompts, retrieval behavior, and output formatting."

This is an inherent constraint of evaluating closed commercial systems and applies equally to any independent evaluation of these tools. Prompt(s) are included in the extended data material.

Reviewer Three:

R3.1: *Originality and significance: if not novel, please include reference*
It is certainly of interest to the readership of the journal. There is an important recent overlapping work https://arxiv.org/pdf/2512.01241, but this is a preprint. The authors of the manuscript under review cite other preprints, and citing and discussing the results of (First, do NOHARM: towards clinically safe large language models D Wu, FN Haredasht, SK Maharaj, P Jain, J Tran, M Gwiazdon, A Rustagi, J Jindal… arXiv preprint arXiv:2512.01241, 2025•arxiv.org) would add to the manuscript. On this NOHARM benchmark a domain specific model does outperform the best performing mass-market model, but only by a very narrow margin. Discussion of the context for the difference between these core results would be interesting for all readers of the manuscript under review.

Response: We thank the reviewer for highlighting this important work. We agree that the NOHARM benchmark provides valuable complementary evidence, particularly given its focus on clinical safety – a dimension that knowledge-based and alignment-based benchmarks may not fully capture. In the revised manuscript, we have added both the citation and a dedicated discussion paragraph addressing NOHARM.

"Additionally, recently proposed safety-focused evaluations of LLM medical recommendations such as the NOHARM framework suggest that knowledge and communication benchmarks may not fully capture clinical risk, and our results should be read with this in mind."

R3.2: *[TITLE] "Generalist Large Language Models Outperform Clinical Tools on Medical Benchmarks". Is the term "Generalist "appropriate here. …Or rather, it is correct and appropriate but also confusing.*

The term has two overlapping meanings and can confuse. All of the models tested in the manuscript study have "Generalist" properties. OpenEvidence and UpToDate Expert AI are" Medically "Generalist", as they take broad questions across all medicine disciplines and all medical stages producing answers that can be graded on these multiple axes. Gemini/ChatGPT/Sonnet are generally-"Generalist".

Both a "Generalist" in their own way.

As a second point, both OpenEvidence and UpToDate Expert AI are LLMs at their core and the original title could confuse the readers - who cold interpret it as "Generalist Large Language Models Outperform" [non-LLM] "Clinical Tools on Medical Benchmarks"

An accurate and more informative title would be:

"Mass-market Large Language Models Outperform Major Commercial Clinical Models on Medical Benchmarks" Tat is 99 characters and may be longer than allowed, but it is not for me to write a title - just to point out limitations of the existing one.

Response: We thank the reviewer for this thoughtful observation. We agree that "Generalist" may carry ambiguity, as both categories of models are indeed generalist within their respective domains, and that the original title could be misread as implying the clinical tools are not LLM-based. In the revised manuscript, we have changed the title to address both concerns:

"General-Purpose Large Language Models Outperform Specialized Clinical AI Tools on Medical Benchmarks"

This revision makes three clarifications that respond directly to the reviewer's points. First, "General-Purpose" replaces "Generalist," distinguishing these models by their intended scope rather than using a term that could apply to medically broad clinical tools as well. Second, "Specialized Clinical AI Tools" replaces "Clinical Tools," making clear that the comparison is between two categories of AI systems rather than between LLMs and non-LLM tools. Third, the framing now emphasizes the contrast in design philosophy (general-purpose vs. specialized) rather than implying a categorical difference in underlying technology.

We appreciate the reviewer's suggested alternative ("Mass-market Large Language Models Outperform Major Commercial Clinical Models on Medical Benchmarks") and agree it captures an important dimension — commercial availability and market positioning. However, we felt "General-Purpose" vs. "Specialized" more precisely describes the technical distinction at issue (broad training vs. domain-specific optimization), which is the core hypothesis being tested. We are open to further refinement if the reviewer or editors prefer alternative phrasing.

R3.3: *[MAIN]. "We assessed two widely deployed clinical AI systems (OpenEvidence and UpToDate Expert AI) against three state-of-the-art generalist LLMs (GPT-5, Gemini 3 Pro, and Claude Sonnet 4.5) using a 1,000-item mini-benchmark combining MedQA (medical knowledge) and HealthBench (clinician-alignment) tasks."*

It is important that the manuscript acknowledges that HealthBench is a system developed by OpenAI. All the authors of the preprint are from OpenAI. The preprint is relatively lacking in information on the development of the "5,000 multi-turn conversations between a model and an individual user or healthcare professional". The preprint states that "HealthBench comprises 5,000 realistic health conversations between a model and an individual user or healthcare professional. Conversations were selected for relevance, realism, and difficulty, and span a wide range of geographies, languages, and healthcare personas. Given a conversation, the task for an LLM is to respond to the last user message." My exploration of the system is limited to the preprint alone, and some general community discussion of it. It may be very good, but it does rely on a relatively small number of physician opinions per rubric, and not highly transparent flexible criterial for the evaluation these rubrics. It is not really known how 'good' the physicians where and how "right" they were in the assessment of rubrics.

In the preprint, OpenAI then use the developed HealthBench to test a series of Mass-market LLMs, and report, that, surprisingly.... their model, GPT-4o is best of all models ... . It is important to acknowledge this linking of HealthBench to one of the tested models in the current manuscript under review - which, is the next generation of models tested in the preprint Arora et al. It is highly highly likely that OpenAI had privilege knowledge and access (including aspects not in the public domain)to HealtBench in the further development of GPT models.

In the limitations section of the current manuscript the risk of access of models to benchmark is discussed - this point about the HeathBench, and specifically that one of the test models is from the same company, needs to be addressed more clearly and directly. More emphasis is also needed on the point about models potentially having access to the benchmarks in their training. This is highly likely to be the case. To some extent, this is unavoidable and to an extent benchmarking mark test studies actually test the "newness" of models with respect to the benchmarks. This clearly benefits the mass market models with the current faster development turnaround than the medical domain fine-tuned models. More acknowledgement of this is needed.

Was "HealthBench Hard" from Arora used in the manuscript study? From the preprint "We also introduce HealthBench Hard, a subset of 1, 000 HealthBench examples chosen because they are difficult for current frontier models. HealthBench Hard is designed to be a difficult target for future models, with many current models achieving a score of zero on it. The procedure for selecting HealthBench Hard examples is described in Appendix C." If not, why was HealthBench Hard" not used. It might provide important insights.

Response: We thank the reviewer for this detailed and important critique. We agree that HealthBench's OpenAI origin must be explicitly acknowledged, particularly given that a GPT model is among the top performers in our evaluation. The revised manuscript now addresses this directly:

"HealthBench is an OpenAI-developed benchmark that relies on a small number of physicians for each rubric, and public documentation provides limited detail on its construction and evaluation. Evaluation of OpenAI models, including the highest-scoring model, GPT-5.2, on HealthBench may be influenced by potential benchmark–developer overlap."

We appreciate the reviewer's additional observations regarding the limited transparency around physician selection and rubric development in the HealthBench preprint, and the fact that prior HealthBench evaluations similarly found OpenAI's own model (GPT-4o) to be the top performer. Even absent direct data leakage, optimization under similar reward signals and instruction-following paradigms could systematically advantage GPT models. We are happy to strengthen this language further if the reviewer considers it necessary.

The reviewer also raises an important structural point about benchmark freshness: frontier models undergo faster development cycles and are more likely to have encountered public benchmarks during training. The revised manuscript acknowledges that frontier models *"also benefit from faster iteration cycles, larger training corpora, and greater alignment than specialist systems"* Crucially, however, the RCQ benchmark (novel, privately held, and impossible to have appeared in any model's training data) produces the same tier structure as MedQA and HealthBench. This convergence across contamination-vulnerable and contamination-proof evaluations, combined with the shift to a multi-model grading panel (GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6) replacing single-model GPT-4.1 scoring, substantially mitigates the concern that our conclusions are artifacts of HealthBench-specific biases.

Regarding HealthBench Hard: we sampled randomly from the full HealthBench rather than the Hard subset, as described in the Online Methods ("filtering to single-turn prompts and randomly sampling 500 items with seed 62"), to provide a representative cross-section of clinically relevant tasks. We agree that HealthBench Hard could yield additional insights, particularly in revealing whether ceiling effects mask meaningful differences among top performers or whether clinical tools fare comparatively better on items designed to challenge frontier systems.

R3.4: *[MAIN] "GPT-5 achieving the highest scores, while OpenEvidence and UpToDate demonstrated deficits in completeness, communication quality" Can more information be provided on the assessment of communication quality. Should this necessarily be high for OpenEvidence and UpToDate Expert AI. These models have a specific target user, who are Health Care Professionals (HCPs). They are not designed to communicate with patients directly. Does the communication metric here only assess the communication with HCPs, or does it assess general or lay communication? If it does, then it is not a fair assessment for the task in hand. Much more context is needed on this metric and its applicability.*

Response: We thank the reviewer for this important clarification. We agree that communication quality must be assessed relative to the intended audience, and that clinical tools designed for HCPs should not be penalized for lacking patient-facing communication features.

The new RCQ benchmark directly addresses this concern by design. The 100 queries were drawn from actual clinician interactions with a HIPAA-compliant GPT instance during routine clinical care at NYU Langone — these are questions physicians asked in the course of their work, not patient-facing or lay prompts. The evaluation was then conducted by 12 U.S. physicians, and the relevant communication dimension is explicitly defined as "Clarity for clinicians," not general or lay communication.

The rubric anchors range from "Confusing/disorganized" (score 1) to "Exceptionally clear and skimmable" (score 4), assessing structure and readability for a physician audience specifically. In other words, both the queries and the evaluation criteria reflect the exact use case these clinical tools are designed for – a clinician seeking actionable medical information. Despite this HCP-targeted framing and the fact that clinicians were grading the model responses, OpenEvidence still scored lowest on this dimension. From our revised manuscript:

"OpenEvidence scored lowest on clarity (mean 2.84), suggesting its weakness was communication, not knowledge."

This indicates that the deficit is not an artifact of evaluating clinical tools against a patient-communication standard, but rather reflects difficulty in presenting information in a format that physicians themselves found clear and actionable during the kind of workflow these tools are marketed to support.

R3.5: [MAIN]. "Frontier generalist models benefit from substantially larger training corpora and more advanced alignment, which may enable them to match or exceed specialist performance depending on the task." Related to Comment 2. Frontier generalist models also potentially benefit from more frequent update iterations. Both GPT5 and Gemini3 are very recent models developed through enormous investment. The phenomena/results described in the manuscript may reflect our current short-term development curve / exponential stage of the development curve of the Frontier generalist models. The rate of increase of the performance of the very large generalist models may slow. Many experts have predicted that it will (or is already doing so). At this point the advantage of domain tunes model may become apparent. Do the authors agree? Some (brief) discussion of this point in the manuscript would make sense.

Response: We agree with the reviewer. The performance advantages we observe may in part reflect the current rapid iteration cycle and massive investment in frontier generalist models — a phase that may not persist indefinitely. If scaling returns diminish, as many experts have predicted, domain-specific tuning and retrieval grounding could become increasingly valuable differentiators.

The revised manuscript addresses this:

"They also benefit from faster iteration cycles, larger training corpora, and greater alignment than specialist systems."

However, we agree that the converse point (that this advantage may be transient) deserves explicit acknowledgment. We propose adding the following to the discussion:

"The observed advantages of frontier general-purpose models may partly reflect the current pace of development and investment in these systems. Should scaling returns diminish, the relative value of domain-specific tuning, curated retrieval, and clinician-in-the-loop optimization may increase. Our results should therefore be interpreted as a snapshot of a rapidly evolving landscape rather than a permanent ordering of approaches."

This framing preserves our empirical findings while acknowledging, as the reviewer rightly notes, that the competitive dynamics between generalist and specialist systems are likely to shift as the field matures.

R3.6: *"Moreover, standardized benchmarks, while allowing quantitative comparison, have inherent shortcomings.10 It is possible that recently trained models or tools have been trained on the benchmarks themselves, which were incidentally swept up in their training data."*

- More detail on these "inherent shortcomings" is needed here, perhaps from reference 10.
- More consideration is needed of "It is possible that recently trained models or tools have been trained on the benchmarks themselves, which were incidentally swept up in their training data.", as mentioned under my Comment 2.

Response: Thank you for pointing this out. In response to the potential of data contamination and the adjacent risks, we have developed a new, private dataset to evaluate the models. This is described in detail in R3.4.

We also note in regard to reference 10:

"Additionally, recently proposed safety-focused evaluations of LLM medical recommendations such as the NOHARM framework suggest that knowledge and communication benchmarks may not fully capture clinical risk, and our results should be read with this in mind."

R3.7: *References: appropriate credit to previous work?*
As discussed under B., There is an important recent overlapping work https://arxiv.org/pdf/2512.01241, but this is a preprint. The authors of the manuscript under review cite other preprints, and citing and discussing the results of (First, do NOHARM: towards clinically safe large language models
D Wu, FN Haredasht, SK Maharaj, P Jain, J Tran, M Gwiazdon, A Rustagi, J Jindal…
arXiv preprint arXiv:2512.01241, 2025•arxiv.org) woudl add to the manuscript.

Response: Please see R3.1.

R3.8: *H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions*

The manuscript is very strong in these areas - More discusssion context on the following preprint would help the reader: https://arxiv.org/pdf/2512.01241, (First, do NOHARM: towards clinically safe large language models D Wu, FN Haredasht, SK Maharaj, P Jain, J Tran, M Gwiazdon, A Rustagi, J Jindal…
arXiv preprint arXiv:2512.01241, 2025•arxiv.org) woudl add to the manuscript.

Response: We are glad to hear that our piece is clear. See R3.1 in response to the recommended reference.
Editor:

E.1: *Thank you for your patience during the review process. Your Brief Communication, "Generalist Large Language Models Outperform Clinical Tools on Medical Benchmarks" has now been seen by 2*

referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Medicine, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

In addition to addressing all reviewer comments, please strengthen the methodological rigor to more robustly support the findings. Of particular importance is resolving benchmark contamination, incorporating a blinded clinician evaluation, and revising the evaluation metrics to ensure they align with the intended use.

Response: Regarding items of particular importance, we now provide a summary of the changes that address these points.

Resolving Benchmark Contamination: We developed the Real Clinical Queries (RCQ) benchmark – an entirely novel, privately held evaluation set comprising 100 de-identified queries drawn from practicing providers interacting with a HIPAA-compliant GPT instance during routine clinical care at NYU Langone. These queries were never publicly available and could not have appeared in any model's training data, providing a contamination-proof evaluation that no prior study of clinical AI tools has attempted. The RCQ results independently replicate the MedQA and HealthBench findings, producing the same two-tier performance structure across all dimensions. This convergence across two public benchmarks and one private benchmark substantially mitigates contamination as an alternative explanation. We additionally expanded the revised manuscript to explicitly acknowledge HealthBench's OpenAI provenance, the risk of benchmark–developer overlap advantaging GPT models, and the structural advantage that faster iteration cycles confer on frontier models with respect to public benchmarks.

Blinded Clinician Evaluation: The RCQ benchmark underwent a rigorous independent physician evaluation of clinical AI tools. Twelve U.S. clinicians evaluated model responses in a fully randomized, blinded design, producing 1,800 model–question annotations. Responses were scored across four clinician-defined dimensions (clinical correctness, completeness, safety/harm avoidance, clarity for clinicians) on a 1–4 Likert scale, with binary flags for harmful content and hallucination. Inter-rater concordance was strong (Kendall's $W = 0.651$, $P = 2.3 \times 10^{-7}$), with all 12 clinicians independently ranking frontier LLMs above clinical tools. Effect sizes were quantified with rank-biserial correlations and 5,000 bootstrap iterations. Cumulative link models with rater fixed effects served as the primary regression framework, with linear mixed models as sensitivity checks. This is, to our knowledge, the first benchmark of clinical AI systems using actual physician queries generated during the course of patient care.

Revising Evaluation Metrics: The revision addresses metric alignment from multiple angles. First, the RCQ was purpose-built to reflect the actual clinical use case: queries came from practicing physicians, the evaluation rubric (Extended Data Fig. 2) explicitly specifies that the audience is HCP unless patient-facing language is requested, and all raters were practicing clinicians evaluating responses they would encounter in their own workflows. Second, safety was assessed through both a dedicated harm avoidance dimension (scored 1–4) and binary flags for harmful content and hallucination, with formal

statistical testing across models (Cochran's Q). Third, refusal behavior was quantified, revealing that UpToDate AI refused 19% of queries — significantly more than all other models. Fourth, we replaced single-model GPT-4.1 grading of HealthBench with panel-majority voting across three frontier LLMs from different developers (GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6) to mitigate single-developer scoring bias. Fifth, we added Google Search AI Overviews as a real-world control, representing the baseline AI interaction available to any clinician and found that clinical tools marketed for clinical decision support performed comparably to this free, auto-enabled search feature.

Response to reviewers

We thank the referees of our submission for their thoughtful and considerate comments. We addressed each point and feel the manuscript is substantially stronger thanks to their insights. Please find below a point-by-point response to the suggestions and criticisms raised by the referees of the second submission and how the changes are incorporated into the revised manuscript. For clarity, we label each reviewer comment as RX.Y , where X is the reviewer number and Y is the number of their comment. We then place our response immediately after each of their comments. Excerpts or quotes that were added to or changed in our manuscript are included in *blue italics*.

Reviewer Two:

R2.1: *The multi-model grading panel is a reasonable fix, but residual circularity remains. Replacing solo GPT-4.1 grading with consensus voting across GPT-5.2, Gemini 3.1 Pro, and Claude Opus 4.6 is better than what they had, but the authors correctly acknowledge that using frontier models as both evaluated systems and judges is still problematic. The RCQ physician evaluation serves as the real escape valve here. The human evaluation should be the primary evidence and HealthBench is supplementary. As Audre Lorde quipped, "The master's tools will never dismantle the master's house." We are being played by the tech companies.*

Response: Thank you for this comment. We agree that, although replacing single-model grading with a multi-model panel reduces bias, it does not eliminate the residual circularity that arises when frontier models are used both as evaluated systems and as judges. We also agree that the blinded physician evaluation on the RCQ benchmark should be regarded as the primary evidence in the manuscript, with HealthBench serving as a supplementary benchmark. To make this clearer, we have added brief language emphasizing the primacy of the RCQ human evaluation and clarifying the limitations of HealthBench, including potential benchmark–developer overlap and judge-model circularity. This framing is consistent with the manuscript's existing emphasis on RCQ as the real-world, contamination-resistant component of the study and on HealthBench as a more limited secondary benchmark.

"HealthBench is an OpenAI-developed benchmark that relies on a small number of physicians for each rubric, and public documentation provides limited detail on its construction and evaluation[5]. Evaluation of OpenAI models, including the highest-scoring model on HealthBench, GPT-5.2, may be influenced by potential benchmark–developer overlap, including potential similarities in training data, optimization objectives, or rubric design. Grading bias is also possible, since frontier models served as both evaluated systems and judges, although we used a multi-model panel to mitigate this effect. ***Accordingly, we view the blinded physician evaluation on the RCQ benchmark as the primary evidence in this study, while HealthBench should be interpreted as supplementary.***"

[5] Arora, R. K. et al. HealthBench: Evaluating large language models towards improved human health. arXiv [cs.CL] (2025) doi:10.48550/ARXIV.2505.08775.

R2.2: *The authors make a fair argument that the asymmetric querying concern is a structural constraint rather than a methodological choice: clinical tools don't offer APIs. The addition of the RCQ benchmark helps because it evaluates all systems on real clinical queries, but the fundamental issue that browser-queried systems face hidden UI constraints remains. How about matching response length or format across systems?*

Response: We agree that the asymmetric querying constraint is a meaningful limitation, but we believe that matching response length or format across systems would itself introduce artifacts. Output structure (including verbosity, use of headers, and formatting) is an integral part of each system's clinical interface and directly affects dimensions like clarity, which our physician raters evaluated. Artificially truncating or reformatting responses would obscure real differences in usability that clinicians encounter in practice. The same logic applies to length: if a model's verbosity detracts from clarity, we believe that it should register as a genuine performance difference rather than be normalized away. Hospital systems and clinicians can opt to not use the default, potentially constrained, UI for the frontier models by using their APIs (and often do), but there is no alternative for the clinical tools that we evaluate.

That said, we appreciate the spirit of the suggestion and have added a sentence in our Online Methods addressing this matter:

"Although matching response length or format across systems would reduce one source of variability, we opted against this approach because output structure, verbosity, and formatting are integral to each system's clinical interface and directly influence dimensions such as clarity; normalizing these features would obscure real usability differences that clinicians encounter in practice."

R2.3: *The revised manuscript reported that no model produced significantly more harmful content or hallucinations (Cochran's Q non-significant). It presented a scoring rubric with anchor points, but not a systematic categorization of error types, a structured error taxonomy. The OpenEvidence clarity finding (mean 2.84) is a nice granular insight, but this was not really a failure mode analysis.*

Response: Thank you for this comment. We agree that the non-significant Cochran's Q result and the OpenEvidence clarity finding do not constitute a systematic failure-mode analysis. To engage with this concern, we reviewed all (optional) rater free-text notes associated with scores of 1 or 2 on any evaluation dimension (excluding refusals) and classified each into one or more

inductively derived error categories: factual error, incomplete clinical content, omission of safety-critical information, disorganized or hard-to-follow reasoning, audience mismatch or verbosity, hallucination or unsupported claim, and failure to answer the question posed (**Extended Data Table 2**). The most frequent failure modes were incomplete clinical content and omission of safety-critical information. Clinical tools — particularly OpenEvidence — were disproportionately represented in the disorganization category, consistent with the clarity deficit reported in the main text. Frontier model errors, when they occurred, were more commonly omissions of completeness rather than factual inaccuracies or safety failures.

We include this taxonomy as supplementary material to provide the reviewer and readers with additional transparency into the nature of low-scoring responses. However, because these categories were derived post hoc from unstructured rater notes rather than from a prospectively designed error annotation protocol, we do not believe it would be appropriate to present them as a formal result in the manuscript. A prospective failure-mode analysis with predefined error categories, review of full response text, and formal inter-annotator agreement would be a valuable follow-up study.

Model	Factual error	Incomplete clinical content	Safety-critical omission	Disorganized / hard to follow	Audience mismatch / verbosity	Hallucination / unsupported claim	Did not answer the question	Total
GPT-5.2	3	8	4	1	4	1	0	21
Gemini 3.1 Pro	1	4	2	0	1	0	0	8
Claude Opus 4.6	3	5	4	3	4	0	0	19
OpenEvidence	4	15	12	13	4	2	2	52
UpToDate Expert AI	2	8	3	1	2	1	3	20
Google AI Overview	7	15	8	0	1	0	2	33
Total	**20**	**55**	**33**	**18**	**16**	**4**	**7**	**153**

Extended Data Table 1. Preliminary error taxonomy of low-scoring responses on the Real Clinical Queries benchmark. Rater free-text notes associated with scores ≤2 on any evaluation dimension (excluding refusals; n = 101 annotated notes) were classified post hoc into seven inductively derived error categories. Leaving a note for a response was optional, and not all low-scores had notes associated with them. Notes could be assigned to more than one category; row totals therefore exceed the number of unique annotations. Color shading indicates model type: blue = frontier LLM, orange = clinical AI tool, green = search-embedded AI.

We also note this in our discussion of results in our revised manuscript:

*"Qualitatively, incomplete clinical content, safety-critical omissions, and disorganized responses were common, particularly for OpenEvidence and Google AI Overview (**Extended Data Table 1**)."*

R2.4: *There's still no cost, latency, or citation quality comparison. These are practical considerations for clinical deployment that benchmark studies typically ignore, and they are not even acknowledged as limitations.*

Response: Thank you for this comment. We have now included a supplemental table detailing the cost comparison between these models (note that the clinical tool models are subscription based and do not charge per token, in the status quo).

Model	**Access Type**	**Cost Structure**
GPT-5.2	API	$1.75 / $14.00 per 1M input/output tokens
Gemini 3.1 Pro Preview	API	$2.00 / $12.00 per 1M input/output tokens
Claude Opus 4.6	API	$5.00 / $25.00 per 1M input/output tokens
OpenEvidence	Browser (free)	Free for verified U.S. clinicians (ad-supported)
UpToDate Expert AI	Browser (subscription)	~$699/year (UpToDate Pro Plus, individual)
Google AI Overview	Browser (free)	Free (integrated into Google Search)

Extended Data Table 2. Cost comparison of evaluated AI systems. Estimated costs for frontier LLMs, clinical AI tools, and Google AI Overview as of February 2026. Frontier LLM

costs reflect standard pay-per-use API pricing and do not include reasoning tokens, search tool surcharges, or caching discounts; costs may differ on third-party platforms (e.g., AWS Bedrock, Google Vertex AI). OpenEvidence and UpToDate Expert AI are subscription- or ad-supported products that do not charge per token; OpenEvidence is free for verified U.S. healthcare professionals, while UpToDate Expert AI requires a Pro Plus (~$699/year) or Enterprise subscription. Google AI Overview is freely available within Google Search. Direct cost-per-query comparisons between per-token and subscription models are inherently limited, as the effective per-query cost of subscription tools depends on usage volume.

For the citation quality and user-facing latency, we have added a sentence into our discussion acknowledging this as an avenue of future work and important factors for clinical deployment.

"Finally, our evaluation did not assess response latency or citation quality. These factors are important for real-world clinical deployment and workflow integration and that may differ substantially between API-accessed frontier models and subscription-based clinical tools (Extended Data Table 2). Future work should systematically and compare these practical dimensions alongside accuracy and safety."

R2.5: *In summary, please clarify inter-rater reliability distinctions, address model version changes between submission rounds, and add at minimum a paragraph engaging with alternative evaluation frameworks beyond industry-created benchmarks.*

Response: Thank you for this summary. We address each point:

First, the inter-rater reliability distinctions are detailed in the Methods. Item-level Krippendorff's α was fair (0.10–0.20), but disagreements were overwhelmingly between adjacent scores (within ±1 agreement: 89–95%). When collapsed to a clinically meaningful acceptable/unacceptable dichotomy, agreement was substantially higher (PABAK = 0.55–0.83), and agreement on binary safety flags was high (PABAK = 0.86–0.95).

Second, regarding model version changes between submission rounds, we used the most up-to-date frontier models when possible. In the revised version (R1 & R2), all frontier models were queried using the same model versions reported in the Methods (GPT-5.2-2025-12-11, Gemini 3.1 Pro Preview, Claude Opus 4.6; all accessed 02/26). Some Clinical tools were initially queried in 09/25–11/25 for the HealthBench & MedQA tasks, and re-queried on 02/26 for the RCQ benchmark.

Third, we have added a short paragraph in the Discussion engaging with alternative evaluation frameworks. Specifically, we discuss the NOHARM framework (Wu et al., 2025), which emphasizes that knowledge and communication benchmarks may not capture clinical risk, and

note the broader concern that industry-created benchmarks may systematically favor the systems developed by their creators. We frame our RCQ benchmark as a partial response to this concern (it is constructed from real clinical queries, evaluated by independent clinicians, and free from training-set contamination) while acknowledging that no single evaluation framework is sufficient. While we would love to add more on this important topic, we note that this is a brief communication under strict word limit requirements.

"Additionally, recently proposed safety-focused evaluations of LLM medical recommendations such as the NOHARM[12] framework suggest that knowledge and communication benchmarks may not fully capture clinical risk. Related work also points to health-system-grounded evaluation frameworks, such as institution-specific operational tasks and prediction settings embedded in local clinical workflows, as an important complement to public, industry-authored benchmarks, because they may better capture whether a model is clinically useful in a given care environment.[13,14]"

[12] Wu, D. et al. First, do NOHARM: towards clinically safe large language models. arXiv [cs.CY] (2025) doi:10.48550/arXiv.2512.01241.
[13] Jiang, L. Y. et al. Generalist foundation models are not clinical enough for hospital operations. arXiv [cs.CL] (2025) doi:10.48550/arXiv.2511.13703.
[14] Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. Nature 619, 357–362 (2023).

Reviewer Three:

I am satisfied that my comments have been reasonably addressed. For R3.3 the authors responded with a reasonable description in the manuscript of the limitations of HealthBench for testing OpenAi models. and stated "We are happy to strengthen this language further if the reviewer considers it necessary." I would feel more comfortable with strengthened language, but I do not ask for another review round just to address this point.

Response: Thank you for your comment. Per the reviewer's suggestion, we have strengthened the language regarding the limitations of HealthBench for evaluating OpenAI models. The revised text now reads:

"HealthBench is an OpenAI-developed benchmark that relies on a small number of physicians for each rubric, and public documentation provides limited detail on its construction and evaluation[5]. Evaluation of OpenAI models, including the highest-scoring model on HealthBench, GPT-5.2, may be influenced by potential benchmark–developer overlap, including potential similarities in training data, optimization objectives, or rubric design. Grading bias is also

possible, since frontier models served as both evaluated systems and judges, although we used a multi-model panel to mitigate this effect. Accordingly, we view the blinded physician evaluation on the RCQ benchmark as the primary evidence in this study, while HealthBench should be interpreted as supplementary."

[5] Arora, R. K. et al. HealthBench: Evaluating large language models towards improved human health. arXiv [cs.CL] (2025) doi:10.48550/ARXIV.2505.08775.