

Assessing Safety and Trustworthiness of Large Language Models in Medicine

Corresponding Author: Dr Zhiyong Lu

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Communications Medicine.

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I thank the authors for the chance to review their paper and their answers to the reviewers' questions. I am overall satisfied with this response. However, I have a few remaining requests:

1. Comparison on Accuracy Index and Safety Index in Fig 3 do not seem fair. MedQA and MedGuard contain different scope of questions, moreover, MedQA is an order of magnitude larger. Is there perhaps a way to compute accuracy on a modified version of MedGuard or pre-select a similar number of closest questions from MedQA for a fairer comparison? Also, p15 mentions Zhang et al., whereas ref 35 is to Jin et al.
2. Please consider the restricted-access release of the benchmark data within PhysioNet secure environment in compliance with MIMIC procedures rather than public release on Github. Github release should contain only the code not the data.
3. Please elaborate on the profile of medical experts, as well as native speakers who verified the benchmark Q&A.
4. Please double-check Supplementary material to ensure table caption and column names are readable. Consider reporting paraphrased data, additional outputs per aspect, as well as highlighting safety compliant / non compliant parts of responses in Supplementary Data.

On a minor note, I have noticed several typos (independently -> independently, sutdy -> study).

Reviewer #2

(Remarks to the Author)

Note: I was looking for a marked up paper draft with changes highlighted, but did not find it in the submission files. This made the work of re-reviewing much more difficult. I encourage the authors to submit annotated documents with changes highlighted in the future to simplify this process.

In general, I appreciate the authors' changes to the manuscript to address comments, especially around clarifying the motivation for the five principles and adding some requested further analysis. I still see the main contribution of this work as the MedGuard benchmark itself, which can be reused. Such a benchmark would be a valuable contribution to the community. As such, I recommend this manuscript be accepted with minor changes, suggested below.

- It would be good to comment on how training on the MedGuard dataset or equivalent might change a model's safety and accuracy performance characteristics. While such experiments may be out of scope for the current paper, which focuses on analysis and benchmarking, it would be a natural extension of this work and is missing from discussion.
- In many Results paragraphs and the Discussion section paragraph starting with "The improvement in different model families also reflects the development focus and prevention by different development groups..." — the text references improvements by aspect which are not obvious to see in Figure 4, which is very complex. I suggest adding numbers to help in the interpretation of results in this paragraph and others like it.
- Nit: The color contrast in several figures is quite low and results are hard to see. The results presented in Figure 4 are also overly complex and hard to interpret. Would suggest remaking these to improve ease of readability.

- Nit: Many typos have been introduced in the edits, e.g., Page 17, “We [further] compute Pearson’s correlation...” and missing articles in the Methods section.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I would like to thank the authors for their thorough responses. I am in general satisfied with those revisions.

I would be grateful though if the authors could update the Data Availability Statement. Also, rather than providing the temporary data-sharing link, I encourage the authors to provide the final GitHub repository with the encrypted file and instructions on how to use it. It would also be helpful to see an updated version of Supplementary Data 1 with paraphrased input for the equity example.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I have reviewed the authors' response, the revised manuscript and the provided code. I am satisfied with the revision and happy to endorse the manuscript.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>



Dear Editor,

We thank the reviewers for their thoughtful re-review of our manuscript, "Ensuring Safety and Trust: Analyzing the Risks of Large Language Models in Medicine." Per their comments, we have implemented the requested adjustments, which we believe have further strengthened the work. Detailed responses to the remaining suggestions are provided below.

Reviewer 1:

I thank the authors for the chance to review their paper and their answers to the reviewers' questions. I am overall satisfied with this response. However, I have a few remaining requests:

1. Comparison on Accuracy Index and Safety Index in Fig 3 do not seem fair. MedQA and MedGuard contain different scope of questions, moreover, MedQA is an order of magnitude larger. Is there perhaps a way to compute accuracy on a modified version of MedGuard or pre-select a similar number of closest questions from MedQA for a fairer comparison? Also, p15 mentions Zhang et al., whereas ref 35 is to Jin et al.

We thank the reviewer for this insightful suggestion. While we agree that aligning the datasets would provide a more direct comparison, modifying MedQA to perfectly match our dataset is unfeasible due to its distinct scope, beyond the specific sub-categories we have already analyzed and presented in Figure 2. Nonetheless, to prevent any potential misinterpretation and to more accurately describe the results, we have renamed the indices in Figure 3 from 'Accuracy Index' and 'Safety Index' to 'MedQA performance' and 'MedGuard performance,' respectively.

Regarding the citation on page 15, we thank the reviewer for pointing this out. We have clarified the text to ensure accuracy: reference 35 (Jin et al.) correctly refers to the dataset, whereas Zhang et al. corresponds to reference 37.

2. Please consider the restricted-access release of the benchmark data within PhysioNet secure environment in compliance with MIMIC procedures rather than public release on Github. Github release should contain only the code not the data.

Thank you for raising this important data-sharing concern. We fully agree with the necessity of restricting access to sensitive datasets. To clarify our current sharing strategy, our GitHub repository does not contain any raw or processed MIMIC data. Instead, we provide an

encrypted file that requires the original MIMIC dataset to act as the decryption key. A user can only reconstruct our benchmark dataset if they already possess authorized, local access to the MIMIC files. We believe this encryption approach strictly prevents unauthorized access and successfully complies with MIMIC redistribution policies.

3. Please elaborate on the profile of medical experts, as well as native speakers who verified the benchmark Q&A.

We have added a more detailed profile of the experts including their native language on page 32.

4. Please double-check Supplementary material to ensure table caption and column names are readable. Consider reporting paraphrased data, additional outputs per aspect, as well as highlighting safety compliant / non compliant parts of responses in Supplementary Data.

We thank the reviewer for these suggestions. We have updated Supplementary Table 2 to ensure the table captions and column names are fully readable. and we have included the code to generate and run the paraphrased data in our GitHub repository. Regarding the suggestion to highlight specific parts of the responses, we believe that the criteria for what constitutes a compliant versus non-compliant response are comprehensively defined in the Methods section.

On a minor note, I have noticed several typos (indipendently -> independently, sutdy -> study).

We have fixed the typos on page 11, 20.

Reviewer #2:

Note: I was looking for a marked up paper draft with changes highlighted, but did not find it in the submission files. This made the work of re-reviewing much more difficult. I encourage the authors to submitted annotated documents with changes highlighted in the future to simplify this process.

The newly submitted version now includes a marked-up manuscript highlighting all the changes made in response to the current feedback.

In general, I appreciate the authors' changes to the manuscript to address comments, especially around clarifying the motivation for the five principles and adding some requested further analysis. I still see the main contribution of this work as the MedGuard benchmark itself, which can be reused. Such a benchmark would be a valuable contribution to the community. As such, I recommend this manuscript be accepted with minor changes, suggested below.

- It would be good to comment on how training on the MedGuard dataset or equivalent might change a model's safety and accuracy performance characteristics. While such experiments may be out of scope for the current paper, which focuses on analysis and benchmarking, it would be a natural extension of this work and is missing from discussion.

We agree that exploring how safety-focused training impacts a model's safety and accuracy performance characteristics is a vital next step for the field. As suggested, we have expanded our discussion (Page 25) to address this potential extension. In the revised text, we clarify that while such training is highly beneficial, our current dataset is limited in size due to the significant expense and time required for expert medical annotation, making it unsuitable for direct model training. We also highlight the methodological risk that training models directly on benchmark-style data risks overfitting, which could distort the dataset's capability to serve as an objective evaluation tool.

- In many Results paragraphs and the Discussion section paragraph starting with "The improvement in different model families also reflects the development focus and prevention by different development groups..." — the text references improvements by aspect which are not obvious to see in Figure 4, which is very complex. I suggest adding numbers to help in the interpretation of results in this paragraph and others like it.

We have added reference to Figure 2 for the quoted statement in the comments on page 23. Please refer to the next response paragraph for adjustment in Figure 4.

- Nit: The color contrast in several figures is quite low and results are hard to see. The results presented in Figure 4 are also overly complex and hard to interpret. Would suggest remaking these to improve ease of readability.

We have updated Figure 4 with clearer coloring and explicit number captions so that it is more readable.

- Nit: Many typos have been introduced in the edits, e.g., Page 17, "We [further] compute Pearson's correlation..." and missing articles in the Methods section.

We have fixed these typos along with the typos mentioned by Reviewer 1.



Dear Editor,

We thank the reviewers for their thoughtful re-review of our manuscript, "Ensuring Safety and Trust: Analyzing the Risks of Large Language Models in Medicine." Per reviewer's comments and editorial instructions, we have implemented the requested adjustments. Detailed responses to the remaining suggestions are provided below.

Reviewer 1:

I would like to thank the authors for their thorough responses. I am in general satisfied with those revisions.

I would be grateful though if the authors could update the Data Availability Statement. Also, rather than providing the temporary data-sharing link, I encourage the authors to provide the final GitHub repository with the encrypted file and instructions on how to use it. It would also be helpful to see an updated version of Supplementary Data 1 with paraphrased input for the equity example.

[Thank you for reviewing our revisions and for the positive feedback.](#)

[We have updated the Data Availability Statement and replaced the temporary data-sharing link with the final GitHub repository link, which now includes the encrypted file and instructions for its use. Additionally, as requested, we have updated Supplementary Data 1 to include the paraphrased input for the equity example.](#)