PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

Title (Provisional)

ChatGPT (GPT-4) versus Doctors on Complex Cases of the Swedish Family Medicine Specialist Exam: An Observational Comparative Study

Authors

Arvidsson, Rasmus; Gunnarsson, Ronny; Entezarjou, Artin; Sundemo, David; Wikberg, Carl

---

VERSION 1 - REVIEW

Reviewer	**1**
Name	**OZTERMELI, Ahmet**
Affiliation	**Gebze Fatih State Hospital, Ortopedics and Traumatology**
Date	**31-Mar-2024**
COI	**No competing interest**

A recent study has been conducted comparing ChatGPT's performance with that of other doctors in the family medicine examination. ChatGPT was found to be less successful compared to human doctors.

1-) When questioning ChatGPT, was a case given with follow-up questions, or was it scored based on what was written at once? Providing a sample question model within the publication would make it more understandable.

2-) Were the evaluators of ChatGPT and human participants' responses specialists in family medicine? Is there a status difference between the human participants taking the exam? It has been stated that correct answers are not available online. Were correct answers determined based on specific textbooks?

3-) How was the selection made for top-tier doctors? It should be more explainable. Are the evaluators also other top-tier doctors?

4-) It has been mentioned that no other comparison has been made regarding examination results between human doctors. However, in the study provided below, a comparison was made based on the scores and rankings of human participants who took the examination. It is suggested that this be discussed in the debate.

https://pubmed.ncbi.nlm.nih.gov/37565917/

5-) What could be the future implications of ChatGPT's clinical usage? Only how its success in exams can be increased has been discussed.

Reviewer	**2**
Name	**Hirosawa, Takanobu**
Affiliation	**Dokkyo Medical University**
Date	**18-Apr-2024**
COI	**N/A**

General comments

==========

I appreciate the opportunity to review your article titled “ChatGPT (GPT-4) versus Doctors on Complex Cases in Family Medicine: An Observational Comparative Study.” The subject matter is undoubtedly compelling; however, the manuscript needs considerable elaboration to enhance its academic value and readability.

Specific comments

==========

Major comments

[Whole Manuscript]

- Clarity of Scope: The connection between the Swedish family medicine specialist exam and the "complex cases" mentioned in the title needs clarification. It's essential to explicitly state why these cases are considered complex within this specific context.

- Terminology Consistency: The manuscript alternates between "ChatGPT-4" and "GPT-4" without clear distinctions. A precise definition of these terms at the beginning of the document will prevent reader confusion. Since the GPT-4 API was used in your methods, consistency with the term "GPT-4" is advisable.

[Title]

- Study Type Clarification: Indicate that the study is preliminary concerning the Swedish family medicine specialist exam. This adjustment will help manage the expectations of the readers about the research's scope and outcomes.

[Abstract]

- Definition of Terms: The term “responses” used in the abstract is vague. Please specify whether these are responses from the AI or the doctors involved in the study.

[Introduction]

- AI Development Background: Providing a detailed historical background of AI development, particularly in generative and multimodal AI, will help situate your study within the larger field of AI research.

- AI in Family Medicine: It would be beneficial to include a discussion on the role and potential implications of AI and generative AI in family medicine. Highlighting previous studies or existing theoretical frameworks would offer a solid foundation for your research.

- Research Questions: Conclude the introduction with a clear articulation of the research question(s) your study aims to address.

[Methods]

- Study Overview: Start the Methods section with a summary of the study’s design and objectives.

- Case Selection Details: Explain how the 48 cases were selected from the Swedish family medicine specialist exam. If these cases span from 2017-2022, provide details on their relevance and focus—whether diagnostic or management-oriented.

- Exam Overview for an International Audience: Provide an overview of the Swedish family medicine specialist exam for readers unfamiliar with it, including its target audience and any restrictions like word limits.

- Collection of Doctors' Responses: Clarify the process for collecting responses from doctors, including recruitment strategies and participant selection criteria.

- Evaluator Selection: Explain the rationale and process for selecting the three medical doctors who evaluated the responses.

[Results]

- Representative Responses: Include examples of responses from a randomly chosen doctor, a top-tier doctor, and GPT-4 to illustrate the findings.

[Discussion]

- Immediate Discussion of Results: Begin the Discussion section by directly addressing the study results.

- Performance Analysis: Discuss potential reasons for GPT-4’s underperformance compared to the doctor groups.

- Limitations: Address significant limitations such as the lack of medical use approval and the absence of specific medical tuning or reinforcement for GPT-4.

==========

Minor comments

[Font Consistency]

- Typography: Ensure uniform font type and size across the manuscript, particularly on lines 27-37 of page 8, to maintain professional presentation standards.

VERSION 1 - AUTHOR RESPONSE

Reviewer 1:

When questioning ChatGPT, was a case given with follow-up questions, or was it scored based on what was written at once? Providing a sample question model within the publication would make it more understandable.

- We have clarified this under "Obtaining GPT-4 responses, group C." We have also included the corresponding GPT-4 responses of the three cases in Supplemental file 1. If the Editor prefers, we could add one example in the main manuscript, but we have not done so now, in order to keep the manuscript concise.

Were the evaluators of ChatGPT and human participants' responses specialists in family medicine? Is there a status difference between the human participants taking the exam? It has been stated that correct answers are not available online. Were correct answers determined based on specific textbooks?

- We have clarified who the reviewers were and elaborated a bit more on the creation of the scoring guide under "Scoring the responses." Additionally, we have clarified who the human participants were in the last paragraph of "Background."

How was the selection made for top-tier doctors? It should be more explainable. Are the evaluators also other top-tier doctors?

- This has now been clarified further under "Sourcing of doctor responses, group A and B."

It has been mentioned that no other comparison has been made regarding examination results between human doctors. However, in the study provided below, a comparison was made based on the scores and rankings of human participants who took the examination. It is suggested that this be discussed in the debate.
https://pubmed.ncbi.nlm.nih.gov/37565917/

- The provided study compares GPT-3.5 with human doctors on a Turkish exam required before doctors begin their residency/specialisation. This exam is in multiple-choice format. GPT-3.5 ranked 1,787th out of 22,214 individuals in its best performance and 4,428th out of 21,476 individuals in its worst performance. Since it is a bit more similar to our study, we have now referenced this study instead of a previously referenced study about GPT performance on a test about cirrhosis and hepatocellular carcinoma in both the background and discussion section. Upon reviewing our own manuscript, we do not find anywhere that we state that "no other comparisons have been made regarding examination results between human doctors and ChatGPT.". In the interest of keeping our manuscript concise, we have not added a new paragraph discussing this study in detail, but it is used as an example of similar studies on tests with multiple choice questions.

What could be the future implications of ChatGPT's clinical usage? Only how its success in exams can be increased has been discussed.

- We have now addressed this under "Implications for current practice and future research."

Reviewer 2:

Clarity of Scope: The connection between the Swedish family medicine specialist exam and the "complex cases" mentioned in the title needs clarification. It's essential to explicitly state why these cases are considered complex within this specific context.

- We have added a better description of the Swedish family medicine specialist exam at the end of the Background section, including the complexity of the cases and the fact that they require comprehensive long-form responses.

Terminology Consistency: The manuscript alternates between "ChatGPT-4" and "GPT-4" without clear distinctions. A precise definition of these terms at the beginning of the document will prevent reader confusion. Since the GPT-4 API was used in your methods, consistency with the term "GPT-4" is advisable.

- We have revised this to consistently use GPT-3.5 and GPT-4.

Study Type Clarification: Indicate that the study is preliminary concerning the Swedish family medicine specialist exam. This adjustment will help manage the expectations of the readers about the research's scope and outcomes.

- We have changed the title to make this clear.

Definition of Terms: The term “responses” used in the abstract is vague. Please specify whether these are responses from the AI or the doctors involved in the study.

- We have improved the wording of the abstract and hope that it is more clear now.

AI Development Background: Providing a detailed historical background of AI development, particularly in generative and multimodal AI, will help situate your study within the larger field of AI research.

- Our first draft of the background section was considerably longer and included more detailed information on the history of AI development, particularly in medicine and generative AI. We decided to condense this section to keep the manuscript concise and focused. We believe the current background sufficiently contextualizes our study while maintaining a good pacing. Therefore, we would like to seek the Editor's opinion before making any further adjustments.

AI in Family Medicine: It would be beneficial to include a discussion on the role and potential implications of AI and generative AI in family medicine. Highlighting previous studies or existing theoretical frameworks would offer a solid foundation for your research.

- Rather than adding more about this in the Introduction/Background, we have now addressed this under "Implications for current practice and future research." Additionally, comparison with previous studies is provided under "Comparison with the existing literature."

Research Questions: Conclude the introduction with a clear articulation of the research question(s) your study aims to address.

- We have added this to the Background section.

Study Overview: Start the Methods section with a summary of the study’s design and objectives.

- Upon review, we believe that the current structure of the Methods section effectively provides a clear overview of the study’s design and objectives right from the start. We are open to adding a summary before the individual subheadings "Study Design" and "Objective and Outcome Measures." However, to avoid redundant repetition, we would like to seek the Editor's opinion before making any further adjustments.

Case Selection Details: Explain how the 48 cases were selected from the Swedish family medicine specialist exam. If these cases span from 2017-2022, provide details on their relevance and focus — whether diagnostic or management-oriented.

- We have made this a lot clearer and included a table describing what topics the cases cover.

Exam Overview for an International Audience: Provide an overview of the Swedish family medicine specialist exam for readers unfamiliar with it, including its target audience and any restrictions like word limits.

- We have provided more information about the Swedish family medicine specialist exam in the background section.

Collection of Doctors' Responses: Clarify the process for collecting responses from doctors, including recruitment strategies and participant selection criteria.

- We have clarified this under "Sourcing of doctor responses, group A and B."

Evaluator Selection: Explain the rationale and process for selecting the three medical doctors who evaluated the responses.

- Two of the evaluators were part of the research group and one was a colleague who volunteered. There was no more sophisticated method for the selection of the evaluators. We have now briefly mentioned this under "Scoring the responses."

Representative Responses: Include examples of responses from a randomly chosen doctor, a top-tier doctor, and GPT-4 to illustrate the findings.

- We have included the corresponding GPT-4 responses to the cases in Supplemental file 1. We do not have the right to redistribute the answers written by the doctors, but the top-tier responses are publicly available, which is mentioned under "Availability of data and materials." As previously mentioned, we are open to including one example with a GPT-4 response in the main manuscript as well, if the editor finds that preferable.

Immediate Discussion of Results: Begin the Discussion section by directly addressing the study results.

- We now address the main results more clearly at the beginning of the discussion.

Performance Analysis: Discuss potential reasons for GPT-4's underperformance compared to the doctor groups.

- We have added a part about ChatGPT not being specifically trained for medical use, which may explain its underperformance.

Limitations: Address significant limitations such as the lack of medical use approval and the absence of specific medical tuning or reinforcement for GPT-4.

- We have added a note about this - the same as mentioned in the previous response.

Thanks again for your valuable feedback!

Best regards,

Rasmus Arvidsson & colleagues

VERSION 2 - REVIEW

Reviewer	**1**
Name	**OZTERMELI, Ahmet**
Affiliation	**Gebze Fatih State Hospital, Ortopedics and Traumatology**
Date	**03-Sep-2024**
COI	**No competing interest**

I want to congratulate the authors for doing such an interesting study

---

Reviewer	**2**
Name	**Hirosawa, Takanobu**
Affiliation	**Dokkyo Medical University**
Date	**27-Aug-2024**
COI	**NA**

Thank you for inviting me to review this manuscript. My expertise in clinical research focusing on digital health solutions positions me to assess the general clinical aspects of the study. However, this manuscript would greatly benefit from an expert review by a specialist

in artificial intelligence, given its central role in the study's methodology and analysis. The manuscript appears to be in a preliminary stage, particularly in its integration and interpretation of AI within the specific context of the Swedish family medicine specialist exam. To elevate the manuscript to a publishable standard, a significant expansion in its scope and depth of analysis is required. This should include a more detailed examination of AI performance nuances and a broader discussion of the implications within the healthcare field.

VERSION 2 - AUTHOR RESPONSE

REVIEWER 1 COMMENT: "I want to congratulate the authors for doing such an interesting study"

AUTHOR RESPONSE: Thank you, we appreciate this remark!

REVIEWER 2 COMMENT: "Thank you for inviting me to review this manuscript. My expertise in clinical research focusing on digital health solutions positions me to assess the general clinical aspects of the study. However, this manuscript would greatly benefit from an expert review by a specialist in artificial intelligence, given its central role in the study's methodology and analysis."

AUTHOR RESPONSE: Whether another review from an expert in AI is needed or not, we leave for the editorial team to decide. We would like to note that our research team does include members and collaborators with expertise in artificial intelligence. While they are not listed as authors on this manuscript, they were actively consulted regarding the study's methodology, ensuring a solid technical foundation for the AI components of the work.

REVIEWER 2 COMMENT: "The manuscript appears to be in a preliminary stage, particularly in its integration and interpretation of AI within the specific context of the Swedish family medicine specialist exam. To elevate the manuscript to a publishable standard, a significant expansion in its scope and depth of analysis is required. This should include a more detailed examination of AI performance nuances and a broader discussion of the implications within the healthcare field."

AUTHOR RESPONSE: Although we believe that the original methodology was appropriate for answering our research question, we have taken the reviewer's feedback into consideration and have performed a more detailed examination of the AI performance nuances within the specific context of the Swedish family medicine specialist exam. Additionally, we have broadened the discussion to reflect the implications of these findings in the healthcare field. In this revision, we aimed to strike a balance between expanding the depth of our analysis and maintaining a focused, concise manuscript. We believe this approach enhances the clarity and impact of the work, while avoiding unnecessary elaboration or digression. Furthermore, we have performed a repeat of the original experiment including the newer GPT-4o version of ChatGPT. This provides a sense of the rapid development of general purpose chatbots. We hope that this version of the manuscript meets the publication standards and that the revisions satisfactorily address the reviewer's concerns. Should further feedback be provided, we kindly request that it include specific and actionable points, as the previous round of feedback, while appreciated, was somewhat open-ended and subject to interpretation. Nevertheless, we believe that the revisions have strengthened the manuscript overall.