Peer Review File
Article information: https://dx.doi.org/10.21037/jmai-23-180

 # Reviewer A

COMMENT 1: This is an important study assessing the utility of LLMs in nuclear medicine.
The introduction is incomplete without citing few examples of how LLMs are used in different field of medicine (lines 92-94). Please consider adding the usefulness of LLM in medicine (https://doi.org/10.1038/s41591-023-02448-8), ophthalmology & optometry (DOI: 10.1111/opo.13207), etc.
REPLY 1: Thank you for providing these useful information and valuable references. We did the additional literature review and provided additional references in the revised manuscript.
CHANGES IN THE TEXT 1: Additional references have been added.

COMMENT 2: Can the authors please classify the 20 questions into any categories of nuclear medicine?
REPLY 2: Of course. This suggestion is very useful to ensure that the question set encompasses all the relevant nuclear medicine topics.
CHANGES IN THE TEXT 2: We have inserted a new column classifying the topics of nuclear medicine into different topics.

COMMENT 3: Please discuss the incorrect responses of LLMs in more detail such that even non-experts can understand how the responses were erroneous, and how they are contradictory.
REPLY 3: We felt deeply appreciated by this considerate comment because it can help people across different disciplines who are readers of the journal JMAI to better understand the context.
CHANGES IN THE TEXT 3: The erroneous text generated from the LLMs have been discussed further and the context has been provided to help it understandable readers in different fields.

COMMENT 4 Conclusion needs to be clearer.
REPLY 4: Thank you for your suggestion. We hope the revised conclusion has become clearer.
CHANGES IN THE TEXT 4: The conclusion has been revised.

 # Reviewer B

COMMENT 5: I present my review in the form of subsections. Please refer to all of them and, if necessary, follow the recommendations/advice contained therein: do you not think that doing statistics or making conclusions on a group of about 20 questions is not a good approach? Why were you limited with the number of questions?

REPLY 5: We limited to the small number of questions due to this report being a pilot test, but further investigation with larger number of questions will surely be beneficial
CHANGES IN THE TEXT 5: We have added the limitation due to the small arbitrary number of questions in the discussion.

COMMENT 6: are you considering extending the study to other scientific disciplines and/or increasing the number of questions in this study?
REPLY 6: Yes, we will be pleased to extend the scope and the number of questions in future study.
CHANGES IN THE TEXT 6: N/A

COMMENT 7: have you tried asking questions in a different way by paraphrasing them?
REPLY 7: Unfortunately, we did not paraphrase the question.
CHANGES IN THE TEXT 7: We have inserted the statement that we did not try paraphrasing in the methods section.

COMMENT 8: 4. did you ask the questions of the Chats several times?
REPLY 8: Unfortunately, we did not.
CHANGES IN THE TEXT 8: We have inserted the statement that we did not try repeat asking the questions in the methods section.

COMMENT 9: I think the moteodlogical approach to this study is mediocre. Creating a survey with about 20 questions is a serious mistake. You should consider a larger number of questions.
REPLY 9: Thank you for your comments. We regret that this brief report is only a pilot study with limited number of questions. We will try to conduct a new experiment with larger number of questions in the future.
CHANGES IN THE TEXT 9: We have added the limitation due to the small arbitrary number of questions in the discussion.

COMMENT 10: Please refer to other studies using ChatGPT in Nuclear Medicine and Radiology and compare your study and results to them:
https://pubmed.ncbi.nlm.nih.gov/37808173/
https://irjnm.tums.ac.ir/article_40129.html
REPLY 10: Thank you for suggesting useful related articles. Those articles have been referred to and discussed in the revised manuscript.
CHANGES IN THE TEXT 10: Those articles have been referred to and discussed in the revised manuscript.

 # Reviewer C

COMMENTS 11: General remarks: The authors have wisely chosen multiple-choice

questions to be able to objectively compare the correctness of the answers. That's a plus point of the analysis. Moreover, it is positive to note that 3 models were compared with each other.
REPLY 11: Thank you for your positive comment.
CHANGES IN THE TEXT 11: N/A

COMMENTS 12: The main limitation, which the authors briefly acknowledge, is the arbitrary choice of topics and small number of questions. Ideally, the authors would have chosen a standardized set of questions, e.g., from a nuclear medicine exam.
REPLY 12: We totally agree with the idea of using the standardized set of question. Unfortunately, we are not allowed to use the summative exam item for the experiment. This set of question was; however, used as a formative quiz for medical students.
CHANGES IN THE TEXT 12: N/A

COMMENTS 12.5: The authors should include more previous publications on LLMs in the context of medicine and medical imaging in the introduction and the discussion and evaluate their results in the light of these other publications (see some suggestions in the comments below).
REPLY 12.5: Thank you for suggesting useful related articles. Those articles have been referred to and discussed in the revised manuscript.
CHANGES IN THE TEXT 12.5: Those articles have been referred to and discussed in the revised manuscript.

COMMENT 14: Abstract Line 52: “However, inconsistencies in responses to five questions, which varied in complexity, highlighted the inherent unpredictability of LLM-generated answers.” The authors may consider using another word than “inconsistency” or better explain the meaning of it. As I understand, it relates to the fact that in some questions, the answers that the LLMs chose differed between the LLMs? If this is what the authors mean, they may use a statement such as “responses differed between the LLMs”. Inconsistency could also relate to a variation in responses from the same LLM to the same question (if the same question was submitted to the LLMs several times). One example of such an analysis of consistency of the same LLM for different trials of the same prompt is pubmed ID 37709536. Such an approach would have been better suited to assess the “consistency” of the LLMs.
REPLY 14: We apologize for using the word “consistency”, which cause a lot of confusion. The consistency we had referred to was the consistency, that is, the same response from the three LLMs. We have revised the manuscript to avoid the ambiguity and mostly used the word accuracy for better clarity.
CHANGES IN THE TEXT 14: We have revised the manuscript to avoid the ambiguity and mostly used the word accuracy for better clarity.

COMMENT 15: Line 51: “[…] with a consistent correctness of 75.0% across all

models." This is the same issue as above: What does "consistent accuracy" mean? Does it mean that 75% of questions were answered correctly by all three LLMs? If yes, the authors may prefer such a description.
REPLY 15: Thank you for the suggestion than help our manuscript avoiding the ambiguity.
CHANGES IN THE TEXT 15: We have revised the text to avoid the ambiguity.

COMMENT 16. In conclusion, this study finds that although LLMs demonstrate a high correct response rate, their reliability in delivering accurate information is limited and unpredictable. Therefore, it advocates for careful application of LLMs in medical contexts.". I agree that it is worth mentioning that the LLMs did not chose the correct answer in all 20 questions. However, the authors should consider a conclusion that is a bit weaker / less strict. I also recommend to modify the two sentences to underline that this analysis only addressed a very specific topic: nuclear medicine (not medicine in general). Notably, the current LLMs have proven that they are generally able to solve medical exams:
https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2806980
https://healthitanalytics.com/news/chatgpt-passes-us-medical-licensing-exam-without-clinician-input
REPLY 16: Thank you for the useful comments.
CHANGES IN THE TEXT 16: We have revised the conclusion to avoid the over-generalization.

COMMENT 17:
- Added: Introduction:6. Lines 92-94: Could you add references to papers that have assessed the accuracy of LLMs in medical context, e.g.:
https://journals.sagepub.com/doi/full/10.1177/08465371231193716 ;
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9947764/ ;
https://pubmed.ncbi.nlm.nih.gov/37709536/ ;
https://pubmed.ncbi.nlm.nih.gov/37014239/
- Added Methods: 7. Curation of test questions: Please provide details about the expertise of the two investigators that developed the questions.
REPLY 17: Thank you for the suggestion. We are pleased to inform you that add the mentioned issues have been added.
CHANGES IN THE TEXT 17: The mentioned issues have been added.

COMMENT 18: 8. Statistical Analysis: Fisher's exact test was used to compare proportions. However, as I understand, these were paired proportions (because the same questions were rated by the different LLMs), which means that McNemar's test should be used instead.
REPLY 18: Thank you for pointing out this critical mistake. We corrected the statistics in the revised manuscript.
CHANGES IN THE TEXT 18: We corrected the statistics in the revised manuscript.

COMMENT 19: Results: 9. Line 138: “The three LLMs provided consistent answers in 15 out of the 20 questions (75.0%) (Table 1).” Please consider a different term instead of “consistent”, such as “In 15 out of 20 questions, all three LLMs chose the correct answer. (see comment above).
REPLY 19: We apologize for using the word consistent ambiguously.
CHANGES IN THE TEXT 19: We have revised the manuscript to avoid the ambiguity and mostly used the word accuracy for better clarity.

COMMENT 20: 10. Lines 140-155 only repeat the content of Table 1 and is difficult to read. This paragraph should be completely removed or shortened considerably.
REPLY 20: Thank you for the suggestion that help improving the conciseness of our manuscript.
CHANGES IN THE TEXT 20: The section has been edited to reduce redundancy.

COMMENT 21: Edited Discussion: 11. Line 226: “As a result, we encourage further investigation into the accuracy of LLMs on open questions related to nuclear medicine.” Please consider discussing pubmed ID 37709536.
REPLY 21: Thank you for suggesting useful related article. The article have been referred to and discussed in the revised manuscript.
CHANGES IN THE TEXT 21: The article have been referred to and discussed in the revised manuscript

COMMENT 22: Table 1: 12. Question 18: As I understand, Tc-99m-DTPA can be used for brain death examination, but it is administered intravenously (brain perfusion) not intrathecal (CSF). In my understanding, only B and C would be correct. Please double-check and, if the authors think that A is also correct, please provide adequate references that support this answer. If A is incorrect, this question should either be modified and submitted to the three LLMs again or completely removed or replaced by a new question.
REPLY 22: Tc-99m DTPA can be used in the indication of brain death, albeit not as accurate as Tc-99m HMPAO (perfusion scan), as in the article: https://doi.org/10.1053/j.semnuclmed.2011.07.007
CHANGES IN THE TEXT 22: N/A

 # Reviewer D

COMMENT 23: This article delves into an interesting subject: the role of LLMs in the medical field. The paper addresses a relatively unexplored domain, LLMs in nuclear medicine. However, many significant methodological aspects of this study merit consideration and clarification.
REPLY 23: Thank you for giving us opportunities.
CHANGES IN THE TEXT 23: N/A

COMMENT 24: The prompting methods used to address LLMs is not clarified.

Moreover, it is not clear if the questions were posed to LLMs only a single time or more than one time. In this latter case did the authors evaluated different LLMs temperature settings and token outputs, that could influence LLMs responses?
REPLY 24: We apologize for not stating clearly on how we prompted the question.
CHANGES IN THE TEXT 24: We revised the text so that it unambiguously state that the questions were prompt to each LLM only once.

COMMENT 25: 2. In the abstract, the authors claim that no previous studies have tested the reliability of Large Language Models (LLMs) in the field of nuclear medicine "none have tested their reliability in the field of nuclear medicine". However, in the discussion, they refer to a study by the Japan Radiology Society. This study included 15 questions related to nuclear medicine. Moreover, even if not many, there are some published papers regarding this topic. Hence, the statement, asserting that "none have tested their reliability in the field of nuclear medicine "appears inaccurate. Lastly the reported percentages of the cited paper in the article are the one pertaining all radiology answers, but in the cited paper the percentages specifically for nuclear medicine were notably different and should be reported.
REPLY 25: We apologize for missing such an important article out. The article and additional information have currently been discussed in the revised manuscript.
CHANGES IN THE TEXT 25: The article and additional information have currently been discussed in the revised manuscript.

COMMENT 26: The statement "The findings showed GPT-3.5, GPT-4, and Bard achieved accuracy rates of 85.0%, 95.0%, and 90.0%, respectively, with a consistent correctness of 75.0% across all models" lacks clarity in defining "consistent correctness." The authors should clarify whether this term denotes self-consistency, indicating that no contradictory statements were given by the same LLM when tested multiple times, or if it represents the percentage of cases where all three LLMs provided the correct and identical response. If the latter, the reported percentage should be 70%, not 75%?
REPLY 26: We apologize for using the word consistent ambiguously.
CHANGES IN THE TEXT 26: We have revised the manuscript to avoid the ambiguity and mostly used the word accuracy for better clarity.

COMMENT 27: Again, when they state in the conclusion "Although they exhibited impressive accuracy, there were still inconsistencies and self-contradictions in their responses". This part should be clarified in the methods and results; how did they evaluate and calculate "self-contradictions"?
REPLY 27: We use the word "self-contradiction" to refer to the occurrence when the LLM chose one choice but state in its explanation that another choice is correct. The example is the answer of Bard to the question regarding neonatal jaundice, which is exemplified in the revised manuscript.
CHANGES IN THE TEXT 27: The example of self-contradictory statement has been provided in the revised manuscript.

COMMENT 28 5. From line 206 to line 213 the authors hypothesized that "inaccurate responses emerged in this study could be attributed to misinterpretation of available information by the LLMs" and later "regarding the cause of false negative F-18 FDG PET/CT, the LLM could be overwhelmed by the sheer number of texts that referred to the low FDG uptake in primary lesion of lung cancer". This latter hypothesis should at least be supported by some bibliography and/or results. Indeed, not knowing on which dataset LLMs was trained on you cannot know the "number of texts that referred to the low FDG uptake in primary lesion of lung cancer" considered by LLMs.
REPLY 28: Thank you for pointing this out. We revised the manuscript to avoid the over-speculation.
CHANGES IN THE TEXT 28: We revised the manuscript to avoid the over-speculation.

COMMENT 29: Minors:
6. They didn't mention what software was employed to perform the statistical analysis
7. No reported confidence intervals
8. Some typos within the text
REPLY 29: Thank you for your suggestion**.**
CHANGES IN THE TEXT 29: The information on the software used in the statistical tests have been added, as well as the typo correction.

COMMENT 30: In conclusion even if there are few papers regarding LLMs and nuclear medicine, the stories seemed very similar to an academical medical Q&A and many other papers have already assessed the good performances of LLMs in this setting (see Med-Palm, Llama and also GPT). This article doesn't add any significant novelty. I suggest the authors to evaluate more complex clinical scenarios and/or to train LLMs with specific nuclear medicine datasets.
REPLY 30: Thank you for valuable insights.
CHANGES IN THE TEXT 30: N/A

 # Reviewer E

COMMENT 31
- The manuscript presents a study assessing the accuracy of three Large Language Models (GPT-3.5, GPT-4, and Google Bard) in answering nuclear medicine questions. Utilizing a 4-choice quiz format with twenty questions, the study finds accuracy rates of 85.0% for GPT-3.5, 95.0% for GPT-4, and 90.0% for Bard. Despite high correct response rates, the study highlights the unpredictability and limitations of LLMs, particularly for complex queries.
- Major points: The method of testing each LLM with twenty nuclear medicine questions is clear. However, the manuscript would benefit from a more detailed description of how these questions were selected and if they represent a comprehensive range of topics within nuclear medicine.

REPLY 31: The revised table is now showed the content areas to confirm that the ranges of topic have been coved.
CHANGES TO THE TEXT 31: The table has been revised.

COMMENT 32:
- The use of a 4-choice single best answer quiz format is an interesting approach, but the paper should discuss how the 'expected responses' were determined. Is there a possibility of bias in these expected answers, and how was this mitigated? The gold standard construction is not clear.
- While the accuracy rates of the LLMs are presented, the manuscript lacks a detailed statistical analysis. Information on the statistical methods used to assess the significance of the results would strengthen the conclusions drawn.
REPLY 32: Thank you for the useful comments. The expected answer for each question is the correct answer as in the single best answer question format. We wished we could have additional statistical analysis but unfortunately the number of question limit by the fact that this is a pilot experiment does not allow that.
CHANGES TO THE TEXT 32: We revised the manuscript for better clarity.

COMMENT 33 The paper notes inconsistencies in responses to five complex questions. A deeper analysis of these inconsistencies would be valuable. For instance, what specific characteristics of these questions led to varied responses, and what does this imply about the limitations of LLMs in handling complex medical queries?
REPLY 33: Unfortunately, we cannot identify the pattern of questions that resulted in incorrect response. The pattern did not seem correlate with the complexity of the question as the questions answered incorrectly span across different levels of Bloom's modified taxonomy levels.
CHANGES TO THE TEXT 33: We revised the manuscript for better clarity.

COMMENT 34: The study finds no significant difference in accuracy among the three LLMs. The manuscript would benefit from a discussion on why this might be the case, considering the different architectures and training data of these models.
REPLY 34: Thank you for this useful suggestion. We would be more confident to discuss the lack of accuracy after testing with larger set of questions. Because this is a pilot study and the sample size was not calculated to accommodate the degree of difference. We refrained ourselves form concluding that the correct response rate is truly non-significant.
CHANGES TO THE TEXT 34: We suggest further experiments with larger set of questions in future studies.

COMMENT 35 5. The conclusion rightly emphasizes the need for careful application of LLMs in medical contexts. Expanding on the potential risks and ethical considerations of using LLMs in nuclear medicine, along with possible safeguards, would make the paper more comprehensive.
REPLY 35: We totally agree that expanding future studies into potential risks and

safeguards.
CHANGES TO THE TEXT 35: We suggested future studies in the discussion section.

COMMENT 36: Minor points:
1. Ensure that the paper includes a thorough literature review, particularly focusing on previous studies that have assessed the reliability of LLMs in other areas of medicine.
2. The manuscript is well-written but would benefit from a thorough proofreading to correct minor grammatical errors and enhance clarity.
3. Future Research Directions: Finally, suggesting areas for future research, such as testing LLMs with more diverse and complex medical datasets or exploring their application in clinical decision support, would be valuable.
REPLY 36: Thank you for these useful comments. These issues have been addressed in the revised manuscript.
CHANGES TO THE TEXT 36: The necessary changes have been addressed in the revised manuscript.