Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-47

Reviewer A

Comment 1: Abstract line 44 should include that images were specifically omitted from the prompts

Reply 1: Added words to Lines 44-45

Changes in text: "...and the presence of an image omitted from the prompt were recorded…"

Comment 2: Lines 111-114, the "hallucinations" are due to more than the technique of supervised learning that is used to train ChatGPT and includes the training data that was used; providing the term "hallucination" can help the readers.

Reply 2: Added hallucinations in line 113 instead of 'content'

Changes in text: "This has led to the creation of factually incorrect 'hallucinations' that sound believable enough to stump experts"

Comment 3: Introduction or conclusion may comment on the different types of material that is commonly found on Step 1 vs 2 vs 3, with step 1 often containing more "minutiae" and step 2 and 3 being more clinically applicable information.

Reply 3: Added information at line 124.

Changes in text: Added "In general, the content in Step 1 consists of basic science content, while Steps 2 and 3 aim to test participants on the clinical applications of knowledge."

Comment 4: What version of ChatGPT was used? ChatGPT3.5? version chatgpt4?

Reply 4: v 3.5

Changes in text: Added "(v3.5) to line 137

Question input and recording: please provide more information (lines 154-159).

Comment 5:- was ChatGPT told whether it was right or wrong after answering each question?

Reply 5: No it wasn't. Added sentance at line 157.

Changes in text: Added "ChatGPT received no instruction on whether or not its response was correct"

Comment 6: - was the same ChatGPT instance used for all questions and answering? (i.e, if the same instance was used, since the LLM has recall memory of the entire conversation and if ChatGPT was corrected, this can affect performance.

Reply 6: No it wasn't. Added sentence at line 157

Changes in text: "Questions were input as new instances each time as to not bias answers from previous entries."

Comment 7: 168-172: statistical analysis. 2 tailed paired can be applied for the comparison of questions without images vs questions with images since the test subject (chatGPt algorithm) is the same; however comparing to the passing score of 60% should use the one-sample t-test.

Reply 7: Ok, I changed methods and then went through and changed the

information in Table 1.
Changes in text: Added "A one sample paired t-test analyzed the performance on question sets compared to the estimated" at line 170. Updated results section (lines 182-183) and table 1.

Comment 8: line 214, where does the number 47% correct come from? should this be 46% from the step 3 subgroup?
Reply 8: The 47% is from the performance on Step 3 questions which was the highest. It isn't a key point the actual percentage so I can take it out for simplicity.
Changes in text: Removed "(at most 47% correct)" from line 214

Comment 9: line 227: please clarify what is meant by the "high-probability selection for the patient demographic". Does this mean ChatGPT based its answer selection based off solely patient's age, ethnicity?
Reply 9: We don't know for certainty that that's how ChatGPT was picking an answer but yes, it would select answers that were common for the demographic but would disregard key findings that would indicate a different answer (eg rash in pediatric patient might have been guessed that it was Hand, foot, and mouth but missed the patient immigrated from a measles endemic country).
Changes in text: None

Comment 10: line 239-240: while ChatGPT's performance is analyzed on the different difficulty questions, please describe the characteristics of questions that ChatGPT frequently answered incorrectly along this treatment algorithm. Were they often "next step in diagnosis" or "next step in treatment" questions? Where along the treatment or diagnostic algorithm did ChatGPT fail? Was there a trend?
Reply 10: It was more in diagnosis (imaging tests/special lab work) next-step style questions. It performed better in identification-style questions. We didn't notice a particular pattern, only that the explanations generally reflected understanding of the clinical picture but failed to order the correct tests to confirm the diagnosis. We didn't notice any one particular pattern of where it missed these questions. It wasn't something we expected to encounter, but more something we realized as we went through it.
Changes in text: None

Comment 11: Limitations have no mentions of hallucinations.
Reply 11: Added a sentence to the limitations, line 266.
Changes in text: ChatGPT's ability to create hallucinations that trick experts should be considered, as well, when analyzing the veracity of its explanations

Comment 12: References: please adhere to the same citation format for all articles (reference 4 vs reference 5, 6, 12 for example)
Reply 12: Went through and made the adjustments.
Changes in text: Changes made to references 5, 6, 12, 14, 15, 20.

Reviewer B
Comment 1: Define AMBOSS with first use.
Reply 1: In the methods section I put a brief description for AMBOSS which is the first introduction to it outside of the abstract.

Changes in Text: None, line 141-143.

Comment 2: Restructure this last sentence "Including deep learning systems that analyze images in conjunction with ChatGPT may improve accuracy and provide a more robust educational tool in dermatology."
Reply 2: I made the change on line 59.
Changes in Text: “Using ChatGPT in conjunction with deep learning systems that include image analysis may improve accuracy and provide a more robust educational tool in dermatology.”

Reviewer C
Thank you for proposing this interesting work on the use of artificial intelligence. In particular this manuscript investigates the ability of GPT chat to correctly answer to standardized questions, with various features, and levels of difficulty, concerning dermatology.
I only suggest few form corrections in the used language:

Key Findings
Comment 1: “Meaning: As artificial intelligence continues to be incorporated into medical education, it is important to understand that it may not apply to all medical disciplines” please, rephrase with verbs.
Reply 1: I am not too sure what you mean but I changed it from passive voice to active voice.
Changes in Text: “As medical education incorporates artificial intelligence, we must consider that it may not similarly apply to all medical disciplines.”

Comment 2: “RESULTS: Overall percent correct was 41% (Step 1 = 41%, Step 2CK = 34%, and Step 3 = 46%).” please, rephrase with the indirect object.
Reply 2 : Changed line 48.
Changes in Text: “ChatGPT answered 41% of all the questions correctly”

Comment 3: “The percent correct was calculated by dividing the number of correct responses by the total number of questions for that level” please, clarify as “the percentage of correct answers…”
Reply 3 : Done.
Changes in Text: “The percentage of correct answers was calculated by dividing the number of correct responses…”

Reviewer D
This is a study that examines how Chat-GPT fares on the dermatology portion of the USMLE via AMBOSS questions. Overall it is well written. I have a few minor revisions:

1) Please state the version of chat GPT was used. It seems It was ChatGPT-3 based on when the study was performed.
Reply: Version 3.5, added

2) Please define/spell out AMBOSS when first introduced in the methods.
Reply: AMBOSS isn’t an acronym but it is the name of a company, but I changed it

to be more clear. Line 141
Change in text: "AMBOSS is an online education company that provides resources such as content and questions for board exams and continuing medical education curriculums accredited through the Accreditation Council for Continuing Medical Education"

3) 154 questions contained images but were included in the analysis. I would exclude these as I would think the images are needed to answer the question
Reply: I think that was part of our analysis and an exciting part of the study was seeing how ChatGPT scored equally on both sets of questions. Also, some of the questions used the images more as a supplementary resource and could be answered without it, but it is hard to know for sure how much a picture would aid in answering correctly so we used all the questions.

4) The authors should discuss other medical studies in which ChatGPT failed the exam as well in the discussion to show that it is not just dermatology that ChatGPT is failing. This highlights the need for ChatGPT to improve in the medical field. In particular a recent Gastroenterology study recently received a lot of media attention (PMID 37212584).
Reply: Yes that was an interesting article because it compared v3.5 to v4.
Change in text: (Line 271) "Incorporating ChatGPT and other LLMs into medical education might not be similar for all medical specialties. We observed in this study that ChatGPT answered correctly fewer dermatology questions as compared to the general USMLE licensing exams observed by prior publications. Another recent study in gastroenterology produced similar findings, highlighting the need for further improvement of ChatGPT in medical fields.27"

5) The discussion should also state how chatboxes like ChatGPT and Google's Bard can be trained for medicine. Perhaps use of medical databases for medical information, or medical journals.
Reply: We hinted at the end of one paragraph about training that maybe medical professionals could be involved more in the LLM learning process. There is an LLM for Pubmed that performed worse than ChatGPT, so I don't know if that would help or not. (https://crfm.stanford.edu/2022/12/15/pubmedgpt.html) I think it is a very interesting question but I am not sure what the best idea would be, and it would be more speculation.
Change in text: None