Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-94

Reviewer A

The authors provide an overview of ChatGPT / GPT-4 with an emphasis on medical image analysis as a context. The piece is interesting but unduly brief and imprecise in its technical descriptions. It requires significant revisions to be of a sufficient standard to publish.

Response: We appreciate your feedback to help optimize this manuscript. Point by point changes have been included below.

Major comments:

1. The description of ChatGPT's architecture reads like an imprecise overview of deep learning more generally. I would suggest the authors steer clear of the exact technicalities or significantly rework this paragraph to improve their accuracy. GPT-3.5 and GPT-4 are transformer-based LLMs which underlie the processing of ChatGPT, which is merely the web-application most users interact with. Further detail may be found here: https://doi.org/10.1038/s41591-023-02448-8

Response: We agree with this comment and have deleted lines 67-70, “The machine learning algorithm behind ChatGPT involves using a neural network to process and analyze large amounts of text data. This neural network is made up of multiple layers of linked nodes that carry out intricate mathematical operations to find patterns and connections in the data.”

2. The description of the training process used to develop GPT-3.5 and GPT-4 is simplistic and incomplete. While pretraining does involve associative learning based on large volumes of text, significant resources are invested in subsequent fine tuning to develop an LLM capable of useful dialogue. Further detail here: https://doi.org/10.1038/s41591-023-02448-8

Response: Added lines 70-73: “Following this process, a sophisticated fine-tuning process occurs including the use of human researchers to provide prompts and responses. Reinforcement learning from human feedback (RLHP) is also conducted, where human graders rank GPT responses to a set of queries.”

Cited: Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8

3. The authors choose to discuss GPT-4 and no other LLMs. At the very least, GPT-3.5 provides useful contrast to GPT-4. Moreover, discussing other options is essential so as

not to confer an unfair advantage on OpenAI. Google's PaLM 2 (and clinically fine tuned Med-PaLM 2) also support multimodal input, and Meta's LLaMA (open source) may follow suit. For further discussion of the potential of multimodal AI in medicine, see https://doi.org/10.1038/s41591-022-01981-2

Response: Added lines 163-168: “It should also be noted that Google recently released an AI chatbot, Bard, that has image analysis capabilities. Meta is also developing its own large language model, LLaMa 2, which may soon also provide image analysis capabilities. Future work will be required to determine which LLM possesses the best image analysis capabilities. In conclusion, although significant advancements and further research is still required, the future of automated medical image analysis is highly promising.”

4. The bottom panel in Figure 5 does not show OCT of any layer of the retina; it shows derived analyses from OCT data but not the scan itself. This is a major mistake and must be corrected.

Response: Thank you for this catch. Figure 5 has now been changed to “The bottom left panel shows derived analyses from OCT data…”.

5. Future directions may draw on more of the available literature base -- an enormous amount of work on LLMs in medicine has been conducted already.

Response: We highly agree with this future direction, and actually plain on doing a literature review of LLMs in medicine!

Reviewer B

Lines 84 and 85: The ability to browse the web has been suspended since 3rd July and GPT-4 currently does not interpret images. This should be updated to say that the study was performed during the introductory phase when GPT-4 had browsing capabilities.

Response: We agree this is important to mention. Added lines 85-87: “However, the ability to browse the web has been suspended since 3rd July and GPT-4 currently does not interpret images. This study was performed during the introductory phase when GPT-4 had browsing capabilities.”

There is no mention of ethical concerns regarding using ChatGPT for such decision making. Please refer to this helpful article which may provide more insight into how Radiology can use ChatGPT and cite it: https://pubmed.ncbi.nlm.nih.gov/37425598/
Additional article that may help: https://jamanetwork.com/journals/jama-health-forum/fullarticle/2805334
Lines 86-88: need citation.
Line 154-155: Worthwhile to mention that cut off training date for ChatGPT is September 2021.

Response: Citation has been added.

Lines 152-156 have been added: “An AI model can remain only as unbiased as the data that it is trained on, so it is essential that diverse datasets are used to minimize the potential of AI bias. Additionally, the cut off training date for ChatGPT is September 2021. As medical knowledge and clinical guidelines are constantly evolving, this can to the generation of outdated recommendations.”

Added lines 161-163: “Due to the black box nature of AI models, interpretation of GPT-4 outputs must be made with caution to ensure that no errors have been made”.

Added lines 171-173: “As LLMs become more prevalent, it is also important to implements mechanisms to ensure equitability and accountability. Open-source codes could allow for routine oversight, increased accountability for data generation standards, and facilitate LLM development.”.