Multimodal ChatGPT-4V for ECG Interpretation: Promise and Limitations

Lingxuan Zhu, Weiming Mou, Keren Wu, Yancheng Lai, Anqi Lin, Tao Yang, Jian Zhang, Peng Luo

Submitted to: Journal of Medical Internet Research
on: November 16, 2023

Table of Contents

Multimodal ChatGPT-4V for ECG Interpretation: Promise and Limitations

Lingxuan Zhu[1, 2*]; Weiming Mou[3*]; Keren Wu[1]; Yancheng Lai[1]; Anqi Lin[1]; Tao Yang[4]; Jian Zhang[1]; Peng Luo[1]

[1]Department of Oncology Zhujiang Hospital, Southern Medical University Guangzhou CN
[2]Department of Etiology and Carcinogenesis National Cancer Center/ National Clinical Research Center for Cancer/Cancer Hospital, Changping laboratory, Chinese Academy of Medical Sciences and Peking Union Medical College Beijing CN
[3]Department of Urology Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine Shanghai CN
[4]Department of Medical Oncology National Cancer Center/National Clinical Research Center for Cancer /Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College Bejing CN
[*]these authors contributed equally

Corresponding Author:
Peng Luo
Department of Oncology
Zhujiang Hospital, Southern Medical University
253 Industrial Avenue
Guangzhou
CN

Abstract

Electrocardiogram (ECG) interpretation is an essential skill in cardiovascular medicine. This study evaluated the capabilities of newly released ChatGPT-4V, a large language model with visual recognition abilities, in interpreting ECG waveforms and answering related multiple-choice questions. A total of 62 ECG-related multiple-choice questions were collected from reputable medical exams. ChatGPT was prompted to answer the questions by analyzing the accompanying ECG images. Requiring at least 1 of 3 responses to be correct, ChatGPT achieved an overall accuracy of 83.87% across all question types. ChatGPT demonstrated significantly lower performance on counting-based questions like calculating QT intervals compared to diagnostic and treatment recommendation questions. The findings indicate that while ChatGPT shows promising potential in ECG interpretation and decision-making, its diagnostic reliability and quantitative analysis abilities need improvement before real clinical use. Further large-scale studies are warranted to fully evaluate ChatGPT's capabilities and track its progress as the model accumulates more medical knowledge through ongoing training. With technological advancements, multimodal AI like ChatGPT may one day play an important role in assisting clinicians with ECG interpretation and cardiovascular care.

(JMIR Preprints 16/11/2023:54607)
DOI: https://doi.org/10.2196/preprints.54607

Original Manuscript

Multimodal ChatGPT-4V for ECG Interpretation: Promise and Limitations

Lingxuan Zhu[1,2†], Weiming Mou [3†], Keren Wu [1], Yancheng Lai [1], Anqi Lin [1], Tao Yang[4], Jian Zhang [1*], Peng Luo [1*]

Article Type: Research Letter

[1] Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, 510282, China.

[2] Department of Etiology and Carcinogenesis, National Cancer Center/ National Clinical Research Center for Cancer/Cancer Hospital, Changping laboratory, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.

[3] Department of Urology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

[4] Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer /Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College

Word count: 800 words

*** Correspondences to:**

Peng Luo and Jian Zhang,

Department of Oncology, Zhujiang Hospital, Southern Medical University, 253 Industrial Avenue, Guangzhou, 510282, Guangdong

Tel: 0086-18588447321 (Peng Luo); 0086-13925091863 (Jian Zhang)

mail: luopeng@smu.edu.cn (Peng Luo); zhangjian@i.smu.edu.cn (Jian Zhang)

†: Lingxuan Zhu and Weiming Mou have contributed equally to this work and share first authorship.

Main

Electrocardiogram (ECG) interpretation is an essential skill in cardiovascular medicine. The rise of artificial intelligence (AI) has led to many attempts to automate ECG interpretations[1]. As a representative of generative AI, ChatGPT has shown promising potential in cardiovascular medicine. For example, it has been found that ChatGPT can provide advice for cardiovascular disease prevention[2] and also pass the AHA's life support exams[3]. However, since early versions of ChatGPT cannot process graphical information, previous studies have failed to assess ChatGPT's ability for ECG interpretation. The newly released ChatGPT-4V(ision) model adds visual recognition capabilities[4], which makes it possible to directly read and interpret ECG waveforms. Based on this, we tried to evaluate the performance of ChatGPT-4V in ECG waveform analysis.

We gathered a set of multiple-choice questions related to ECG waveform interpretation from various question banks, including the American Heart Association's Advanced Cardiovascular Life Support (ACLS) exam (February 2016), USMLE sample questions, AMBOSS's USMLE practice questions, and the Certified EKG Technician (CET) practice exam. We evaluated the ChatGPT-4V's responses using the official reference answers as a standard to ensure the reliability of the study. A total of 62 ECG-related questions were included. When screening the questions, as long as the question involved ECG image interpretations, we included this test question in the study and did not perform additional exclusions. Apart from questions focused on ECG diagnosis, some also assessed the ability to determine further treatment plans based on ECGs and corresponding clinical scenarios. We uploaded ECG images to ChatGPT at the original resolution provided in the test questions. These images were unedited, and no additional information was provided to ChatGPT to maintain consistency with the original test questions. ChatGPT was prompted to answer the questions by analyzing the accompanying ECG images and we told it that it was participating in a diagnostic challenge as a representative of AI to avoid it refusing to make a diagnosis (See Supplementary Material). The prompt does not contain any hints about the correct answer. We asked each question

to ChatGPT 3 times to collect 3 responses per question to mitigate the effect of randomness in ChatGPT on our evaluation. Fisher's exact test was used to compare ChatGPT's accuracy in answering different types of questions. To further confirm whether ChatGPT could make accurate diagnoses without relying on options, 19 questions of the diagnostic type that purely examined ECG interpretation without needing integration of clinical history from the prompts were converted to open-ended questions. ChatGPT was then prompted to provide a diagnosis after reading the ECG without options.

The study included 62 questions: 26 for diagnosis, 29 for treatment, and 7 for counting tasks like QT interval length calculation. We evaluated ChatGPT's accuracy using 3 standards: getting at least 1, 2, or 3 correct answers out of 3 attempts. The overall accuracy was 83.87% for at least 1 correct, 70.97% for at least 2, and 53.23% for 3 out of 3 correct (Supplementary Figure 1). There were significant accuracy differences across question types with 1 or 2 correct responses (Table 1, p-values: 0.015 and 0.009), but there is no difference when all 3 responses were required to be correct (p=0.085). Using correct at least 2 times as the standard, the accuracy rates were 65.38% for Diagnosis questions, 86.21% for Treatment Recommendation questions, and 28.57% for Counting questions. Subgroup analysis showed lower accuracy in counting-type questions compared to diagnostic and treatment-related questions, both for requiring one and two correct responses ($P<0.05$, Table 1). Treatment recommendation questions had higher accuracy than other types when at least two correct responses were needed (Table 1). ChatGPT performed poorly in diagnosing ECGs without options, making the correct ECG diagnosis in only 7 out of the responses, which suggests that the diagnostic ability of the current version of ChatGPT for ECG is only possible when a limited range of options is provided. In addition, we found that incorrect responses were related to specific functionalities of ChatGPT-4V. For instance, the insufficient ability of GPT-4V to count parameters such as PR intervals could lead to errors in diagnostic and therapeutic questions. The inadequacy of GPT-4V in integrating electrocardiogram parameters could result in nonspecific diagnoses; for

example, it may diagnose myocardial infarction but fail to combine various parameters to determine the specific location of the infarction.

Our findings indicate that while ChatGPT-4V can analyze ECGs to some extent and can even make treatment decisions based on the EGC, its diagnostic stability and reliability need further improvement for real clinical application. Meanwhile, when comparing the performance of ChatGPT in answering different question types, we found that ChatGPT had significantly lower accuracy on counting-based questions. This suggests that ChatGPT still has limitations in making precise quantitative measurements of the ECG.

It should be noted that the model was not specifically trained on ECG data. We expect ChatGPT-4 may perform better on ECG interpretation as it accumulates more data and training. As a general model, ChatGPT-4's capabilities are not limited to just correctly diagnosing ECGs. Its performance on ECG treatment recommendation questions highlights the potential for applying ChatGPT-4 in medical decision making. By leveraging ChatGPT-4V's abilities to analyze free text and images, we can directly generate management recommendations based on patient data and ECG waveforms, improving healthcare efficiency. While current bedside cardiac monitors can only warn for simple issues such as abnormal heart rhythms or atrial fibrillation, models like ChatGPT-4V could serve as 24/7 "attending physicians" that monitor and analyze ECGs of critically ill patients in the future, capturing low frequency but important ECG abnormalities and promptly detecting condition changes to recommend interventions timely. In addition, ChatGPT can be used to help medical trainees to learn ECG learning for or as an automated second reader to identify high risk diagnoses.

Our study provides a first look at the state-of-the-art ChatGPT-4V model’s capabilities in ECG interpretation. While early results are promising, also highlight current limitations of the model. With further technological developments, multimodal generative AI like ChatGPT may one day play an important role in clinical ECG interpretation and cardiovascular care. Further research will require larger-scale validation to fully evaluate ChatGPT-4V’s ability in ECG interpretation. We believe the

rapid development of large language models will lead to even more exciting progress.

Statements & Declarations

Funding

Not applicable.

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

Not applicable.

Data Availability

The data that support the findings of this study are available on request from the corresponding author upon reasonable request.

Figure legends.

Table 1. Accuracy of the Multimodal ChatGPT-4V Model for Different Types of Question. Fisher's exact test was used to compare the accuracy of ChatGPT in answering different types of questions. If there was a statistically significant difference, subgroup analysis using Fisher's exact test was further performed to respectively compare the accuracy of each type with the other two types. The *fisher.test* function in the stats package of R (4.2.3 version) was used.

Supplementary Material. Prompt used for this study.

Supplementary Figure 1. Accuracy of the Multimodal ChatGPT-4V model in answering ECG interpretation multiple-choice questions. From left to right shows the number of correct responses among 3 attempts for each question. The accuracy rates with at least 1, 2, and 3 correct responses are annotated on the right from bottom to top. Different shapes represent question types. ChatGPT's responses were collected from October 4 to 8, 2023. The ggplot2 R package was used for visualization

Reference

1. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol. 2021;18:465–78.

2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023;

3. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. Resuscitation. 2023;188:109783.

4. Yang Z, Li L, Lin K, Wang J, Lin C-C, Liu Z, et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision) [Internet]. 2023 [cited 2023 Oct 21]. Available from: https://arxiv.org/abs/2309.17421v2

Supplementary Files

Multimedia Appendixes

Table 1. Accuracy of the Multimodal ChatGPT-4V Model for Different Types of Question. Fisher's exact test was used to compare the accuracy of ChatGPT in answering different types of questions. If there was a statistically significant difference, subgroup analysis using Fisher's exact test was further performed to respectively compare the accuracy of each type with the other two types. The fisher.test function in the stats package of R (4.2.3 version) was used.
URL: http://asset.jmir.pub/assets/14ff142dcc2e2febea8b1db66da5533f.docx

Supplementary Materials. Prompts used for this study.
URL: http://asset.jmir.pub/assets/0cda20aad78708fe6e636605c78aaef7.docx

Supplementary Figure 1. Accuracy of the Multimodal ChatGPT-4V model in answering ECG interpretation multiple-choice questions. From left to right shows the number of correct responses among 3 attempts for each question. The accuracy rates with at least 1, 2, and 3 correct responses are annotated on the right from bottom to top. Different shapes represent question types. ChatGPT's responses were collected from October 4 to 8, 2023. The ggplot2 R package was used for visualization.
URL: http://asset.jmir.pub/assets/fa1342d0c5bbca7d56fc10aee26b11c0.png