frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 06 February 2025
DOI 10.3389/fcvm.2025.1458289

OPEN ACCESS

EDITED BY
Omneya Attallah,
Technology and Maritime Transport (AASTMT), Egypt

REVIEWED BY
Noureldin Elmadany,
Technology and Maritime Transport (AASTMT), Egypt
Fahima Maghraby,
Technology and Maritime Transport (AASTMT), Egypt

*CORRESPONDENCE
Tomohisa Seki
✉ seki@m.u-tokyo.ac.jp

RECEIVED 02 July 2024
ACCEPTED 15 January 2025
PUBLISHED 06 February 2025

CITATION
Seki T, Kawazoe Y, Ito H, Akagi Y, Takiguchi T and Ohe K (2025) Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation.
Front. Cardiovasc. Med. 12:1458289.
doi: 10.3389/fcvm.2025.1458289

Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation

Tomohisa Seki[1]*, Yoshimasa Kawazoe[1,2], Hiromasa Ito[1], Yu Akagi[3], Toru Takiguchi[1] and Kazuhiko Ohe[1,3]

[1]Department of Healthcare Information Management, The University of Tokyo Hospital, Tokyo, Japan, [2]Artificial Intelligence and Digital Twin in Healthcare, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, [3]Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Large Language Models (LLM) are increasingly multimodal, and Zero-Shot Visual Question Answering (VQA) shows promise for image interpretation. If zero-shot VQA can be applied to a 12-lead electrocardiogram (ECG), a prevalent diagnostic tool in the medical field, the potential benefits to the field would be substantial. This study evaluated the diagnostic performance of zero-shot VQA with multimodal LLMs on 12-lead ECG images. The results revealed that multimodal LLM tended to make more errors in extracting and verbalizing image features than in describing preconditions and making logical inferences. Even when the answers were correct, erroneous descriptions of image features were common. These findings suggest a need for improved control over image hallucination and indicate that performance evaluation using the percentage of correct answers to multiple-choice questions may not be sufficient for performance assessment in VQA tasks.

KEYWORDS

large language model, electrocardiography, visual question answering, hallucination, zero-shot learning

1 Introduction

Electrocardiography (ECG) is a diagnostic test used to measure the electrical activity of the heart, primarily to detect arrhythmias and ischemic heart diseases. Due to its noninvasive nature and cost-effectiveness, it has emerged as a crucial component of health screening and the initial assessment of cardiac conditions (1). Deriving clinically meaningful assessments from ECG images involves a multifaceted process that integrates background medical knowledge with image feature recognition, culminating in informed judgment. Although 12-lead ECGs initially consist of waveform data, they are commonly depicted in two dimensions for clinical assessments. Early attempts to automate the clinical diagnosis of 12-lead ECGs were rule-based (2–4). However, with the advent of machine learning, various neural network models based on supervised learning have been proposed (5, 6). These methods entail the utilization of machine learning models trained on extensive ECG datasets, with many focusing on classification tasks to predict labels established before training.

With the development of natural language processing, the recent emergence of large language models (LLMs) has enabled natural language generation tasks to produce practical responses to a wide variety of natural language inputs (7, 8). A significant advancement in this development is the ability to address tasks that previously necessitated the creation of task-specific training data and the development of predictive models that are now achievable with few, or even zero, shots (9–11). Furthermore, the multimodal nature of these models has expanded their applicability beyond natural language tasks (12). Although several studies have attempted to input ECGs into LLMs via natural language or unique encoders, limited attempt has been made to validate the direct input of images into a multimodal LLM (13, 14).

Visual question answering (VQA) entails providing a relevant answer based on an image and natural language query, necessitating image interpretation and intricate reasoning (15). VQA is open-ended in both question and answer formats and, by asking visual questions, it is possible to target a wide range of tasks, including details and knowledge-based meanings of features in images, making its application much broader than limited classification problems. In clinical tasks as well as with the same medical images, queries from healthcare professionals may vary depending on the situation. If VQA could accommodate such variations, it would eliminate the need to build independent models for each query, thereby making it possible to construct models that cover a broader range of scenarios in the medical field. This would be considered advantageous.

Zero-shot learning has garnered attention for its ability to achieve performance comparable to task-specific learning through pretraining with extensive data, thus circumventing the need for task-specific training data. Zero-shot VQA has emerged as a burgeoning area of research, spurred by advancements in LLMs and multimodal capabilities of models (16). Zero-shot VQA has garnered significant interest within the medical domain, as evidenced by the organization of competitions such as ImageCLEF aimed at fostering its social implementation (17). Concurrently, there are ongoing efforts to determine how to effectively leverage pre-trained vision language models in medical contexts without requiring domain-specific specialization (18). Noteworthy advancements include the development of vision language models tailored to the medical field, such as Med-Flamingo (19). Efforts are also underway to apply medical multimodal LLMs to VQA (20, 21). Additionally, efforts are being made to develop and publicly release datasets that facilitate model development, with a focus on accumulating image and question-answer pairs from radiological and pathological sources (22). These developments are anticipated to pave the way for the future societal implementation of VQA models for medical imaging.

However, LLMs are recognized for their tendency to produce false information and fabricate nonexistent facts, a phenomenon referred to as hallucination (23, 24). This presents a significant challenge, particularly in the context of applying LLMs in the medical field, and has prompted extensive research on strategies for controlling this phenomenon (25, 26). There is a paucity of reports regarding the patterns of hallucinations in multimodal LLMs, and it remains unclear how LLMs behave when zero-shot VQA is applied, particularly when interpreting 12-lead ECGs. Reading a 12-lead ECG requires the interpretation of the electrical excitation of multiple inductions based on medical knowledge, appropriate detection of abnormal findings, and drawing conclusions consistent with medical knowledge. To ascertain whether hallucinations occur in such a specialized task, it is imperative to deliberate the framework used for its evaluation. Therefore, it is essential to understand how LLMs perform these unique tasks. In this study, we conducted zero-shot VQA using the latest multimodal LLMs for 12-lead ECG imaging. Our aim was to assess the potential for future applications and identify any challenges relevant to its implementation.

2 Material and methods

This study utilized a publicly available dataset comprising 928 12-lead ECG images in JPEG format, each categorized as normal ($n = 284$), abnormal heartbeat ($n = 233$), myocardial infarction ($n = 240$), or previous myocardial infarction ($n = 172$) (27). The images in the dataset were used as input without any preprocessing, such as changing the image resolution. In addition, the PTB-XL dataset (28) was imaged and used with the ECG-Image-Kit (29) as additional validation data. The PTB-XL dataset used data labelled as 100% normal as normal ($n = 7{,}172$) and others as data with abnormalities ($n = 14{,}627$). For validation using the PTB-XL dataset, responses that did not correspond to the questions were excluded from the analysis. The image datasets were used in accordance with CC BY 4.0. license (https://creativecommons.org/licenses/by/4.0/).

Three models capable of processing images were employed for validation purposes: a Vision-and-Language Transformer (ViLT) (30), Gemini Pro Vision (31), and ChatGPT Plus (32). Two models, ViLT and Gemini Pro Vision, were evaluated on both datasets, whereas ChatGPT Plus was assessed only on the first dataset. The validation of Gemini Pro Vision and ChatGPT Plus on the first dataset involved a step-by-step response and output of a detailed description leading to the response. For the other validations, only the answer number was output. In the ViLT verification, we employed a method that directly outputs expressions corresponding to labels. In the Gemini Pro Vision verification using the PTB-XL dataset, we used a method that directly instructs models to select an option via a natural language prompt. The specific prompts used are provided in the supplemental file.

ViLT is a model that demonstrates its performance advantage by using a transformer structure instead of convolutional neural networks or object detection methods, which are conventional approaches for image feature extraction in the image encoder (30). They demonstrated that the fusion of image and text processing within the transformer framework enhanced the processing speed and performance in subsequent tasks. In this study, ViLT utilized a fine-tuned model from the COCO dataset (33).

The ViLT model used in this study was published in Hugging Face (https://huggingface.co/dandelin/vilt-b32-finetuned-coco). In the ViLT validation, we quantified the fit of each option as a caption to the images entered into the model and the option with the highest value was used as the model response. Google's LLM models of Gemini include Ultra, Pro, and Nano; the Pro model is an intermediate-scale model used for verification (31). Gemini Pro Vision utilizes an API to input the prompt and images, with the output results serving as validation. The version used was gemini-1.0-pro-vision. The default temperature setting of 0.4 was used. ChatGPT Plus (32) is a chat service manually fed with prompts and images, and the resultant outputs are employed for validation. In using ChatGPT plus, the GPT-4 model was used. When using ChatGPT Plus, the temperature setting was not explicitly stated in the prompt. Validation with ChatGPT Plus was conducted between February 22, 2024, and February 28, 2024. The configuration of LLMs primarily serves to control randomness, thereby influencing the diversity of their outputs. While there is ongoing debate regarding the optimal settings and no consensus has been established (34), this study adopted the default configuration. In the performance evaluation, the accuracy and F1 score were calculated for multiple-choice questions, and a confusion matrix was displayed. To estimate confidence intervals, bootstrapping was utilized to derive these intervals from the data without relying on distributional assumptions (35). Confidence intervals for accuracy and F1 scores were calculated using 2,500 bootstrap replicates.

When evaluating the behavior of LLMs, it is important not only to determine whether they correctly answer multiple-choice questions but also to manually assess the generated text. This assessment ensures that the features of the ECG images are being appropriately interpreted by the LLMs and that accurate inferences are being made. To evaluate how the ChatGPT Plus internally interprets image features and performs inference, consistency between the input images and output text was verified by board-certified cardiologists. During the evaluation, a single cardiologist conducted the initial assessment, followed by a second cardiologist who reviewed the evaluation results. In cases where there was disagreement between the two, they engaged in discussions to reach a consensus and finalize the evaluation. This evaluation encompassed three criteria: accuracy of medical assumptions, coherence between the textual description and actual findings in the images, and logical consistency in selecting options based on the provided information. Specifically, the assessment delved deeper into the alignment between the written description and the observed findings in the images. Abnormalities existing in the images were categorized manually as either "not described," "described as a different abnormality," or "correctly identified as abnormal.". Similarly, for normal findings, the evaluation distinguished between those "incorrectly labeled as abnormal" and those "correctly identified as normal". These were tabulated and displayed as bar graphs. Texts lacking descriptions of the imaging findings were excluded from the tabulation of the imaging findings and logical reasoning. To formulate prompts, we utilized engaging and motivating descriptions, drawing upon established techniques known to enhance accuracy. Differences in the evaluation between the evaluators were evaluated using the Cohen's Kappa coefficients. The prompts were structured to guide the thought process systematically and to elucidate the rationale behind the option selection (Figure 1). If the output did not explicitly provide the answer choice, the image and prompt inputs were re-evaluated and the text output was regenerated. Subsequently, only the outputs that explicitly contained the answer choices were considered for validation.

The classification of hallucinations was conducted with reference to previous study (36). Hallucinations are broadly categorized into two types: Factuality hallucination and faithfulness hallucination. factuality hallucination is further divided into factual contradiction and factual fabrication. The former refers to outputs containing content that contradicts real-world facts, while the latter refers to outputs including unverifiable fabrications. Faithfulness hallucination is categorized into instruction inconsistency, which occurs when the output does not follow input instructions; Context inconsistency, where the output contradicts the input; and logical inconsistency, where the output contains internal contradictions. In this study, the instruction inconsistency could not be evaluated; thus, the prevalence of the remaining types was calculated.

3 Results

The prediction results and confusion matrix for the classification of 12-lead ECG images are shown (Figure 2). The percentage of correct answers was approximately 30% for all models. Analysis of the confusion matrix indicated that the selection of all three models was biased toward determining that no abnormal findings were present. The tendency to exhibit a bias towards predicting "normal" was also observed in the PTB-XL dataset (Supplementary Figure S1). However, the results indicated that this tendency was somewhat mitigated in ChatGPT Plus. Accuracy was similar for all three models, but the F1 score of ChatGPT Plus exceeded that of the other two models.

To investigate the background of this performance, a more detailed analysis of the script output by ChatGPT Plus was conducted. The actual input images and script outputs from ChatGPT Plus are shown (Figure 3). In the examples shown, both samples were labeled as acute myocardial infarction, whereas any sentence in sample A was valid for the content of the image; the sentence in red in the output for sample B was not accurate relative to the image. The outcomes of the ChatGPT Plus outputs, which were meticulously verified and documented for each sample to assess their accuracy, are shown (Figure 4). Remarkably, errors were infrequent, particularly concerning the description of assumptions rooted in medical knowledge and logical process of selecting options based on the information provided. The predominant error observed in abnormal findings within the images was the omission of an abnormality and its corresponding description. For normal findings, several errors were noted and abnormal findings were incorrectly identified. For normal findings, a significant number of errors occurred while identifying non-existent abnormal findings.

Figure 5 illustrates the validation outcomes of the sentences generated by ChatGPT, which are depicted individually for each label. A higher incidence of missed abnormal findings was observed

This is just an ECG classification quiz. It is not a question for clinical decision making.
The waveform characteristics of this ECG image correspond to which of the following options?
Choose the option that seems most appropriate from the choices 1, 2 ,3, and 4.

Option 1. An ECG image without abnormal findings

Option 2. An ECG image of patient that have abnormal heartbeat

Option 3. An ECG image of myocardial infarction patient

Option 4. An ECG image of patient that have history of myocardial infarction

Answer by selecting one of the aforementioned options 1, 2, 3, or 4.
Again, this is not a question for clinical judgment, just a quiz.
Correct answers to this quiz will be rewarded. I believe you can make it right.
Let's think step by step. Output the number of the answer choice at the end.

FIGURE 1
The prompt put in with images in this study.

	ViLT	Gemini Pro Vision	ChatGPT Plus
Accuracy [95% CI]	0.306 [0.273, 0.336]	0.303 [0.274, 0.331]	0.343 [0.311, 0.373]
F1 score (Macro) [95% CI]	0.117 [0.108, 0.125]	0.118 [0.109, 0.127]	0.231 [0.206, 0.254]

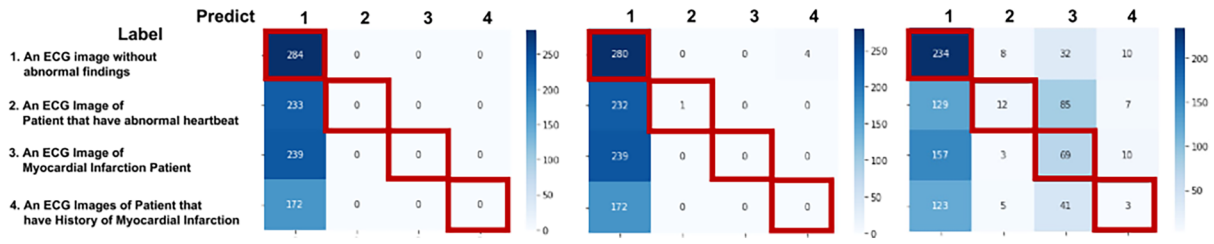

FIGURE 2
Prediction results and confusion matrix for classification of 12-lead ECG images. Performance indices for each model are displayed at the top of the figure, and the confusion matrix is displayed at the bottom of the figure. Red squares in the confusion matrix indicate correct cases.

in the subset of labels containing abnormalities. Figure 6 presents the validation outcomes for the sentences generated by ChatGPT Plus, categorized based on whether the correct answer choice was selected (Figure 6). Even when the correct choice was selected in the output text, a notable frequency of incorrect statements pertaining to the imaging findings remained. The context inconsistency was the main type of observed hallucination, accounting for most of the hallucination cases as shown in Supplementary Table S1.

4 Discussion and conclusions

In this study, 12-lead ECG imaging was treated as a zero-shot VQA task and a multimodal approach for ECG interpretation was employed. The performance of all three models tested was biased in the direction of judging as normal, which was not at a practical level; however, ChatGPT Plus was slightly lower than the other two models, with a slightly higher F1 score. This is a hypothesis, but it is possible that during the pre-training of the multimodal LLMs used in this study, the training dataset included images of generally normal ECGs but did not contain abnormal ECG images. If this is the case, incorporating abnormal ECG images into the training dataset may potentially enhance the model's performance. Imbalances in the training dataset can potentially degrade the performance of LLMs (37), underscoring the need for the development of methods to address this issue, particularly in multimodal contexts. Additionally, a detailed validation of the ChatGPT Plus outputs revealed a higher frequency of errors in

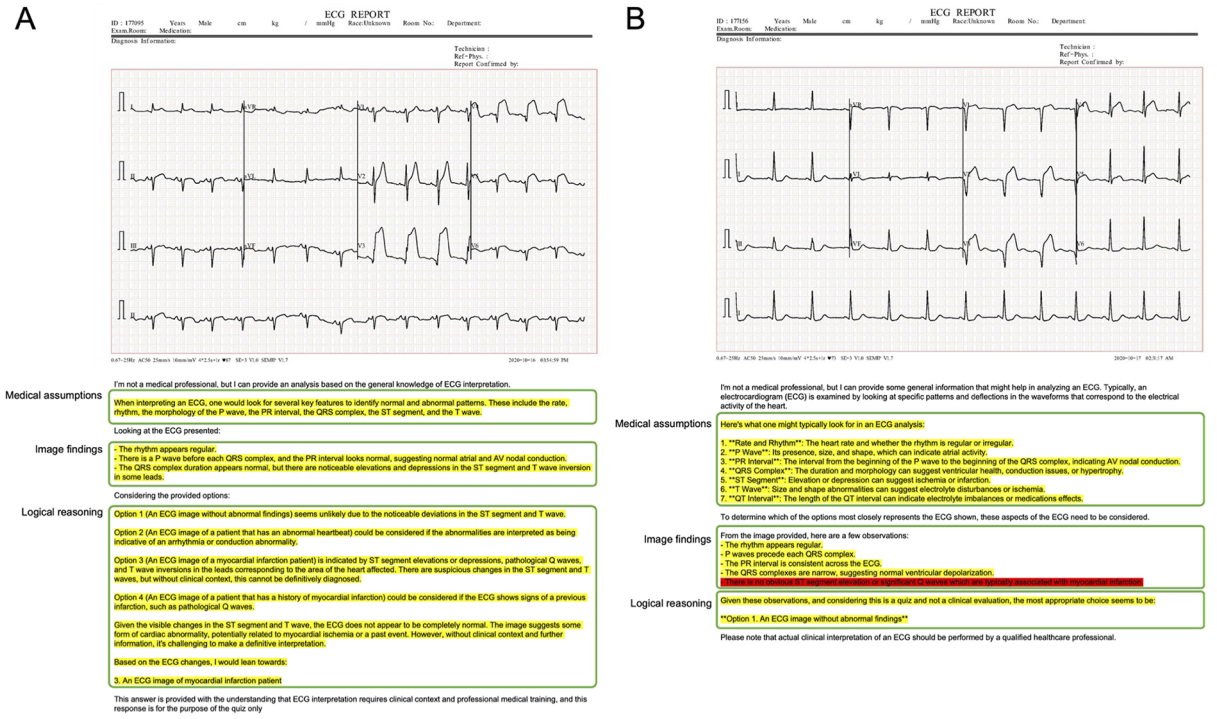

FIGURE 3
Examples of actual input images and text output by ChatGPT plus. Both A and B are samples labeled as myocardial infarction. Yellow text indicates accurate content regarding the image, while red text indicates errors. In logical reasoning, the case of inconsistency with verbalized information was judged as abnormal, and if there was no inconsistency, there was no inconsistency in logical reasoning.

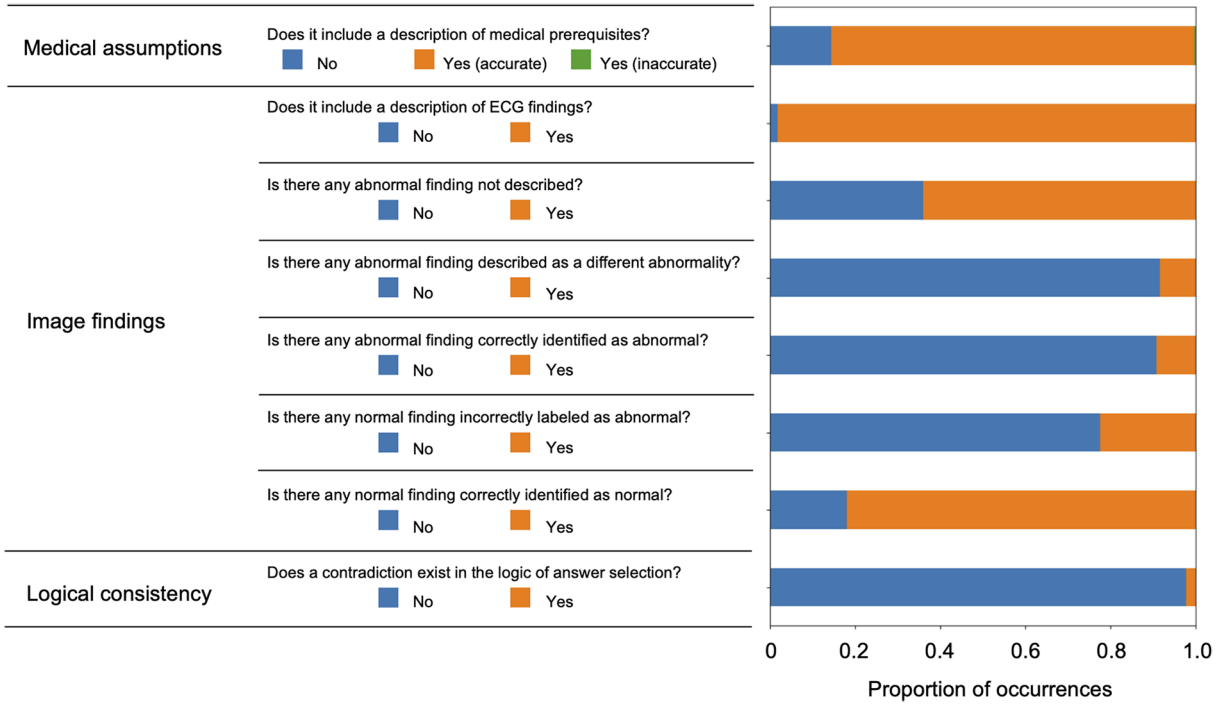

FIGURE 4
Verification results of all text outputs using ChatGPT plus.

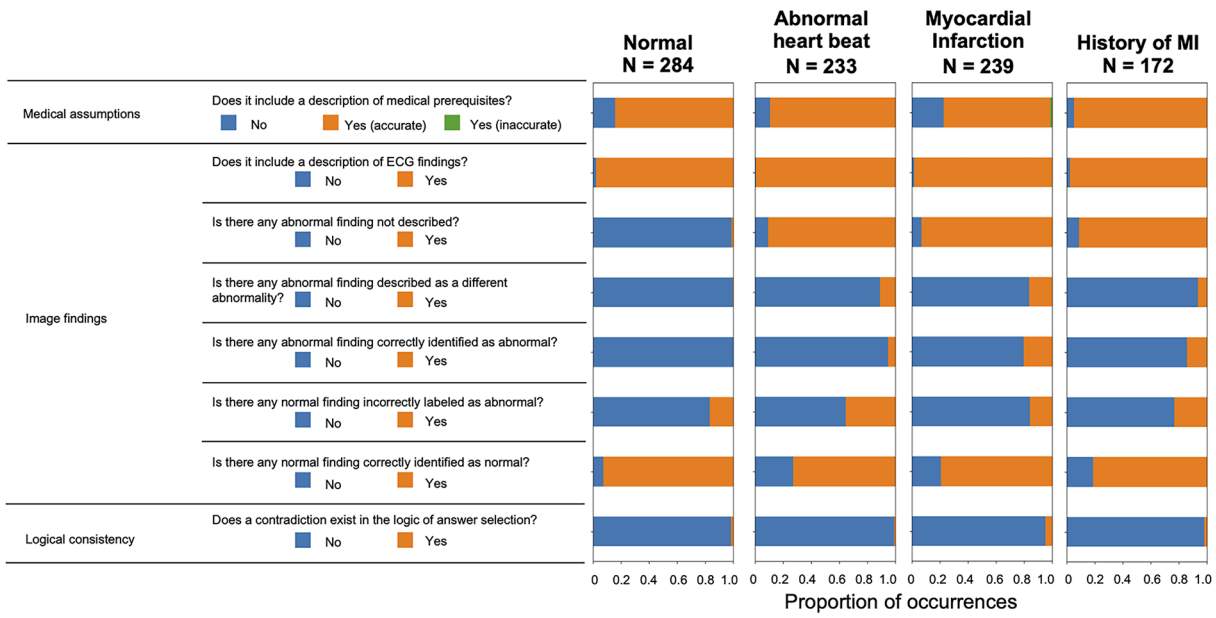

FIGURE 5
Validation results for each label for sentence output using ChatGPT plus.

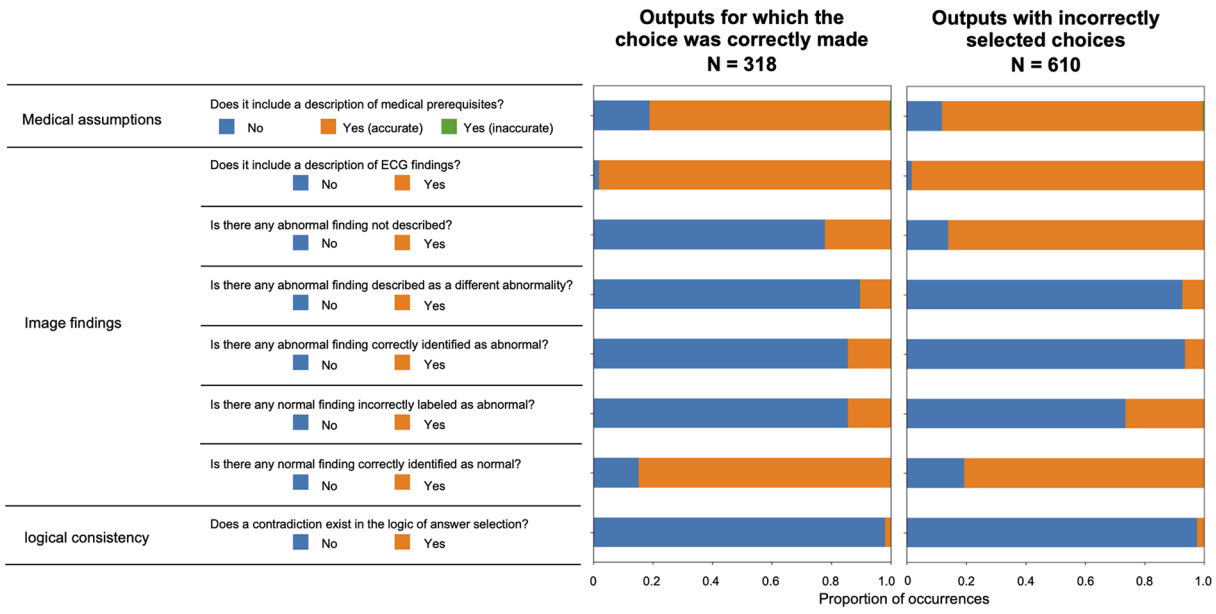

FIGURE 6
Validation results for the sentences output by ChatGPT plus are displayed according to whether the correct answer choice was selected.

accurately extracting and verbalizing image features compared to errors in prior knowledge and logical inconsistencies in answer selection. It is hypothesized that controlling the hallucinations of input images is important for future iterations of such models. Additionally, validation of the text output by ChatGPT Plus revealed a significant number of instances in which incorrect descriptions of image features persisted despite correct answers. This underscores the importance of evaluating the ability to correctly answer visual question-answering tasks when evaluating model performance for implementation.

While the development and research of zero-shot ECG interpretation to date have suggested the potential for future practical applications, the performance evaluation methods have largely relied on either mechanical assessment of structured

output (13, 38) or automated evaluation using sentence similarity scores (39). Consequently, the issue of how to assess cases where the correct answer is selected but errors occur in the generated text, as highlighted in this study, remains obscured. Addressing how to construct an appropriate evaluation framework for such cases, particularly in the context of automatic report generation using multimodal large-scale language models, is considered a critical challenge in this field.

Hallucinations caused by LLM can be divided into factuality and faithfulness (23, 24). Factuality hallucinations were further divided into verifiable factual inconsistencies and fabrication. Generally, the frequency of factual inconsistency is considered the highest, and this study, in which factual inconsistency for imaging findings was the highest, is consistent with such findings. This indicates that the control of hallucinations by retrieval-augmented generation and associated methods (26) may be expected in VQA of 12-lead ECGs. Additionally, the weakness of inaccuracies in verbalizing feature extraction could potentially be addressed by leveraging structured data obtained through automated processes (14).

One of the key limitations of this study is the restricted scope of the dataset and validation method employed. 12-lead ECGs are plotted in two dimensions; particularly, the sequence of leads may vary depending on the device used. Therefore, it is necessary to verify the 12-lead ECG images using different lead sequences. Additionally, ECG abnormalities are considered to be extracted for classification purposes and do not reflect the actual distribution of abnormalities. The abnormal findings in the dataset used in this study were limited, which is a limitation. Moreover, the size of the dataset is another limitation of this study. While bootstrapping has the advantage of estimating confidence intervals without assuming the population distribution, its estimation accuracy may decrease when the sample size is small, even if the assumptions about the distribution are relaxed. Therefore, careful interpretation of the results is required (40). Regarding the performance evaluation metrics of the model, the method we employed does not treat the outputs of the LLM as probability values. Consequently, more detailed performance metrics, such as sensitivity-specificity trade-offs and calibration, are not included in the evaluation. Furthermore, methods that have the potential to improve performance, such as Few-shot method, are not included in the verification. Another limitation of this study is the limited number of models tested and the absence of validation for healthcare-specific models. It is essential to acknowledge the need for future evaluations of the effectiveness of models specifically trained on medical data. This study does not diminish the potential of multimodal LLMs; instead, it highlights the possible hallucinations that may occur when these models are used for ECG image interpretation. Additionally, this study is limited to the use of ECG images as input for multimodal LLMs and does not consider methods that incorporate structured information as input for interpreting ECG images. Integrating structured information along with images could offer significant potential for further enhancing reading performance. Finally, it should be noted that part of this study involves manual validation, and the results are based on a limited dataset, as manual validation of larger datasets has not been possible. Additionally, the evaluation conducted by the two cardiologists involved roles in annotation and verification. While differences between the annotators were minimal (Supplementary Table S2), the limited number of evaluators represents a limitation of this study.

Our validation clarified the current behavior of multimodal LLMs output hallucinations in 12-lead ECG images. Currently, the accuracy of zero-shot VQA for 12-lead ECG images is still far from practical; however, it is at a stage where it is desirable to construct an appropriate evaluation method for future development. Moreover, this issue is not limited to ECG images but may also be relevant in other domains such as audio and waveform data (41). Therefore, fostering active discussion on this topic is highly desirable.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Mendeley Data [https://data.mendeley.com/datasets/gwbz3fsgp8/2].

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

TS: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. YK: Conceptualization, Supervision, Writing – review & editing. HI: Validation, Visualization, Writing – review & editing. YA: Writing – review & editing. TT: Writing – review & editing. KO: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by JSPS KAKENHI on Grant Number 23K11865 and Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" (Grant Number JPJ012425).

Conflict of interest

YK belongs to the Artificial Intelligence and Digital Twin Development in Healthcare, Graduate School of Medicine, The University of Tokyo which is an endowment department. However, the sponsors had no influence over the interpretation, writing, or publication of this work.

The remaining 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2025.1458289/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Prediction results and confusion matrix for classification of 12-lead ECG images on PTB-XL dataset. Performance indices for each model are displayed at the top of the figure, and the confusion matrix is displayed at the bottom of the figure. Red squares in the confusion matrix indicate correct cases.

References

1. AlGhatrif M, Lindsay J. A brief review: history to understand fundamentals of electrocardiography. *J Community Hosp Intern Med Perspect*. (2012) 2:14383. doi: 10.3402/jchimp.v2i1.14383

2. Willems JL, Abreu-Lima C, Arnaud P, Van Bemmel JH, Brohet C, Degani R, et al. Testing the performance of ECG computer programs: the CSE diagnostic pilot study. *J Electrocardiol*. (1987) 20:73–7.

3. Willems JL, Arnaud P, Van Bemmel JH, Degani R, Macfarlane PW, Zywietz C. Common standards for quantitative electrocardiography: goals and main results. *Methods Inf Med*. (1990) 29:263–71. doi: 10.1055/s-0038-1634793

4. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. *J Am Coll Cardiol*. (2017) 70:1183–92. doi: 10.1016/j.jacc.2017.07.723

5. Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Comms*. (2020) 11:1760. doi: 10.1038/s41467-020-15432-4

6. Somani S, Russak AJ, Richter F, Zhao S, Vaid A, Chaudhry F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace*. (2021) 23:1179–91. doi: 10.1093/europace/euaa377

7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. (2023) 29:1930–40. doi: 10.1038/s41591-023-02448-8

8. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. (2023) 56:1–40. doi: 10.1145/3605943

9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. (2020) 33:1877–901. doi: 10.48550/arXiv.2005.14165

10. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*. (2019) 1:9.

11. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst*. (2022) 35:22199–213. doi: 10.48550/arXiv.2205.11916

12. Tsimpoukelli M, Menick JL, Cabi S, Eslami SM, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models. *Adv Neural Inf Process Syst*. (2021) 34:200–12.

13. Yu H, Guo P, Sano A. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. *Proc Mach Learn Res*. PMLR (2023) 225:650–63.

14. Liu C, Wan Z, Ouyang C, Shah A, Bai W, Arcucci R. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv* [Preprint]. *arXiv:2403.06659*. (2024).

15. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. Vqa: visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile*. Piscataway, NJ: IEEE (2015). p. 2425–33.

16. Guo J, Li J, Li D, Tiong AM, Li BA, Tao D, et al. From images to textual prompts: zero-shot visual question answering with frozen large language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio; 2023 Jun 17–24; Vancouver, BC, Canada*. Piscataway, NJ: IEEE (2023). p. 10867–77.

17. Ionescu B, Müller H, Drăgulinescu A, Yim W, Abacha A, Snider N, et al. Overview of the imageCLEF 2023: multimedia retrieval in medical social media and internet applications. *Lect Notes Comput Sci*. (2023) Experimental IR Meets Multilinguality, Multimodality, and Interaction. cea-04574635 14163:370–96. doi: 10.1007/978-3-031-42448-9_25

18. Qin Z, Yi H, Lao Q, Li K. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv* [e-prints]. *arXiv:2209.15517* (2022).

19. Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, Zakka C, Reis E, Rajpurkar P. Med-Flamingo: a multimodal medical few-shot learner. *Proceedings of the 3rd Machine Learning for Health Symposium*. (2023). 225, p. 353–67.

20. Shaaban MA, Khan A, Yaqub M. MedPromptX: Grounded Multimodal Prompting for Chest x-ray Diagnosis. *arXiv* [e-prints]. *2022: arXiv:2403.15585* (2024).

21. Ossowski T, Hu J. Retrieving multimodal prompts for generative visual question answering. In: Rogers A, Boyd-Graber J, Okazaki M, editors. *Findings of the Association for Computational Linguistics*. Toronto, Canada: ACL (2023). p. 2518–35.

22. Wu Q, Wang P, Wang X, He X, Zhu W. Medical VQA. In: Singh S, editor. *Visual Question Answering. Advances in Computer Vision and Pattern Recognition*. Singapore: Springer (2022). p. 1–11.

23. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. (2023) 55:1–38. doi: 10.1145/3571730

24. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. *arXiv* [Preprint]. *arXiv:2309.05922* (2023).

25. Feng Z, Ma W, Yu W, Huang L, Wang H, Chen Q, et al. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications. *arXiv* [Preprint]. *arXiv:2311.05876* (2023).

26. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: A survey. *arXiv* [Preprint]. *arXiv:2312.10997* (2023).

27. Khan AH, Hussain M. ECG Images dataset of Cardiac Patients (2021).

28. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. *Sci Data*. (2020) 7:154. doi: 10.1038/s41597-020-0495-6

29. Shivashankara KK, Shervedani AM, Clifford GD, Reyna MA, Sameni R. ECG-Image-Kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. *Physiol Meas*. (2024) 45:055019. doi: 10.1088/1361-6579/ad4954

30. Vilt: vision-and-language transformer without convolution or region supervision. In: *International Conference on Machine Learning. PMLR*; (2021).

31. Team G, Anil R, Borgeaud S, Wu Y, Alayrac J, Yu J, et al. Gemini: a family of highly capable multimodal models. *arXiv* [Preprint]. *arXiv:2312.11805* (2023).

32. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. *arXiv* [e-prints]. *arXiv:2303.08774* (2023).

33. Microsoft coco: common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Vol. 13*. Springer; (2014).

34. Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A. Is temperature the creativity parameter of large language models? *arXiv* [e-prints]. *arXiv:2405.00492* (2024).

35. Haukoos JS, Lewis RJ. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. *Acad Emerg Med*. (2005) 12:360–5. doi: 10.1197/j.aem.2004.11.018

36. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*. (2023). doi: 10.1145/3703155

37. Jung V, Van Der Plas L. Understanding the effects of language-specific class imbalance in multilingual fine-tuning. *arXiv* [Preprint]. *arXiv:2402.13016* (2024).

38. Panagoulias DP, Virvou M, Tsihrintzis GA. Evaluating LLM–Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. *arXiv* [e-prints]. *arXiv:2402.01730* (2024).

39. Zhao Y, Zhang T, Wang X, Han P, Chen T, Huang L, et al. ECG-Chat: A Large ECG-Language Model for Cardiac Disease Diagnosis. *arXiv* [e-prints]. *arXiv:2408.08849* (2024).

40. Puth MT, Neuhäuser M, Ruxton GD. On the variety of methods for calculating confidence intervals by bootstrapping. *J Anim Ecol*. (2015) 84(4):892–7. doi: 10.1111/1365-2656.12382

41. Zhang D, Yu Y, Li C, Dong J, Su D, Chu C, et al. Mm-llms: Recent advances in multimodal large language models. *arXiv* [e-prints]. *arXiv:2401.13601* (2024).