THE ANATOLIAN JOURNAL OF CARDIOLOGY

Official journal of the TURKISH SOCIETY OF CARDIOLOGY

Large Language Models in Intracardiac Electrogram Interpretation: A New Frontier in Cardiac Diagnostics for Pacemaker Patients

ORIGINAL INVESTIGATION

ABSTRACT

Background: Interpreting intracardiac electrograms (EGMs) requires expertise that many cardiologists lack. Artificial intelligence models like ChatGPT-4o may improve diagnostic accuracy. This study evaluates ChatGPT-4o's performance in EGM interpretation across 4 scenarios (A-D) with increasing contextual information.

Methods: Twenty EGM cases from *The EHRA Book of Pacemaker, ICD, and CRT Troubleshooting* were analyzed using ChatGPT-4o. Ten predefined features were assessed in Scenarios A and B, while Scenarios C and D required 20 correct responses per scenario across all cases. Performance was evaluated over 2 months using McNemar's test, Cohen's Kappa, and Prevalence- and Bias-Adjusted Kappa (PABAK).

Results: Providing clinical context enhanced ChatGPT-4o's accuracy, improving from 57% (Scenario A) to 66% (Scenario B). "No Answer" rates decreased from 19.5% to 8%, while false responses increased from 8.5% to 11%, suggesting occasional misinterpretation. Agreement in Scenario A showed high reliability for atrial activity ($\kappa = 0.7$) and synchronization ($\kappa = 0.7$), but poor for chamber ($\kappa = -0.26$). In Scenario B, understanding achieved near-perfect agreement (Prevalence-Adjustment and Bias-Adjustment Kappa (PABAK) = 1), while ventricular activity remained unreliable ($\kappa = -0.11$). In Scenarios C (30%) and D (25%), accuracy was lower, and agreement between baseline and second-month responses remained fair ($\kappa = 0.285$ and 0.3, respectively), indicating limited consistency in complex decision-making tasks.

Conclusion: This study provides the fir t systematic evaluation of ChatGPT-4o in EGM interpretation, demonstrating promising accuracy and reliability in structured tasks. While the model integrated contextual data well, its adaptability to complex cases was limited. Further optimization and validation are needed before clinical use.

Keywords: Artificial intelligence, large language models, intracardiac electrograms, pacemaker, ChatGPT-4o

INTRODUCTION

Intracardiac electrograms (EGMs) offer a highly detailed view of cardiac electrical activity, serving as a critical tool in the management of pacemaker (PM) patients. However, their intricate nature often necessitates specialized expertise, which may not always be readily available among clinicians. This gap in knowledge and expertise poses a signifi ant challenge to achieving accurate and timely diagnoses, potentially impacting patient outcomes. Moreover, the global surge in cardiac implantable electronic device (CIED) procedures has placed increasing strain on healthcare systems, both in terms of clinical capacity and financial resources. These combined factors underscore the urgent need for innovative solutions to streamline EGM interpretation and improve the efficiency of CIED management in modern healthcare settings.

Approximately, 25% of patients miss follow-ups within the fir t year, with access challenges particularly affecting elderly, disabled, and rural populations. Remote device management for CIED, including routine remote follow-up and event-triggered remote monitoring (RM), has improved access and follow-up compliance. The coronavirus disease 2019 (COVID-19) pandemic underscored the importance

Serdar Bozyel[1]

Ahmet Berk Duman[1]

Şadiye Nur Dalgıç[1]

Abdülcebar Şipal[1]

Faysal Şaylık[2]

Şükriye Ebru Gölcük Önder[1]

Metin Çağdaş[1]

Tümer Erdem Güler[1]

Tolga Aksu[3]

Ulas Bağcı[4]

Nurgül Keser[5]

[1]*Department of Cardiology, Health Sciences University, Kocaeli City Hospital, Kocaeli, Türkiye*
[2]*Health Sciences University, Van Training and Research Hospital, Van, Türkiye*
[3]*İstanbul Aydın University, Medical Park Florya Hospital, İstanbul, Türkiye*
[4]*Department of Radiology, Machine & Hybrid Intelligence Lab, Northwestern University, Chicago, IL, USA*
[5]*Department of Cardiology, Health Sciences University, Sultan Abdulhamid Han Training and Research Hospital, İstanbul, Türkiye*

Corresponding author:
Serdar Bozyel
✉ drserdarbozyel@gmail.com

Received: February 15, 2025
Accepted: May 20, 2025
Available Online Date: July 4, 2025

Cite this article as: Bozyel S, Duman AB, Dalgıç ŞN, et al. Large language models in intracardiac electrogram interpretation: A new frontier in cardiac diagnostics for pacemaker patients. *Anatol J Cardiol.* 2025;XX(X):1-10.

DOI:10.14744/AnatolJCardiol.2025.5238

of remote healthcare for overcoming logistical barriers.[1-3] Studies like TRUST and CONNECT demonstrate RM's effectiveness in detecting arrhythmias, often matching or surpassing in-person visits.[4,5] However, current systems rely on rigid algorithms that lack real-time contextual adaptability, leading to delays, increased workloads, and potential diagnostic errors.

Artificial intelligence (AI), particularly large language models (LLMs), holds transformative potential for addressing these challenges. While extensively used in electrocardiogram (ECG) interpretation, LLMs' application in analyzing the more complex EGMs remains underexplored.[6-9] Artificial intelligence tools like ChatGPT-4o can automate the interpretation of such datasets, identifying patterns beyond the scope of conventional algorithms. With its ability to integrate textual and visual data, ChatGPT-4o emerges as a valuable diagnostic aid, especially in resource-limited settings lacking specialized expertise.

This study evaluates ChatGPT-4o's ability to interpret intracardiac EGMs from PM patients using a scenario-based framework. The assessment focuses on its accuracy, reliability, and adaptability across varying clinical complexities. Findings provide insights into how LLMs could be integrated into clinical workfl ws to support diagnostics and improve patient care.

METHODS

Twenty PM EGM examples were sourced from *The European Heart Rhythm Association (EHRA) Book of Pacemaker, ICD, and CRT Troubleshooting*, a key reference for cardiologists specializing in cardiac devices.[10] This resource, commissioned by the EHRA and developed under the guidance of the EHRA Education Committee, features tests structured as multiple-choice questions with correct answers and brief explanations. Each case-based question comprises 4 sections:

1. **Case description/device parameters:** This part outlines the patient's clinical condition and the device parameters.
2. **EGM:** This section presents the device EGM record that serves as the foundation of the case.
3. **Question sentence.**
4. **Multiple-choice questions.**

The fir t 20 EGM cases were selected sequentially without specific cri eria, avoiding randomization or stratifi ation by difficult , which may impact generalizability. ChatGPT-4o, utilizing image processing, interpreted these cases across 4 scenarios of increasing complexity, ranging from isolated EGM analysis to clinical context and multiple-choice formats.

Ten predefined features were assessed in Scenarios A and B, while Scenarios C and D required 20 correct responses per scenario across all cases.

Four scenarios based on the same case were presented to the model at 1-week intervals:

1. Scenario A involves "only the visualized EGM data'' without patient history or device parameters (Supplementary Figure 1A).
2. Scenario B adds "clinical context and device parameters'' to aid interpretation (Supplementary Figure 1B).
3. Scenario C further includes "a specific question'' requiring an answer based on the provided data (Supplementary Figure 1C).
4. Scenario D incorporates all previous elements but requires selecting the correct answer from "multiple-choice options'' (Supplementary Figure 1D).

Intracardiac electrograms images were presented in their original format, as extracted from the source material, without modifi ations such as preprocessing, resizing, or fil ering. The resolution and clarity of these images were consistent with those in the EHRA book. This ensured that the model's performance was evaluated under realistic conditions.

The procedures detailed above were conducted again over a second month, with each scenario spaced 1 week apart, to evaluate the consistency of the model's responses. This repeated evaluation allowed for an assessment of intra-model variability and potential learning effects across sessions.

Evaluation Method

In the fir t 2 scenarios, an assessment was conducted to determine whether the features in the EGM were accurately identified. This analysis was based on 10 specific criteria (Table 1). ChatGPT-4o's responses to 20 EGM cases were evaluated using a structured accuracy assessment:

1. False answer.
2. True answer.
3. No answer (N/A).
4. No malfunction or pseudomalfunction detected (N/R)

In the third and fourth scenarios, the assessment aimed to determine if the model (ChatGPT-4o) could accurately give or select the correct answer from the given choices (False or True). Scenarios C and D were evaluated based on whether the model selected the correct answer for each case. This grading system ensured a detailed and structured analysis of the model's performance in interpreting EGM data. The responses to all 4 scenarios were evaluated by 2 independent

HIGHLIGHTS

- ChatGPT-4o was systematically evaluated for the fir t time in EGM interpretation and demonstrated promising accuracy rates.
- The model's accuracy was 57% when using only EGM data, increasing to 66% when clinical context and device parameters were included.
- While high reliability was observed in critical variables, the model's consistency remained limited, particularly in complex decision-making processes and certain rhythm parameters.
- Advanced optimization and large-scale validation studies are required for the integration of large language models (LLMs) into clinical practice.

Table 1. Key Evaluation Criteria for Electrogram (Intracardiac Electrograms) Interpretation

1) Understanding	Can it understand that it is an electrocardiogram trace?
2) Marker annotations	Can it interprete the specific ann tations markers ?
3) Atrial Activity	Can it provide an accurate description of atrial activation?
4) Ventricular activity	Can it provide an accurate description of ventricular activation?
5) Chamber	Can it identify how many chambers the device has?
6) Pacing mode	Can it identify the device pacing mode?
7) Timing Intervals	Can it identify the timing intervals between events ?
8) Malfunction	Can it accurately identify the existing malfunction?
9) Pseudomalfunction	Can it accurately identify the existing pseudomalfunction?
10) Synchronizations	Can it offer any insights into the atrioventricular relationship?

cardiac device specialists. In cases of disagreement, a third expert was consulted. The inter-rater reliability between evaluators was not quantified, but consensus-based adjudication was employed to resolve discrepancies.

ChatGPT-4o

ChatGPT-4o, developed by OpenAI, is an advanced LLM designed to process and generate human-like text. Built on deep learning techniques, it supports a wide range of tasks, including natural language understanding, text generation, and image analysis. Its multimodal capabilities allow it to integrate and analyze both visual and textual data, making it particularly suitable for complex, data-rich applications.

In this study, ChatGPT-4o was tasked with interpreting EGM visuals alongside accompanying textual information, such as clinical context and device parameters. This dual-input approach leveraged the model's ability to synthesize diverse data types, providing a comprehensive framework for evaluating diagnostic accuracy. By combining visual signal interpretation with textual analysis, ChatGPT-4o demonstrated its potential as a versatile tool for improving the interpretation of EGMs and supporting clinical workfl ws.

Repeated testing over a second month was performed to evaluate intra-model consistency. As ChatGPT-4o does not retain memory between sessions, this design does not reflect model learning or adaptation. While the model itself was the subject of analysis, it was not used to generate any scientific content or interpret study results. Language polishing was performed with AI support, but all intellectual and analytical contributions were made by the authors.

Statistical Analysis

All statistical analyses were conducted using Python version 3.1.4 (Python Software Foundation, USA). The distributions of responses in Scenarios A and B were presented with the percentages of answers (True, False, Non-Relevant, No Answer) to evaluate the model's baseline performance across categories. Responses were categorized using the classifi ation system described previously. For demonstration of percentages, the authors used pie charts and stacked bar charts. For comparisons and agreement analyses of answers between scenarios, true responses were compared with the other responses. Percentage differences between Scenario A and Scenario B were calculated for each variable, and the results were visualized using a heatmap to provide a clear comparison of performance variations across categories.

Because the answers were paired (dependent) and categorical, McNemar's test was used for comparisons between Scenarios A-B and C-D, as well as between baseline and second-month percentages. Agreement between baseline and second-month responses was assessed using Cohen's Kappa and PABAK (Prevalence-Adjustment and Bias-Adjustment Kappa) analyses, with interpretations provided to classify the degree of agreement (e.g., poor, slight, moderate, substantial, and perfect).

It was determined that Cohen's Kappa has a prominent limitation in data with different prevalences (e.g. very high or very low) and in non-balanced data.[11,12] Therefore, it is advised to use PABAK or Gwet's AC1 in such cases. Since the authors' data has a non-balanced distribution between answer groups, the authors aimed to demonstrate in this study that Cohen's Kappa has signifi ant limitations when applied to imbalanced data, which can lead to incorrect results, and to show the superiority of using PABAK in such cases instead. These methods allowed for a comprehensive assessment of performance consistency and reliability across scenarios and time points. All figu es were reviewed during manuscript preparation to ensure clarity, consistency in axis labels, and alignment between visual content and descriptive text.

This study employed the AI-based LLM ChatGPT-4o (OpenAI) for the interpretation of EGMs. Additionally, AI-assisted tools were used for language editing to improve the clarity and readability of the manuscript. However, all scientific content, analysis, and conclusions were generated by the authors without AI influen e.

Transparency Statement

ChatGPT-4o was evaluated solely as the subject of analysis in this study. It was not used to generate, interpret, or revise any scientific content related to study design, data analysis, or conclusions. All scientific reasoning, methodology development, and diagnostic interpretation were performed independently by the authors. Minor language editing was conducted using external AI-based proofreading tools (ChatGPT-4o and Grammarly), limited to stylistic and grammatical refineme t only.

RESULTS

Scenarios A and B: Distribution of Answers and Variable Performance

The distributions of answers across Scenarios A and B highlight key differences in performance and engagement. In Scenario

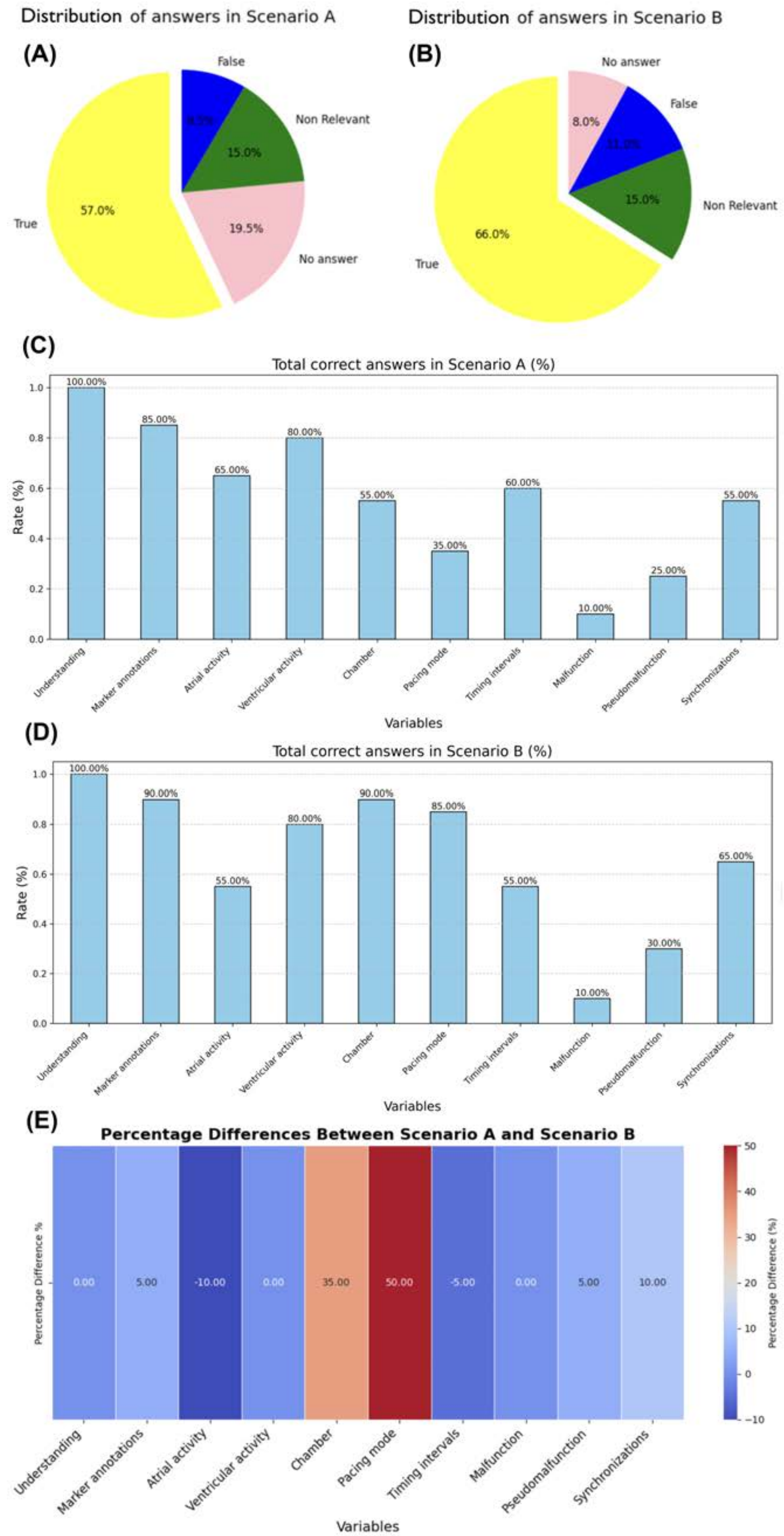


Figure 1. Performance metrics across scenarios. (A) Distribution of answers in scenario A. (B) Distribution of answers in scenario B. (C) Distribution of correct answers across key variables in Scenario A. (D) Distribution of correct answers across key variables in Scenario B. (E) Accuracy Improvement (Heatmap) Between Scenario A and Scenario B across features.

A, the majority of responses (57%) were True, while 19.5% were categorized as "No Answer," 15% as "Non-Relevant," and 8.5% as "False." Conversely, in Scenario B, the percentage of True answers improved to 66%. The proportion of "No Answer" responses decreased to 8%, while "False" responses increased slightly to 11% (Figures 1A and B). These visual distributions clearly illustrate how the inclusion of clinical context in Scenario B not only improved accuracy but also reduced uncertainty, as reflec ed by the decline in "No Answer" rates.

Figures 1C and D illustrate the distribution of correct responses across key variables for Scenarios A and B, with statistical comparisons summarized in Table 2. The understanding variable maintained perfect accuracy (100%) across both scenarios, while Marker Annotations improved slightly from 85% to 90%, though this was not statistically signifi ant. The most notable improvements were observed in pacing mode, which increased from 35% to 85% (P = .002), and chamber, which rose from 55% to 90% (P = .039) in Scenario B. Persistent challenges were noted in malfunction (remaining at 10%) and pseudomalfunction, which increased slightly from 25% to 30%, both without statistical signifi ance. Overall, the most pronounced and persistent disparity was observed in pacing mode, while other features showed either stability or minor, non-signifi ant variations. Figure 1A-D illustrate how contextual enrichment improved the distribution of response types and diagnostic accuracy across key variables, particularly by reducing uncertainty and increasing correct classifi ations.

The percentage differences between Scenarios A and B are visualized in Figure 1E. Pacing mode and chamber demonstrated the largest gains, with improvements of +50% and +35%, respectively. The most prominent improvements in diagnostic accuracy were observed in pacing mode (+50%) and chamber identifi ation (+35%), providing key insight into the model's enhanced performance when contextual information is included. In contrast, atrial activity exhibited a decline of −10%, though this was not statistically signifi ant (P = .5). Other variables, such as marker annotations (+5%) and synchronizations (+10%), showed modest improvements. The heatmap format in Figure 1E enables a side-by-side comparison of variable-specific accuracy changes, highlighting the model's strengths and limitations across diagnostic domains.

Accuracy Comparison Between Scenarios C and D

In Scenario C, 30% of responses were True, while in Scenario D, this rate decreased to 25% (Figure 2). Statistical analysis indicates that this difference is not signifi ant (P = 1). Both scenarios demonstrate low accuracy, highlighting the need for further improvement. In the second month, accuracy in Scenario C was 30% and increased to 50% in Scenario D (P = .125), indicating no signifi ant difference or improvement between the scenarios (Table 3).

Figure 3 demonstrates ChatGPT's ability to interpret an EGM, identifying ventricular oversensing characterized by irregular R-R intervals and pacing inhibition in a Ventricular Demand Pacing with Dual atrial sensing (VDD) PM.

Agreement Metrics for Baseline and Second-Month Responses

Tables 4 and 5 summarize the agreement metrics comparing baseline and second-month responses in Scenarios A and B across 10 key features.

In Scenario A, agreement levels ranged from poor to substantial, with substantial agreement observed for atrial activity (kappa = 0.7) and synchronizations (kappa = 0.7), indicating high reliability. Moderate agreement was noted for pseudomalfunction (kappa = 0.47), while fair or slight agreement was seen for timing intervals (kappa = 0.4) and Pacing Mode (kappa = 0.08). Poor agreement was identified for chamber (kappa = −0.26) and ventricular activity (kappa = −0.25).

In Scenario B, agreement ranged from poor to almost perfect, with understanding (PABAK = 1) achieving almost perfect agreement, reflecting high consistency. Moderate agreement was noted for pseudomalfunction (kappa = 0.6) and timing intervals (kappa = 0.4), while fair agreement was observed for pacing mode (kappa = 0.3) and synchronizations

Table 2. Differences in Accuracy Rates Between Scenarios A and B at Baseline and Second Month

Key Features	Scenario A Baseline	Scenario B Baseline	P	Scenario A Second Month	Scenario B Second Month	P
Understanding	1	1	Nan	0.95	1	
Marker annotations	0.85	0.9	1	0.95	0.8	.25
Atrial activity	0.65	0.55	.5	0.5	0.55	1
Ventricular activity	0.8	0.8	1	0.8	0.8	1
Chamber	0.55	0.9	.039	0.75	0.9	.453
Pacing mode	0.35	0.85	.002	0.1	0.6	.002
Timing intervals	0.6	0.55	1	0.5	0.5	1
Malfunction	0.1	0.1	1	0	0.1	
Pseudomalfunction	0.25	0.3	1	0.25	0.15	.625
Synchronizations	0.55	0.65	.625	0.4	0.55	.453

Most diagnostic errors in Scenario A were concentrated in pacing mode identifi ation, which showed the lowest baseline accuracy (35%) and the most statistically signifi ant improvement in Scenario B (P = .002). Errors related to marker interpretation were also observed but were less pronounced.

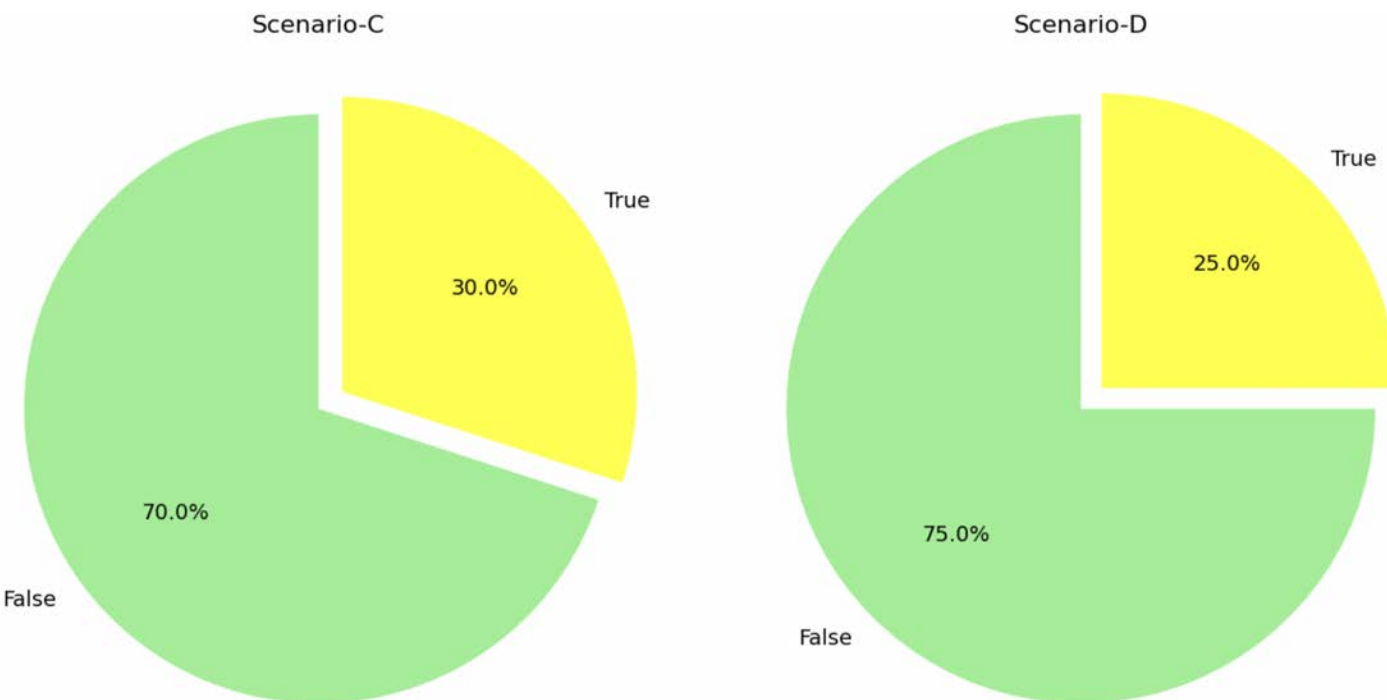


Figure 2. Comparison of true and false response counts in Scenarios C and D.

(kappa = 0.3). Poor agreement was evident for ventricular activity (kappa = −0.11).

Overall, Scenarios A and B highlighted strong reliability in certain features, such as understanding and atrial activity, but also signifi ant variability in features like chamber and ventricular activity.

Table 6 summarizes the agreement metrics for baseline and second-month responses in Scenarios C and D. Both scenarios demonstrated 'Fair agreement,' with a Cohen's kappa value of 0.285 for Scenario C and 0.3 for Scenario D. These findings indicate moderate alignment between observers, suggesting no substantial variability across the evaluations.

DISCUSSION

Our study evaluates the performance of ChatGPT-4o in interpreting intracardiac EGMs and provides important findings on how the model can support physicians working with cardiac devices. The accuracy rate in Scenario A was 57%, increasing to 66% in Scenario B when additional contextual information and device parameters were provided. This improvement highlights the role of enriched contextual data in enhancing the model's diagnostic performance. Additionally, the proportion of "No Answer" responses decreased from 19.5% to 8%, indicating the model's increased ability to generate responses when given additional information. However, the rise in "False" responses (from 8.5% to 11%) suggests that the model sometimes misinterprets contextual data, leading to incorrect predictions.

Table 3. Differences in Accuracy Rates Between Scenarios C and D at Baseline and Second Month

	Scenario C Total True Answers (Accuracy Rate)	Scenario D Total True Answers (Accuracy Rate)	*P*
Baseline	6 (0.3)	5 (0.25)	1
Second month	6 (0.3)	10 (0.5)	.125

In Scenarios C and D, diagnostic errors primarily resulted from failures to correlate EGM signals with device algorithm behavior—particularly in identifying pacing inhibition, mode switching, and atrial undersensing—highlighting the model's challenges in temporal and logic-based reasoning.

The model demonstrated high accuracy in specific variables, particularly in pacing mode, which reached 85% in Scenario B, suggesting that contextual support can significantly enhance diagnostic precision. However, in the second-month evaluation, the accuracy rate declined to 60%, indicating potential consistency issues in long-term performance. Similarly, the chamber variable showed a signifi ant improvement from 55% to 90% at baseline, but by the second month, this difference was no longer statistically significant (*P* = .453). These findings suggest that while the model benefits from additional contextual information, its ability to retain and consistently apply this knowledge over time remains limited.

The model's low accuracy in more complex tasks, particularly in Scenarios C and D (30% and 25%, respectively, at baseline), is noteworthy. This decline in performance can be attributed to several factors, including the requirement for direct answer selection and the complexity of the clinical context. Despite the addition of visual EGM data, text-based clinical information, and specific questions, accuracy in Scenario C remained stagnant at 30% (6 correct answers) between the baseline and second-month evaluations. While Scenario D demonstrated an improvement to 50% (10 correct answers) in the second month, this change was not statistically signifi ant (P = .125). The combination of visual and text-based data, along with specific questions and potentially similar answer options, may have increased the cognitive load on the model, making it more challenging to synthesize and prioritize relevant information efficie tly. These findings highlight the need for enhanced multimodal learning strategies that can improve the model's ability to extract critical insights from complex datasets.

Introduction to the case

An 80-year-old woman implanted with a VDD pacemaker for complete AV block had recently undergone a generator replacement. She reported malaise, while carrying her shopping. The device settings are shown in Table 9.1.

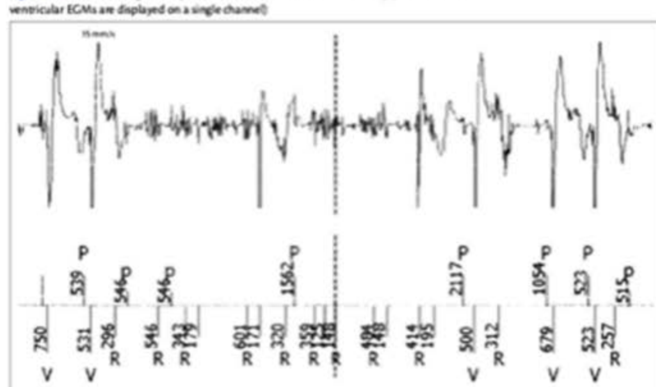
Table 9.1 Device settings

Mode	VDD
Baseline rate	50bpm
UTR	120bpm
Atrial sensing (bipolar)	0.2mV
Ventricular sensing (unipolar)	2.5mV
VP (bipolar)	2.5V at 0.35ms
AV delay (adaptive)	80–125ms
Post-ventricular atrial blanking (PVAB)	150ms

Upon device interrogation, multiple ventricular high-rate episodes were retrieved from the device memory, one of which (lasting 10s) is shown in Figure 9.1.

Question

Figure 9.1 Ventricular high-rate episode retrieved from the device memory (the atrial and ventricular EGMs are displayed on a single channel)



What is the cause of the patient's symptoms?

A Ventricular tachycardia (VT)
B AF
C Ventricular non-capture
D Ventricular oversensing

Clinical Case Summary

An 80-year-old woman with a VDD pacemaker for complete AV block reported malaise during physical activity after generator replacement. Device settings: VDD mode, baseline rate: 50 bpm, upper tracking rate (UTR): 120 bpm, ventricular sensing threshold: 2.5 mV (unipolar), post-ventricular atrial blanking (PVAB): 150 ms.

EGM Analysis

1. **Irregular R-R Intervals**: The tracing shows irregular and rapid ventricular sensing events with R-R intervals as short as 320 ms (187.5 bpm), exceeding the UTR. These irregularities are inconsistent with ventricular tachycardia (VT), which typically produces regular intervals.

2. **Pacing Inhibition**: The absence of pacing spikes indicates that the device inhibited pacing due to ventricular events it sensed as intrinsic activity.

3. **Ventricular Oversensing**: The unipolar ventricular sensing threshold (2.5 mV) increases the likelihood of oversensing far-field signals or noise. Oversensing leads to inappropriate inhibition of ventricular pacing, explaining the patient's symptoms (e.g., bradycardia).

Reasoning for Correct Answer

A. **Ventricular tachycardia (VT)**: Unlikely due to the irregular nature of R-R intervals.
B. **Atrial fibrillation (AF)**: The atrial sensing channel shows no irregularities.
C. **Ventricular non-capture**: There are no pacing spikes, indicating the issue is pacing inhibition, not non-capture.
D. **Ventricular oversensing**: Correct. The irregular ventricular sensing patterns and pacing inhibition are consistent with this diagnosis.

Conclusion

The cause of the patient's symptoms is ventricular oversensing. This results in pacing inhibition and bradycardia, aligning with the clinical scenario and EGM findings.

Figure 3. Example of ChatGPT-assisted EGM interpretation. This figu e illustrates ChatGPT's interpretation of an EGM showing ventricular oversensing, characterized by irregular R-R intervals and pacing inhibition in a patient with a VDD PM.

Table 4. Agreement Metrics Comparing Baseline and Second-month Responses in Scenario A across 10 Key Features

Key Features	McNemar *P*	Cohen's Kappa	PABAK	Kappa Interpretation	PABAK Interpretation
Understanding	N/A	0	0.9	Slight agreement	Almost perfect agreement
Marker annotations	0.479500122	0.459459459	0.8	Moderate agreement	Substantial agreement
Atrial activity	0.248213079	0.7	0.7	Substantial agreement	Substantial agreement
Ventricular activity	0.72367361	−0.25	0.2	Poor agreement	Slight agreement
Chamber	0.386476231	−0.263157895	-0.2	Poor agreement	Poor agreement
Pacing mode	0.130570018	0.0789473684	0.3	Slight agreement	Fair agreement
Timing intervals	0.683091398	0.4	0.4	Fair agreement	Fair agreement
Malfunction	N/A	0	0.8	Slight agreement	Substantial agreement
Pseudomalfunction	0.617075077	0.466666667	0.6	Moderate agreement	Moderate agreement
Synchronizations	0.248213079	0.705882353	0.7	Substantial agreement	Substantial agreement

PABAK, prevalence-adjustment and bias-adjustment kappa.

Table 5. Agreement Metrics Comparing Baseline and Second-Month Responses in Scenario B across 10 Key Features

Key Features	McNemar *P*	Cohen's Kappa	PABAK	Kappa Interpretation	PABAK Interpretation
Understanding	N/A	1	1	Almost perfect agreement	Almost perfect agreement
Marker annotations	0.479500122	0.615384615	0.8	Substantial agreement	Substantial agreement
Atrial activity	0.683091398	0.393939394	0.4	Fair agreement	Fair agreement
Ventricular activity	0.479500122	0.6875	0.8	Substantial agreement	Substantial agreement
Chamber	0.617075077	−0.111111111	0.6	Poor agreement	Moderate agreement
Pacing mode	0.130570018	0.186046512	0.3	Slight agreement	Fair agreement
Timing intervals	1	0.,3	0.3	Fair agreement	Fair agreement
Malfunction	0.479500122	0.444444444	0.8	Moderate agreement	Substantial agreement
Pseudomalfunction	0.37109337	0.305555556	0.5	Fair agreement	Moderate agreement
Synchronizations	0.683091398	0.381443299	0.4	Fair agreement	Fair agreement

PABAK, prevalence-adjustment and bias-adjustment kappa.

In Scenario D, the model exhibited persistent difficulties in EGM interpretation, particularly in timing analysis, sensing detection, and PM algorithm recognition. It frequently misclassified pacing behaviors, such as AV crosstalk, mode switching, and noise reversion, leading to incorrect conclusions about pacing inhibition and atrial tracking. Additionally, sensing issues, particularly ventricular and atrial undersensing, resulted in misdiagnoses of AV block, device malfunction, or loss of capture. The model also struggled with pacing mode classifi ation and threshold determination, affecting its ability to assess capture consistency and pacing behavior. These findings suggest that the model still lacks the ability to correlate programmed device settings with real-time EGM findings and distinguish between similar pacing abnormalities. While its capacity to process EGM data is evident, improving its integration of device-specific algorithms and refining structured decision pathways are essential to optimize its accuracy in complex pacing scenarios. Taken together, these results indicate that despite encouraging performance in simpler settings, the model's current reliability remains insufficie t for clinical decision support—particularly in complex or high-risk pacing scenarios. Further refineme t and device-specific training are likely required for safe and effective clinical deployment.

The reliance on physician-centered approaches for EGM interpretation in PM evaluations poses challenges such as variability in accuracy, time constraints, and potential delays in patient care; however, the adoption of remote device management for CIEDs, has signifi antly improved access and follow-up compliance. Remote monitoring systems are recommended to reduce in-offi e visits and extend follow-up intervals to up to 24 months. However, RM faces several challenges, including staff shortages, organizational inefficiencie , and data overload. PMs generate a high volume of non-urgent alerts, many of which result from false positives due to arrhythmia misclassifi ation, while Implantable Cardioverter Defibrilla ors (ICDs) produce fewer but

Table 6. Agreement Metrics Comparing Baseline and Second-Month Responses in Scenarios C and D

	McNemar *P*	Cohen's Kappa	PABAK	Kappa Interpretation	PABAK Interpretation
Scenario C	.683091398	0.285714286	0.4	Fair agreement	Fair agreement
Scenario D	1	0.30000000000000004	0,3	Fair agreement	Fair agreement

PABAK, prevalence-adjustment and bias-adjustment kappa.

Table 7. Comparative Overview of Recent Artificial I telligence/Large Language Models-Based Electrocardiogram and Intracardiac Electrograms Interpretation Studies

Study	Data Type	Input Modality	Diagnostic Focus	Model Used	Key Outcome
Günay et al, 2024	Surface ECG case texts	Text-only	ECG diagnosis (MCQ-based)	ChatGPT (GPT-4)	90.8% accuracy on 40 ECG case questions
Pang et al, 2023	Surface ECG signals	Signal → Text (via CSSOTP)	Arrhythmia classifi ation	CSSOTP-based LLM	96.2% accuracy on public ECG dataset
Ferreira et al, 2021	CIED reports	OCR + Text	Report-based event and parameter extraction	Custom AI/NLP	Processed 30 CIED reports in under 5 minutes
Current study	Intracardiac EGMs (image) + clinical data	Visual + Textual	Layered EGM interpretation (4 scenarios)	ChatGPT-4o	Accuracy improved from 57% to 66% with context; intra-model agreement assessed

predominantly critical alerts, often caused by oversensing or lead noise.[13-15] These inefficiencies increase the workload for clinicians, highlighting the need for AI-driven solutions that can automate data processing, prioritize alerts, and enhance diagnostic accuracy to improve the efficiency f RM.

Although previous studies have demonstrated the potential of AI and LLMs in cardiac diagnostics, their approaches and focuses vary signifi antly (Table 7). Ferreira et al[6] developed software utilizing optical character recognition (OCR) and natural language processing (NLP) to analyze CIED reports, extracting patient details, device parameters, and event data. While their system processed 30 reports in under 5 minutes, it focused solely on text-based report analysis rather than direct signal interpretation, limiting its diagnostic relevance.[6] Similarly, Pang et al[7] introduced a framework that transforms surface ECG signals into textual patterns for arrhythmia classifi ation, achieving an accuracy of 96.20% on a public dataset. However, their study focused on surface ECGs rather than intracardiac EGMs and did not assess diagnostic reliability in real-world clinical settings.

Günay et al[9] evaluated GPT-4-based ChatGPT for ECG interpretation using 40 multiple-choice questions derived from clinical cases, reporting an impressive 90.8% accuracy. However, since their model relied on textual descriptions rather than direct analysis of ECG images, its applicability in visual signal interpretation remained limited.[8] Another study trained LLMs on ECG-report alignment tasks to improve heart failure (HF) risk prediction from 12-lead ECG data. By correlating ECG signals with corresponding clinical reports, the model effectively identified risk markers, offering promising insights for long-term risk assessment. However, this approach focused on population-level predictions rather than immediate diagnostic applications, making it less relevant to real-time EGM interpretation.[9]

Our study is the fir t to evaluate ChatGPT-4o's performance in interpreting intracardiac EGMs across varying clinical contexts, highlighting its potential as a diagnostic aid. The model demonstrated reasonable accuracy and reliability, particularly in visualized EGM analysis, providing a strong foundation for addressing diagnostic gaps in cardiology. The authors' approach differs from previous studies by integrating both visual and contextual data, assessing the model's adaptability across diverse clinical settings, and identifying critical areas for improvement.

AI-assisted EGM analysis should integrate both visual and structured text-based data to enhance accuracy. The ideal format—raw waveforms vs. structured parameters—remains uncertain, as different manufacturers use distinct sensing and pacing algorithms. Instead of a universal model, fine-tuned models per manufacturer may improve reliability. AI's scope is also crucial—should it analyze full device data or focus on specific tasks like capture loss detection? While full autonomy is a long-term goal, a structured approach incorporating contextual data (e.g., ECGs, patient exams, imaging) may enhance interpretation. A recent study have explored the role of extended passive and active EGM recordings in optimizing device diagnostics, highlighting the importance of systematic data processing for improved detection of anomalies.[16] In future clinical workfl ws, models like ChatGPT-4o could be integrated into remote monitoring systems to assist in alert triage or used as point-of-care diagnostic aids during ambulatory device evaluations. These applications would benefit from further optimization, including manufacturer-specific fine-tuning and structured feedback loops, *to ensure safe and effective deployment*. Even at a basic level, AI-driven anomaly detection can support patient safety and clinical workfl ws.

Our design also has methodological implications. First, the EGM cases were selected sequentially from the EHRA book without randomization or balancing by difficult . Although this approach ensured real-world representativeness, it may have introduced selection bias and limited generalizability. Second, the repetition of the same cases over time was intended solely to assess intra-model decision consistency. Since ChatGPT-4o lacks session memory, this process cannot be interpreted as fine-tuning or learning. These design elements should be refined in future research. This study evaluated the model without fine-tuning, yet it demonstrated self-correction with feedback, highlighting its potential for structured training (Supplementary File 1). With further optimization, AI models could surpass human performance, particularly in electrophysiology and cardiac device troubleshooting, where expertise varies. For LLMs to integrate into clinical workfl ws, continuous learning and real-time clinician feedback are essential. Artificial intelligence-driven adaptive systems can refine performance, enhance RM, and automate data interpretation to reduce diagnostic errors. Future research should focus on fine-tu -ing models with diverse datasets to improve generalizability.

As AI advances, its role in cardiology is expected to transform diagnostics, optimize precision, and improve patient outcomes.

Study Limitations

This study has several limitations that warrant consideration. The model was evaluated on a relatively small dataset of 20 EGM cases which may not fully capture the variability and complexity of real-world clinical scenarios. Moreover, it was not tested in real-time clinical settings, where factors such as noise, incomplete data, and variations across device types could signifi antly impact performance. The dataset also lacked representation of all PM brands and models, limiting the model's generalizability to a wider range of devices. Furthermore, the absence of direct comparison with human experts restricts the ability to benchmark its diagnostic utility effectively. Another limitation is the sequential selection of EGM cases without randomization or difficulty stratifi ation, which may affect generalizability. Repeated testing assessed decision stability, not learning, as ChatGPT-4o lacks memory. Finally, while AI tools aided in language editing, all scientifi content and interpretations were author-generated.

CONCLUSION

This study represents the fir t systematic evaluation of ChatGPT-4o's ability to interpret intracardiac EGMs, demonstrating promising potential as a diagnostic support tool. The model exhibited reasonable accuracy and reliability, particularly in scenarios involving visualized EGM data. However, further optimization and large-scale validation are necessary to fully integrate LLM-based models into clinical practice.

Ethics Committee Approval: Ethical approval was not required for this study, as it involved no human participants or patient data. The EGM cases used for analysis were obtained from a publicly available textbook (The EHRA Book of Pacemaker, ICD, and CRT Troubleshooting), consisting of anonymized, educational material. The study involved evaluating the interpretive performance of a large language model (ChatGPT-4o) on these standard cases, without generating new clinical data.

Informed Consent: Verbal/Written informed consent was not required, as no human participants or patient data were involved in this study. The electrogram (EGM) cases used were anonymized, educational examples obtained from a publicly available textbook (The EHRA Book of Pacemaker, ICD, and CRT Troubleshooting).

Peer-review: Externally peer reviewed.

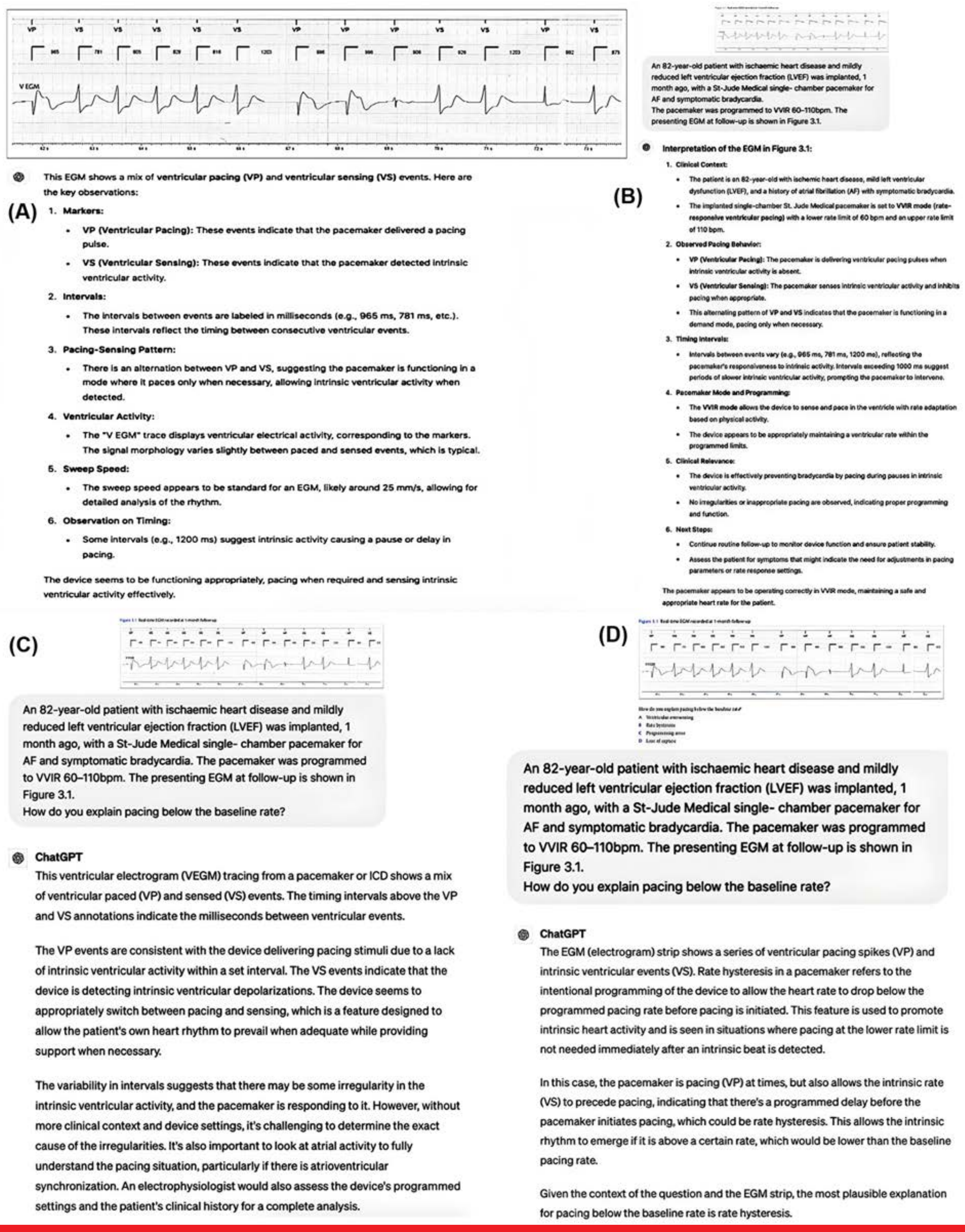
Author Contributions: S.B.: Conception, Design, Supervision, Writer, Materials, Data Collection, and/or Processing; A.B.D.: Resource, Data Collection, and/or Processing; N.D.: Resource, Literature Review; Abdülcebbar Şipal: Resource, and Literature Review; Ş.E.Ö.: Materials, Analysis, and/or Interpretation; M.Ç.: Design, Data Collection, and/or Processing; T.E.G.: Design, Analysis, and/or Interpretation; T.A.: Critical Review; U.B.: Conception and Critical Review; N.K.: Conception and Critical Review.

Declaration of Interests: The authors have no conflicts of interest to declare.

Funding: The authors declare that this study received no financial support.

REFERENCES

1. O'Shea CJ, Middeldorp ME, Hendriks JM, et al. Remote monitoring alert burden: an analysis of transmission in >26,000 patients. *JACC Clin Electrophysiol*. 2021;7(2):226-234. [CrossRef]
2. Hess PL, Mi X, Curtis LH, Wilkoff BL, Hegland DD, Al-Khatib SM. Follow-up of patients with new cardiovascular implantable electronic devices: is adherence to the experts' recommendations associated with improved outcomes? *Heart Rhythm*. 2013;10(8):1127-1133. [CrossRef]
3. Çay S, Duygu KB, Bozyel S. Remote follow-up/monitoring of cardiac implantable electronic devices. *Turk Kardiyol Dern Ars*. Baskıdaki Makaleler: TKDA-34561. 2025;53(1):53-61. [CrossRef]
4. Varma N, Epstein AE, Irimpen A, Schweikert R, Love C, TRUST Investigators. Effi acy and safety of automatic remote monitoring for implantable cardioverter-defibrilla or follow-up: the Lumos-T Safely Reduces Routine Offi e Device Follow-up (TRUST) trial. *Circulation*. 2010;122(4):325-332. [CrossRef]
5. Crossley GH, Boyle A, Vitense H, Chang Y, Mead RH. Connect Investigators. The CONNECT (Clinical Evaluation of Remote Notifi ation to Reduce Time to Clinical Decision) trial: the value of wireless remote monitoring with automatic clinician alerts. *J Am Coll Cardiol*. 2011;57(10):1181-1189. [CrossRef]
6. Ferreira HDS, Ferrer F, Cabanelas N, et al. Artificial intelligence: an attempt to automate remote device follow-up. *EP Europace*. 2021;23(Suppl 3):euab116.517. [CrossRef]
7. Pang X, Zhang Y, Li J, Wang Z. Innovative application of large language models for ECG analysis: a CSSOTP-based feature extraction approach. *J Biomed Signal Process*. 2023;12(3):123-135.
8. Chen C, et al. Large Language Model-Informed ECG Dual Attention Network for Heart Failure Risk Prediction. arXiv Preprint ArXiv:2403.10581. 2024.
9. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and Chat-GPT in electrocardiography assessment. *Am J Emerg Med*. 2024;80:51-60. [CrossRef]
10. Burri H, Israel C, Deharo JC, eds. *The EHRA Book of Pacemaker, ICD, and CRT Troubleshooting: Case-Based Learning with Multiple Choice Questions*. 1st ed. Oxford University Press. 2015. ISBN: 9780198727774.
11. Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol*. 2009;9:5. [CrossRef]
12. Tan KS, Yeh YC, Adusumilli PS, Travis WD. Quantifying interrater agreement and reliability between thoracic pathologists: paradoxical behavior of Cohen's kappa in the presence of a high prevalence of the histopathologic feature in lung cancer. *JTO Clin Res Rep*. 2024;5(1):100618. [CrossRef]
13. Glikson M, Nielsen JC, Kronborg MB, et al. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy. *Eur Heart J*. 2021;42(35):3427-3520. [CrossRef]
14. Ferrick AM, Raj SR, Deneke T, et al. 2023 HRS/EHRA/APHRS/LAHRS expert consensus statement on practical management of the remote device clinic. *Heart Rhythm*. 2023;20(9):e92-e144. [CrossRef]
15. Hindricks G, Varma N, Kacet S, et al. Daily remote monitoring of implantable cardioverter-defibrilla ors: insights from the pooled patient-level data from three randomized controlled trials (IN-TIME, ECOST, TRUST). *Eur Heart J*. 2017;38(22):1749-1755. [CrossRef]
16. Quinonero C, Strik M, Catalan PA, et al. Improving diagnostic yield for analyzing periodic electrograms in the remote detection of pacemaker lead issues. *Sensors (Basel)*. 2025;25(3):656. [CrossRef]

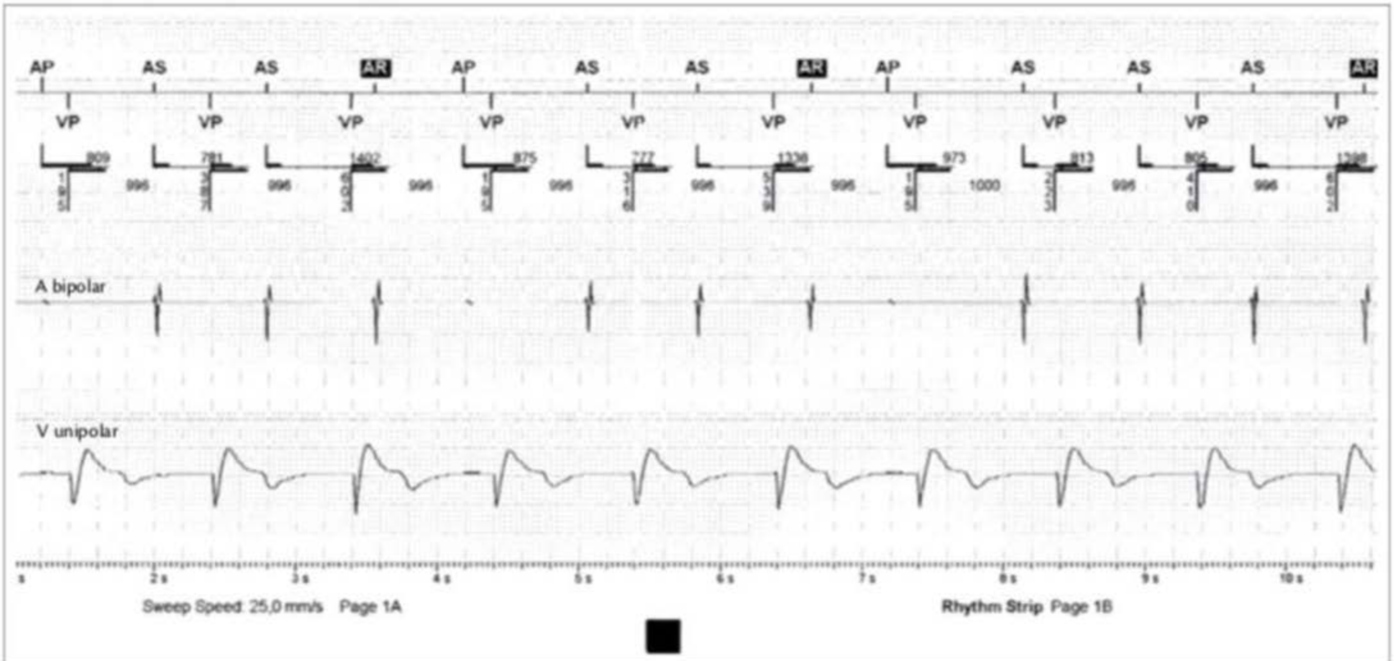


Supplementary Figure 1 . ChatGPT's Analysis of Pacemaker EGMs Across Scenarios A to D: Interpretation and Observations (Scenario A: Isolated EGM interpretation, Scenario B: EGM with clinical context, Scenario C: EGM with clinical context and targeted question, Scenario D: EGM with clinical context, targeted question, and multiple-choice options).

Supplementary File 1. Egm Interpretation: Example Of Model Self-Correction And Structured Learning Potential.

Question

Figure 1.1 Real-time EGM recording



In what pacing mode is the device programmed?

A VDI

B DDI

C VDD

D DDD

EGM ERROR ANALYSIS REPORT

Introduction

This report outlines a sequence of events in which an initial incorrect response was provided during the analysis of a pacemaker EGM tracing. The error was subsequently identified and corrected after a detailed review. The process highlights the importance of recognizing atrioventricular (AV) relationships in interpreting pacemaker modes and provides actionable insights for improvement.

Case Summary

A question was presented regarding the pacing mode demonstrated in an EGM tracing. The options provided were:

A. VDI
B. DDI
C. VDD
D. DDD

The correct answer was B. DDI, as the tracing indicated dual-chamber sensing with inhibition but without tracking atrial-sensed events (P-synchronous pacing). Initially, however, the response identified the pacing mode as D. DDD, which was incorrect.

ERROR ANALYSIS

Upon review, the following factors contributed to the initial incorrect response:

1. Misinterpretation of AV Relationship

The initial analysis incorrectly assumed the presence of P-wave tracking, a hallmark of the DDD mode. However, the tracing demonstrated that ventricular pacing was independent of atrial-sensed events, indicating the absence of P-synchronous pacing, a defining eature of the DDI mode.

2. Bias Toward Commonly Used Modes

The initial response was influen ed by the fact that DDD is the most commonly used dual-chamber pacing mode. This bias led to an assumption rather than a thorough evaluation of the specific t acing provided.

3. Overlooking Key Details in the Tracing
Critical features of the tracing, such as the consistent pacing rate determined by the lower rate limit (LRL) and the lack of atrial-sensed event tracking, were not adequately considered.

Correction and Insights

After being prompted to re-evaluate the tracing, a detailed review was conducted, leading to the identifi ation of the following key observations:

1. Absence of P-Synchronous Pacing
The ventricular pacing rate remained fi ed at the LRL, irrespective of atrial-sensed events, confirming that P-wave tracking was not present.

2. Inhibition Without Tracking
The tracing demonstrated that ventricular pacing was inhibited by intrinsic ventricular events but not triggered by atrial-sensed events, consistent with the DDI mode.

3. Clinical Context Alignment
The DDI mode is particularly suitable for patients with paroxysmal atrial arrhythmias, such as atrial fibrillation, where tracking atrial events could exacerbate ventricular pacing.

These observations confirmed that the correct answer was B. DDI, aligning with the clinical scenario and the EGM tracing provided.

Lessons Learned

This process underscored the following key takeaways:

1. Systematic Analysis of AV Relationships
Accurate determination of pacing modes requires careful evaluation of the relationship between atrial and ventricular events, particularly the presence or absence of P-wave tracking.

2. Avoiding Assumptions
Commonly used pacing modes should not be assumed without detailed verifi ation of the EGM data.

3. Structured Approach to EGM Interpretation
A systematic methodology, including the identifi ation of sensing and pacing patterns, AV delays, and response to intrinsic events, is critical for accurate analysis.

CONCLUSION

The error in the initial response and its subsequent correction highlight the importance of meticulous analysis and continuous learning in the interpretation of pacemaker EGM tracings. This experience serves as a valuable example for refining analysis skills and enhancing diagnostic accuracy in similar scenarios.