American Journal of Emergency Medicine 95 (2025) 220–226

Contents lists available at ScienceDirect

American Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/ajem

Task-specific versus general-purpose AI models in ECG analysis: A comparative study with emergency medicine specialists

Ertugrul Altinbilek [a,*], Adem Az, MD [b], Ozgur Sogut [b], Yunus Dogan, MD [b], Tarik Akdemir, MD [b], Erdal Belen [c], Halil Ibrahim Biter [c], Tugay Saricicek, MD [b], Mehmet Ozcomlekci, MD [b], Nurbaki Kilic, MD [b]

[a] University of Health Sciences, Şişli Hamidiye Etfal Training and Research Hospital, Department of Emergency Medicine, Istanbul, Turkey
[b] University of Health Sciences, Haseki Training and Research Hospital, Department of Emergency Medicine, Istanbul, Turkey
[c] University of Health Sciences, Haseki Training and Research Hospital, Department of Cardiology, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 8 May 2025
Received in revised form 26 June 2025
Accepted 27 June 2025

Keywords:
Electrocardiogram
Artificial intelligence
Emergency medicine
Diagnostic accuracy
Domain-specific AI

ABSTRACT

Purpose: To evaluate and compare the diagnostic accuracy of three Artificial intelligence (AI) models—GPT-4o, Canva-GPT, and ECG Reader-GPT—against emergency medicine specialists (EMSs) in electrocardiogram (ECG) interpretation using a standardized and validated test set.

Methods: In this prospective diagnostic accuracy study, 50 ECG questions were selected from the reference text 150 ECG Cases. Thirty EMSs completed the test once; each AI model was evaluated on the same test set daily over 30 consecutive days. Diagnostic accuracy was compared across predefined ECG subcategories and clinical case types.

Results: EMSs achieved the highest overall diagnostic accuracy (median: 41.5; IQR: 37.0–43.0), followed closely by ECG Reader-GPT (median: 39.5; IQR: 39.0–41.0), with no statistically significant difference between them ($p = 0.530$). ECG Reader-GPT significantly outperformed both GPT-4o and Canva-GPT across all case categories ($p < 0.001$). Subgroup analysis revealed that ECG Reader-GPT performed comparably to EMSs in identifying ischemic syndromes, channelopathies and genetic syndromes, and normal ECGs (all $p < 0.05$); it surpassed them in interpreting rhythm disorders ($p = 0.007$).

Conclusion: ECG Reader-GPT, a customized AI model for ECG interpretation, demonstrated diagnostic accuracy comparable to experienced EMSs and significantly outperformed general-purpose models across all ECG subcategories. These findings highlight the value of domain specialization in developing clinically effective AI tools.
© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

The integration of open-source artificial intelligence (AI) models into clinical practice is rapidly expanding, offering physicians enhanced support in diagnosis, treatment planning, and patient communication. AI-based clinical decision support systems have shown promising results in improving diagnostic accuracy and operational efficiency, particularly in fields such as radiology, pathology, and dermatology [1]. More recently, large language models (LLMs), such as ChatGPT, have been applied to tasks involving diagnostic reasoning, clinical summarization, and physician–patient communication [2,3].

Despite the increasing adoption of LLMs in healthcare, their application in emergency settings remains largely unvalidated, particularly for image-based diagnostic tasks such as electrocardiogram (ECG) analysis. Existing evidence is methodologically limited by small sample sizes, narrow diagnostic spectrums, a lack of domain-specific test cases, and limited external validation under real-world conditions [4-6]. Günay et al. reported that GPT-4 outperformed both emergency medicine specialists (EMSs) and cardiologists in answering ECG questions derived from a standardized textbook [4]. However, a subsequent study evaluating GPT-4o, an updated multimodal version, found that it underperformed relative to both EMSs and cardiologists when tested on the same dataset [5]. These discrepancies can be partially attributed to differences in test design and limitations in sample diversity, but more importantly, to the use of different input modalities.

Conventional LLMs, such as ChatGPT, are designed for text-based tasks involving language comprehension, generation, and reasoning, but they lack the intrinsic ability to process visual data [7]. In contrast, vision–language models (VLMs), such as GPT-4o [8] and Canva-GPT [9], are capable of integrating visual and textual inputs, thereby

* Corresponding author.
E-mail addresses: ertugrulaltinbilek@gmail.com (E. Altinbilek), ozgur.sogut@sbu.edu.tr (O. Sogut).

https://doi.org/10.1016/j.ajem.2025.06.068
0735-6757/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

expanding the potential scope of AI applications to image-based diagnostic tasks. Canva-GPT, for example, has demonstrated competence in interpreting visual content and may thus hold promise for ECG analysis. Nevertheless, neither GPT-4o nor Canva-GPT was specifically trained for ECG interpretation. In contrast, customized GPTs created through OpenAI's platform—such as ECG Reader-GPT [10]—have emerged for ECG analysis. These models are tailored versions of ChatGPT, configured through specific instructions and potentially curated datasets, but without additional supervised model training or architectural modifications. Recent commentaries have emphasized the importance of model specialization and domain-specific training to ensure diagnostic reliability and clinical applicability [11].

This study evaluated the diagnostic accuracy of three AI models —GPT-4o, Canva-GPT, and ECG Reader-GPT—in ECG analysis and compared their performance with that of EMSs using a standardized and validated dataset. We tested the hypothesis that customized AI models for ECG interpretation can achieve diagnostic accuracy levels comparable to those of human experts in ECG interpretation.

2. Methods

2.1. Study design and setting

This single-center, prospective, observational, cross-sectional diagnostic accuracy study was conducted in accordance with the 2024 Declaration of Helsinki. The study protocol was approved by the Institutional Review Board of Haseki Training and Research Hospital, Istanbul, Turkey (approval no. 30–2025). Written informed consent was obtained from all participants prior to enrollment.

Fifty multiple-choice ECG questions were selected and adapted from the reference textbook 150 ECG Cases [12]. Each question included one correct answer and four distractors, and it was categorized into one of six diagnostic groups: rhythm disorders ($n = 12$), ischemic syndromes (n = 12), conduction disturbances (n = 12), channelopathies and genetic syndromes ($n = 5$), normal ECGs (n = 5), and other cardiac pathologies ($n = 4$). Question selection and validation were performed by two senior experts—one professor of emergency medicine and one professor of cardiology. Based on expert consensus, 30 cases were classified as routine ECG presentations commonly encountered in daily clinical practice, whereas the remaining 20 were considered diagnostically challenging. Additionally, 30 ECGs were categorized as potentially life-threatening and 20 as non-life-threatening but clinically significant.

Life-threatening ECG cases were defined by expert consensus and aligned with current emergency cardiology guidelines. These included ST-elevation myocardial infarction (STEMI) or equivalent ischemic syndromes ($n = 10$), malignant arrhythmias (e.g., ventricular tachycardia or fibrillation) ($n = 8$), high-grade atrioventricular blocks requiring urgent intervention (n = 8), high-risk channelopathies and genetic syndromes ($n = 5$), and acute pulmonary embolism (n = 1). The remaining 20 ECGs, although not immediately life-threatening, were still considered clinically relevant.

2.2. Data collection

For human participants, the test was administered using a secure, web-based interface (Google Forms); all questions were presented in each participant's native language. For AI models, the same test set was delivered in English, based on pilot testing that demonstrated improved accuracy and consistency with English-language prompts.

A pilot trial was conducted prior to the main study to determine the optimal method of presenting ECG questions to the AI models. When the questions were presented in a single batch, performance was suboptimal; however, presenting each question individually resulted in higher diagnostic accuracy. Consequently, in the main study, each question was delivered separately to the AI models to maximize performance.

All AI models received the same inputs through a standardized protocol. ECG images were directly uploaded to the model interfaces, along with accompanying patient history, clinical context, and multiple-choice answer options. Fig. 1 illustrates an example of the question format presented to the AI models. Prompts were kept identical across all models, and no model-specific tuning or optimization was applied.

To ensure performance stability and reproducibility, each AI model completed the 50-question test once daily for 30 consecutive days; the question order was randomized to minimize learning effects. In contrast, EMSs completed the test once, simulating a typical real-world clinical evaluation. All responses were independently assessed by a researcher who was blinded to both group assignment and the study hypothesis.

2.3. Study groups and participants

The study included 30 EMSs certified by the Turkish Ministry of Health, each with 5–10 years of clinical experience. All participants were informed of the nature, objectives, and procedures of the study prior to participation.

Three GPT-based AI models were compared with EMSs. The first was GPT-4o, a general-purpose VLM [8]. The second was Canva-GPT, optimized for visual input processing [9]. The third was ECG Reader-GPT, a customized GPT configured via OpenAI's custom GPT platform for ECG interpretation. Although its task-specific configuration suggests enhanced performance in this domain, it represents a tailored instance of ChatGPT rather than a separately trained domain-specific model. Detailed information regarding its customization methodology and any additional datasets used has not been publicly disclosed [10].

2.4. Clinical outcomes

The primary outcome of the study was the diagnostic accuracy of each AI model and the EMS group in analyzing 50 standardized ECG cases. Accuracy was defined as the proportion of correctly answered questions.

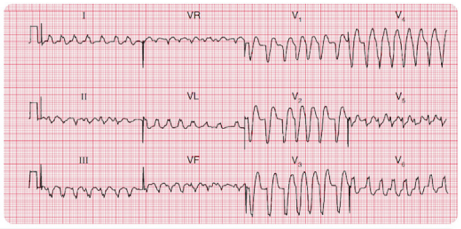
Secondary outcomes included subgroup comparisons based on diagnostic complexity (routine vs. challenging ECGs), clinical severity (life-threatening vs. non-life-threatening cases), and diagnostic subcategories (e.g., rhythm disorders, ischemic syndromes, conduction disturbances).

2.5. Statistical analyses

An a priori power analysis using G*Power ($\alpha = 0.05$, power = 0.99, effect size = 1.75) indicated that a minimum of 12 participants in the EMS group and 12 repeated trials per AI model (36 in total) were sufficient to detect statistically significant differences across the four groups [7].

Data were analyzed using SPSS software (version 28.0 for Windows; IBM Corp., Armonk, NY, USA). Descriptive statistics for numerical variables were presented as median and interquartile range (IQR), and minimum–maximum values. The distribution of variables was assessed using the Kolmogorov–Smirnov test. Given that the data did not follow a normal distribution, non-parametric statistical methods were used.

Comparisons of dependent quantitative variables among the four study groups (EMSs, GPT-4o, Canva-GPT, and ECG Reader-GPT) were conducted using the Friedman test. Post hoc pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction to control for the risk of type I error due to multiple comparisons. Additionally, effect sizes (r) were calculated for all pairwise comparisons. A p-value of <0.05 was considered statistically significant.



An 80-year-old man presented to the hospital with sudden-onset palpitations and shortness of breath. A cardiac murmur suggestive of congestive heart failure and aortic regurgitation was noted. Based on the ECG obtained, what is your most likely diagnosis for this patient?

a) Ventricular tachycardia
b) Wolff-Parkinson-White syndrome
c) Atrial fibrillation + right bundle branch block
d) Torsades de pointes
e) Atrial fibrillation + left bundle branch block

ECG Interpretation (Clinical Level)

The ECG shows the following key features:

- **Irregularly irregular rhythm**: There is no consistent P wave before each QRS complex, and the RR intervals are highly variable.
- **Wide QRS complexes**: The QRS duration is prolonged (>120 ms), and the complexes exhibit a broad, notched morphology.
- **Absence of P waves**: P waves are not discernible, consistent with atrial fibrillation.
- **Bundle branch block morphology**: The QRS complexes are wide and show a **"M"-shaped** or **broad monophasic R waves** in leads V5–V6, and **deep, wide S waves** in V1—these findings are characteristic of a **left bundle branch block (LBBB)**.

Most Likely Diagnosis:

e) Atrial fibrillation + left bundle branch block

Fig. 1. Sample ECG interpretation by ECG Reader-GPT.
Caption: This figure shows an ECG interpretation by ECG Reader-GPT in an 80-year-old man presenting with acute palpitations and dyspnea. The model accurately identified atrial fibrillation with left bundle branch block (LBBB), demonstrating recognition of key features including an irregular rhythm, absence of P waves, and broad, notched QRS complexes.

3. Results

Table 1 presents the diagnostic accuracy of EMSs and the three AI models—GPT-4o, Canva-GPT, and ECG Reader-GPT—in interpreting all ECG cases. EMSs achieved the highest overall diagnostic accuracy (median: 41.5; IQR: 37.0–43.0), significantly outperforming GPT-4o (median: 26.0; IQR: 23.0–27.0) and Canva-GPT (median: 29.0; IQR: 27.0–30.0) (both $p < 0.001$; $r = 0.593$). ECG Reader-GPT (median: 39.5; IQR: 39.0–41.0) also demonstrated significantly higher accuracy than both general-purpose models (both $p < 0.001$; $r = 0.593$) but did not significantly differ from EMSs ($p = 0.530$; $r = 0.096$).

In the subset of 30 routine ECG cases, EMSs again achieved the highest accuracy (median: 29.0; IQR: 26.5–30.0), followed by ECG Reader-GPT (median: 23.0; IQR: 22.0–25.0), Canva-GPT (median: 16.5; IQR: 15.0–17.0), and GPT-4o (median: 15.0; IQR: 13.8–16.0). All pairwise comparisons between groups showed statistically significant differences ($p < 0.01$; $r = 0.710$). For the subset of 20 diagnostically challenging ECG cases, ECG Reader-GPT outperformed all other groups (median: 16.0; IQR: 16.0–17.0), significantly exceeding the performance of EMSs (median: 12.0; IQR: 10.8–14.0), Canva-GPT (median: 12.0; IQR: 11.8–13.3), and GPT-4o (median: 10.0; IQR: 8.0–11.0) ($p < 0.01$; $r = 0.826$). The performance of EMSs and Canva-GPT was comparable ($p = 0.914$; $r = 0.029$), although both significantly outperformed GPT-4o ($p < 0.01$; $r = 0.945$).

In the subset of 30 life-threatening ECG cases, EMSs achieved a higher accuracy (median: 24.0; IQR: 21.0–25.3) than GPT-4o (median: 18.0; IQR: 16.0–19.0) and Canva-GPT (median: 19.0; IQR: 18.0–20.0) ($p < 0.001$; $r = 0.774$); they demonstrated performance comparable to that of ECG Reader-GPT (median: 25.0; IQR: 23.0–26.0) ($p = 0.071$; $r = 0.359$). In the subset of 20 non-life-threatening but clinically significant ECG cases, EMSs again achieved the highest accuracy (median: 17.0; IQR: 16.0–18.0), followed by ECG Reader-GPT (median: 15.0; IQR: 13.0–17.0), Canva-GPT (median: 9.0; IQR: 9.0–10.0), and GPT-4o (median: 8.0; IQR: 6.0–9.0); all intergroup differences were statistically significant ($p < 0.01$; $r = 0.728$–0.917).

Table 2 and Fig. 2 present a subgroup analysis based on clinical diagnostic categories. For the interpretation of 5 normal ECGs, both ECG Reader-GPT (median: 4.0; IQR: 3.0–4.0) and EMSs (median: 3.5; IQR: 3.0–4.0) significantly outperformed the general-purpose AI models ($p < 0.001$; $r = 1.945$). Regarding 12 cases of conduction disturbances,

Table 1
Comparison of diagnostic accuracy of AI models and emergency medicine specialists across ECG case types and clinical scenarios.[a,b,c]

Characteristic		GPT-4o	Canva-GPT	ECG Reader-GPT	EMS	p
Case difficulty						
Routine ECG cases	Median	15.0 [a*,bc***]	16.5 [bc***]	23.0 [c***]	29.0	< 0.001
(*n* = 30)	IQR (25–75)	13.8–16.0	15,0–17,0	22.0–25.0	26.5–30.0	
	Min. – max.	12.0–19.0	14.0–19.0	21.0–28.0	25.0–30.0	
More challenging ECG cases	Median	10.0 [ab***,c**]	12.0 [b***]	16.0	12.0 [b***]	< 0.001
(*n* = 20)	IQR (25–75)	8.0–11.0	11.8–13.3	16.0–17.0	10.8–14.0	
	Min. – max.	6.0–14.0	9.0–14.0	14.0–19.0	6.0–17.0	
Total ECG cases	Median	26.0 [abc***]	29.0 [bc***]	39.5	41.5	< 0.001
(*n* = 50)	IQR (25–75)	23.0–27.0	27.0–30.0	39.0–41.0	37.0–43.0	
	Min. – max.	20.0–31.0	24.0–33.0	35.0–45.0	33.0–46.0	
Clinical scenarios						
Life-threatening ECG cases	Median	18.0 [abc***]	19.0 [bc***]	25.0	24.0	< 0.001
(n = 30)	IQR (25–75)	16.0–19.0	18.0–20.0	23.0–26.0	21.0–25.3	
	Min. – max.	13.0–21.0	15.0–23.0	22.0–28.0	17.0–27.0	
Non-life-threatening ECG cases	Median	8.0 [abc***]	9.0 [bc***]	15.0 [c***]	17.0	< 0.001
(n = 20)	IQR (25–75)	6.0–9.0	9.0–10.0	13.0–17.0	16.0–18.0	
	Min. – max.	5.0–11.0	7.0–11.0	13.0–19.0	12.0–19.0	

Note: Data are presented as median and interquartile range (IQR), minimum (min), and maximum (max).
* Intergroup comparisons were conducted using the Friedman test; post hoc pairwise comparisons were performed using the Wilcoxon signed-rank test.
*$p \le 0.05$, **$p \le 0.01$, and ***$p \le 0.001$.
Abbreviations: EMS, emergency medicine specialist; ECG, electrocardiography; GPT-4o, general-purpose vision–language AI model; Canva-GPT, multimodal AI model optimized for visual input; ECG Reader-GPT, customized AI models for ECG interpretation.
[a] Significant differences vs. Canva-GPT.
[b] vs. ECG Reader-GPT.
[c] vs. EMSs.

EMSs achieved the highest diagnostic accuracy (median: 10.0; IQR: 9.0–11.0), followed by ECG Reader-GPT (median: 7.5; IQR: 7.0–9.0), Canva-GPT (median: 7.0; IQR: 6.0–8.0), and GPT-4o (median: 5.0; IQR: 4.0–6.0); all pairwise differences were statistically significant ($p < 0.05$; $r = 0.636$–1.173).

In 12 cases of ischemic syndromes, ECG Reader-GPT (median: 10.0; IQR: 10.0–11.0) and EMSs (median: 10.0; IQR: 9.0–11.0) again significantly outperformed both Canva-GPT and GPT-4o ($p < 0.001$; $r = 1.173$); there was no significant difference between ECG Reader-GPT and EMSs ($p = 0.313$; $r = 0.304$). A similar pattern was observed in 5 cases of channelopathies and genetic syndromes, such that ECG Reader-GPT (median: 4.0; IQR: 3.0–4.0) and EMSs (median: 5.0; IQR: 3.0–5.0) performed significantly better than both general-purpose models ($p < 0.001$; $r = 1.740$). Concerning 12 cases of rhythm disorders, ECG Reader-GPT achieved the highest diagnostic accuracy (median: 10.0; IQR: 10.0–11.0), outperforming all other groups, including EMSs (median: 9.0; IQR: 9.0–10.0); these differences were statistically significant ($p < 0.01$; $r = 0.853$–1.230).

4. Discussion

The role of AI in critical care diagnostics—particularly its accuracy and applicability—remains an area of intense interest and active research. This study adds to the growing body of evidence by systematically comparing the diagnostic performance of general-purpose versus customized AI models for ECG interpretation. Our findings demonstrate

Table 2
Comparison of diagnostic accuracy of AI models and emergency medicine specialists across ECG subcategories.[a,b,c]

Characteristic		GPT-4o	Canva-GPT	ECG Reader-GPT	EMS	p*
Rhythm disorders	Median	8.0 [abc***]	9.0 [b***,c**]	10.0	9.0 [b*]	< 0.001
(n = 12)	IQR (25–75)	6.8–9.0	8.0–9.0	10.0–11.0	9.0–10.0	
	Min. – max.	5.0–10.0	7.0–10.0	8.0–12.0	6.0–11.0	
Ischemic syndromes	Median	7.0 [bc***]	6.0 [bc***]	10.0	10.0	< 0.001
(n = 12)	IQR (25–75)	5.0–8.0	6.0–7.0	10.0–11.0	9.0–11.0	
	Min. – max.	4.0–9.0	3.0–9.0	8.0–12.0	6.0–12.0	
Conduction disturbances	Median	5.0 [abc***]	7.0 [b*,c***]	7.5 [c***]	10.0	< 0.001
(n = 12)	IQR (25–75)	4.0–6.0	6.0–8.0	7.0–9.0	9.0–11.0	
	Min. – max.	2.0–10.0	4.0–10.0	5.0–12.0	6.0–12.0	
Channelopathies and genetic syndromes (*n* = 5)	Median	3.0 [bc***]	3.0 [bc***]	4.0	5.0	< 0.001
	IQR (25–75)	2.0–3.0	2.8–3.0	3.0–4.0	3.0–5.0	
	Min. – max.	1.0–4.0	1.0–4.0	2.0–5.0	3.0–5.0	
Normal ECGs	Median	1.0 [bc***]	1.0 [bc***]	4.0	3.5	< 0.001
(n = 5)	IQR (25–75)	0.0–1.0	1.0–1.0	3.0–4.0	3.0–4.0	
	Min. – max.	0.0–3.0	0.0–2.0	2.0–5.0	2.0–5.0	
Other cardiac pathologies	Median	2.0 [abc***]	3.0 [bc***]	4.0	3.0 [c***]	< 0.001
(n = 4)	IQR (25–75)	1.0–3.0	3.0–3.0	4.0–4.0	3.0–4.0	
	Min. – max.	1.0–4.0	2.0–3.0	3.0–4.0	2.0–4.0	

Note: Data are presented as median and interquartile range (IQR), minimum (min), and maximum (max).
* Intergroup comparisons were conducted using the Friedman test; post hoc pairwise comparisons were performed using the Wilcoxon signed-rank test.
*$p \le 0.05$, **$p \le 0.01$, and ***$p \le 0.001$.
Abbreviations: EMS, emergency medicine specialist; ECG, electrocardiography; GPT-4o, general-purpose vision–language AI model; Canva-GPT, multimodal AI model optimized for visual input; ECG Reader-GPT, customized AI models for ECG interpretation.
[a] Significant differences vs. Canva-GPT.
[b] vs. ECG Reader-GPT.
[c] vs. EMSs.

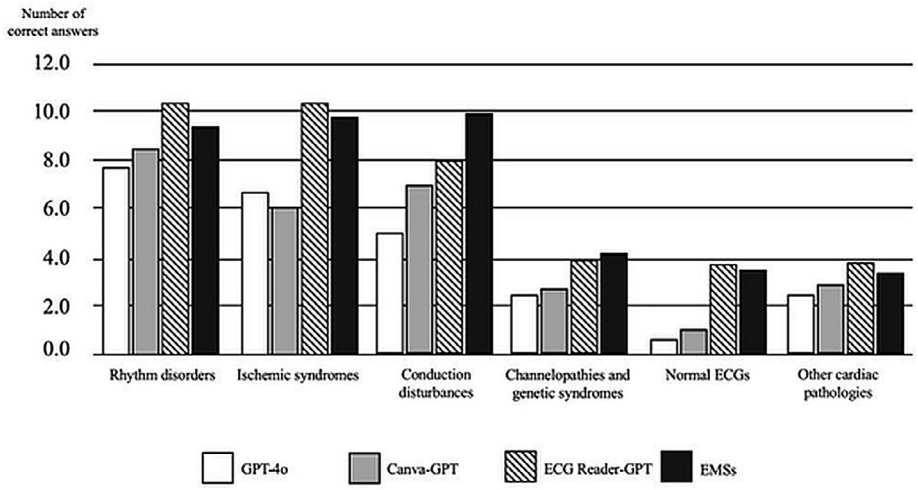


Fig. 2. Diagnostic accuracy of AI models and emergency medicine specialists across ECG subcategories.
Caption: ECG Reader-GPT consistently outperformed GPT-4o and Canva-GPT across all subcategories and demonstrated comparable or superior performance to EMSs in rhythm disorders, ischemic syndromes, channelopathies and genetic syndromes, and other cardiac pathologies.
Abbreviations: EMS, emergency medicine specialist; ECG, electrocardiography; GPT-4o, general-purpose vision–language AI model; Canva-GPT, multimodal AI model optimized for visual input; ECG Reader-GPT, customized AI models for ECG interpretation.
The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see: http://www.textcheck.com/certificate/gB03f8.

that customized AI models for ECG interpretation, particularly ECG Reader-GPT, significantly outperform general-purpose VLMs in ECG analysis. ECG Reader-GPT exhibited diagnostic accuracy comparable to, and in some cases exceeding, that of EMSs, especially in complex and life-threatening ECG cases. These results highlight its potential utility in time-sensitive, high-risk clinical environments.

Previous studies have assessed the diagnostic capabilities of general-purpose LLMs in ECG analysis, often in comparison to human experts. Günay et al. reported that GPT-4 outperformed both EMSs and cardiologists in answering multiple-choice ECG questions [4]. However, these cases relied solely on textual descriptions, inherently favoring LLMs' strengths in language processing. In a follow-up study, the same authors evaluated GPT-4o using visual ECG inputs and found that its performance declined, falling below that of EMSs and cardiologists in routine cases [5]. These contrasting outcomes highlight the importance of input modality and task specificity in evaluating AI models. Although LLMs may excel in text-based tasks, they face challenges when required to process and interpret visual clinical data [13].

Supporting this observation, Zhu et al. found that ChatGPT-4 V performed moderately well on multiple-choice ECG questions but struggled with open-ended tasks requiring visual waveform interpretation and quantitative measurement [14]. Similarly, a diagnostic accuracy study of GPT-4 V across various visual clinical inputs revealed a marked performance decline relative to text-based tasks, emphasizing its limited capacity to generalize from linguistic to image-based reasoning [15]. In agreement with these findings, our study showed that the general-purpose VLMs, GPT-4o and Canva-GPT, performed significantly worse than EMSs in visually based ECG analysis. Specifically, GPT-4o achieved a diagnostic accuracy of approximately 50 %, whereas Canva-GPT reached 58 %. These results likely reflect both the absence of ECG-specific training and the broader limitations of general-purpose models in image-driven diagnostic contexts.

In contrast, customized AI models for ECG interpretation, such as ECG Reader-GPT, demonstrated obviously superior diagnostic accuracy. Whereas GPT-4o and Canva-GPT attained accuracies of roughly 50 % and 58 %, respectively, ECG Reader-GPT accurately interpreted nearly 80 % of the ECG cases. These findings highlight the importance of task-specific training and model specialization in developing clinically effective medical AI. Consistent with our results, Chang et al. reported that ECG-focused machine learning models outperformed general-purpose counterparts in diagnostic accuracy and clinical applicability within cardiology tasks [7].

Furthermore, our analysis revealed that ECG Reader-GPT consistently outperformed GPT-4o and Canva-GPT across all ECG subcategories. Its advantage was especially pronounced in complex categories such as ischemic syndromes, channelopathies, and genetic syndromes—conditions requiring nuanced waveform interpretation beyond basic quantitative analysis. This superiority likely arises from ECG Reader-GPT's training on curated, task-specific datasets, enabling it to recognize clinically relevant patterns often overlooked by general-purpose models.

ECG Reader-GPT achieved diagnostic performance comparable to EMSs, particularly in high-risk and complex cases. It matched EMS accuracy in detecting ischemic syndromes, channelopathies and genetic syndromes, and normal ECGs; it exceeded EMS performance in identifying rhythm disorders. These findings support the growing role of customized AI models for ECG interpretation as a reliable adjunct in acute care, where rapid and accurate interpretation is essential. Although not a replacement for human expertise, such models may help reduce diagnostic delays, support less experienced clinicians, and promote consistent care in resource-constrained settings.

Our results align with recent studies highlighting the promise of domain-specific AI in clinical diagnostics. Tison et al. found that machine learning–based ECG interpretation approaches human-level accuracy in detecting a broad spectrum of cardiac pathologies [16]. Similarly, Strodthoff et al. reported that AI-enhanced ECG systems demonstrated performance comparable to that of clinicians across diverse diagnostic categories in emergency settings, reinforcing the value of task-specific models in high-stakes environments [17]. In parallel, recent work has demonstrated the potential of machine learning–based approaches in broader cardiovascular contexts, including the prediction of obstructive coronary artery disease using treadmill ECG waveform features [20], mortality risk estimation in acute pulmonary embolism with deep learning [21], and the application of AI tools in the diagnosis and management of coronary artery disease and atrial fibrillation [22]. Nevertheless, it remains essential that AI systems serve as supportive tools, rather than primary decision-makers, in clinical practice [18,19].

This study had several limitations. First, the diagnostic task was restricted to a multiple-choice format. Although this enabled standardized comparisons between AI models and clinicians, it limited the scope for open-ended diagnostic reasoning. Neither group was able to provide justifications or elaborate on their choices, potentially constraining the demonstration of their full interpretive capabilities. For example, a model might have correctly identified a STEMI and suggested appropriate clinical management, but if it failed to specify the precise STEMI subtype, the response was deemed incorrect. Additionally, as the original power analysis was designed for comparisons based on total ECG diagnostic accuracy, results from smaller subsets should be interpreted with caution. Although we calculated effect size estimates for pairwise comparisons, some values exceeded 1.0, particularly within ECG subcategories with limited sample sizes. This overestimation reflects statistical inflation rather than the actual magnitude of effect and is likely due to insufficient statistical power in these subgroups. Second, while AI models completed the test over 30 repeated trials, EMSs participated in a single test session. This introduces an imbalance in data granularity and may favor AI performance through averaging effects. Future research should aim to mitigate this bias by increasing the human sample size or incorporating repeated human testing to allow more direct comparisons. Third, the ECG cases were adapted from a widely used reference textbook. Although this ensured consistency and clinical validity, it introduces the possibility that AI models—particularly those pretrained on large-scale, publicly available datasets—may have encountered similar content during training, potentially inflating performance. We cannot exclude the possibility that questions and answers similar or identical to those in our validation set were incorporated as part of its configuration or input examples. This may have contributed to its high diagnostic accuracy and represents a potential source of bias that should be addressed in future external validations. Fourth, a key limitation was the lack of transparency surrounding the development of ECG Reader-GPT. Detailed information regarding its model architecture, training data sources, and optimization methodologies has not been disclosed. This opacity hinders assessments of the model's generalizability, interpretability, and susceptibility to biases. Domain-specific AI models for clinical use should prioritize transparency, including the open sharing of training datasets, model architectures, and evaluation metrics, to enable independent validation and enhance trust in AI-assisted diagnostics. Finally, all ECGs in this study were presented as clean, high-resolution digital tracings. In contrast, real-world clinical environments often involve suboptimal recordings affected by artifacts, noise, or incomplete data. Therefore, further external validation is required to determine the robustness and applicability of these AI models under real-world conditions.

5. Conclusions

This study demonstrated that customized AI models for ECG interpretation—particularly ECG Reader-GPT—substantially outperformed general-purpose VLMs, such as GPT-4o and Canva-GPT, across all ECG subcategories. ECG Reader-GPT achieved higher overall diagnostic accuracy and demonstrated performance comparable to that of experienced EMSs. It exceeded EMS performance in interpreting rhythm disorders and performed similarly in complex and life-threatening cases, highlighting its potential utility in high-stakes clinical environments.

These findings underscore the importance of task-specific training and model specialization in the development of clinically effective AI tools. Although AI should currently function as an adjunct rather than a replacement for human expertise, customized AI models for ECG interpretation such as ECG Reader-GPT may offer valuable opportunities to enhance diagnostic accuracy, reduce inter-clinician variability, and support decision-making in time-sensitive emergency care settings.

Declaration of generative AI in scientific writing

The authors did not use a generative artificial intelligence (AI) tool or service to assist with preparation or editing of this work. The author(s) take full responsibility for the content of this publication.

Authorship contribution

EA, AA, OS, YD, and TA conceived the study and designed the trial. EA, AA, OS, EB, and HIB supervised the conduct of the trial and data collection. EB, TS, MO, and NK undertook the recruitment of participating patients and managed the data, including quality control. HIB, YD, and TA provided statistical advice on study design and analyzed the data; AA chaired the data oversight committee. EA, AA, and OS drafted the manuscript, and all authors contributed substantially to its revision. EA takes responsibility for the paper as a whole.

CRediT authorship contribution statement

Ertugrul Altinbilek: Writing – review & editing, Writing – original draft, Supervision, Project administration, Conceptualization. **Adem Az:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Ozgur Sogut:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization. **Yunus Dogan:** Visualization, Software, Resources, Investigation, Formal analysis, Data curation. **Tarik Akdemir:** Visualization, Validation, Software, Resources, Data curation, Conceptualization. **Erdal Belen:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation, Data curation. **Halil Ibrahim Biter:** Validation, Supervision, Software, Investigation, Data curation. **Tugay Saricicek:** Validation, Software, Methodology, Formal analysis, Data curation. **Mehmet Ozcomlekci:** Visualization, Software, Resources, Funding acquisition, Formal analysis, Data curation. **Nurbaki Kilic:** Visualization, Software, Resources, Data curation, Conceptualization.

Ethical approval

The study protocol was approved by the Institutional Review Board of Haseki Training and Research Hospital, Istanbul, Turkey (approval no. 30–2025).

Funding

The author(s) received no financial support for this work, authorship, and/or publication.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

None.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article.

References

[1] Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024;20(1):26.e1–5.
[2] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.

[3] Mello-Thoms C, Mello CAB. Clinical applications of artificial intelligence in radiology. Br J Radiol. 2023;96(1150):20221031.
[4] Günay S, Öztürk A, Özerol H, et al. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med. 2024;80:51–60.
[5] Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. Am J Emerg Med. 2024;84:68–73.
[6] Borown T, Mann B, Ryder N, et al. Language models are few-shot learners. NeurIPS. 2020;33:1877–901.
[7] Chang KC, Hsieh PH, Wu MY, et al. Usefulness of machine learning-based detection and classification of cardiac arrhythmias with 12-Lead electrocardiograms. Can J Cardiol. 2021;37(1):94–104.
[8] OpenAI. GPT-4o. OpenAI [Internet]. San Francisco (CA): OpenAI; 2024. [cited 2025 Apr 18]. Available from:. https://openai.com/index/hello-gpt-4o/.
[9] OpenAI. Canva-GPT [Internet]. San Francisco (CA): OpenAI; 2024. [cited 2025 Apr 18]. Available from: https://chatgpt.com/g/g-alKfVrz9K-canva.
[10] OpenAI. ECG Reader-GPT [Internet]. San Francisco (CA): OpenAI; 2024. [cited 2025 Apr 18]. Available from: https://chatgpt.com/g/g-hKnYr7yJl-ecg-reader.
[11] Az A. From memory to mastery: optimizing AI models for ECG diagnostics in clinical practice. Am J Emerg Med. 2024;86:161.
[12] Hampton J, Adlam D, Hampton J. 150 ECG cases (5th edition). Amsterdam: Elsevier Publishers; 2019..
[13] Penny P, Bane R, Riddle V. Advancements in AI medical education: assessing ChatGPT's performance on USMLE-style questions across topics and difficulty levels. Cureus. 2024;16(12):e76309.
[14] Zhu L, Mou W, Wu K, et al. Multimodal ChatGPT-4V for electrocardiogram interpretation: promise and limitations. J Med Internet Res. 2024;26:e54607.
[15] Hirosawa T, Harada Y, Tokumasu K, et al. Evaluating ChatGPT-4's diagnostic accuracy: impact of visual data integration. JMIR Med Inform. 2024;12:e55627.
[16] Tison GH, Zhang J, Delling FN, et al. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. Circ Cardiovasc Qual Outcomes. 2019;12(9):e005289.
[17] Strodthoff N, Lopez Alcaraz JM, Haverkamp W. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care. Eur Heart J Digit Health. 2024;5(4):454–60.
[18] Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel). 2024;11(4):337.
[19] Mennella C, Maniscalco U, De Pietro G, et al. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon. 2024;10(4):e26297.
[20] Yilmaz A, Hayıroğlu Mİ, Salturk S, et al. Machine learning approach on high risk treadmill exercise test to predict obstructive coronary artery disease by using P, QRS, and T waves' features. Curr Probl Cardiol. 2023;48(2):101482.
[21] Cicek V, Orhan AL, Saylik F, et al. Predicting short-term mortality in patients with acute pulmonary embolism with deep learning. Circ J. 2025;89(5):602–11.
[22] Hayıroğlu Mİ, Altay S. The role of artificial intelligence in coronary artery disease and atrial fibrillation. Balkan Med J. 2023;40(3):151–2.