THE ANATOLIAN JOURNAL OF CARDIOLOGY

Official journal of the TURKISH SOCIETY OF CARDIOLOGY

Translating Multimodal Intelligence into Cardiac Diagnostics: A Critical Perspective on Large Language Model–Assisted Electrogram Interpretation

LETTER TO THE EDITOR

To the Editor,

We read with great interest the study by Bozyel et al,[1] exploring scenario-based evaluation of ChatGPT-4o for intracardiac electrogram (EGM) interpretation in pacemaker patients. The staged design—from isolated signals to multiple-choice decisions—captures real diagnostic gradients, and the pairing of visual EGM inputs with device context mirrors clinical workflows in cardiac implantable electronic device (CIED) care. Repeating the experiments over 2 months and using consensus adjudication are additional strengths that facilitate a structured appraisal of repeatability. Several methodological choices, however, may constrain clinical translation.

First, the reference standard is derived from The European Heart Rhythm Association case book answers. These keys are optimized for teaching rather than for adjudicating device-specific algorithms across manufacturers.[2] Without an explicit device-vendor ground truth (e.g., programmer logs, marker channels, and algorithm state), the study risks construct drift—particularly for pacing mode, atrioventricular relationships, and pseudomalfunction, where small labeling nuances alter clinical action.

Second, the variable set includes broad constructs (e.g., "understanding") alongside technical items (e.g., "timing intervals"). Collapsing heterogeneous targets into a single accuracy figure obscures domain-specific failure modes.[3] A per-case error taxonomy with clinically anchored severities (benign vs. action-triggering mistakes such as pacing inhibition or oversensing) would reveal whether observed gains translate into safer decisions. Likewise, the "No Answer/Non-Relevant" categories may dilute misclassification rates; a pre-specified handling plan (penalization or imputation) is needed to avoid optimistic accuracy.

Third, the statistical framework mixes raw accuracy with Cohen's Kappa and Prevalence- and Bias-Adjusted Kappa across multiple features and scenarios without interval estimates or multiplicity control. Given known prevalence effects on agreement metrics, reporting CIs, decision-relevant thresholds, and a correction plan for multiple comparisons would prevent over-interpretation.[4] Calibration analyses (e.g., Brier score for probabilistic outputs or thresholded decision curves) are also needed if the goal is clinical support.

Fourth, experimental control and reproducibility require fuller disclosure. Prompt templates, system parameters (temperature, top-p), image fidelity (resolution, compression), and any pre-processing materially affect multimodal performance.[5] Without these details, replicability and fair benchmarking against electrophysiologists under time constraints remain uncertain.

To advance clinical usefulness, future work may: (i) use programmer-verified ground truth spanning major vendors and modes; (ii) define primary endpoints tied to patient management (alert triage yield in remote monitoring, detection of pacing

Shyam Sundar Sah
Abhishek Kumbhalwar

[1]*Dr. D. Y. Patil Medical College Hospital and Research Centre, Dr. D. Y. Patil Vidyapeeth (Deemed-to-be-University), Pimpri, Pune, Maharashtra, India*
[2]*Dr. D. Y. Patil Dental College and Hospital, Dr. D. Y. Patil Vidyapeeth (Deemed-to-be-University), Pimpri, Pune, Maharashtra, India*

Corresponding author:
Shyam Sundar Sah
shyam.sundar.sah@proton.me

Cite this article as: Sah SS, Kumbhalwar A. Translating multimodal intelligence into cardiac diagnostics: A critical perspective on large language model–assisted electrogram interpretation. *Anatol J Cardiol.* 2025;XX(X):X-X.

DOI:10.14744/AnatolJCardiol.2025.5957

inhibition/oversensing); (iii) compare against device specialists with timed reads; (iv) report per-phenotype performance with severity weighting; (v) pre-register analysis plans with CIs and correction for multiplicity; and (vi) explore manufacturer-specific fine-tuning and human-in-the-loop deployment for CIED remote monitoring. Such steps would clarify whether the observed gains in context-rich scenarios can meaningfully reduce clinician workload while maintaining safety.

In conclusion, while large language models show potential for assisting in intracardiac electrogram interpretation, their current performance remains exploratory. Robust validation with real device data, clinical benchmarks, and reproducible methods will be essential before integration into routine cardiac diagnostics.

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

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

REFERENCES

1. 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;29(10):533-542. [CrossRef]
2. Arias MA. The EHRA book of interventional electrophysiology. *Rev Española Cardiol* [Internet]. Elsevier; Amsterdam. 2017;70(10):889-890. [CrossRef]
3. Orouji S, Liu MC, Korem T, Peters MAK. Domain adaptation in small-scale and heterogeneous biological datasets. *Sci Adv*. 2024;10(51):eadp6040. [CrossRef]
4. Turovsky YaA, Borzunov SV, Vahtin AA. An Algorithm for Correction of Statistical Estimations Taking into Account the Effect of Multiple Comparisons based on Test Results Grouping. *Prin*. 2022;13(3):148-152. [CrossRef]
5. Wanaskar K, Jena G, Eirinaki M. Multimodal benchmarking and recommendation of text-to-image generation models. In: 11th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService); Tucson, AZ, USA. Piscataway (NJ): IEEE; 2025. [CrossRef]