EmERGE: A Multimodal LLM-based Evidence-mediated ECG Report Generation and Evaluation Approach

Keyu Yao, Yuchen Liu, Xiaohan Wang, and Gang Shen✉

Huazhong University of Science and Technology, Wuhan 430074, China
gang_shen@hust.edu.cn

Abstract. Multimodal large language models (MLLMs) have significantly advanced the automatic generation of electrocardiogram (ECG) reports. However, current approaches handle diagnosis as a simple signal-to-text mapping task, overlook clinical reasoning, and thus suffer from hallucination and poor interpretability. We propose an evidence-mediated ECG report generation and evaluation (EmERGE) approach that reframes the task as structured causal reasoning. It incorporates six clinically validated ECG features, such as rhythms and rates, P waves, and P-R intervals. These features act as explicit mediators, enabling MLLMs to perform diagnostic conditional reasoning using interpretable electrophysiological evidence. Experiments demonstrate that EmERGE substantially improves ECG report quality, outperforming state-of-the-art grounded report generation methods by 13% in clinical efficacy and 9% in interpretability. To verify whether diagnostic conclusions rely on causal evidence, we developed a rule-guided counterfactual evaluation protocol. The protocol applies signal-level interventions to clinically relevant leads and waveform segments to validate the link between selective evidence and diagnostic outcomes. EmERGE establishes an evidence-driven, causality-focused approach, providing a viable solution for real-world clinical applications. The source code and prompts are available at https://github.com/hustselab511/EmERGE.git.

Keywords: ECG report generation · Multimodal large language models · Evidence-mediated Reasoning · Counterfactual Evaluation.

1 Introduction

Millions of electrocardiogram (ECG) examinations are conducted every day, as ECG charts are crucial in screening cardiovascular diseases, making accurate diagnoses, and monitoring patients over the long term [1,2]. In clinical practice, ECG interpretations are typically presented as free-text reports that include electrophysiological findings, clinical impressions, and follow-up recommendations [3]. Automating ECG report generation is an essential task in medical image computing, as it significantly enhances the efficiency, consistency, and accessibility of cardiac care [4,5].

Early automated ECG interpretation relied on rule-based systems and signal processing techniques, mapping handcrafted features (such as heart rate, waveform amplitude, and interval measurements) to predefined diagnostic labels [6,7]. Although straightforward and generally reliable in most scenarios, their reliance on predefined rules makes them susceptible to variations in signal quality. Consequently, these systems struggle to generate semantically coherent, high-quality textual reports, particularly when confronting severe waveform noise or complex complications [8,9]. In contrast, deep learning enables models to predict diagnostic results directly from ECG waveforms, reducing the need for manual feature engineering and demonstrating significant advantages in automated report generation [10,11,12]. Recent advances in multimodal large language models (MLLMs) have transformed this field [13,14]. The up-to-date systems combine multimodal data, including time-series signals and ECG images, to create textual reports. Models such as ECG-Chat [15], PULSE [16], and GEM [17] confirm that MLLMs can generate fluent, clinically compliant reports. A significant problem that remains underinvestigated is whether these generated diagnostic conclusions are rooted in valid ECG evidence.

In clinical practice, electrocardiologists follow a structured reasoning process: they extract electrophysiological information from signals and then derive diagnostic conclusions [18]. However, most existing models focus on optimizing end-to-end objectives by directly mapping multimodal ECG inputs to text, without investigating the underlying evidence. As a result, these models may produce diagnostically correct reports that are inconsistent with the evidence due to dataset biases, language priors, or co-occurrence statistics [19]. This issue becomes particularly evident in complex cases, where superficial accuracy can obscure flawed clinical reasoning. As shown in Fig 1, electrocardiologists decom-

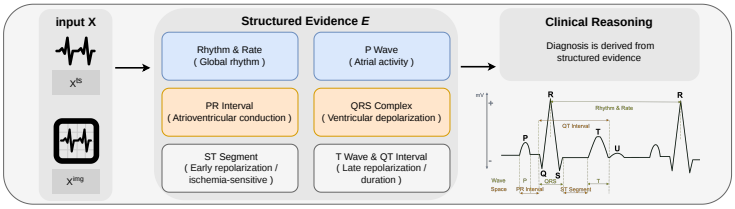

Fig. 1. Clinically defined evidence in a heartbeat on the ECG chart.

pose raw waveforms into six clinically defined evidence dimensions [20,21]. These dimensions represent electrophysiological evidence rather than disease labels and form the foundation for downstream diagnostic reasoning [18].

This study focuses on the automated generation of ECG reports guided by evidence and employs a behavioral-level validation method. We introduce structured electrophysiological evidence as a mediator between the inputs and the reports, thereby emphasizing the importance of evidence in the diagnostic process. Additionally, we assess the reports using rule-based signal-level counterfactual perturbations to determine whether the diagnostic outputs respond selectively to changes relevant to the evidence while remaining unaffected by irrelevant

variations [22]. EmERGE can be seamlessly integrated into existing ECG report generation methods based on MLLM. In our experiments, while keeping the encoder fixed, we employed strong baseline methods, PULSE and GEM, as our backbones, referred to as EmERGE (P) and EmERGE (G), respectively.

This work provides a foundation for building and evaluating trustworthy automated ECG interpretation systems, with the following key contributions:
1. We model six clinically established ECG phenomena as mediators, enabling evidence-driven reasoning and improving the interpretability of language models.
2. We design a rule-guided counterfactual evaluation protocol to validate causal consistency of the reports through signal-level interventions.

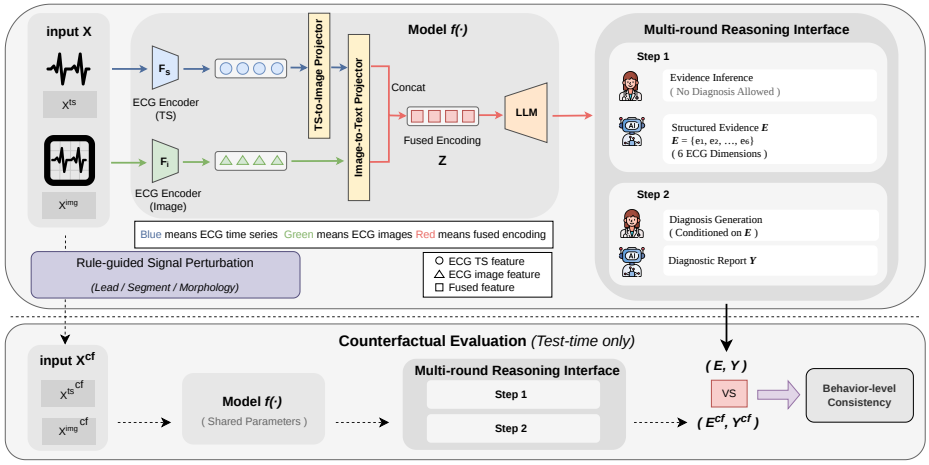

Fig. 2. The proposed EmERGE pipeline.

2 Method

2.1 Problem Definition and Overall Pipeline

Given an input $X = (X^s, X^i)$, where $X^s \in \mathbb{R}^{12\times T}$ denotes the 12-lead temporal signal ($T = 5000$, 10s $\times$ 500 Hz) and X^i denotes its image, our objective is to generate a structured evidence representation E and a diagnostic report text Y. Replacing the end-to-end mapping $X \to Y$ in the existing works, we explicitly introduce a structured evidence variable E to characterize the generation process. In the causality paths (Eq. (1)), E acts as a mediator for diagnostic reasoning, ensuring the transparency and auditability of the report.

$$X \longrightarrow E \longrightarrow Y \tag{1}$$

In the evidence space $\mathcal{E} = \{E_{\mathrm{RR}}, E_{\mathrm{P}}, E_{\mathrm{PR}}, E_{\mathrm{QRS}}, E_{\mathrm{ST}}, E_{\text{T-QT}}\}$, each dimension is a textual description of numerical or lead-level attributes such as amplitude, duration, and morphological keywords, with labels of *Normal, Abnormal*, or *Borderline* [23].

The design of the EmERGE pipeline targets two key goals: an interpretable intermediate state and an intervenable interface. Fig. 2 illustrates the proposed

evidence-mediated ECG report generation and evaluation pipeline. EmERGE follows three sequential stages: (i) **structured evidence derivation** from ECG signals to extract six clinically defined phenotypes, (ii) **evidence-conditioned report generation** that explicitly grounds diagnostic statements on evidence, and (iii) **behavioral-level validation** via physiology-guided signal perturbations.

2.2 Two-Step Evidence-mediated Report Generation

We adopt a two-step generation strategy: $E = f(Z; P_1), Y = f(Z, E; P_2)$. Here, Z denotes ECG encoder outputs, f denotes EmERGE, and P_1 and P_2 are prompt templates for six-dimensional evidence and grounded reports, respectively.

To prevent post-hoc rationalization by first generating diagnoses and then supplementing evidence, Step 1 is strictly restricted to output only E. Step 2 generates the report text under conditioning on evidence E.

The training objective in Eq. (2) combines evidence and reports supervision losses using the weight λ (empirically set to 1.0). In Eq. (2), y^* is the ground-truth report, E^* denotes the six-dimensional evidence supervision signal, derived from joint alignment of machine measurements and annotated text. This objective forces the model to first learn stable representations of the evidence.

$$\begin{gathered} \mathcal{L} = \mathcal{L}\mathrm{e} + \lambda \mathcal{L}\mathrm{r} \\ \mathcal{L}\mathrm{e} = \sum_{i=1}^{6} \log p\big(E_{it}^* \big| E_{i<t}^*, Z\big), \quad \mathcal{L}\mathrm{r} = -\sum_{t} \log p\big(y_t^* \big| y_{<t}^*, Z, E^*\big) \end{gathered} \tag{2}$$

2.3 Rule-guided Signal-level Counterfactual Evaluation

Clinical diagnosis follows an evidence-driven chain: analyzing electrophysiological features before reaching judgments. Robust models ought to respond only to clinically meaningful signal variations and stay unaffected by irrelevant noise. Therefore, EmERGE adopts a rule-based signal-level counterfactual perturbation for behavioral evaluation. Our counterfactual intervention rules target designated ECG leads and waveform segments, replacing them with heartbeat-matched but semantically distinct ones (followed by signal smoothing). These rules follow established ECG interpretation guidelines [24,25] and dataset-specific disease descriptions. We divide interventions into *targeted* and *non-targeted* ones based on their expected diagnostic impacts. These rules serve solely for performance evaluation, rather than model training or inference.

Given an ECG input X, we construct a counterfactual signal X^{cf} by applying localized perturbations to targeted waveform segments after R-peak alignment, and obtain $(E, Y) = f(X)$, $(E^{\mathrm{cf}}, Y^{cf}) = f(X^{cf})$.

We quantify evidence–diagnosis consistency using three behavioral metrics.
Targeted Sensitivity: $\mathrm{TS} = \frac{1}{|S^{TS}|} \sum_{i \in S^{TS}} \mathbb{I}\big[\hat{E}_i^{\,cf} \neq \hat{E}_i\big]$;
Non-target Invariance: $\mathrm{NI} = \frac{1}{|S^{NI}|} \sum_{j \in S^{NI}} \mathbb{I}\big[\hat{E}_j^{\,cf} = \hat{E}_j\big]$, where S^{TS} and S^{NI} denote two sets constructed by targeted and non-target interventions, respectively;
Diagnosis Response Consistency: $\mathrm{DRC} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\big[\mathrm{sign}\big(\Delta_D^{(n)}\big) = \mathrm{sign}\big(\Delta_E^{(n)}\big)\big]$,

testing whether the changes of diagnosis encoding Δ_D and evidence Δ_E are in the same direction).

Here, the indicator function $\mathbb{I}(\cdot)$ counts the TRUE cases. In deterministic counterfactual settings, DRC corresponds to a reversibility criterion for diagnosis retraction or restoration.

3 Experiment Results

3.1 Datasets, Evaluation Metrics, and Implementation Details

Dataset We evaluate on MIMIC-IV (in-domain)[26] and PTB-XL [3] (out-of-domain), following the GEM data split (2,381 samples) and using the official PTB-XL test set (2,041 samples). For training, we randomly sample 12,000 MIMIC-IV recordings and 6,000 PTB-XL recordings, (PTB-XL exclusively for evidence extraction).

Ground-Truth Construction We generate structured six-dimensional ECG evidence annotations using GPT-4o and machine measurements extracted by featureDB [27,28], following clinical definitions. These annotations supervise evidence modeling. For fair comparison with GEM, we strictly use its provided ground truth.

Evaluation Protocols We use three groups of metrics for model evaluation:
(1) *LLM-based clinical evaluation* assesses clinical coverage, coherence, and evidence grounding of generated reports. Following the GEM protocol, we employ an LLM to rate the generated reports on a scale of 0 to 100 across eight specific dimensions (e.g., Diagnosis Accuracy, Analysis Completeness).
(2) *Clinical efficacy metrics* verify whether generated reports contain diagnostically discriminative information. Based on the SCP label system, we design a prompt rather than standard regular expressions to instruct an LLM to map the generated grounding reports to specific diagnostic labels. Then, we compare the results and the ground truth for precision, recall, and F1-score. They serve as auxiliary indicators of valid information extraction.
(3) *Counterfactual behavior analysis* evaluates evidence-diagnosis consistency using physiology-guided signal perturbations. It examines whether diagnosis selectively changes in response to evidence-relevant perturbations while remaining invariant to irrelevant variations.

Model Implementation All models are trained on 6 NVIDIA L40S GPUs (48 GB) with batch size of 48 for 2 epochs, using a CLIP as vision encoder and ECG-Chat as ECG encoder. Other settings follow GEM.

3.2 Main Results

LLM-based Clinical Evaluation Table 1 summarizes the LLM-based evaluation results on the MIMIC-IV dataset. The values of Acc_D, Com, Rel, Cov, Acc_L, Gro, Evi, and Fid represent diagnostic accuracy, analysis completeness, analysis relevance, lead assessment coverage, lead assessment accuracy, ECG feature grounding, evidence-based reasoning, diagnostic fidelity, respectively. Our method consistently outperforms all baseline models under both in-domain and out-of-domain settings.

Table 1. In-domain and out-of-domain LLM-based evaluation results (in %) on the MIMIC-IV and PTB-XL datasets.

Dataset	Method	Acc_D ↑	Com↑	Rel ↑	Cov↑	Acc_L ↑	Gro↑	Evi↑	Fid↑
MIMIC-IV	MEIT [11]	40.51	6.43	64.07	1.25	31.92	3.38	6.84	33.33
	ECG-Chat [15]	18.45	21.95	57.08	2.66	21.49	15.19	19.52	41.85
	PULSE [16]	69.86	36.47	77.19	21.00	57.15	38.55	54.91	69.47
	GEM [17]	74.23	69.89	78.57	81.44	56.99	73.42	72.91	80.79
	EmERGE (P)	79.96	77.93	84.66	82.99	65.03	78.72	78.23	85.59
	EmERGE (G)	77.81	78.46	85.26	81.51	64.09	77.18	78.35	85.00
PTB-XL	MEIT	47.39	9.92	65.48	0.61	37.10	5.23	8.28	44.01
	ECG-Chat	13.88	21.56	49.53	5.66	16.31	13.95	16.8	40.87
	PULSE	46.68	38.84	63.93	19.68	40.02	32.73	41.87	60.84
	GEM	63.75	68.23	79.21	81.96	52.20	68.59	68.92	80.48
	EmERGE (P)	65.44	77.60	84.30	76.27	56.69	71.22	72.32	83.08
	EmERGE (G)	69.25	79.07	85.97	79.61	58.71	73.85	75.11	84.26

On the MIMIC-IV dataset, our model achieves the highest diagnostic accuracy and diagnostic fidelity, significantly surpassing the GEM model while also improving evidence-oriented metrics such as ECG feature grounding and evidence-based reasoning. Early methods, including MEIT and ECG-Chat, exhibit limitations in diagnostic accuracy and weak feature anchoring capability, indicating their reliance on surface-level linguistic patterns. On the PTB-XL dataset, the performance gap is more pronounced under domain shift. While all models experience performance degradation, our method still maintains outstanding diagnostic accuracy and evidential fidelity, demonstrating superior robustness over GEM and PULSE.

Table 2. In-domain and out-of-domain CE evaluation results (in %) on the MIMIC-IV and PTB-XL datasets.

	MIMIC-IV						PTB-XL					
	Pre_M ↑	Re_M ↑	$F1_M$ ↑	Pre_m ↑	Rec_m ↑	$F1_m$ ↑	Pre_M ↑	Re_M ↑	$F1_M$ ↑	Pre_m ↑	Rec_m ↑	$F1_m$ ↑
ECG-Chat [15]	20.18	21.10	10.93	28.65	23.43	25.78	20.26	35.98	22.86	26.20	31.92	28.78
MEIT [11]	17.21	21.42	11.70	29.32	22.81	25.66	8.91	20.00	9.90	44.54	35.12	39.27
DERI [12]	43.30	45.30	43.50	-	-	-	-	-	-	-	-	-
PULSE [16]	61.57	62.25	57.95	59.25	64.38	61.71	42.53	58.42	35.76	32.45	47.53	38.57
GEM [17]	59.36	71.36	64.01	59.40	75.26	66.40	60.29	60.39	57.05	59.47	63.95	61.63
EmERGE (P)	74.31	74.60	73.80	74.85	76.44	75.63	68.03	58.80	61.32	69.46	64.68	66.99
EmERGE (G)	71.39	71.90	71.54	73.09	74.48	73.78	66.64	66.25	64.88	68.45	68.16	68.31

CE-based Diagnostic Accuracy In addition to the LLM-based evaluation, we compute clinically evidenced diagnostic metrics derived from the generated reports. The metrics Pre, Rec and $F1$ represent precision, recall and F1-score, while the subscriptions M and m denote macro-level and micro-level, respectively. As shown in Table 2, our method consistently achieves higher macro and micro F1 scores than other baselines under both in-domain and out-of-domain settings. The improvement in macro F1 is particularly evident, demonstrating that our method enhances the diagnostic extractability of rare conditions, rather than merely boosting performance for dominant classes.

3.3 Ablation Study

Table 3 reports the ablation results on the MIMIC-IV test set. V1 represents evidence used only in training, and V2 implies no evidence in training nor inference. Removing evidence in either stage leads to a clear performance drop. V1

Table 3. Ablation results (F1-scores in %)

	EmERGE (P)	EmERGE (G)	V1 (G)	V2 (G)	V1 (P)	V2 (P)
$F1_M \uparrow$	73.80	71.54	67.52	66.12	61.83	60.67
$F1_m \uparrow$	75.63	73.78	70.51	69.64	64.35	63.14

reduces macro F1 from 71.54 to 67.52 and micro F1 from 73.78 to 70.51, indicating that single-pass generation fails to maintain stable intermediate reasoning and weakens evidence–diagnosis alignment. Removing the evidence representation causes an even larger degradation (macro F1 66.12, micro F1 69.64). For both PULSE and GEM, the ablated EmERGE variant V1 outperforms these direct end-to-end mapping backbones, confirming that performance gains stem from the evidence-mediated reasoning.

3.4 Counterfactual Behavior Analysis

We evaluate whether diagnostic conclusions are causally grounded in ECG evidence using physiology-guided signal-level counterfactual interventions. For each disease, we construct targeted and non-targeted rules: (i) causally relevant perturbations that modify evidence required for the target diagnosis, and (ii) irrelevant perturbations that alter non-diagnostic signal components while preserving key evidence.

Table 4 reports counterfactual behavior results across four typical diseases. EmERGE consistently performs better than GEM on most metrics, except when GEM produces very low OR rates (e.g., TS for ASMI). In particular, EmERGE shows substantially stronger selective responses to evidence-relevant changes while maintaining stability under irrelevant perturbations, confirming improved evidence-level causal consistency.

Table 4. Counterfactual analysis results (OR represents the original recall)

	LVH			**ASMI**			**IMI**			**LAFB**		
	OR↑	TS↑	NI↑	OR↑	TS↑	NI↑	OR↑	TS↑	NI↑	OR↑	TS↑	NI↑
GEM	0.210	0.309	0.881	0.110	1.000	0.182	0.200	0.825	0.850	0.530	0.962	0.896
EmERGE (G)	0.500	0.420	0.840	0.635	0.906	0.890	0.475	0.853	0.905	0.920	0.929	0.957

Table 5 shows the deterministic counterfactual results on LAFB, where pathological and normal segments can be explicitly removed or injected. Both models achieve perfect pre-intervention precision; however, GEM exhibits low etiological accuracy (0.009) and weak post-intervention reversibility (0.377). In contrast, EmERGE attains much higher recall, etiological accuracy, and correct diagnosis retraction after evidence removal.

Overall, these results demonstrate that explicitly modeling ECG evidence enables diagnostic behavior that is responsive to causal signal changes and robust

Table 5. Deterministic counterfactual intervention results on LAFB cases

	Precision↑	Recall↑	Etiological Accuracy↑	DRC (Reversibility)↑
GEM	1.000	0.450	0.009	0.377
EmERGE (G)	1.000	0.860	0.623	0.717

Table 6. Cardiologist Evaluation Results

	Analytical Relevance	Analytical Accuracy	Analytical Completeness	Reasoning Quality	Findings Novelty	Clinical Value	Overall Satisfaction
GPT-4o	4.69	4.23	4.34	4.58	3.83	4.52	4.38
DeepSeek-R1	4.85	4.23	4.55	4.69	4.22	4.61	4.44
EmERGE (G)	4.41	3.78	4.15	4.10	3.64	4.02	3.92

to irrelevant variations, a property not captured by standard accuracy-based metrics.

3.5 Cardiologist Evaluation and Case study

To complement automated metrics and mitigate the limitations of LLM-based evaluation, we conduct human expert assessment and qualitative case analysis following the evaluation protocol adopted in GEM.

Cardiologist Evaluation Results. As EmERGE builds on top of LLM, we invite professionals to verify the quality of silver labels used for model training. We randomly select 200 samples from the MIMIC-IV dataset and test them on EmERGE, GPT-4o, and DeepSeek-R1. Four cardiologists evaluate generated reports across seven dimensions by the GEM criterion with five-point scoring. Average scores reflect overall diagnostic accuracy and the reasoning validity. We clarify that GPT-4o and DeepSeek-R1 are not baselines for comparison: they serve only to verify the reliability of supervision signals, including grounded diagnostic descriptions and clinical evidence supporting EmERGE training. Table 6 shows that EmERGE obtains consistently high scores. The results from GPT-4o indicate that the supervision signals meet clinical standards and contain no systematic bias.

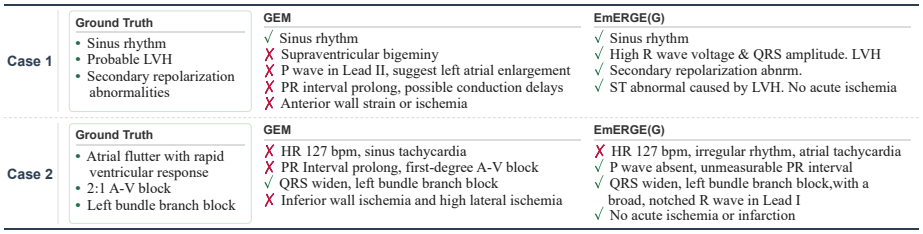

Fig. 3. Comparison with GEM for ECG-grounding examples.

Case Study. Fig. 3 presents example cases for a qualitative comparison of model behavior in challenging clinical scenarios. In Case 1, EmERGE attributes secondary ST–T abnormalities to ventricular hypertrophy and correctly avoids the spurious ischemic diagnoses made by GEM. In Case 2, despite incomplete morphology, EmERGE ensures clinical safety by identifying atrial tachycardia

without making unwarranted claims of infarction. These cases demonstrate that explicitly mediating diagnoses through structured ECG evidence reduces inaccurate findings and supports more cautious clinical decision-making.

4 Conclusion

This work introduces EmERGE, an evidence-mediated framework for ECG report generation. By encoding clinically validated ECG evidence as intermediate representations and assessing model behavior through physiology-guided signal-level counterfactuals, EmERGE enhances diagnostic fidelity, evidence grounding, and cross-domain robustness while remaining backbone-agnostic. Current limitations include partial dependence on LLM-based supervision and evaluation, ambiguous clinical validity of signal-level interventions, and incomplete coverage of complex ECG phenotypes across evidence dimensions. Our future work will target extensive clinical validation on large datasets and fine-grained evidence.

Acknowledgments. This work was supported by the Natural Science Foundation of China (NSFC) under grant 62476105, and Hubei Provincial Technology Innovation Program (Major Science and Technology Project) under grant 2025BCA002. Keyu Yao and Yuchen Liu contributed to this work equally.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. G. A. Roth, G. A. Mensah, and et al, "Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, p. 2982–3021, 12 2020.
2. Y. Birnbaum, K. Nikus, P. Kligfield, M. Fiol, J. A. Barrabés, A. Sionis, O. Pahlm, J. G. Niebla, and A. B. de Luna, "The role of the ecg in diagnosis, risk estimation, and catheterization laboratory activation in patients with acute coronary syndromes: A consensus document," *Annals of Noninvasive Electrocardiology*, vol. 19, pp. 412–425, 09 2014.
3. P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, 05 2020.
4. K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Reviews Cardiology*, vol. 18, pp. 465–478, 02 2021.
5. M. Y. Ansari, M. Yaqoob, M. Ishaq, E. F. Flushing, I. A. c. Mangalote, S. P. Dakua, O. Aboumarzouk, R. Righetti, and M. Qaraqe, "A survey of transformers and large language models for ecg diagnosis: advances, challenges, and future directions," *Artificial Intelligence Review*, vol. 58, 06 2025.
6. J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, pp. 230–236, 03 1985.
7. J. L. Willems, C. Abreu-Lima, P. Arnaud, J. H. van Bemmel, C. Brohet, R. Degani, B. Denis, J. Gehring, I. Graham, G. van Herpen, H. Machado, P. W. Macfarlane, J. Michaelis, S. D. Moulopoulos, P. Rubel, and C. Zywietz, "The diagnostic performance of computer programs for the interpretation of electrocardiograms," *New England Journal of Medicine*, vol. 325, pp. 1767–1773, 12 1991.

8. A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr, *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nature communications*, vol. 11, no. 1, p. 1760, 2020.
9. J. Schläpfer and H. J. Wellens, "Computer-interpreted electrocardiograms: benefits and limitations," *Journal of the American College of Cardiology*, vol. 70, no. 9, pp. 1183–1192, 2017.
10. A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, p. 65–69, 01 2019.
11. Z. Wan, C. Liu, X. Wang, C. Tao, H. Shen, J. Xiong, R. Arcucci, H. Yao, and M. Zhang, "Meit: Multimodal electrocardiogram instruction tuning on large language models for report generation," in *Findings of the association for computational linguistics: ACL 2025*, pp. 14510–14527, 2025.
12. J. Chen, X. Dong, W. Wang, S. Zhou, L. Yu, and X. Hu, "Deri: Cross-modal ecg representation learning with deep ecg-report interaction," *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4824–4832, 09 2025.
13. J. Chen, Y. Liang, and J. Ge, "Artificial intelligence large language models in cardiology," *Reviews in Cardiovascular Medicine*, vol. 26, 07 2025.
14. M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, p. 259–265, 04 2023.
15. Y. Zhao, J. Kang, T. Zhang, P. Han, and T. Chen, "Ecg-chat: A large ecg-language model for cardiac disease diagnosis," *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 06 2025.
16. R. Liu, Y. Bai, X. Yue, and P. Zhang, "Teaching multimodal llms to comprehend 12-lead electrocardiographic images," *npj Digital Medicine*, 2026.
17. X. Lan, F. Wu, K. He, Q. Zhao, S. Hong, and M. Feng, "Gem: Empowering mllm for grounded ecg understanding with time series and images," *Advances in Neural Information Processing Systems*, vol. 38, pp. 94421–94455, 2026.
18. P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, G. van Herpen, J. A. Kors, P. Macfarlane, D. M. Mirvis, O. Pahlm, P. Rautaharju, and G. S. Wagner, "Recommendations for the standardization and interpretation of the electrocardiogram," *Circulation*, vol. 115, pp. 1306–1324, 03 2007.
19. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.
20. J. W. Mason, E. W. Hancock, and L. S. Gettes, "Recommendations for the standardization and interpretation of the electrocardiogram: Part ii: Electrocardiography diagnostic statement list a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology," *Journal of the American College of Cardiology*, vol. 49, no. 10, pp. 1128–1135, 2007.
21. P. M. Rautaharju, B. Surawicz, and L. S. Gettes, "Aha/accf/hrs recommendations for the standardization and interpretation of the electrocardiogram: Part iv: The st segment, t and u waves, and the qt interval a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on

clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology," *Journal of the American College of Cardiology*, vol. 53, p. 982–991, 03 2009.
22. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
23. Y. Xu, Y. Ru, D. Zhang, Y. Liu, and Z. Sun, "Contextualizing borderline ecg analysis via multi-modal feature extraction and large language model inference," *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 06 2025.
24. Philips Medical Systems, "Philips 12-lead algorithm physician's guide," 2003.
25. B. Surawicz, R. Childers, B. J. Deal, and L. S. Gettes, "Aha/accf/hrs recommendations for the standardization and interpretation of the electrocardiogram," *Circulation*, vol. 119, 03 2009.
26. B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J. W. Waks, P. Eslami, T. Carbonati, A. Chaudhari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, and S. Horng, "MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset," *PhysioNet*, Sept. 2023. Version 1.0.
27. S. Hong, M. Wu, Y. Zhou, Q. Wang, J. Shang, H. Li, and J. Xie, "Encase: An ensemble classifier for ecg classification using expert features and deep neural networks," 09 2017.
28. S. Hong, Y. Zhou, M. Wu, J. Shang, Q. Wang, H. Li, and J. Xie, "Combining deep neural networks and engineered features for cardiac arrhythmia detection from ecg recordings," *Physiological Measurement*, vol. 40, p. 054009, 06 2019.