Electrocardiogram Report Generation and Question Answering via Retrieval-Augmented Self-Supervised Modeling

Jialu Tang[1] Tong Xia[2] Yuan Lu[1] Cecilia Mascolo[2]
Aaqib Saeed[1]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands
[2]University of Cambridge, Cambridge, United Kingdom

Abstract

Interpreting electrocardiograms (ECGs) and generating comprehensive reports remain challenging tasks in cardiology, often requiring specialized expertise and significant time investment. To address these critical issues, we propose ECG-ReGen, a retrieval-based approach for ECG-to-text report generation and question answering. Our method leverages a self-supervised learning for the ECG encoder, enabling efficient similarity searches and report retrieval. By combining pre-training with dynamic retrieval and Large Language Model (LLM)-based refinement, ECG-ReGen effectively analyzes ECG data and answers related queries, with the potential of improving patient care. Experiments conducted on the PTB-XL and MIMIC-IV-ECG datasets demonstrate superior performance in both in-domain and cross-domain scenarios for report generation. Furthermore, our approach exhibits competitive performance on ECG-QA dataset compared to fully supervised methods when utilizing off-the-shelf LLMs for zero-shot question answering. This approach, effectively combining self-supervised encoder and LLMs, offers a scalable and efficient solution for accurate ECG interpretation, holding significant potential to enhance clinical decision-making.

1 Introduction

Electrocardiograms (ECGs) are non-invasive, cost-effective diagnostic tools that play a crucial role in detecting cardiac arrhythmias in clinical practice. While numerous studies have demonstrated the effectiveness of machine learning models in predicting arrhythmia types from ECGs Mincholé and Rodriguez [2019], the tasks of interpreting ECGs, generating illustrative reports, and answering patient questions remain largely under-explored Moor et al. [2023]. These tasks present significant challenges, as they are time-consuming and require specialized expertise, even for experienced cardiologists. Moreover, they pose unique difficulties for machine learning models, demanding fine-grained feature extraction and the ability to generate cross-modality outputs (i.e., converting signals into coherent textual descriptions).

Recent advancements in large language models (LLMs) have shown promise in medical image interpretation, particularly for generating radiology reports from chest X-rays Chen et al. [2020], Yan et al. [2023]. However, their application to ECG analysis remains largely unexplored, despite ECGs' critical role in cardiology. ECGs, as time-series biosignals, present unique challenges compared to static medical images. Converting ECG data into LLM-compatible features for report generation is complex and data-intensive, as LLMs are not designed for processing physiological signals directly Thirunavukarasu et al. [2023]. Concerns also persist about the efficiency and generalizability of such

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

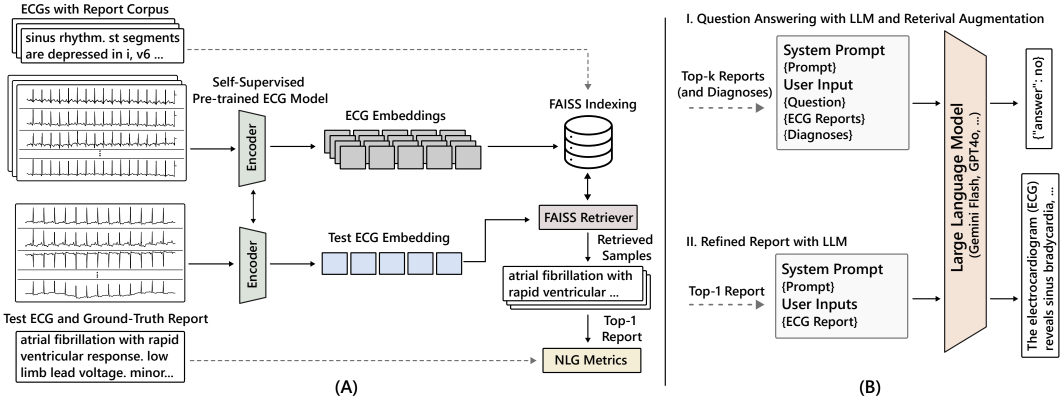

ECGs with Report Corpus
sinus rhythm. st segments are depressed in i, v6 ...
Self-Supervised Pre-trained ECG Model
Encoder
ECG Embeddings
FAISS Indexing
Encoder
Test ECG Embedding
FAISS Retriever
Retrieved Samples
atrial fibrillation with rapid ventricular ...
Top-1 Report
Test ECG and Ground-Truth Report
atrial fibrillation with rapid ventricular response. low limb lead voltage. minor...
NLG Metrics
(A)
I. Question Answering with LLM and Reterival Augmentation
Top-k Reports (and Diagnoses)
System Prompt
{Prompt}
User Input
{Question}
{ECG Reports}
{Diagnoses}
Large Language Model
(Gemini Flash, GPT4o, ...)
{"answer": no}
II. Refined Report with LLM
Top-1 Report
System Prompt
{Prompt}
User Inputs
{ECG Report}
The electrocardiogram (ECG) reveals sinus bradycardia, ...
(B)

Figure 1: Overview of ECG-ReGen leveraging self-supervised model to address report generation via retrieval and QA with LLMs.

models to unseen cardiac conditions or diverse patient populations Mittermaier et al. [2023]. These challenges highlight the need for robust approaches that can effectively bridge ECG signal processing with LLMs' natural language capabilities.

In this paper, we introduce a novel retrieval-based approach for ECG-to-text report generation and question answering (QA), pioneering its application in this domain. Unlike traditional task-specific, fully supervised learning methods Oh et al. [2024], Wan et al. [2024], our approach leverages similarity search to interpret new ECGs by referencing similar samples in the dataset, enhancing efficiency and explainability. While such methods have proven successful in chest X-ray report generation Endo et al. [2021], applying them to ECGs presents unique challenges due to the complex nature of ECG data. The key challenge lies in learning useful representations for similarity measures in the feature space, particularly difficult for multi-lead, long time-series ECG data where some cardiac conditions manifest as subtle waveform changes. Additionally, the brevity and diversity of ECG report phrases add complexity to the task.

To address these challenges, we propose a novel approach that pre-trains an ECG encoder to learn generalizable embeddings for efficient similarity searches. Our encoder is trained in a self-supervised manner, integrating ECG signals with their corresponding textual reports. During inference, we retrieve relevant reports based on the nearest neighbors of an ECG embedding, using the top-1 most similar report as the generated report. This retrieval-based method underpins our question-answering system, where we feed the top-k retrieved reports and their diagnosis labels into an LLM for zero-shot question answering. The LLM processes this information to provide accurate and contextually relevant responses to queries. Our approach combines similarity-based retrieval with the natural language understanding capabilities of LLMs, offering potentially more grounded and interpretable results. We validate our approach using two prominent ECG datasets: PTB-XL and MIMIC-IV-ECG. The retrieval-based method demonstrates superior performance in both settings, underlining its robustness and generalizability. Additionally, we employ off-the-shelf LLMs for zero-shot question answering, achieving competitive results compared to fully supervised approaches without task-specific fine-tuning.

2 Method

We frame ECG-to-text report generation and question answering as a retrieval augmented generation task using a report corpus R. Our approach combines a self-supervised ECG model with efficient similarity search to generate reports and answer ECG-related questions. The method consists of four stages: (1) self-supervised pre-training, (2) embedding generation and indexing, (3) report retrieval and refinement, and (4) zero-shot LLM-based question-answering. Figure 1 illustrates our proposed technique.

We pre-train a multi-modal model to learn joint representations of ECG signals and their textual reports. Let $X = x_1, x_2, ..., x_n$ be a set of ECGs, where $x_i \in \mathbb{R}^{L \times C}$ represents an ECG with L time steps and C channels. The corresponding reports are denoted as $R = r_1, r_2, ..., r_n$. We use a masked autoencoder-based self-supervised learning approach similar to Chen et al. [2022], combining three loss terms: masked language modeling (MLM), masked ECG modeling (MEM), and ECG-text matching (ETM), as $L = L_{\text{MLM}} + L_{\text{MEM}} + L_{\text{ETM}}$. The ETM task is a binary classification problem to determine if ECG-text pairs are semantically aligned.

Our multi-modal Transformer-based architecture Vaswani et al. [2017] for ECGs and texts comprises: 1) separate uni-modal encoders, 2) a multi-modal fusion module, and 3) separate uni-modal decoders for pre-training tasks. The text encoder uses a pre-trained BERT model Devlin [2018]. The ECG encoder applies 1D convolutional layers to extract local features from signal before feeding them into a transformer. We follow the configurations from Chen et al. [2022] for other architectural aspects. The model is pre-trained using above defined MLM, MEM, and ETM tasks.

After pre-training, we generate embeddings for all ECG samples using the ECG encoder $f(\cdot)$: $z_i = f(x_i), z_i \in \mathbb{R}^d$, where d is the embedding dimension. We build a FAISS Douze et al. [2024] index for efficient similarity search (Figure 1.A). During inference, for a new ECG x_t sample, we compute $z_t = f(x_t)$ and retrieve k nearest neighbors using FAISS. Let $N_k(z_t) = (i_j, d_j)|j = 1, ..., k$ denote the set of k nearest neighbors, where i_j is the index of the j-th neighbor and d_j is the distance. We then retrieve the corresponding reports: $R_{\text{retrieved}} = r_{i_j}|(i_j, d_j) \in N_k(z_t)$.

For a test ECG embedding e_t, we assign the closest retrieved sample's report as an initial prediction, that can be optionally refined by an LLM. For zero-shot ECG question answering (Figure 1.B), we use the k retrieved reports and their diagnoses labels as a way to perform in-context learning. We concatenate these as: $r_{\text{concat}} = r_{i_1} \oplus l_{i_1} \oplus r_{i_2} \oplus l_{i_2} \oplus \cdots \oplus r_{i_k} \oplus l_{i_k}$, where $r_{i_j} \in R_{\text{retrieved}}$, l_{i_j} is the j-th report's diagnoses label, and $\oplus$ denotes concatenation. This concatenated input, the question, and a system prompt are passed to the LLM, instructing it to leverage the provided data to answer questions about the test ECG in a zero-shot manner.

For pretraining, we use the PTB-XL dataset Wagner et al. [2020] with 75% and 15% masking for MEM and MLM tasks, respectively. We use a batch size of 32 and a learning rate of 5×10^{-5}, keeping other hyperparameters consistent with Chen et al. [2022].We apply global max pooling over ECG encoder's output to get fixed-dimensional embeddings (z, $d = 768$). During indexing and retrieval, embeddings are L2 normalized. We set $k = 1$ for report generation and $k = 3$ for question-answering.Our ECG-QA LLM leverages GPT-4o (mini)OpenAI [2024], Gemini-Flash1.5Reid et al. [2024], and Llama3-70B Dubey et al. [2024] due to their cost-effectiveness. We set temperature $= 1$ and max_tokens $= 256$. Prompt details are provided in Appendix B.

3 Experiments

Datasets. We evaluate our report generation method on the PTB-XL Wagner et al. [2020] and MIMIC-IV-ECG Gow et al. [2023] datasets, both of which provide 12-lead ECG signals and corresponding textual reports. PTB-XL offers extensive cardiologist-annotated auxiliary information, including interpretive summaries, diagnostic assessments, likelihood estimates, and signal characteristics. MIMIC-IV-ECG provides machine-generated reports covering various conditions, which we combine into a single unified report. From this dataset, we randomly select 5k samples from the test set for cross-domain evaluation. For question answering, we utilize the ECG-QA Oh et al. [2024] dataset, which comprises curated questions about key ECG aspects. We experiment with verify, choose, and query question types, utilizing the provided train, test, and validation splits based on patient IDs to ensure no overlap among sets.

Baseline and Evaluation Metrics. We evaluate our approach using standard metrics for language fluency and accuracy. For report generation, we employ NLG metrics: BLEU-1,2,3,4 Papineni et al. [2002], BERTScore Zhang et al. [2019], Meteor Denkowski and Lavie [2011], and Rouge Lin [2004] to assess lexical and semantic similarity. We compare our method against: a) randomly selected reports, b) the most frequent report, c) vanilla transformer d) R2GenCMN Chen et al. [2020] adapted for ECG signals with randomly initialized (R) and pre-trained (P) ECG encoders. For zero-shot ECG-QA, we use exact match accuracy, comparing against six supervised baselines from Oh et al. [2024] specifically trained for QA task.

Results. The main experimental results on report generation task are presented in Table 1. The results demonstrate the effectiveness of ECG-ReGen across all evaluated metrics. On the PTB-XL dataset, it achieves substantial improvements over the R2GenCMN technique with pre-trained model, with a

Table 1: Performance comparison of various methods for ECG report generation task.

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BERTScore	Meteor	Rouge
PTB-XL (in-domain)	Random	0.152	0.091	0.060	0.049	0.582	0.226	0.234
	Common	0.212	0.150	0.134	0.132	0.663	0.317	0.385
	Transformer	0.273	0.179	0.119	0.093	0.626	0.292	0.303
	R2GenCMN (R) Chen et al. [2021]	0.337	0.249	0.188	0.159	0.687	0.362	0.392
	R2GenCMN (P) Chen et al. [2021]	0.393	0.298	0.232	0.196	0.710	0.412	0.443
	ECG-ReGen (Ours)	0.801	0.768	0.737	0.700	0.920	0.836	0.836
MIMIC-IV-ECG (cross-domain)	Random	0.113	0.055	0.031	0.023	0.579	0.185	0.201
	Common	0.127	0.048	0.038	0.033	0.651	0.283	0.399
	Transformer	0.229	0.137	0.082	0.054	0.598	0.250	0.240
	R2GenCMN (R) Chen et al. [2021]	0.305	0.215	0.133	0.084	0.677	0.346	0.378
	R2GenCMN (P) Chen et al. [2021]	0.325	0.224	0.141	0.091	0.672	0.360	0.383
	ECG-ReGen (Ours)	0.348	0.271	0.212	0.182	0.714	0.419	0.439

Table 2: Evaluation of ECG-ReGen paired with an off-the-shelf LLM on the ECG-QA dataset against supervised approaches.

Method	Model	S-Verify	S-Choose	S-Query
per Q-type majority	-	67.7	31.2	23.2
Supervised Oh et al. [2024]	M3AE	74.6	57.1	41.0
	MedViLL	73.9	54.1	40.4
	Fusion Transformer	72.1	46.4	37.4
	Blind Transformer	67.7	31.0	24.0
	Deaf Transformer	67.3	31.4	27.0
ECG-ReGen (Ours)	Gemini Flash 1.5	72.54	58.52	32.57
	GPT-4o mini	72.79	58.66	30.56
	Llama3-70B	71.99	54.83	32.02

BLEU-4 score of 0.700 compared to 0.196, and a BERTScore of 0.920 versus 0.710. This significant performance gain highlights the superiority of our retrieval-augmented approach in capturing and generating accurate ECG reports. The cross-domain evaluation on MIMIC-IV-ECG shows that our method maintains its superior performance, albeit with a smaller margin, achieving a BLEU-4 score of 0.182 and a BERTScore of 0.714. This robustness in cross-domain scenarios underscores the generalizability of our method.

Notably, ECG-ReGen consistently outperforms both random and common baselines, as well as the R2GenCMN variants, across all metrics on both datasets. The high ROUGE and METEOR scores further indicate that our method generates reports with better content coverage and semantic similarity to ground truth reports. Further, Figure 2 and 3 in Appendix A show qualitative results of retrieving similar examples with reports that closely match the ground truth, capturing key diagnostic features. These results collectively demonstrate the efficacy and simplicity of leveraging self-supervised representations for retrieval-based report generation. Our approach not only produces high-quality reports but also enables transparency by allowing clinicians to inspect and compare the generated report with similar examples.

Table 2 showcases our ECG-ReGen approach's performance on three ECG question-answering tasks. Operating in a zero-shot setting, our method demonstrates competitive performance against supervised approaches. ECG-ReGen, paired with Gemini Flash 1.5 and GPT-4o mini, achieves top scores on the S-Choose task (58.52% and 58.66%), surpassing all supervised models. For S-Verify, our approach performs comparably to the best supervised model (M3AE), with scores ranging from 71.99% to 72.79%. While showing lower performance on the S-Query task, it still outperforms the per Q-type majority baseline and some supervised models. These results are notable given our method's zero-shot nature, requiring no task-specific fine-tuning. The small performance gap between different LLMs suggests that the choice of language model may not be crucial when the retrieved samples are similar to the test case, highlighting the effectiveness of our retrieval-augmented approach in leveraging off-the-shelf language models for ECG analysis.

4 Conclusions

This work introduces a novel retrieval-based method for ECG report generation and question answering, leveraging self-supervised pre-training, efficient similarity search, and LLM-powered zero-shot question answering. Our approach demonstrates superior performance in both in-domain and cross-domain evaluations for report generation task, showcasing improved efficiency, inherent explainability, and enhanced generalizability. By integrating LLMs for zero-shot QA, we further augment the system's capabilities, offering a promising avenue for accurate ECG interpretation with potential benefits for cardiologist workflow and patient care.

References

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, 2021.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Ana Mincholé and Blanca Rodriguez. Artificial intelligence for the electrocardiogram. *Nature medicine*, 25(1):22–23, 2019.

Mirja Mittermaier, Marium M Raza, and Joseph C Kvedar. Bias in ai-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine*, 6(1):113, 2023.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36, 2024.

OpenAI. Gpt-4o (mini). `https://openai.com`, 2024. [Large language model].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

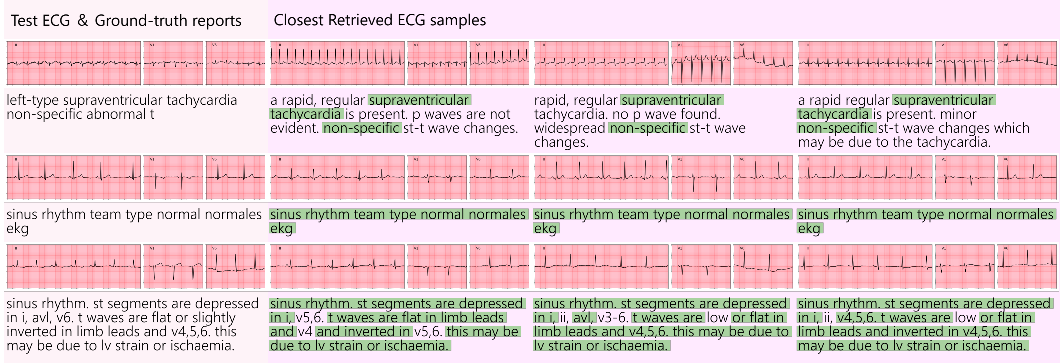
Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Zhenwu Peng, Jie Fu, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. Electrocardiogram instruction tuning for report generation. *arXiv preprint arXiv:2403.04945*, 2024.

Benjamin Yan, Ruochen Liu, David E Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P O'Connell, Agustina Saenz, Pranav Rajpurkar, et al. Style-aware radiology report generation with radgraph and few-shot prompting. *arXiv preprint arXiv:2310.17811*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

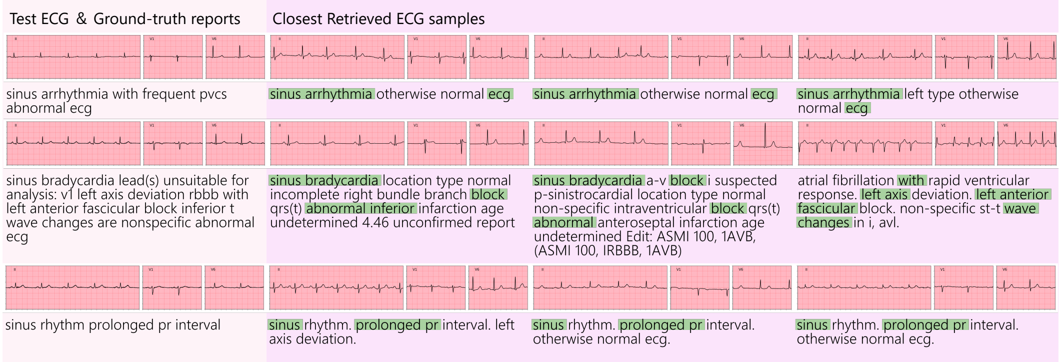
A Qualitative Analysis

We conduct a qualitative analysis of ground-truth and the top three retrieved reports along with their ECG signals.



Test ECG & Ground-truth reports	Closest Retrieved ECG samples		
left-type supraventricular tachycardia non-specific abnormal t	a rapid, regular supraventricular tachycardia is present. p waves are not evident. non-specific st-t wave changes.	rapid, regular supraventricular tachycardia. no p wave found. widespread non-specific st-t wave changes.	a rapid regular supraventricular tachycardia is present. minor non-specific st-t wave changes which may be due to the tachycardia.
sinus rhythm team type normal normales ekg	sinus rhythm team type normal normales ekg	sinus rhythm team type normal normales ekg	sinus rhythm team type normal normales ekg
sinus rhythm. st segments are depressed in i, avl, v6. t waves are flat or slightly inverted in limb leads and v4,5,6. this may be due to lv strain or ischaemia.	sinus rhythm. st segments are depressed in i, v5,6. t waves are flat in limb leads and v4 and inverted in v5,6. this may be due to lv strain or ischaemia.	sinus rhythm. st segments are depressed in i, ii, avl, v3-6. t waves are low or flat in limb leads and v4,5,6. this may be due to lv strain or ischaemia.	sinus rhythm. st segments are depressed in i, ii, v4,5,6. t waves are low or flat in limb leads and inverted in v4,5,6. this may be due to lv strain or ischaemia.

Figure 2: PTB-XL



Test ECG & Ground-truth reports	Closest Retrieved ECG samples		
sinus arrhythmia with frequent pvcs abnormal ecg	sinus arrhythmia otherwise normal ecg	sinus arrhythmia otherwise normal ecg	sinus arrhythmia left type otherwise normal ecg
sinus bradycardia lead(s) unsuitable for analysis: v1 left axis deviation rbbb with left anterior fascicular block inferior t wave changes are nonspecific abnormal ecg	sinus bradycardia location type normal incomplete right bundle branch block qrs(t) abnormal inferior infarction age undetermined 4.46 unconfirmed report	sinus bradycardia a-v block i suspected p-sinistrocardial location type normal non-specific intraventricular block qrs(t) abnormal anteroseptal infarction age undetermined Edit: ASMI 100, 1AVB, (ASMI 100, IRBBB, 1AVB)	atrial fibrillation with rapid ventricular response. left axis deviation. left anterior fascicular block. non-specific st-t wave changes in i, avl.
sinus rhythm prolonged pr interval	sinus rhythm. prolonged pr interval. left axis deviation.	sinus rhythm. prolonged pr interval. otherwise normal ecg.	sinus rhythm. prolonged pr interval. otherwise normal ecg.

Figure 3: MIMIC-IV ECG

B LLM System Prompt

For *question prompts*, we use variations for different answer formats: "Answer should be only yes and no" (single-verify), "Choose correct answer from the options provided below. If none... answer will be 'none'.If both conditions are present then provide both options as an answer." (single-choose), and a similar prompt for single-query allowing multiple answer selections. Answer options are always provided within the prompt.

Prompt for *Zero-Shot* ECG-QA Task.

Given the closest ECG retrieved reports and diagnoses to the test ECG as discovered by a multimodal model. Your job is to analyze the report and only answer the question. Output should be JSON of the following structure: {'answer': ...}.
Question Specific Prompt: `${question_prompt}`
Think step-by-step to generate an answer without any explanation.
ECG Reports: `${reports}`
Diagnoses: `${diagnoses}`
Question: `${question}`