

# Q-HEART: ECG Question Answering via Knowledge-Informed Multimodal LLMs

Hung Manh Pham<sup>a</sup>, Jialu Tang<sup>b</sup>, Aaqib Saeed<sup>b</sup> and Dong Ma<sup>a,\*</sup>

<sup>a</sup>Singapore Management University  
<sup>b</sup>Eindhoven University of Technology

**Abstract.** Electrocardiography (ECG) offers critical cardiovascular insights, such as identifying arrhythmias and myocardial ischemia, but enabling automated systems to answer complex clinical questions directly from ECG signals (ECG-QA) remains a significant challenge. Current approaches often lack robust multimodal reasoning capabilities or rely on generic architectures ill-suited for the nuances of physiological signals. We introduce Q-HEART, a novel multimodal framework designed to bridge this gap. Q-HEART leverages a powerful, adapted ECG encoder and integrates its representations with textual information via a specialized ECG-aware transformer-based mapping layer. Furthermore, Q-HEART leverages dynamic prompting and retrieval of relevant historical clinical reports to guide tuning the language model toward knowledge-aware ECG reasoning. Extensive evaluations on the benchmark ECG-QA dataset show Q-HEART achieves state-of-the-art performance, outperforming existing methods by a 4% improvement in exact match accuracy. Our work demonstrates the effectiveness of combining domain-specific architectural adaptations with knowledge-augmented LLM instruction tuning for complex physiological ECG analysis, paving the way for more capable and potentially interpretable clinical patient care systems.

## 1 Introduction

Electrocardiograms (ECGs) are non-invasive, cost-effective diagnostic tools that play a pivotal role in detecting cardiac arrhythmias and other cardiovascular abnormalities in clinical practice [40, 20, 15, 25]. A standard clinical ECG typically involves 12 leads, each capturing electrical activity from different anatomical perspectives of the heart, offering a comprehensive view of cardiac function. Recently, single or few-lead ECG acquisition has also become commonly used in wearables, facilitating large-scale, continuous monitoring in both clinical and consumer settings. As the volume and accessibility of ECG data increase, there is a pressing demand for intelligent systems that can interpret these data and assist clinicians or patients in understanding cardiac health. Accordingly, the need for advanced systems that can respond to clinical questions about ECGs has become more essential. ECG-based question answering (ECG-QA) aims to address this challenge by enabling automated systems to generate precise and clinically relevant answers to questions derived from ECG data.

In parallel with these developments, large language models (LLMs) have rapidly emerged as foundational tools in artificial intelligence,

demonstrating impressive capabilities in text generation, understanding, and reasoning. By leveraging their vast parameter spaces and pretraining on diverse datasets, LLMs can encode a wealth of knowledge, enabling them to handle complex reasoning and produce human-like responses. Recent models such as Med-Gemini [28], BioMedGPT [17], and AMIE [23] have achieved state-of-the-art performance across a variety of medical tasks. While LLMs have primarily been applied to textual data, the integration of non-linguistic inputs (particularly information-rich physiological signals like ECGs) is an emerging research direction. This leads directly to the central question of our work: *can LLMs be effectively adapted to ground their reasoning in ECG signals to accurately answer clinical questions in an ECG-QA setting?*

Despite the promise of LLMs in healthcare, their specific application to interactive ECG-QA remains nascent, facing distinct and underexplored challenges. Specifically, existing efforts have mainly focused on related applications such as ECG report generation [41, 32, 30] and retrieval-augmented diagnosis [36], which, although valuable, fall short of enabling fine-grained, interactive question answering. Early works like ECG-QA [22] pioneered this direction by curating a benchmark dataset and corresponding baselines, but showing limitations or underperformance in the weak ECG encoder and LLM usage, hinder their reasoning capabilities. Recent approaches, such as ECG-LM [35] and Med-Gemini [28], have advanced ECG-text alignment through multimodal architectures. Nonetheless, they often require specialized training data (e.g., personalized information or GPT-generated templates), depend heavily on large proprietary models, and do not fully leverage efficient instruction tuning, automatic retrieval augmentation, or ECG lead-specific knowledge integration to achieve better multimodal reasoning in clinical QA settings.

In this paper, we introduce Q-HEART, a novel LLM-based multimodal framework designed to address the emerging task of ECG-based question answering (ECG-QA). Q-HEART integrates ECG signal representations with textual embeddings through a specially designed transformer mapping layer, enabling accurate and contextually relevant answers to clinical questions. By leveraging carefully crafted instruction tuning, supported by retrieval-based historical reports from a large-scale database with dynamic prompting techniques, our approach enhances multimodal alignment, reasoning capabilities, and shows potential interpretability in ECG-QA tasks. We evaluate Q-HEART on the benchmark dataset, achieving state-of-the-art performance in exact-match accuracy and significant improvements in natural language generation metrics. The contributions of this work can be summarized as follows:

\* Corresponding Author. Email: dongma@smu.edu.sg  
 Preprint

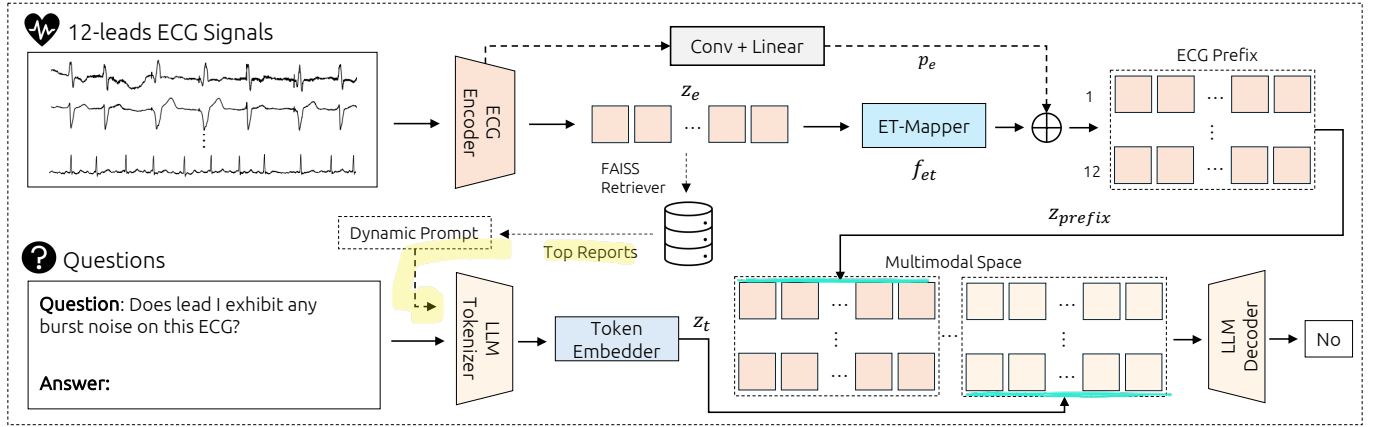


Figure 1: Overview of our proposed Q-HEART framework.

- We present Q-HEART, a novel framework for ECG question answering (ECG-QA) that synergistically integrates adapted deep ECG encoding, specialized multimodal mapping, and knowledge-informed LLM instruction tuning. Q-HEART establishes a new state-of-the-art on the benchmark ECG-QA task.
- We design the ET-Mapper, a novel ECG-aware transformer mapping layer, specifically developed to effectively bridge structural ECG signal representations with the LLM’s textual embedding space, noticeably enhancing multimodal alignment.
- We incorporate retrieval augmentation using historical clinical reports, sourced via efficient similarity search, to inject crucial contextual knowledge into the LLM’s reasoning process for ECG-QA, notably without requiring patient-specific data or restrictive templates.
- We introduce dynamic prompting strategies for instruction tuning, including randomized answer option shuffling and selective use of retrieved reports, demonstrably improving model robustness and overall performance.

## 2 Related Works

### 2.1 ECG Representation Learning

Over the years, the development of ECG representation learning has played a crucial role in enabling a wide range of downstream tasks such as arrhythmia classification and clinical report generation. In the context of ECG-based question answering, learning robust and expressive signal representations is also essential, as it directly affects the model’s ability to reason over physiological patterns and provide accurate answers. In this section, we present prior approaches based on self-supervised learning (SSL), which have proven to be highly effective in medical domains where labeled data is often scarce or expensive to obtain. We focus on two primary streams of work: unimodal ECG representation learning and multimodal ECG-text learning.

Unimodal ECG representation learning focuses on extracting informative signal-level features directly from raw ECG waveforms, without relying on auxiliary modalities. One line of early SSL research approaches this by training autoencoder-based models to reconstruct masked or corrupted input signals, thereby encouraging the model to capture essential morphological patterns and temporal dependencies [40, 9, 20]. In contrast, another branch of SSL methods emphasizes contrastive objectives, aiming to maximize the agreement be-

tween augmented views of the same ECG instance. These approaches employ domain-specific augmentation strategies. For instance, [5] introduces physiologically motivated transformations such as vectorcardiogram (VCG) projections to capture the three-dimensional spatiotemporal characteristics of cardiac activity. Additionally, [21] proposes to use contrastive-based pretraining, Contrastive Multi-Segment Coding (CMSC) [12], and random lead masking to simulate diverse local and global lead configurations.

With the rise of multimodal learning, recent attempts have focused on aligning ECG signals with textual data, enabling zero-shot interaction across modalities for tasks like cardiac condition classification or ECG-report retrieval. Several works adopt contrastive learning frameworks such as ETP [16], MERL [15], ESI [37], and others [13], with common architectures like ResNet or ConvNeXt for ECG and BERT-based models for text encoding, which demonstrate effective cross-modal alignment. More recently, C-MELT [25] introduces a hybrid self-supervised approach that hybridly combines contrastive learning with masked autoencoding, achieving state-of-the-art performance across a range of downstream ECG diagnosis tasks.

### 2.2 ECG Report Generation

Although our focus is on ECG-based question-answering, related works in text generation, such as diagnosis reports, share some similar goals of interpreting ECG data for clinical insights. Accordingly, several recent approaches have emerged in this domain. For example, [1] proposes an encoder-decoder framework for ECG report generation, where ECG signals are processed using a ResNet encoder and decoded via LSTM or Transformer to produce clinical interpretations. [26] further uses a multimodal ECG interpretation framework that encodes ECG signals as images and aligns them with diagnostic reports using a vision-language learning paradigm. Next, a recent work, ECG-Chat [41] introduces a multimodal framework leveraging contrastive learning to align ECG waveforms with diagnostic reports. This model is supported by a large instruction dataset and an innovative LaTeX report pipeline to improve user clarity and usability. Similarly, MEIT [32] proposes large language instruction-tuned models to address the complexities of ECG report generation. By introducing an attention-based fusion mechanism for ECG and text, MEIT achieves effective representational alignment between ECG signals and text. Finally, [30] presents ECG-ReGen, a retrieval-augmented framework that combines ECG encoders with LLMs for zero-shot report generation.

### 2.3 ECG Question Answering

ECG-QA is a growing field that offers meaningful benefits, such as improving clinical decision-making and enhancing interactive patient care. Although it is still in its early stages, several studies have demonstrated promising results. For example, ECG-QA work [22] is the first significant contribution to this domain, comprising a novel dataset with ECG recordings, questions, and answers across diverse types, such as single-verify, single-choose, and single-query formats. The study also provides benchmarking results using various models in a classification manner (e.g., leveraging M<sup>3</sup>AE [2], SE-WRN[7], and common OpenAI model APIs), highlighting the potential for advancing ECG question-answering research. Following this, ECG-LM [35] also built upon the ECG-QA dataset by developing a multimodal large language model that integrates textual and ECG signal data. The framework employs a ResNET-18-based ECG encoder [11] and usage of BioMedGPT-LM-7B [17] for text generation given detailed prompts, which overall shows better performance than the previous ECG-QA baselines. However, it relies on rich patient-related metadata (e.g., age, gender, weight, height) and exact clinical diagnosis reports from current examined datasets.

Further, Med-Gemini [28] presents a comprehensive multimodal large language model capable of handling diverse medical tasks, including ECG-based QA. Although Med-Gemini shows state-of-the-art performance with its proposed instructed Gemini model, its methodology remains vague, especially for ECG embedding processing and ECG-QA instruction tuning, with limited implementation details disclosed. More recent works, like [33, 29] show their interesting ideas on ECG question-answering developments, such as a novel diverse dataset for evaluating diagnostic capabilities in ECG interpretation or a multimodal meta-learning framework for few-shot ECG QA, although they focus on their specific tasks and evaluations.

Despite progress on the task, existing works face key limitations. First, they often rely on relatively weak ECG encoders, limiting physiological generalization. Second, the existing ECG-QA models lack development with ECG-aware knowledge in aligning ECG and text with a mapping architecture, which could strengthen multimodal ECG-text learning and decoding. Lastly, current approaches remain largely unexplored in the available usages of historical reports and textual diagnoses in LLM instruction tuning, leaving gaps in interpretability and critical reasoning capabilities for clinical trust.

## 3 Methods

Our approach, Q-HEART, introduces a multimodal framework specifically designed to enable LLMs to perform complex question-answering directly from ECG signals. As illustrated in Figure 1, the core idea is to bridge the gap between physiological signal processing and natural language reasoning. First, Q-HEART processes the input ECG using a powerful, adapted deep encoder (Section 3.1) to capture condensed and informative ECG features. Simultaneously, to provide essential clinical context, relevant historical reports associated with similar ECGs are retrieved from a large-scale database (also Section 3.1). The crucial step of multimodal integration is handled by our proposed ET-Mapper (Section 3.2), an ECG-aware transformer architecture that transforms the ECG embeddings into a format compatible with the LLM’s textual input space. These aligned ECG representations, along with the retrieved reports and the user’s question, are then formulated into a dynamic instruction prompt (Section 3.3). Finally, an instruction-tuned LLM decoder processes this fused input to generate a contextually relevant and accurate answer. The following sections delve into each key component of our model.

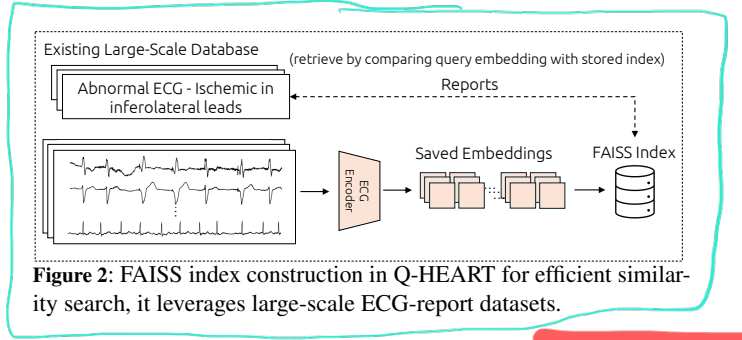


Figure 2: FAISS index construction in Q-HEART for efficient similarity search, it leverages large-scale ECG-report datasets.

### 3.1 ECG Encoder and Report Retrieval

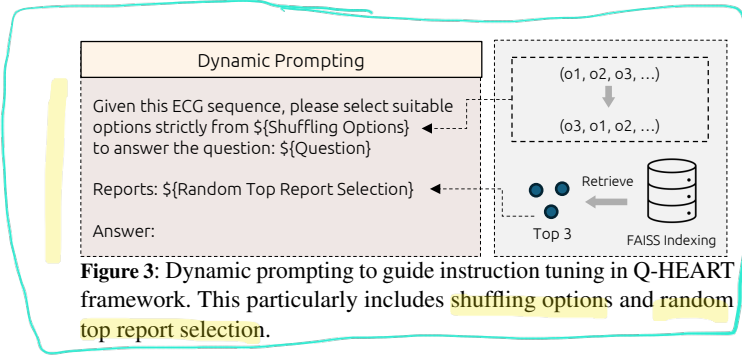
The ECG encoder in our model (denoted as  $f_e$ ) is adapted from C-MELT’s ECG encoder architecture [25], designed for robust cross-modal ECG-text representation learning, which has been shown to perform notably well on various ECG downstream tasks, covering more than 100 cardiac conditions. Their ECG encoder employs multiple convolutional layers with GELU activation and group normalization for efficient initial feature extraction. These are followed by positional encoding and 8 transformer layers, enabling robust sequential learning of ECG signals. Ultimately, the features are projected into a  $d = 768$  dimensional space for efficient downstream tasks. Specifically for ECG signals, their encoder utilizes masking strategies such as random lead masking and segment dropout. Overall, their framework is enriched by contrastive modeling tasks in a multimodal learning manner using SigLIP-based loss [38] and search-based negative sampling. In our pretraining step, we adopt and train the same design but extend it to 12 transformer layers. We also incorporate random lead masking with a high probability ( $p = 0.65$ ) of masking each lead, motivated by the need that many questions pertain to a single or a few leads, encouraging the model to focus more on learning arbitrary lead-specific representations useful for downstream tasks. Ultimately, the trained ECG encoder is incorporated into the Q-HEART framework as in Figure 1. We present the performance of the adapted ECG encoder (referred to as C-MELTv2) in Section 4.4.1.

To enable flexible clinical context and interpretability in the ECG-QA task, we propose an automatic information retrieval-based augmentation mechanism that consists of two stages: index construction and query-time retrieval. During the index construction phase, we leverage a large-scale ECG-report paired dataset such as MIMIC-IV ECG [6]. Each ECG signal in the dataset is encoded using our ECG encoder  $f_e$ , and the resulting embeddings are stored in a FAISS index [4], with each entry linked to its corresponding clinical report (Figure 2). This enables scalable similarity search across a vast corpus of ECG data. At query time, the input ECG  $x_e$  is encoded into a dense vector  $z_e = f_e(x_e) \in \mathbb{R}^d$ , which is then used to retrieve the top-3 most similar embeddings based on cosine similarity. The textual reports associated with these retrieved entries (denoted  $\mathcal{R}_3$ ) are returned and incorporated into the dynamic prompting strategy. This design not only strengthens the contextual grounding of the LLM but also introduces external clinical knowledge that enhances answer accuracy and model interpretability.

### 3.2 ET Mapper

To bridge the modality gap between ECG signals and textual data, we introduce a transformer-based mapping called ET-Mapper. This component is responsible for transforming the  $z_e$  into a prefix ECG embedding that is compatible with LLM token embeddings  $z_t$ . Formally, given an ECG embedding  $z_e$ , the ET-Mapper learns a mapping





function  $f_{et} : \mathbb{R}^d \rightarrow \mathbb{R}^{c \times d'}$  that projects  $\mathbf{z}_e$  into a sequence of  $c = 12$  channel embeddings, each of dimension  $d'$  (dimension of LLM token embeddings). The choice of  $c = 12$  directly corresponds to the 12 standard leads in a clinical ECG, which aims to produce the prefix tokens that structurally reflect the inherent spatial diversity of ECG signals. More specifically, the transformation begins by projecting  $\mathbf{z}_e$  through a fully connected linear layer, which reshapes it into an initial prefix sequence of shape  $\mathbb{R}^{c \times d'}$ . This sequence is then refined by a lightweight transformer architecture comprising  $L = 2$  layers with  $H = 4$  attention heads. The transformer enables the deeper modeling of ECG features within the prefix sequence, allowing the final output  $\mathbf{z}_{et} = f_{et}(\mathbf{z}_e)$  to serve ECG-text alignment during LLM-based decoding.

However, at this stage, the lead positional information from individual ECG leads turns out to be unclear (due to the nature of the global characteristics from  $\mathbf{z}_e$ ). Since clinical questions might target specific leads of the ECG signal, it is essential to retain and inject lead-aware spatial structure into the prefix representation. To address this, we incorporate an auxiliary skip connection from the ECG encoder’s positional embedding layer into the ET-Mapper. Specifically, we extract the positional lead representation from the encoder (denoted as  $\mathbf{p}_e \in \mathbb{R}^{c \times d'}$ ) via a lightweight convolutional layer followed by a linear projection. The result  $\mathbf{p}_e$  is then added element-wise to  $\mathbf{z}_{et}$ , and the final mapped prefix embedding is updated as  $\mathbf{z}_{prefix} = \mathbf{z}_{et} + \mathbf{p}_e$ . This additive fusion enriches the ECG prefix embedding with localized lead-awareness, ensuring that the language model receives input conditioned not only on the overall ECG context but also on structurally lead-specific cues.

### 3.3 Instruction Tuning

To empower Q-HEART with relevant reasoning and strong clinical text generation capabilities, we adopt instruction tuning on a pre-trained large language model, specifically leveraging the LLaMA-Instruct 1B [18]. Despite its compact size, LLaMA-Instruct offers robust general-purpose instruction-following skills, making it a suitable foundation for our resource-efficient setup. Our tuning specifically follows a simple yet effective prompting strategy as illustrated in Figure 3, with an emphasis on two dynamic components that vary across training steps (denoted *DP*): (1) To prevent position bias and encourage robust understanding of candidate options, we randomly shuffle the order of multiple-choice options ( $o$ ) during training. This ensures the model learns to focus on the semantic content rather than memorizing positional patterns. We present our strategies to specify the possible options for the task in Table 1. (2) We randomly select one of the top-3 retrieved reports  $\mathcal{R}_3$  (as described in Section 3.1), at every training step. This promotes diverse evidence grounding and prevents overfitting to fixed report-query pairs.

Technically, during the forward pass, the prompt-formatted question is first tokenized into a sequence  $\mathbf{x}_t$ , which is then passed through the LLM’s token embedding layer  $f_t$  to obtain token embeddings  $\mathbf{z}_t = f_t(\mathbf{x}_t) \in \mathbb{R}^{l \times d'}$ , where  $l$  denotes the number of tokens. Subsequently, we concatenate those text embeddings with previously-defined ECG prefix as a single embedding  $\mathbf{z}_{fuse} = \text{concat}(\mathbf{z}_{prefix}, \mathbf{z}_t) \in \mathbb{R}^{(c+l) \times d'}$ . The fused embedding is subsequently fed into the LLM decoder to perform the task of answer generation. Here, we fine-tune instead of freezing LLaMA-Instruct 1B, as it naturally exhibits general-context reasoning capabilities but might lack rich domain-specific grounding in biomedical scenarios such as ECG reports and cardiology-related QA tasks. Here, we use LoRA [8] method with  $r = 8$ ,  $\alpha = 32$ , and dropout = 0.1 settings, while employing standard cross-entropy loss in autoregressive language modeling.

## 4 Experiments

### 4.1 Datasets

In our study, we use four public datasets, including (1) MIMIC-IV-ECG v1.0 [6], PTB-XL [31], (3) CODE-15 [27], and (4) ECG-QA [22]. Here, ECG-QA is the main dataset to fine-tune our proposed model on the question-answering task as the main experiment. Below, we briefly describe each dataset and its specific role in our framework.

**MIMIC-IV-ECG v1.0.** This includes 800,035 paired samples from 161,352 unique subjects with numerous 10-second ECG recordings sampled at 500 Hz and the corresponding text reports. We use this large-scale dataset to pretrain ECG-text multimodal model and build FAISS indexing construction, which eventually aims for an effective ECG encoder and contextual report retrieval, respectively, for the ECG-QA task.

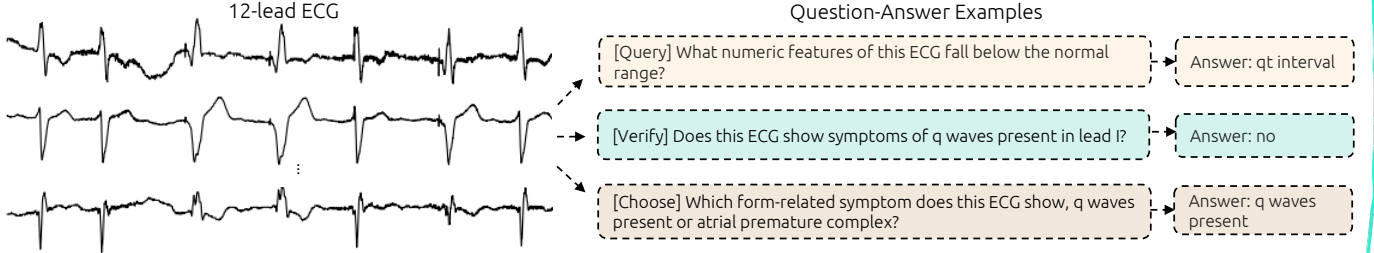
**PTB-XL.** The PTB-XL dataset includes 21,837 ECG signals collected from 18,885 patients. Each sample has a 12-lead ECG recording sampled at 500 Hz over 10 seconds and corresponding cardiac labels. Particularly, this dataset is commonly split into four sub-datasets (super - 5 classes, sub - 23 classes, form - 19 classes, and rhythm - 12 classes). Following the prior works [25, 15, 31], we have processed the test set of each group (2158, 2158, 880, 2098 samples, respectively), which can be used to validate the zero-shot performance of the trained ECG encoder.

**CODE-15.** We also validate the trained ECG encoder by using the test dataset of the CODE-15 dataset. Specifically, it contains 827 12-lead ECG samples of various lengths and a sampling rate of 400 Hz, covering 6 classes. Here, we ensure the processed signals are resampled at 500 Hz and have a length of 10 seconds.

**ECG-QA.** Finally, this specialized dataset comprises curated questions and corresponding answers about various crucial cardiac aspects on the PTB-XL ECG samples. We focus on single-verify, choose, and query question types, yielding 159306, 31137, and 41093 samples for train, valid, and test sets. We present a more detailed description of this dataset as shown in Table 2. Particularly, we notice that many questions contains the word "lead" (approximately 60%), which shows that this ECG-QA dataset also focuses on medical patterns of certain leads, rather than just giving attention to overall 12-lead ECGs as global diagnoses. This motivates us to put more care into the ECG positional encoding, as designed in Section 3.2. Finally, we also provide an example of an ECG sample with its corresponding questions and answers in Figure 4.

**Table 1: Construction rules for candidate answer options based on question type and content.** These options are appended to the prompt during instruction tuning. Note that `<attribute list>` is provided in the dataset itself and also used by other existing methods.

Question Type	Condition	Candidate Options
Single-Query	Starts with "What leads"	lead I, lead II, lead III, lead aVR, lead aVL, lead aVF, lead V1, lead V2, lead V3, lead V4, lead V5, lead V6
	Starts with "What numeric features"	rr interval, p duration, pr interval, qrs duration, qt interval, qt corrected
	Starts with "What range"	below the normal range, within the normal range, above the normal range
	Other cases	<code>&lt;attribute list&gt;</code> , none
Single-Verify	Any	yes, no, not sure
Single-Choose	Any	<code>&lt;attribute list&gt;</code> , none



**Figure 4: Examples of three question-answer pairs for a given 12-lead ECG signal.**

**Table 2: ECG-QA data description** across three subsets (train, valid, test). For each set, we show the number of samples for each question type and, at the end, the percentage of questions that have 'lead' care out of the total number of questions.

Set	Query	Verify	Choose	'Lead' Question (%)
Train	46737	62554	50015	60%
Valid	11334	10718	9085	57%
Test	18157	13081	9855	58%

they utilize large language models, and we categorize them into four groups accordingly. (1) *API*: These methods leverage off-the-shelf LLMs, including SE-WRN [42] paired with GPT-4, GPT-3.5-turbo, and text-davinci-003 [22] and ECG-ReGen [30], where ECG signals are first converted into descriptive prompts and then fed into LLMs for answer prediction. (2) *None*: The methods do not employ LLMs at all and instead form the ECG-QA task as a classification task following strategies introduced in [22], such as M<sup>3</sup>AE [2], MedViLL [19], the Fusion Transformer, the Blind Transformer (text-only), and the Deaf Transformer (ECG-only). (3) *Frozen*: Med-Gemini [28] and ECG-LM [35], which align ECG signals and textual inputs within a shared space while retaining the original LLM parameters. (4) *Tuning*: including our proposed Q-HEART and also Med-Gemini in its fine-tuned variant, adapting LLM parameters via instruction tuning to improve performance.

## 4.2 Training Details

For the ECG encoder pretraining, we follow the training configurations in C-MELT report. For Q-HEART in detail, the model is fine-tuned using the AdamW optimizer with a learning rate of  $5 \times 10^{-5}$ , employing a cosine learning rate scheduler with a 0.1 warmup ratio. The optimizer is configured with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$ , an epsilon value of  $1 \times 10^{-6}$ , and a weight decay of 0.01. The model is trained for only 5 epochs with a batch size of 32 on a single NVIDIA H100-80GB GPU. During training, the best-performing model corresponding to the lowest validation loss is saved for later evaluation.

## 4.3 Baselines and Metrics

### 4.3.1 Baselines

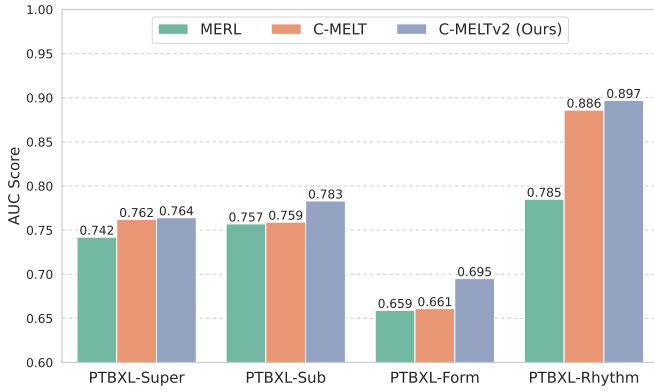
To comprehensively evaluate the performance of our proposed approach, we benchmark Q-HEART against a diverse set of baseline models spanning multiple design paradigms.

First, for validating the trained ECG encoder, we compare it with two state-of-the-art ECG encoders, *i.e.*, C-MELT [25] and MERL [15], which are well-pretrained on large-scale ECG-text datasets and have demonstrated strong performance across a variety of downstream cardiovascular tasks. These comparisons help isolate the contribution of our ECG representation module and justify its integration within our broader pipeline.

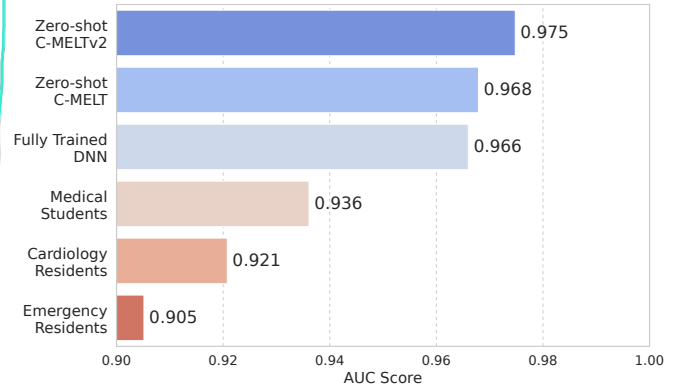
For the main task of ECG-QA, the evaluation is carried out under two regimes: using the entire testing data of ECG-QA dataset, and using only 10% of it to assess robustness in low-resource settings [22]. More importantly, we note that the baselines differ in how

### 4.3.2 Evaluation Metrics

In our work, we use six primary metrics for evaluation: Macro AUC score (0-1), Exact-Match Accuracy (EM-Acc, 0-1), BERTScore [39], ROUGE [14], METEOR [3], and BLEU-1 [24]. Here, we use the macro AUC score in validating ECG encoder effectiveness, as followed by prior works [25, 15], while EM-Acc serves as the main performance metric for comparing our model with existing works [22, 35, 28], as they only report this standard metric with strict correctness of the generated answers by comparing to the ground truth. However, EM-Acc alone may not fully capture the model's reasoning quality or partial correctness, especially in complex query-type questions. For instance, a response that conveys the correct diagnosis using slightly different phrasing would still be penalized under EM-Acc, despite being clinically valid. Therefore, we complement this metric with natural language generation (NLG) metrics (BERTScore, ROUGE, METEOR, and BLEU-1) which offer deeper insight into the semantic alignment, lexical overlap, and fluency of generated answers. We use these metrics in our ablation studies, where finer-grained differences in model outputs need to be interpreted beyond binary correctness.



**Figure 5:** Zero-shot performance of different ECG encoders on the PTB-XL dataset.



**Figure 6:** Performance of human experts vs. C-MELTv2 (ours) and other ECG encoders on the CODE-15 dataset.

## 4.4 Quantitative Results

### 4.4.1 ECG Encoder Validation

It is essential to have a strong ECG encoder to extract informative representations for the downstream task of multimodal ECG-based question answering. In this section, we conduct a comprehensive validation of our encoder’s representational quality. As described in section 3.1, we trained entirely from scratch without relying on any pretrained weights, so this necessitates a verification of the downstream performance. Specifically, we design zero-shot classification experiments, following [25] on the CODE-15 and PTB-XL datasets to test generalization without any fine-tuning. While evaluating on the PTB-XL dataset is important because it represents the signal data distribution that is used in the ECG-QA dataset, CODE-15 also provides useful insights as it allows us to compare the performance between models and human experts.

As can be seen in Figure 5, our trained ECG encoder (C-MELTv2) achieves the best performance, averaging of 78.5% (AUC) over four PTB-XL sub-datasets, higher than the original C-MELT and MERL by nearly 2% and 5%, respectively. These improvements not only validate the original design from C-MELT but also benefit from enhancements such as a stronger transformer decoder and lead masking strategies. C-MELTv2 also shows its superior performance in the CODE-15 dataset evaluation, as shown in Figure 6. Specifically, its zero-shot result is slightly better than the original C-MELT’s and fully trained DNN model [27], but still show clear improvement (4%-7%) over human experts like medical students or in-domain residents. This further confirms the practical clinical relevance of our encoder in assisting or augmenting expert decision-making in real-world scenarios.

### 4.4.2 ECG Question Answering Performance

We present comprehensive benchmarking results comparing our proposed model, Q-HEART, with a variety of baseline approaches on the ECG-QA task, as shown in Table 3. Overall, Q-HEART outperforms all existing baselines across both evaluation settings, using the full test set and the 10% subset, achieving up to 4% and 3% higher average exact-match accuracy than the second-best models, respectively. This demonstrates Q-HEART’s robustness in both low-resource and full-data scenarios. When analyzing individual question types, Q-HEART consistently achieves the highest accuracy on both the verify and choose question types, which necessarily demand precise decision-making and fine-grained understanding. Notably, in the verify case under full test evaluation, Q-HEART reaches nearly

91% accuracy, showing substantial gains over M<sup>3</sup>AE (74.6%) and ECG-ReGen (72.5%). These improvements underscore the efficacy of our ECG-to-prefix alignment and the impact of dynamic prompting during proposed instruction tuning in enhancing the model’s capacity to process structured clinical queries.

Regarding the query questions, which remain the most challenging for all models, primarily due to their open-ended nature with a diversity of answer choices, Q-HEART still performs competitively. Accordingly, in the full test set evaluation, while Q-HEART surpasses methods such as Blind Transformer, Deaf Transformer, and ECG-ReGen, it is slightly outperformed by approaches like M<sup>3</sup>AE, MedViLL, and Fusion Transformer. However, it is important to emphasize that these competing models frame the task as a multi-answer classification problem [22], which reduces the complexity of the output space by relying on predefined label sets. In contrast, Q-HEART is designed for a more demanding generative modeling setting, where the model must produce free-form text responses conditioned on ECG and prompts. This setup requires more nuanced reasoning and sophisticated language generation, demonstrating the broader applicability and flexibility of our framework in real-world clinical question-answering scenarios. That being said, the performance gap in the query category is likely attributable to our use of a relatively compact and general-purpose model, LLaMA-Instruct 1B, which might lack prior exposure to biomedical texts or domain-specific fine-tuning. Alternatively, some counterpart models utilize larger or domain-specialized language backbones, such as BioMedGPT-LM-7B (in ECG-LM), which may provide stronger generative capabilities in complex ECG reasoning scenarios.

### 4.4.3 Ablation Studies

To gain deeper insights into our Q-HEART model, we conduct ablation studies detailed in Table 4. These experiments systematically omit or change a specific component, allowing a determination of their impact on overall performance. Overall, we can see that each component of Q-HEART contributes meaningfully to its final result. For example, when removing dynamic prompting (DP), the performance slightly drops, approximately 1–2% in EM-Acc and other metrics, while BERTScore remains relatively unchanged. This is expected, as techniques like option shuffling and random report selection contribute more to training robustness than to the overall semantic fidelity of the generated answers, which is what BERTScore primarily measures.

Following that, ablation results reveal more substantial performance



**Table 3:** Comparison of different LLM-based approaches on three question types: Verify, Choose, and Query (report in exact-match accuracy using the same test data configuration of the ECG-QA dataset).

Method	Test Data	LLM Usage	Verify	Choose	Query	Avg
SE-WRN+gpt-4[22]	10%	API	0.710	0.481	0.357	0.516
SE-WRN+gpt-3.5-turbo[22]	10%	API	0.693	0.361	0.311	0.455
SE-WRN+text-davinci-003[22]	10%	API	0.750	0.378	0.360	0.496
M <sup>3</sup> AE[2]	10%	None	0.760	0.582	0.400	0.581
ECG-LM[35]	10%	Frozen	0.758	0.574	0.399	0.577
Med-Gemini[28]	10%	Frozen	–	–	–	0.577
Q-HEART (Ours)	10%	Tuning	0.913	0.626	0.345	0.628
ECG-ReGen[30]	All	API	0.725	0.585	0.326	0.545
M <sup>3</sup> AE[2]	All	None	0.746	0.571	0.410	0.576
MedViLL[19]	All	None	0.739	0.541	0.404	0.561
Fusion Transformer	All	None	0.721	0.464	0.374	0.519
Blind Transformer (Seeing questions only)	All	None	0.677	0.310	0.240	0.409
Deaf Transformer (Seeing ECGs only)	All	None	0.673	0.314	0.270	0.419
Med-Gemini[28]	All	Tuning	–	–	–	0.584
Q-HEART (Ours)	All	Tuning	0.909	0.603	0.329	0.614

**Table 4:** Impact of model components (report average of three question types). We evaluate the models on 10% of testing data using the five metrics: EM-Acc, BERTScore, ROUGE, METEOR, and BLEU-1.

Setting	EM-Acc	BERTScore	ROUGE	METEOR	BLEU-1
W/o DP	0.616	0.844	0.714	0.495	0.729
W/o Retrieval Report	0.605	0.826	0.694	0.471	0.716
W/o ET-Mapper	0.600	0.833	0.701	0.486	0.726
W/o Pos-Encoder	0.582	0.815	0.669	0.461	0.696
Frozen LLM	0.583	0.820	0.677	0.453	0.694
Frozen ECG Encoder	0.587	0.829	0.687	0.481	0.715
Q-HEART (Ours)	0.628	0.845	0.727	0.502	0.744

degradation when removing key components related to modality alignment and clinical grounding. Specifically, removing the retrieved clinical report (W/o Retrieval Report) results in an average performance drop of approximately 2–3% across most metrics, underscoring the importance of incorporating external domain-specific knowledge to enrich the model’s contextual understanding. Similarly, eliminating the ET-Mapper (W/o ET-Mapper), which is responsible for transforming ECG embeddings into LLM-compatible prefix representations, yields a comparable decline in performance. This confirms the critical role of this module in bridging the modality gap between ECG signal features and language token representations, facilitating effective multimodal fusion during decoding.

Notably, a pronounced degradation is also observed when omitting the positional layer (W/o Pos-Encoder), which removes the skip connection that conveys lead-specific information from the ECG encoder into the transformer mapper. This configuration leads to a drop of up to 4% in EM-Acc and also substantial decreases in BERTScore (3%) and other generative metrics such as ROUGE (6%), METEOR (4%), and BLEU-1 (5%). The observed decline suggests that preserving positional leads, particularly in 12-lead ECGs, where each lead contributes unique spatial insights, is crucial for maintaining the semantic integrity of the generated answers. Without this lead-aware encoding, the model may struggle to differentiate among lead-specific features, potentially resulting in overfitting or entangled representations during prefix generation.

Similarly, we observe that freezing either the LLM or the ECG encoder leads to substantial performance degradation, with EM-Acc dropping by approximately 4%, BERTScore decreasing by 2%, and other generation-focused metrics such as BLEU-1, ROUGE, and METEOR reduced by 3–5%. In particular, when the LLaMA-Instruct 1B model is frozen, it may struggle to adapt to the characteristics of ECG-derived inputs and the nuances of clinical language. Likewise, tuning

the ECG encoder to adapt to the PTB-XL domain, especially when it was previously pretrained on a different dataset such as MIMIC IV ECG, also proves beneficial to the overall performance in the target ECG domain.

## 5 Discussion

Q-HEART has demonstrated strong performance in the ECG-QA task, highlighting its potential to bridge physiological signals and language understanding for clinical applications. To further advance the development of ECG-based question-answering systems, we outline several future directions that emerge from our analysis: (1) First, we observe that questions requiring numerical reasoning, particularly within the query-type category, remain challenging. This is especially evident when accurate responses depend on precise physiological metrics from raw signals, such as P wave duration, QRS interval, or QT interval. This limitation arises from the fact that the ECG encoder is not explicitly optimized to extract such fine-grained clinical features, and LLMs alone lack the inductive bias to infer low-level signal semantics from high-dimensional embeddings. To overcome this, future work could incorporate conventional ECG-derived features (e.g., heart rate, heart rate variability, noise indicators, or various interval durations) into the prompt context [36], enhancing the model’s capacity for grounded clinical reasoning. (2) Second, our usage of the lightweight LLM due to resource constraints, while efficient, may limit generalization in biomedical domains due to its lack of medical pretraining. We believe that integrating larger, medically fine-tuned LLMs could further enhance answer reasoning capabilities in ECG-focused QA tasks. (3) Lastly, we highlight the potential of knowledge graph integration[34, 10], where graph-structured clinical entities (e.g., Arrhythmia) and relations (e.g., Atrial Fibrillation shares some patterns with Atrial Flutter) are introduced to inject textual knowledge into ECG features, possibly improving question-answering performance and its interpretability.

## 6 Conclusion

We present Q-HEART, a novel multimodal LLM framework designed for ECG-based question answering. Q-HEART seamlessly integrates ECG signals with textual inputs through an instruction-tuned language model, supported by an ECG-aware transformer-based mapping module that aligns ECG embeddings with text representations. To enhance contextual reasoning and improve response quality, our approach

leverages clinically relevant information retrieved from similar ECG cases and integrates it into dynamically customized prompts, tailored to each input. Through extensive evaluations on the benchmark dataset, Q-HEART achieves state-of-the-art performance in exact-match accuracy and shows strong competitiveness across generative evaluation metrics. These results highlight the model’s capability for nuanced clinical reasoning and emphasize its potential to facilitate more interpretable and effective ECG understanding in real-world healthcare applications.

## References

- [1] A. Bleich, A. Linnemann, B. H. Diem, and T. O. Conrad. Automated medical report generation for ecg data: Bridging medical text and signal processing with deep learning. *arXiv preprint arXiv:2412.04067*, 2024.
- [2] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. 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.
- [3] M. Denkowski and A. Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL <https://aclanthology.org/W11-2107>.
- [4] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.
- [5] B. Gopal, R. Han, G. Raghupathi, A. Ng, G. Tison, and P. Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine learning for health*, pages 156–167. PMLR, 2021.
- [6] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, S. Berkowitz, D. Moukheiber, P. Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.
- [7] H. Han, S. Park, S. Min, H.-S. Choi, E. Kim, H. Kim, S. Park, J. Kim, J. Park, J. An, et al. Towards high generalization performance on electrocardiogram classification. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021.
- [8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [9] R. Hu, J. Chen, and L. Zhou. Spatiotemporal self-supervised representation learning from multi-lead ecg signals. *Biomedical Signal Processing and Control*, 84:104772, 2023.
- [10] X. Hu, L. Gu, K. Kobayashi, L. Liu, M. Zhang, T. Harada, R. M. Summers, and Y. Zhu. Interpretable medical image visual question answering via multi-modal relationship graph learning. *Medical Image Analysis*, 97:103279, 2024.
- [11] E. Jing, H. Zhang, Z. Li, Y. Liu, Z. Ji, and I. Ganchev. Ecg heartbeat classification based on an improved resnet-18 model. *Computational and Mathematical Methods in Medicine*, 2021(1):6649970, 2021.
- [12] D. Kiyasseh, T. Zhu, and D. A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.
- [13] S. K. Lalam, H. K. Kunderu, S. Ghosh, H. Kumar, S. Awasthi, A. Prasad, F. Lopez-Jimenez, Z. I. Attia, S. Asirvatham, P. Friedman, et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023.
- [14] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [15] C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, and R. Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In *Forty-first International Conference on Machine Learning*.
- [16] C. Liu, Z. Wan, S. Cheng, M. Zhang, and R. Arcucci. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8230–8234. IEEE, 2024.
- [17] Y. Luo, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao, and Z. Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.
- [18] Meta AI. Llama 3.2 1b instruct, 2024. URL <https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>.
- [19] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.
- [20] Y. Na, M. Park, Y. Tae, and S. Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.
- [21] J. Oh, H. Chung, J.-m. Kwon, D.-g. Hong, and E. Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pages 338–353. PMLR, 2022.
- [22] J. Oh, G. Lee, S. Bae, J.-m. Kwon, and E. Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 66277–66288. Curran Associates, Inc., 2023.
- [23] J. W. O’Sullivan, A. Palepu, K. Saab, W.-H. Weng, Y. Cheng, E. Chu, Y. Desai, A. Elezaby, D. S. Kim, R. Lan, et al. Towards democratization of subspecialty medical expertise. *arXiv preprint arXiv:2410.03741*, 2024.
- [24] K. Papineni, S. Roukos, T. Ward, and W.-J. 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.
- [25] M. Pham, A. Saeed, and D. Ma. C-melt: Contrastive enhanced masked auto-encoders for ecg-language pre-training. *arXiv preprint arXiv:2410.02131*, 2024.
- [26] J. Qiu, J. Zhu, S. Liu, W. Han, J. Zhang, C. Duan, M. A. Rosenberg, E. Liu, D. Weber, and D. Zhao. Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report. In *Machine Learning for Health (ML4H)*, pages 480–497. PMLR, 2023.
- [27] 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*, 11(1):1760, 2020.
- [28] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.
- [29] J. Tang, T. Xia, Y. Lu, C. Mascolo, and A. Saeed. Electrocardiogram-language model for few-shot question answering with meta learning. *arXiv preprint arXiv:2410.14464*, 2024.
- [30] J. Tang, T. Xia, Y. Lu, C. Mascolo, and A. Saeed. Electrocardiogram report generation and question answering via retrieval-augmented self-supervised modeling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- [31] 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*, 7(1):1–15, 2020.
- [32] Z. Wan, C. Liu, X. Wang, C. Tao, H. Shen, Z. Peng, J. Fu, R. Arcucci, H. Yao, and M. Zhang. Meit: Multi-modal electrocardiogram instruction tuning on large language models for report generation. *arXiv preprint arXiv:2403.04945*, 2024.
- [33] X. Wang, J. Kang, P. Han, Y. Zhao, Q. Liu, L. He, L. Zhang, L. Dai, Y. Wang, and J. Tao. Ecg-expert-qa: A benchmark for evaluating medical large language models in heart disease diagnosis. *arXiv preprint arXiv:2502.17475*, 2025.
- [34] Q. Yan, J. Duan, and J. Wang. Multi-modal concept alignment pre-training for generative medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5378–5389, 2024.
- [35] K. Yang, M. Hong, J. Zhang, Y. Luo, Y. Su, O. Zhang, X. Yu, J. Zhou, L. Yang, M. Qian, et al. Ecg-lm: Understanding electrocardiogram with large language model. *Health Data Science*, 2024.
- [36] H. Yu, P. Guo, and A. Sano. Zero-shot ecg diagnosis with large language models and retrieval-augmented generation. In *Machine Learning for Health (ML4H)*, pages 650–663. PMLR, 2023.
- [37] H. Yu, P. Guo, and A. Sano. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text. *arXiv preprint arXiv:2405.19366*, 2024.
- [38] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
- [39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [40] W. Zhang, L. Yang, S. Geng, and S. Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE*



*Transactions on Neural Networks and Learning Systems*, 2023.

- [41] Y. Zhao, T. Zhang, X. Wang, P. Han, T. Chen, L. Huang, Y. Jin, and J. Kang. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. *arXiv preprint arXiv:2408.08849*, 2024.
- [42] X. Zhong, O. Gong, W. Huang, L. Li, and H. Xia. Squeeze-and-excitation wide residual networks in image classification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 395–399. IEEE, 2019.