Proceedings of Machine Learning Research 225:650–663, 2023 Machine Learning for Health (ML4H) 2023

Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation

Han Yu HY29@RICE.EDU
Department of Electrical and Computer Engineering, Rice University

Peikun Guo PG34@RICE.EDU
Department of Computer Science, Rice University

Akane Sano AKANE.SANO@RICE.EDU
Department of Electrical and Computer Engineering, Rice University

Abstract

Recently, Large Language Models (LLMs) have become essential players in the deep learning domain. While their capabilities are evident across various textual tasks, this study aims to bridge the gap and explore the potential of leveraging LLMs in diagnosing cardiac diseases and sleep apnea from Electrocardiography (ECG). Earlier work touched on converting ECG signals into text for LLMs, but a comprehensive LLM-based approach for dealing with more complicated symptoms remains relatively unexplored. To investigate the ECG diagnosis with an LLM-based approach, our research introduces a zero-shot retrieval-augmented diagnosis technique. We have built databases filled with specific domain knowledge for cardiac symptom and sleep apnea diagnosis, which encourages the LLMs from merely relying on the inherent LLM knowledge to a more holistic pipeline from carefully crafting prompts and infusing expert knowledge to guide LLMs. We evaluate the proposed approach on two datasets for diagnosing arrhythmia and sleep apnea, respectively. The evaluation results indicate that our zero-shot approach not only surpasses previous few-shot LLM-based methods but is also competitive with supervised learning techniques fully trained on extensive datasets.

Keywords: Electrocardiogram (ECG), Large Language Model (LLM), Retrieval-Augmented Generation (RAG), Zero-Shot Learning, Arrhythmia, Apnea

1. Introduction

The rapid advancements in large language models (LLMs) have established them as a paramount subset of deep learning algorithms. With exponentially increasing parameters and expansive training datasets, LLMs can encode a massive amount of textual information, thereby improving their proficiency in understanding and generating human-like text (Touvron et al., 2023; OpenAI, 2023; Chowdhery et al., 2022). While many researchers have harnessed LLMs for tasks such as dialogue and question-answering across varied domains including generation (Rozière et al., 2023), finance (Wu et al., 2023), and health (Singhal et al., 2023), the emphasis has primarily been on textual data. However, data in the real world is diverse and extends beyond just text. For various applications, we have other formats of data such as image, audio, and clinical physiological sequences such as electrocardiogram (ECG). This observation leads us to a pivotal research question: Can LLMs be adeptly utilized for analyzing clinical physiological data, specifically in this paper, ECG to diagnose medical conditions?

Inspired by the potential of integrating various modalities into LLMs, researchers have investigated the processing of non-linguistic inputs including images (Chen et al., 2023), videos (Zhang et al., 2023), speech (Chen et al., 2023), and human activities (Girdhar et al., 2023) using LLMs. Although ECG is crucial for cardiac disease diagnosis and numerous auto-diagnosis approaches using deep learning methods have been developed (Liu et al., 2021b; Pyakillya et al., 2017; Sannino and De Pietro, 2018), the application of LLMs in the ECG diagnosis domain is still in its early stages. Li et al. (2023) introduced a method of converting physiological signals, including 1-lead ECGs, into textual descriptions for LLMs to infer health-related question-answering systems through few-shot fine-tuning. However, the LLM-based diagnosis system for more intricate symptoms such as arrhythmia remains under-investigated.

Moreover, previous LLM-based studies inferred ECG conditions by fine-tuning the prompts that perform the diagnosis process primarily using the intrinsic knowledge of the LLMs. As highlighted by researchers, the vast pre-training corpora of LLMs can sometimes lead to biased information or hallucinations, which poses risks for many domain-specific tasks (Lewis et al., 2020; Liu et al., 2023). In an ECG diagnosis system, it is essential to avoid such biased and misleading results. On the other hand, the acquisition of diagnostic labels for ECGs is usually expensive as they require the input of clinical professionals. Thus, developing an auto-diagnosis approach even without training samples can benefit a larger population.

To address these challenges, this study proposes a zero-shot retrieval-augmented diagnosis method employing LLMs. We build databases to manage documents with domain knowledge related to selected sources, e.g., textbooks and papers. Consequently, instead of depending on the existing knowledge of the LLMs, we augment the whole process including prompt preparation and answer generation by introducing expert domain knowledge to help LLMs understand the problem thoroughly. The feature selection and the prompt engineering processes are steered by the domain knowledge stored in the built database. Also, we augment the final prompts by retrieving relevant documents of the observed ECG abnormalities for more accurate diagnosis.

Our contributions can be summarized as follows:

- We introduce a retrieval-augmented ECG analysis model that integrates feature extraction and prompt design, and leverages domain expertise for more accurate model inferences.
- We conduct comprehensive evaluations for diagnosing arrhythmia and sleep apnea. Our technique demonstrates superior performance in a zero-shot learning context, not only surpassing the few-shot learning methods of previous research but also rivaling fully supervised methods.
- While previous studies have explored the use of few-shot tuned LLMs for basic information extraction from ECG signals, to the best of our knowledge, we are the first to utilize LLMs for the analysis of ECGs in relation to cardiac diseases without any tuning.

2. Related Work

2.1. ECG Diagnosis with Deep Learning

The success of deep learning approaches in various domains has naturally led to their exploration in ECG diagnosis applications (Liu et al., 2021b; Pyakillya et al., 2017; Sannino and De Pietro, 2018; Wagner et al., 2020; Śmigiel et al., 2021; Mostafa et al., 2019). For example, Śmigiel et al. (2021) utilized a 1D-CNN model on processed ECG data, enriched with entropy information, to distinguish patient arrhythmia classes, achieving an AUC score of 0.91 across five distinct classes. Mostafa et al. (2019) curated an exhaustive survey of deep learning techniques targeted at ECGs for sleep apnea detection, emphasizing models including convolutional neural network (CNN) and recurrent neural network (RNN) that have surpassed 90% accuracy on dedicated datasets.

However, the cost-intensive nature of clinical annotation has motivated researchers to study pre-training methods, which were designed to reduce the dependency on the labeled ECG sequences (Sarkar and Etemad, 2020; Mehari and Strodthoff, 2022; Oh et al., 2022). For instance, Mehari and Strodthoff (2022) leveraged prominent self-supervised learning paradigms such as SimCLR (Chen et al., 2020b), BYOL (Grill et al., 2020), and CPC (Oord et al., 2018) for 12-lead ECG pre-training. The preliminary phase accentuated the models' robustness, resulting in a 2% AUC score uplift against their purely supervised counterparts for 5-class arrhythmia classification. Nevertheless, even though pre-training strategies generally provide insights into performance improvement and the decreasing demands of the required clinical annotations, these methods usually face challenges in generalization. For example, Mehari and Strodthoff (2022) learned the ECG embedding from the 12-lead 10-second sequences with a sampling rate of 500 Hz, which can trigger distribution shifts on the ECG sequences under different conditions. This limitation raises the complexity when dealing with general ECG applications.

2.2. Large Language Models (LLMs)

LLMs are a class of language models that are designed in large-scale transformer-based architectures and pre-trained on massive corpora of general textual data, such as GPT-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2022), LLaMA2 (Touvron et al., 2023), etc. LLMs have demonstrated extreme capabilities in understanding natural language and inferring complex tasks via text generation.

Beyond their foundational capabilities in textual data, there has been a rising interest in adapting LLMs to other data types, such as images (Lu et al., 2019; Tsimpoukelli et al., 2021) and tabular structures (Liu et al., 2021a). While some of these techniques involve adjustments to the input/output layers, thereby potentially incurring the "catastrophic forgetting" issue (Chen et al., 2020a), others, like the approach suggested by Dinh et al. (2022), proposed a solution for the conversion of feature spaces into pure textual data for LLM interpretation, followed by task-specific fine-tuning. Notably, Li et al. (2023) employed this strategy on physiological time-series data, including ECG sequences. They converted the ECG sequences into textual descriptions with a series of numeric inter-beat-intervals (IBIs) in milliseconds and fine-tuned the prompts for detecting conditions such as heart rates and Sinus rhythms with promising results. Such adaptations have yielded encouraging outcomes in few-shot learning compared to 1D-CNN-based few-shot learning from scratch. However, the detection of more complicated ECG symptoms such as arrhythmias, remains unexplored. Further, the domain-specific knowledge learned within LLMs can sometimes be a limiting factor. For instance, the dataset used for training LLaMA, with 65% of its data sourced from the generalized common crawl corpus (Touvron et al., 2023), might pose risks of incomplete or skewed clinical interpretations.

To prevent LLMs from hallucinating and generating biased texts, some researchers have pivoted towards retrieval-augmented generation approaches. These methods empower LLMs to merge input prompts with trusted, domain-specific knowledge sources, thereby facilitating more accurate information understanding and generation (Lewis et al., 2020; Liu et al., 2023). Inspired by these methods, we propose our strategy for solving clinical applications by engineering the ECG data into textual prompts and inferences with the combined domain knowledge.

3. Objective & Datasets

In this study, we explore the diagnosis of arrhythmia with PTB-XL(+) datasets(Wagner et al., 2020; Strodthoff et al., 2023) and sleep apnea with Apnea-ECG dataset (Penzel et al., 2000). The summary of the two datasets can be found in Table 1.

3.1. Arrhythmia Diagnosis

Our study examines the proposed methods on the PTB-XL+ and PTB-XL datasets for examining arrhythmia diagnosis. The PTB-XL dataset is a large dataset containing 21,837 clinical 12-lead ECG records from 18,885 patients of 10-second length, where 52% are male and 48% are female with ages ranging from 0 to 95 years (median 62 and interquartile range of 22). There are two sampling rates: 100 and 500 Hz, available in the dataset. The raw ECG data are annotated by two cardiologists into five major categories, including normal ECG (NORM), myocardial infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). The PTB-XL+ dataset covers algorithm-extracted features on the ECG sequences, such as durations, amplitudes, on/off-sets of segments, fiducial points, median beats, etc. The datasets contain a comprehensive collection of many different co-occurring pathologies and a large proportion of healthy control samples. To ensure a fair comparison of machine learning algorithms trained on the dataset, we follow the recommended splits of training and test sets. However, given that our proposed method employs a zero-shot approach, we do not use the training samples to fine-tune the models.

3.2. Sleep Apnea Detection

While sleep apnea is widely detected by analyzing polysomnography (PSG), researchers also investigated the potential of using ECG to diagnose the apnea, for example using the Apnea-ECG dataset (Penzel et al., 2000). The database can be accessed through Physionet (Goldberger et al., 2000). The dataset contains 70 records of ECG recorded at a sampling rate of 100 Hz without features extracted, 35 of which are used for training and 35 for testing. The duration of the records ranges from almost 7 hours to nearly 10 hours. Labels indicating the presence or absence of sleep apnea are assigned to each minute of the recordings. Consequently, we segment the ECG recordings into one-minute intervals, which result in 6000 data points for each segment. There are 17233 training samples and 17010 samples for the test set with a non-apnea to apnea sample ratio of 61.49% to 38.51%.

Table 1: Summary of datasets used in this study

Dataset	**PTB-XL+**	**Apnea-ECG**
Tasks	Arrhythmia Diagnosis	Sleep Apnea Detection
Total Records	21,837	34,243 (split from 70)
Annotations	5 categories of arrhythmia	Presence/Absence of sleep apnea
Data Types	Raw ECG & Fiducial Annotations	Raw ECG
Test Set Size	2203	17,010

4. Methods

In this section, we introduce the methods for feature extraction, prompt preparation, and retrieval-augmented model inference. To ensure clarity, we illustrate our framework using the PTB-XL+ dataset (Strodthoff et al., 2023). This approach is exemplified in the context of arrhythmia diagnosis, as demonstrated in Figure 1.

4.1. Construct Database of Domain Knowledge

To enhance our understanding of ECG in arrhythmia diagnosis, we construct a local vector database with guidance from two published books: (1) *ECG Workout: Exercises In Arrhythmia Interpretation* by Huff (2006) and (2) *12-Lead ECG: The Art of Interpretation* by Garcia (2015). For diagnosing sleep apnea with Apnea-ECG dataset, we prepare the vector database by encoding apnea-related textbook (Randerath et al., 2006) and papers (Almazaydeh et al., 2012; Drinnan et al., 2000; McNames and Fraser, 2000; Zywietz et al., 2004). For both datasets, we utilize the *text-embedding-ada-002* embedding extraction API (OpenAI, 2023) and manage the extracted embedding using the *Chroma* database tool in conjunction with the *LangChain* Python library (Mendable, 2023). This setup facilitates the search and retrieval of related text from the embedding space with appropriate prompts.

4.2. Feature Extraction and Prompts Preparation

Prompts are crucial for guiding LLMs to generate relevant responses, especially for models that are not further fine-tuned (frozen LLMs). To transform ECG into effective prompts, we first extract hand-crated features from sequences and engineer the selected features into prompts. In the case of the PTB-XL+ dataset, comprehensive features including the detailed fiducial information are annotated by the original authors using both commercial and open-source algorithms. Thus, we leverage the calculated features rather than engineering features from scratch on the PTB-XL+ dataset.

For the Apnea-ECG dataset, which does not cover pre-annotated fiducial points or specific ECG features, we use a Python library *NeuroKit2* (Makowski et al., 2021) to detect the fiducial points and extract features such as heart rate variability and spectral power.

4.2.1. Retrieval-augmented Feature Selection

While extracting features from original ECGs typically involves universal elements such as waveforms and amplitudes of fiducial points and intervals, the large number of diverse features across ECG leads presents a challenge. However, overloading LLMs with an extensive array of comprehensive features for reasoning and inference might not only exceed input length restrictions of LLMs but also may introduce redundant information, which potentially hinders accurate diagnosis.

To mitigate this, we intend to refine our feature extraction strategy by looking up the domain-specific databases and extracting crucial insights from two leading ECG textbooks. Our strategy involves querying targeted questions pertaining to the interpretation of specific arrhythmia types, such as ST/T segment change (STTC), myocardial infarction (MI), conduction disturbance (CD), and hypertrophy (HYP). This method enables us to identify and focus on the most relevant features for each diagnostic category, which provides LLMs with clinically related data and helps avoid information overload. Consequently, We extract features with the queried diagnosis guidance and check 15 different fiducial points and segments across different leads such as QRS complex, T wave, P wave, PR segment, RS segment, etc. For example, we look for the J-point amplitude for ST-segment elevation and depression and the ratio of

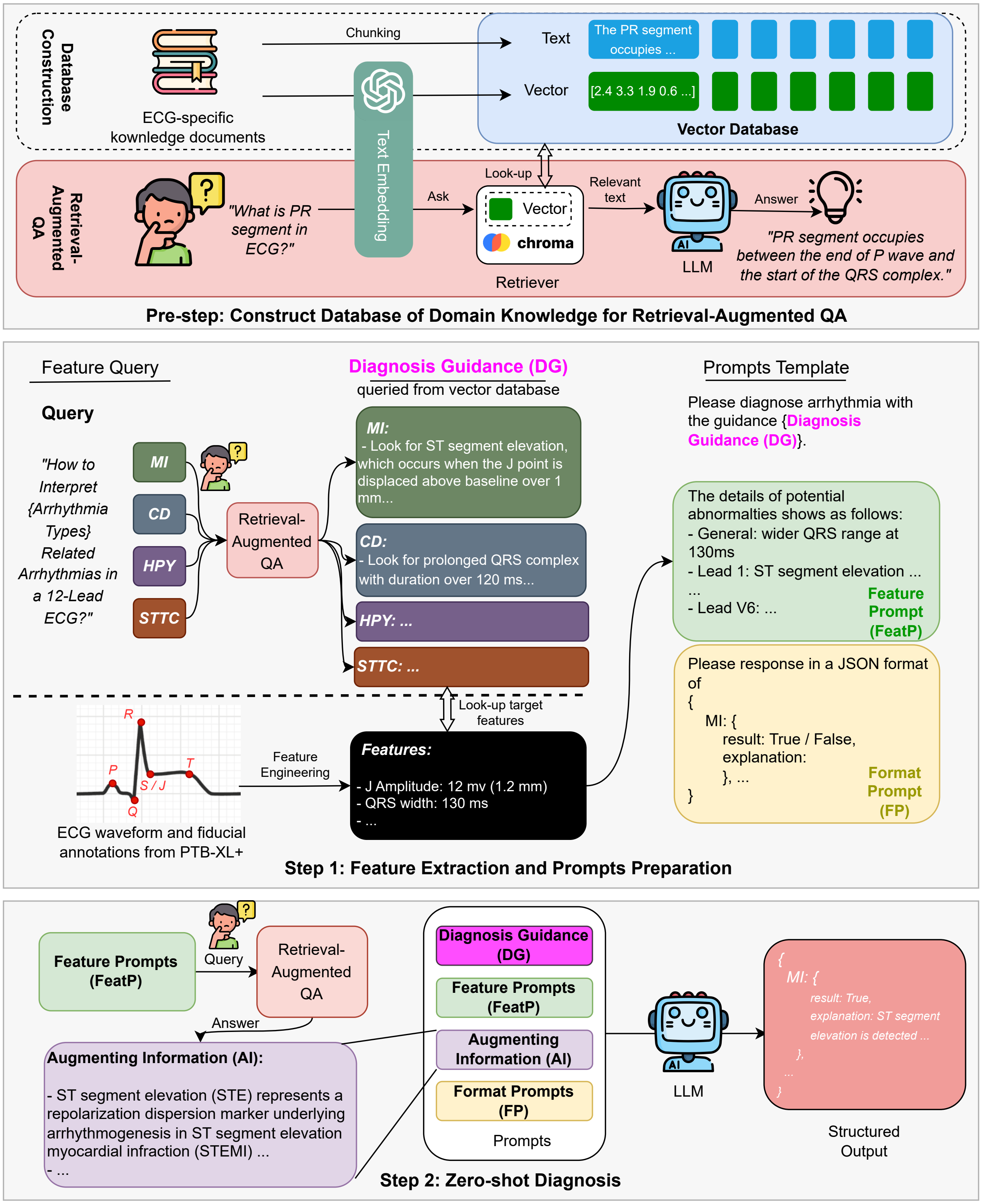

Figure 1: The overall framework of the proposed method includes constructing database of domain knowledge, feature extraction & prompts preparation, and zero-shot retrieval-augmented inference for arrhythmia diagnosis as an example.

R / S amplitude for the waveform of RS complexes, etc. We conduct the same procedures for the querying guidance of diagnosing sleep apnea. In the absence of crafted features within the original Apnea-ECG dataset, we extract 8 features using the NeuroKit2 library (Makowski et al., 2021), which focus on the heart rate variability analysis and spectral analysis, especially for depression and elevation of heart rate and spectral power in the very low frequency (VLF) band (0.01-0.04 Hz). The extracted features cover the average heart rate, variability of R-R intervals, elevation of spectral power in the VLF band, power in both the low-frequency (LF) and high-frequency (HF) bands, as well as the ratio of power between LF and HF bands.

4.2.2. Prompts Preparation

With the guidance queried from the domain-specific databases and the correspondingly engineered features, we structure the prompts for LLMs in three folds:

Diagnosis Guidance: We begin by integrating the insights previously queried from the textbooks, which cover the essential information on interpreting specific arrhythmia types or sleep apnea detection. LLMs can receive this information along with the input prompts.

Feature Prompt: Next, we incorporate detailed ECG information highlighting potential abnormalities that are converted from the features we extracted. This information is organized into two main categories including general information and lead-wise information:

- General Information: This covers general insights into the ECG, such as the QRS duration, providing an overview of the ECG and anomalies.
- Lead-wise Information: Since abnormalities can present differently across the 12 leads, we integrate specific information for each lead, such as the waveform of the P and T waves. This ensures that the LLMs can discern and diagnose conditions that might be prominent in one lead but subtle or absent in other leads.

Format Prompt: We also introduce format prompts that would guide the LLMs to produce structured responses for easy post-processing. To be more specific, we ask the LLMs to respond in a structured JSON format with each arrhythmia type as the first-layer key followed by boolean diagnosis results and reasoning explanation.

4.3. Zero-shot Generation

Leveraging the engineered prompts, we employ a retrieval-augmented generation for zero-shot inference. The initial step involves utilizing the detailed ECG information prompts as a querying mechanism. We can retrieve relevant textual information on the target ECG samples as the augmenting information by querying the potential ECG abnormalities. Differing from the diagnosis guidance, the augmenting information is queried based on the features we extracted in prompts, e.g., ST segment elevation and prolonged QRS complex. This step aims to retrieve information derived from specific features so that provides a more detailed context for these abnormalities. For example, by querying keywords of *"ST segment elevation"*, the augmenting information covers *"This can be indicative of myocardial injury or infarction (heart attack). However, it can also be caused by other conditions such as coronary artery spasm, acute pericarditis, ventricular aneurysm, early repolarization pattern, hyperkalemia, or hypothermia..."*. Then, we concatenate the augmenting information with the original prompts to finalize our input prompts for LLMs for more comprehensive insights. Finally, we leverage LLMs to directly understand and infer the prompts without training or fine-tuning.

In essence, we aim to make the LLMs consistently augmented with domain-specific insights, guaranteeing that the outputs are precise and reflect a deep-rooted understanding of the ECG condition and diagnosis.

4.3.1. Frozen Large Language Models

In this study, we leverage both the open-source model such as LLaMA2 and the closed-source GPT-3.5 models in zero-shot inference.

LLaMA2: LLaMA2 is an LLM developed by Meta AI. LLaMA2 has 7 billion to 70 billion parameters, and it can be used for a variety of tasks, such as dialogue and question-answering. It has been shown to outperform other open-source LLMs on many benchmarks. LLaMA2 is available for free for research and commercial use. Due to the constraints on the computational resources, we employ only the 7B and 13B versions.

GPT-3.5: GPT-3.5 is an LLM developed by OpenAI. GPT-3.5 has 175 billion parameters, which makes it one of the largest LLMs ever created. It can be used for a variety of tasks. It has been shown to outperform other LLMs on many benchmarks. Generally, GPT-3.5 is accessible via API calls.

5. Experiments

5.1. Methods of Evaluation

For our evaluations, we utilize a diverse set of models, ranging from traditional supervised methods to state-of-the-art LLMs. The selected models are as follows:

5.1.1. Supervised Method

We implement a 1D-CNN model as the supervised baseline. Following the prior studies (Strodthoff et al., 2023), the 1D-CNN kernel is designed to capture the temporal patterns in the ECG sequences, making it suitable for tasks that require understanding the sequential nature of the ECGs. With the supervised baseline, we implement the full training strategy, which leverages all the available training samples to help models learn useful parameters from scratch.

5.1.2. Numerical prompts with few-shot tuning

We implement the method from the prior LLM-based study (Li et al., 2023) as a part of the evaluation for performance comparisons. This method converts ECG signals into a textual sequence of IBIs, e.g., *"Identify the average heart rate from given interbeat interval sequence 896,1192,592,1024,1072,808,888 …"*, which shows promising results in detecting heart rates and Sinus rhythms in a 25-shot training setting. Due to the discrepancy between tasks of detecting heart rhythms and detecting cardiac diseases, we reproduce their approach by converting the ECGs into sequences of IBI numbers and enhancing the prompts by covering the location and amplitudes of fiducial points including P, T, Q, R, and S for each lead. Also, randomly sampled 25 ECGs are used as the training sample following the few-shot learning scheme. We employ LoRA (Hu et al., 2021), which is an efficient fine-tuning method widely applied for LLMs, in fitting the training data.

5.1.3. Retrieval-augmented Generation

To evaluate the proposed method, we use only the test sets for examining the performance. The preparation including vector database and prompt engineering is described in Section 4.

5.2. Evaluation Results

5.2.1. Arrhythmia Diagnosis

Table 2 shows the evaluation performances on the arrhythmia diagnosis. The performances are evaluated in metrics of accuracy rate, macro precision, macro recall, and macro f1 score across all the classes. The GPT-3.5 model outperformed the open-source LLaMA2 models in all metrics. When comparing our proposed zero-shot retrieval-augmented strategy with the supervised learning method, we observed superior performances in accuracy rate, macro precision, and macro F1 scores in our proposed method; whereas the supervised method shows a higher macro recall score compared to the proposed method. This result suggests that our proposed approach can be effective in detecting arrhythmia even without leveraging any training samples.

The class-wise diagnostic performance offers insights into the efficacy and potential limitations of our zero-shot retrieval-augmented approach using the GPT-3.5 model. A deeper dive into the results, as presented in Table 3, reveals patterns in diagnosis across various classes of arrhythmia. CD, HYP, and MI detection show high precision scores, indicating that once these conditions are detected, the false detecting rate remains relatively low. While precision is promising in certain classes, there have been instances where conditions were not detected and were instead misclassified as normal ECGs. This could be caused by the fact that the engineered features and prompts might not have captured comprehensive nuances associated with certain arrhythmia types. In addition, the detecting performance for STTC is relatively lower compared to the other arrhythmia classes.

The explanations generated by the LLM also provide some insights into our error analysis. For all samples incorrectly identified as HYP by LLMs, the explanations cite that the ECG matches the Sokolov-Lyon criteria for diagnosing HYP by checking the R waves in lead V1/V2 and S waves in lead V5/V6, even when HYP was not identified in the human-annotated labels. Such inconsistencies might stem

Table 2: The evaluation results of arrhythmia diagnosis in metrics of accuracy rate, macro precision, macro recall, and macro F1 score. *Few-shot TNP*: few-shot textual numeric prompts. *Zero-shot RAG*: zero-shot retrieval-augmented generation. Bold represents the highest performances in the evaluation set.

Method	Model	Training	Accuracy	Macro Precision	Macro Recall	Macro F1
Supervised	1D-CNN	17441	0.748	0.708	**0.643**	0.660
Few-shot TNP (Li et al., 2023)	LLaMA2-7B	25	0.417	0.391	0.277	0.357
	LLaMA2-13B	25	0.422	0.401	0.294	0.348
Zero-shot RAG (Ours)	LLaMA2-7B	0	0.714	0.765	0.548	0.617
	LLaMA2-13B	0	0.726	0.770	0.561	0.622
	GPT-3.5	0	**0.757**	**0.791**	0.616	**0.669**

Table 3: Class-wise performances for the zero-shot retrieval-augmented method

Class	Samples	Precision	Recall	F1 Score
NORM	912	0.54	0.79	0.61
CD	473	0.93	0.61	0.77
HYP	243	0.91	0.55	0.70
MI	415	0.80	0.63	0.70
STTC	516	0.77	0.50	0.58

from information loss during the signal filtering process or flawed fiducial point annotations. The errors from signal processing can directly affect the precision of prompts. On the other hand, currently detecting STTC majorly depends on abnormalities observed in the T wave and the duration of PT. Another challenge arises when trying to precisely describe complex waveform patterns in textual data, such as the varying waveform morphology in real ECGs.

5.2.2. Sleep Apnea Diagnosis

Table 4 displays the performance of the examined method in diagnosing sleep apnea. This table reveals that the supervised learning method excels in terms of accuracy and precision scores. In contrast, the proposed method using the GPT-3.5 model delivers the highest recall and F1 scores. Similar to our findings in arrhythmia diagnosis, the numeric prompts with the few-shot tuning method yield less-than-ideal results for the apnea task.

Despite our method showing promise in recall rates, the LLM-based approaches produce a comparatively low precision score when compared with the supervised learning method. We find that this disparity may arise from signal quality and prompt engineering precision. Our prompts, engineered from features crafted based on R-R intervals extracted by software, are susceptible to signal noises. By combining the error analysis with signal quality check (Zhao and Zhang, 2018), we find the average precision scores on test sequences in "excellent" quality (6.38% of all test sequences) are 6.4% higher than ECGs in "barely acceptable" quality (74.21% of all test sequences). Additionally, ECG processing software can mis-detect R peaks in some sequences, resulting in extended intervals that manifest as confusing features, even when the original signal is normal. Among the false-positive samples that are detected as ECG with apnea, 72.1% highlights either high VLF power or significant heart rate variability.

5.3. The Contribution of Each Step in Proposed Framework

We conduct an ablation study to assess the contribution of each component of our proposed LLM-based retrieval-augmented method. The input prompt for the proposed method consists of diagnosis guidance, feature prompts, augmenting information, and format prompts, which build a comprehensive understanding of ECG signals and diagnostic information. Among these prompt components, the feature and format prompts are not removable as they function essentially in describing the ECGs and generating processable output, respectively. Thus, we evaluated the performance of removing diagnosis guidance or augmenting information in prompts on the PTB-XL+ dataset to understand the impacts of these components. Table 5 shows the performances of removing specific components in prompts on the PTB-XL+ dataset with GPT-3.5. From the table, we can see

Table 4: The evaluation results of sleep apnea diagnosis in metrics of accuracy rate, precision, recall, and F1 score. *Few-shot TNP*: few-shot textual numeric prompts. *Zero-shot RAG*: zero-shot retrieval-augmented generation. Bold represents the highest performances in the evaluation set.

Method	Model	Training	Accuracy	Precision	Recall	F1
Supervised	1D-CNN	17233	**0.821**	**0.804**	0.843	0.787
Few-shot TNP (Li et al., 2023)	LLaMA2-7B	25	0.675	0.492	0.535	0.504
	LLaMA2-13B	25	0.691	0.512	0.562	0.522
Zero-shot RAG (Ours)	LLaMA2-7B	0	0.753	0.710	0.855	0.758
	LLaMA2-13B	0	0.772	0.728	0.859	0.770
	GPT-3.5	0	0.804	0.763	**0.910**	**0.801**

When the Diagnosis Guidance (DG) component is removed, the F1 score drops from 0.669 to 0.593, indicating a decrease of 0.076. Removing the Augmenting Information (AI) component results in a smaller decrease in the F1 score from 0.669 to 0.628. When both the DG and AI components are removed, the F1 score drops significantly to 0.571, which is a decrease of 0.098 from the full method. This suggests that both components contribute to the overall performance, with the DG component being more critical in the model performance than the AI component.

Table 5: Performances in arrhythmia diagnosis on PTB-XL+ dataset with GPT-3.5. Macro-F1 score is used as the evaluation metric. *Diff.* indicates the performance differences after removing prompts components.

Removed Prompts	**F1**	**Diff.**
None (Full Method)	0.669	-
Diagnosis Guidance (DG)	0.593	$\downarrow$ 0.076
Augmenting Information (AI)	0.628	$\downarrow$ 0.041
DG & AI	0.571	$\downarrow$ 0.098

6. Discussion & Conclusion

This study underscores the potential and limitations of leveraging advanced language models, such as LLaMA2 and GPT-3.5, for complex medical diagnostic tasks such as arrhythmia and sleep apnea detection. Our zero-shot retrieval-augmented approach demonstrated promising performances, even when no training samples were used, highlighting its applicability in scenarios where labeled data is scarce or expensive to obtain. The proposed approach outperformed the few-shot LLM-based approach in a prior study and even achieved competitive results on fully trained supervised learning methods. While the efficacy of the retrieval-augmented generation was showcased using ECG data, its potential extends further. We believe that this technique can be effectively applied to an array of physiological signals such as photoplethysmogram (PPG) and electrodermal activity (EDA). Furthermore, this methodology can be adapted into a multimodal system to tackle more intricate diagnostic tasks and insights.

Nevertheless, our study has limitations. First, our current prompts may not capture all the nuances of signals that might cause misdiagnosis, for example, the STTC arrhythmia as demonstrated in the previous section. Second, our current feature and prompt design are currently based on books and datasets. To address these issues and leverage more capabilities of the proposed method, we will (1) prioritize designing more robust pre-processing algorithms or prompts engineering methods that make prompts more comprehensive with the information in the original signal and (2) inject clinical domain knowledge to design features and prompts by collaborating with clinicians.

Acknowledgments

This work is supported by NSF #2047296 and Rice University.

References

Laiali Almazaydeh, Khaled Elleithy, and Miad Faezipour. Detection of obstructive sleep apnea through ecg signal features. In *2012 IEEE International Conference on Electro/Information Technology*, pages 1–6. IEEE, 2012.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv preprint arXiv:2004.12651*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

M Drinnan, J Allen, P Langley, and A Murray. Detection of sleep apnoea from frequency analysis of heart rate variability. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 259–262. IEEE, 2000.

Tomas B Garcia. *12-lead ECG: The art of interpretation*. Jones & Bartlett Learning, 2015.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jane Huff. *ECG workout: Exercises in arrhythmia interpretation*. Lippincott Williams & Wilkins, 2006.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*, 2023.

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*, 2023.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*, 2021a.

Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021b.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

JN McNames and AM Fraser. Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram. In *Computers in Cardiology 2000. vol. 27 (Cat. 00CH37163)*, pages 749–752. IEEE, 2000.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141: 105114, 2022.

Mendable. Langchain: A framework for developing applications powered by language models. Software framework, 2023. Retrieved from https://github.com/mendable/langchain.

Sheikh Shanawaz Mostafa, Fábio Mendonça, Antonio G. Ravelo-García, and Fernando Morgado-Dias. A systematic review of detecting sleep apnea using deep learning. *Sensors*, 19(22):4934, 2019.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward 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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Gpt-4 technical report, 2023.

OpenAI. Embedding - openai. Software API, 2023. Retrieved from https://platform.openai.com/docs/guides/embeddings.

Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

Boris Pyakillya, Natasha Kazachenko, and Nikolay Mikhailovsky. Deep learning for ecg classification. In *Journal of physics: conference series*, volume 913, page 012004. IOP Publishing, 2017.

Winfried J Randerath, Bernd M Sanner, and Virend K Somers. *Sleep apnea: current diagnosis and treatment*, volume 35. Karger Medical and Scientific Publishers, 2006.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Giovanna Sannino and Giuseppe De Pietro. A deep learning approach for ecg-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems*, 86:446–455, 2018.

Pritam Sarkar and Ali Etemad. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3): 1541–1554, 2020.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Sandra Śmigiel, Krzysztof Pałczyński, and Damian Ledziński. Ecg signal classification using deep learning techniques based on the ptb-xl dataset. *Entropy*, 23(9):1121, 2021.

Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip J Aston, Ashish Sundar, Claus Graff, Jørgen K Kanters, Wilhelm Haverkamp, Olaf Dössel, Axel Loewe, et al. Ptb-xl+, a comprehensive electrocardiographic feature dataset. *Scientific Data*, 10(1):279, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

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):154, 2020.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Zhidong Zhao and Yefei Zhang. Sqi quality evaluation mechanism of single-lead ecg signal based on simple heuristic fusion and fuzzy comprehensive evaluation. *Frontiers in physiology*, 9:727, 2018.

CW Zywietz, V Von Einem, B Widiger, and G Joseph. Ecg analysis for sleep apnea detection. *Methods of information in medicine*, 43(01):56–59, 2004.

Appendix A. Examples of Prompts and Outputs

In this section, we present examples of using LLM to diagnose ECG. The prompts will be shown in each prompt component.

Example 1: Normal ECG in PTB-XL+.

Prompts:

Identify the types of arrhythmia in the ECG signal with diagnostic guidance and the extracted features.

The diagnostic guidance you should follow is detailed below. Additionally, consider the supplemental information from textbooks regarding the detected features. Please be careful about the features in the different leads.

Diagnosis Guidance:

When diagnosing a Myocardial Infarction (MI), various ECG changes must be considered. The ST segment elevation is a critical indicator that signals myocardial injury. For instance, if the ST elevation is observed in leads II, III, and aVF, an inferior MI is suggested, whereas ST elevation in leads V2 to V4 points to an anterior MI. Additionally, ST depressions opposite the infarct area, known as reciprocal changes, are also significant. T wave abnormalities are another aspect, where inverted or sharply peaked "tombstone" T waves can be seen in the affected leads. Lastly, the presence of Q waves, which are pathological, indicates a transmural MI and will appear in the corresponding leads of the infarct area.

Conduction disturbances in the heart manifest through various changes in the ECG. A QRS complex that is wider than 0.12 seconds is indicative of a disturbance. Specifically, an RSR' or rSR' pattern in lead V1 suggests a right bundle branch block (RBBB), whereas a wide S wave or notched R wave in lead V6 indicates a left bundle branch block (LBBB). In lead III, multiple peaked QRS complexes may show localized intraventricular conduction delays. Additionally, a QRS complex wider than 0.12 seconds without the specific characteristics of LBBB or RBBB points to a generalized intraventricular conduction delay (IVCD).

Hypertrophy within the heart can be detected by assessing certain ECG features. Left Ventricular Hypertrophy (LVH) is characterized by tall R waves in leads I and V5-V6, coupled with deep S waves in V1-V2. A sum greater than 35 mm of the S wave depth in V2 and the R wave height in V5 is indicative of LVH. Right Ventricular Hypertrophy (RVH) is suggested by increased R wave amplitude in V1 and a deep S wave in V6, with an R:S ratio greater than 1 in V1. Atrial enlargement is also identifiable; left atrial enlargement shows as a broad and notched P wave in lead II (P-mitrale), while right atrial enlargement presents as a tall and peaked P wave in the same lead (P-pulmonale).

(Only when there is no other diagnosable arrhythmia.) STTC is only significant when there is no other arrhythmia types, otherwise, abnormalities are diagnostic rather than STTC. Analyzing ST/T changes on an ECG requires a lookout for any obvious abnormalities. For example, wide and bizarre complexes interspersed with normal cycles may hint at ventricular tachycardia. ST depression and T waves that deviate from the normal direction are indicative of ischemia. ST elevation, which may or may not be accompanied by changes in T waves, denotes injury. The presence of pathological Q waves, especially in leads other than V1, is significant. Moreover, symmetrical or asymmetrical changes in the ST segment and T waves provide in- sights into the underlying pathology. Clear signs of ischemia or infarction are regional T wave abnormalities, ST segment changes, and the appearance of Q waves.

Feature Prompts:

General: N/A Leads: -I: N/A, -II: N/A, -III: Notched R Wave. -aVF: N/A, -aVR: N/A, -aVL: N/A, -V1: N/A, -V2: N/A, -V3: N/A, -V4: N/A, -V5: N/A, -V6: N/A

Augmenting Information:

Supplemental information on the ECG features: The notched R wave in lead III is not a cause for concern. It is just a part of the complex and does not indicate any abnormality.

Format Prompts:

Based on the above guidance and considering the supplemental information, please categorize the ECG into the following arrhythmia categories: - Normal ECG (NORM) - ST/T Change (STTC) - Myocardial Infarction (MI) - Conduction Disturbance (CD) - Hypertrophy (HYP).Please output in a Json format of NORM:, MI:, STTC:, CD:, HYP: with the explanation of each type. For example: {NORM: {Result: True, Explanation: ...}}

Note: please directly provide the results in a JSON format with no other explanation.

Response:

NORM: - Result: true, - Explanation: "The only feature detected is a notched R wave in lead III, which, according to the supplemental information, is not indicative of any abnormality."

STTC: Result: false, Explanation: "There are no ST depressions, T wave inversions, or any other ST/T changes mentioned in the ECG features."

MI: - Result: false, - Explanation: "There is no evidence of ST segment elevation, pathological Q waves, or reciprocal changes in any of the leads provided."

......

Example 2: ECG with both CD and MI in PTB-XL+.

The starter prompt and diagnostic guidance is the same as in example 1 ......

Format Prompts:

The following information describe a 12-lead ECG in features:

General: Long PR interval of 214.0 ms, which might indicate a conduction disturbance or ST/T change. Prolonged QT interval detected with a corrected QT of 469.0 ms. Leads: -I: Inverted T wave is approximated at -1.31 mm. Inverted T wave detected, which might suggest ST/T change if no other arrtythmia diagnosed. -II: Abnormal RS complex might indicate Left anterior fascicular block (LAFB) as condution disturbance. -III: ST segment elevation is approximated at 1.26 mm. Abnormal RS complex might indicate Left anterior fascicular block (LAFB) as condution disturbance. -aVF: Abnormal RS complex might indicate Left anterior fascicular block (LAFB) as condution disturbance. -aVR: N/A -aVL: Inverted T wave is approximated at -2.14 mm. Prolonged R peak time, might indicate condution disturbance. Inverted T wave detected, which might suggest ST/T change if no other arrtythmia diagnosed. -V1: ST segment elevation is approximated at 2.19 mm. Deeper Q wave is approximated at amplitude of -22.99 mm. Q wave can be seen in lead V1, which might indicate an abnormal pattern. -V2: ST segment elevation is approximated at 3.95 mm. Abnormal T wave detected, which might suggest ST/T change if no other artythmia diagnosed. -V3: ST segment elevation is approximated at 2.63 mm. -V4: N/A -V5: N/A -V6: N/A

Augmenting Information:

Supplemental information on the ECG features: Based on the information provided, the ECG shows a long PR interval, which may indicate a conduction disturbance. The ECG also shows ST segment elevation in leads III, aVF, V1, V2, and V3, which may suggest ST/T changes. Inverted T waves are seen in leads I and aVL, which may also indicate ST/T changes. Abnormal RS complexes in leads II, III, and aVF may suggest left anterior fascicular block as a conduction disturbance. The ECG also shows a prolonged QT interval. Overall, the ECG findings suggest possible conduction disturbances, hypertrophy, and ST/T changes.

The format prompts are the same as in example 1 ......

Response:

NORM: - Result: false, - Explanation: "The ECG shows significant abnormalities including ST segment elevation, inverted T waves, and abnormal RS complexes, indicating the presence of arrhythmias and excluding a normal ECG."

STTC: Result: false, Explanation: "Diagnosable change in ST and T waves has been detected."

CD: - Result: true, - Explanation: "There is no evidence of ST segment elevation, pathological Q waves, or reciprocal changes in any of the leads provided."

MI: - Result: true, - Explanation: "Abnormal RS complexes in leads II, III, and aVF suggest a left anterior fascicular block. Prolonged PR and QT intervals also indicate conduction disturbances."

HYP: - Result: false, - Explanation: "There is no direct evidence of hypertrophy such as tall R waves in leads V5-V6 or deep S waves in V1-V2 for LVH, or increased R wave amplitude in V1 and deep S wave in V6 for RVH."