RESEARCH-ARTICLE

ECG-Doctor: An Interpretable Multimodal ECG Diagnosis Framework Based on Large Language Models

DONGSHENG TIAN, University of Science and Technology of China, Hefei, Anhui, China

JUNZHE JIANG, University of Science and Technology of China, Hefei, Anhui, China

KAI ZHANG, University of Science and Technology of China, Hefei, Anhui, China

CHANGCHUN LIU, University of Science and Technology of China, Hefei, Anhui, China

YU YUAN, University of Science and Technology of China, Hefei, Anhui, China

MIN GAO, The First Affiliated Hospital of USTC, Hefei, Anhui, China

View all

Published: 10 November 2025

Citation in BibTeX format

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
November 10 - 14, 2025
Seoul, Republic of Korea

Conference Sponsors:
SIGWEB
SIGIR

CIKM '25: Proceedings of the 34th ACM International Conference on Information and Knowledge Management (November 2025)
https://doi.org/10.1145/3746252.3761082
ISBN: 9798400720406

ECG-Doctor: An Interpretable Multimodal ECG Diagnosis Framework Based on Large Language Models

Dongsheng Tian
University of Science and Technology of China
Hefei, China
ds_tian@mail.ustc.edu.cn

Junzhe Jiang
University of Science and Technology of China
Hefei, China
jzjiang@mail.ustc.edu.cn

Kai Zhang†
University of Science and Technology of China
Hefei, China
sa517494@mail.ustc.edu.cn

Changchun Liu
University of Science and Technology of China
Hefei, China
changchun_liu@mail.ustc.edu.cn

Yu Yuan
University of Science and Technology of China
Hefei, China
yyhappier@mail.ustc.edu.cn

Min Gao
The First Affiliated Hospital of University of Science and Technology of China
Hefei, China
Gm197866@mail.ustc.edu.cn

Enhong Chen†
University of Science and Technology of China
Hefei, China
cheneh@ustc.edu.cn

Abstract

Electrocardiogram (ECG) diagnosis aims to automatically classify ECG recordings into clinically meaningful categories, playing a vital role in medical decision-making. Deep learning methods, while promising, demand extensive annotated data and lack interpretability. Large Language Models (LLMs) offer potential in low-data scenarios and generating interpretable outputs, yet their application to ECG diagnosis, especially leveraging multimodal data (e.g., raw signals, derived features, and clinical knowledge), remains under-explored. To address these challenges, we propose **ECG-Doctor**, an interpretable and multimodal ECG diagnosis framework based on LLMs. ECG-Doctor comprises four key components: (1) **ECG Knowledge Acquisition Module**, which integrates external medical knowledge and Chain-of-Thought (CoT) reasoning to address the inability of LLMs to follow standardized ECG diagnostic procedures; (2) **ECG Feature Extraction Module**, which incorporates domain knowledge to overcome LLMs' limitations in comprehensively understanding structured ECG features; (3) **ECG Waveform Analysis Module**, which introduces time-series ECG models to equip LLMs with the capability to interpret and reason over raw ECG signal morphologies; (4) **KNN-based ECG Retrieval Module**, which retrieves the top-k most similar ECG samples and guides LLMs through in-context learning (ICL), enabling them to differentiate and learn from variations across ECGs. The outputs of these modules are aggregated and provided to the LLM as diagnostic context, enabling ICL to perform comprehensive ECG diagnosis. This design effectively simulates the diagnostic reasoning process of experienced electrocardiologists. Extensive experiments on the PTB-XL dataset demonstrate that ECG-Doctor is compatible with various LLMs and consistently outperforms existing baselines at both 100 Hz and 500 Hz sampling rates, showcasing its strong versatility and robustness. Furthermore, ECG-Doctor provides well-grounded diagnostic explanations, highlighting its superior interpretability.

† Corresponding authors.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761082

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; • **Applied computing** → *Health care information systems*.

Keywords

ECG Diagnosis and Classification, Large Language Model, KNN-based ECG Retrieval, Chain of Thought, In-Context Learning

ACM Reference Format:
Dongsheng Tian, Junzhe Jiang, Kai Zhang†, Changchun Liu, Yu Yuan, Min Gao, and Enhong Chen†. 2025. ECG-Doctor: An Interpretable Multimodal ECG Diagnosis Framework Based on Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761082

1 Introduction

The electrocardiogram (ECG) is a non-invasive, cost-effective diagnostic tool that records cardiac electrophysiological activity and

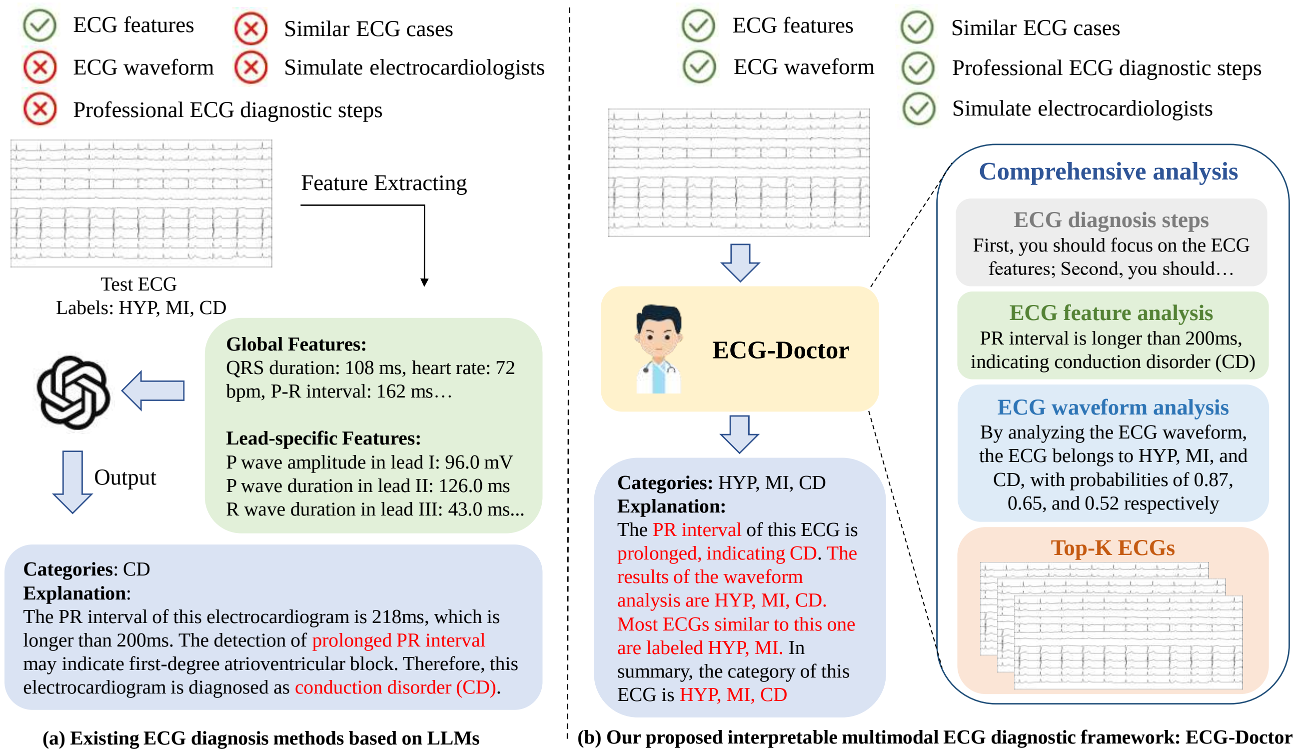

Figure 1: Illustration of the existing LLMs-based ECG diagnosis methods and our proposed interpretable multimodal ECG diagnostic framework, ECG-Doctor. For the same ECG, existing LLMs-based methods only diagnose one of the categories, while ECG-Doctor makes a comprehensive and accurate diagnosis.

plays a vital role in modern clinical diagnosis [2, 4]. While deep learning-based ECG methods have shown strong performance, they rely on large volumes of annotated data, which are costly to obtain. Moreover, their poor interpretability, due to an inability to produce coherent and human-understandable textual explanations, renders them black boxes, which limits trust and hinders their adoption in real-world clinical practice.

In recent years, Large Language Models (LLMs) have been widely adopted across various research domains, including finance [65] and healthcare [55], owing to their powerful natural language processing capabilities. Notably, the successful application of LLMs to medical imaging tasks, such as chest X-ray interpretation [7], has sparked growing interest in extending their use to ECG diagnosis. Several studies have explored this emerging direction. For example, Yu et al. [69] constructed a knowledge base from ECG textbooks, enabling LLMs to retrieve relevant clinical information and perform ECG diagnosis in conjunction with features extracted from ECG signals. Liu et al. [34] introduced SignalGPT, a framework that integrates LLMs with signal classifiers to analyze multiple types of biosignals, including ECG, and generate descriptive textual reports. The application of LLMs to ECG diagnosis holds significant promise. It not only enables competitive diagnostic performance but also addresses the challenges of limited high-quality labeled data and enhances interpretability, thereby paving the way for broader adoption in clinical practice.

However, current methods that leverage LLMs for ECG diagnosis still face significant limitations. First, these approaches fail to comprehensively utilize the diverse and rich information of ECG signals, thereby falling short in comprehensively simulating the diagnostic process of experienced electrocardiologists. For instance, Liu et al. [35] employed LLMs to extract disease subtypes and attribute-level information to enhance clinical semantics, but did not incorporate signal-level ECG features. Conversely, Yu et al. [69] focused on analyzing ECG features using LLMs but overlooked the waveform-level information. Second, although K-nearest neighbor (KNN) similarity retrieval has been successfully applied to ECG-related question answering tasks [46, 56], this technique has not yet been integrated into ECG diagnosis itself. As a result, existing methods are unable to compare and contrast different ECGs effectively, limiting their capacity for analogy-based learning [34]. Finally, current LLM-based ECG diagnostic frameworks do not align with standard clinical workflows and therefore cannot replicate the structured reasoning process followed by trained cardiologists [34, 69]. As illustrated in Figure 1(a), for the ECG instance labeled with hypertrophy (HYP), conduction disturbance (CD), and myocardial infarction (MI), existing methods, rely solely on ECG feature analysis. These approaches do not adhere to standard clinical diagnostic workflows and fail to emulate the reasoning process of trained electrocardiologists. Consequently, they produce incomplete diagnostic results, identifying only a single abnormality.

Inspired by the diagnostic procedures of professional electrocardiologists [22, 28], we propose **ECG-Doctor**, an interpretable multimodal ECG diagnosis framework grounded in LLMs, to overcome the limitations of existing LLM-based methods in capturing comprehensive diagnostic insights. ECG-Doctor comprises four key modules: (1) **ECG Knowledge Acquisition Module**. This component integrates external domain knowledge with Chain-of-Thought (CoT) reasoning [64] to ensure that LLMs adhere to professional ECG diagnostic workflows; (2) **ECG Feature Extraction Module**. It leverages both demographic information and structured ECG features of the patient. External knowledge is incorporated to enhance the LLMs' understanding of these clinical features; (3) **ECG Waveform Analysis Module**. This module integrates time-series ECG models and processes raw ECG waveforms to guide LLMs in comprehending the temporal patterns inherent in ECG signals; (4) **KNN-Based ECG Retrieval Module**. By retrieving the top-k most similar ECG cases using KNN algorithm, this module facilitates comparative analysis and enhances the LLMs' diagnostic reasoning through In-Context Learning (ICL) [5].

The outputs of these modules are collectively provided to the LLM as diagnostic context, enabling comprehensive ECG diagnosis through ICL. This design allows the LLM to effectively integrate diverse sources of ECG information and more faithfully emulate the diagnostic reasoning and decision-making processes of experienced electrocardiologists. As shown in Figure 1(b), ECG-Doctor performs a comprehensive analysis that integrates ECG features, waveform data, and retrieval of similar cases. This approach follows established diagnostic procedures and effectively models the reasoning process of experienced electrocardiologists. As a result, ECG-Doctor generates a more accurate and complete diagnosis, successfully identifying all three disease categories.

Our contributions are summarized as follows:

- We propose ECG-Doctor, the first interpretable multimodal ECG diagnosis framework based on LLMs that simulates the cognitive and diagnostic reasoning of electrocardiologists by integrating CoT, ICL, and heterogeneous ECG data.
- ECG-Doctor introduces a KNN-based retrieval module that guides LLMs in ICL through similar ECG cases, enabling comparative analysis. To our knowledge, it is the first to combine KNN retrieval, CoT and ICL for ECG diagnosis.
- ECG-Doctor supports diverse LLMs and ECG time-series models, offering strong versatility across systems. To the best of our knowledge, it is the most comprehensive framework for LLM-based ECG diagnosis to date.
- Extensive experiments on the PTB-XL dataset demonstrate that ECG-Doctor not only outperforms all baseline methods, but also exhibits strong robustness and interpretability.

2 Related Work

2.1 ECG Diagnosis

Electrocardiogram (ECG) diagnosis involves the automatic classification of ECG signals into categories such as hypertrophy, myocardial infarction, and other cardiac conditions. Traditional approaches predominantly leverage deep learning techniques [38, 47, 51], treating ECG signals as time series data [18, 42, 60]. Convolutional Neural Networks (CNNs) [19, 21, 63, 73] and Recurrent Neural Networks (RNNs) [9, 20] have been widely adopted for this task [3, 54, 66, 76, 80], achieving promising performance. However, these methods require large volumes of labeled data, which are costly and labor-intensive to obtain in clinical settings [12, 41].

To address this limitation, ECG self-supervised learning (eSSL) approaches have been proposed to learn meaningful representations from unlabeled ECG signals [10, 23, 29, 44, 52]. Mehari and Strodthoff [40] applied general SSL frameworks such as SimCLR [6], BYOL [16], and CPC [45] to ECG pre-training. Comparative eSSL methods like CLOCS [27] and ASTCL [62] enhance performance by leveraging temporal-spatial invariance and adversarial learning, respectively. Generative approaches [25, 43, 53, 75] instead reconstruct masked ECG segments to learn latent representations. Despite their success, these eSSL methods primarily focus on ECG signals alone, neglecting multimodal data that could enrich semantic understanding. To overcome this, recent studies have explored multimodal self-supervised learning by aligning ECG signals with other modalities [24, 32, 49, 71, 74]. For example, Lalam et al. [30] aligned ECG signals with electronic health records (EHRs), while Liu et al. [35] aligned ECGs with their corresponding textual reports to improve representation learning.

2.2 Large Language Models

LLMs are typically built upon the Transformer architecture, pre-trained on massive text corpora using unsupervised learning, and exhibit strong capabilities in natural language understanding and generation. In recent years, LLMs have achieved remarkable advances across diverse domains, spanning motion generation [79], material synthesis [17], computer-aided design [72], drug discovery [67], and Chinese spelling check [11, 33, 36], among others [77].

Beyond these applications, LLMs have also gained increasing traction in the medical and healthcare domains. Ren et al. [50] introduced Healthcare Copilot, a framework that leverages LLMs for medical consultations. Kim et al. [26] proposed Health-LLM, which utilizes prompt-based or fine-tuned LLMs for health prediction from wearable sensor data. Liu et al. [37] designed a clinical evaluation framework combining general-purpose LLMs with Retrieval-Augmented Generation (RAG) [13, 68] to assess their medical reasoning capabilities. In another study, Liu et al. [39] applied LLMs to cuffless blood pressure estimation using wearable biosignals. Chen et al. [7] proposed AdaMatch-Cyclic for bidirectional generation between chest X-rays and corresponding diagnostic reports. Tang et al. [56] further extended the use of LLMs to ECG report generation and question-answering tasks, highlighting their potential in cardiology-related applications.

2.3 ECG Diagnosis Based on LLMs

Recent research has increasingly explored leveraging the natural language generation and reasoning capabilities of LLMs for ECG diagnosis. Yu et al. [70] employed RAG to generate textual descriptions of ECG signals, aligning these with the raw signals for pre-training. Tian et al. [59] enhanced ECG reports by incorporating medical term explanations and background knowledge, then aligned the original ECG signals, enhanced reports, and structured labels for contrastive learning. Liu et al. [35] aligned ECG signals with clinical reports, generated category-level descriptions using

RAG, and computed similarity between ECG embeddings and category embeddings for classification. Yu et al. [69] constructed a knowledge base from ECG textbooks, retrieved relevant knowledge, and combined it with signal-extracted features for diagnosis using LLMs. Liu et al. [34] introduced SignalGPT, which integrates LLMs with signal classifiers to interpret various biosignals, including ECG, and generate textual diagnostic reports.

Despite these advancements, most existing LLM-based ECG diagnostic methods remain limited in scope. They primarily employ LLMs for tasks such as RAG [31] and data augmentation to enrich clinical semantics, or rely solely on features extracted from ECG signals for diagnosis. These approaches lack adherence to professional diagnostic workflows, fail to integrate multiple modalities of diagnostic information, and ultimately fall short of emulating the comprehensive reasoning process employed by expert electrocardiologists in real-world clinical practice.

3 Method

3.1 Overview

ECG-Doctor is an interpretable multimodal framework for ECG diagnosis, built upon LLMs. By integrating both signal-based and text-based modalities, ECG-Doctor effectively leverages diverse sources of information and closely simulates the diagnostic reasoning process of experienced cardiologists. The framework requires only a limited amount of labeled data to achieve accurate and efficient ECG diagnosis while maintaining strong interpretability.

The overall architecture of ECG-Doctor is illustrated in Figure 2. The process begins with the ECG knowledge acquisition module, which retrieves domain-specific knowledge, namely the diagnosis guideline, and standardized ECG diagnostic procedures from an external knowledge base. For each test ECG signal, three modules are applied in parallel: the ECG feature extraction module, the ECG waveform analysis module, and the KNN-based ECG retrieval module. The ECG feature extraction module first derives global and lead-specific features from the ECG signal. These features are then processed by an LLM, which selects those most relevant to the diagnosis of five target categories, identifies potential abnormalities, and interprets the features in accordance with the diagnosis guideline. The ECG waveform analysis module extracts morphological characteristics from the signal and predicts diagnostic categories along with the associated probabilities. The KNN-based ECG retrieval module encodes the input signal and retrieves the top-k most similar ECGs from a pre-built embedding database. The cosine similarities and diagnostic labels of these retrieved cases are used as references to support ICL [5]. Finally, the outputs from the above modules are aggregated and fed into the LLM. Leveraging techniques such as CoT prompting [64] and ICL, the LLM performs a comprehensive and interpretable diagnosis by adhering to both the diagnosis guideline and standardized ECG diagnostic procedures.

3.2 ECG Knowledge Acquisition Module

In our experiments, we utilize open-source LLMs, such as Llama-3.1-8B-Instruct [15]. Compared to commercial models like GPT-4 [1] and Gemini 1.5 Flash [57], these open-source LLMs are trained on smaller corpora and contain significantly fewer parameters. As a result, the amount of medical knowledge embedded within these models is limited. Directly relying on their internal knowledge for ECG diagnosis may lead to hallucinations, which can degrade performance and pose serious safety risks in clinical applications.

To address these issues, we introduce the ECG knowledge acquisition module. In this module, GPT-4 is employed as an external knowledge source. We construct prompts to retrieve knowledge closely related to the diagnosis of five ECG categories: Conduction Disturbance (CD), Hypertrophy (HYP), Myocardial Infarction (MI), Normal ECG (NORM), and ST/T Change (STTC). The information is consolidated and referred to as the diagnosis guideline. In addition, we extract standardized ECG diagnostic procedures from the same knowledge source. By combining this information with CoT prompting [64], we guide the open-source LLMs to follow logical reasoning paths during diagnosis. All retrieved knowledge is reviewed and validated by professional electrocardiologists before being used to guide the LLMs. This process mitigates the limitations of relying solely on the LLMs' inherent knowledge and significantly reduces the likelihood of hallucinations, thereby improving the reliability and accuracy of ECG diagnosis.

3.3 ECG Feature Extraction Module

The primary function of the ECG feature extraction module is to extract both global and lead-specific features from each test ECG instance, identify potential abnormalities, and thereby fully leverage ECG feature information. This process emulates how electrocardiologists apply their clinical expertise to identify abnormal features.

For each ECG instance, we first obtain patient demographic features from the PTB-XL dataset, including age, gender, height, and weight, as these attributes are clinically relevant to ECG interpretation. For example, normative ranges for certain ECG features such as heart rate and PR interval vary significantly across age groups. We then extract the global features and the 12-lead-specific features from the ECG signal, such as the QT interval, P-wave morphology in lead II, and R-wave amplitude in lead V5. The extraction order follows the conventional clinical reading sequence used by electrocardiologists: global features first, followed by features from the limb leads (I, II, III, aVR, aVL, aVF), and then features from the precordial leads (V1–V6). For each test ECG sample, the extracted features comprise demographic information, global features, and lead-specific features. Due to the large number of extracted features, directly inputting all of them into the LLM may exceed its context length and introduce redundant information, as many features are not directly relevant to diagnosing the five target categories: CD, HYP, MI, NORM, and STTC. To address this, we employ the LLM to select and retain only those features most pertinent to diagnosing these categories. The LLM then identifies potential abnormalities based on the diagnosis guideline.

3.4 ECG Waveform Analysis Module

The waveform analysis module extracts local features from ECG signals for classification, emulating the diagnostic process of cardiologists who assess waveform morphology to detect abnormalities.

We adopt the KED model [59], which comprises four components: an ECG signal encoder, a knowledge encoder, a label query network (LQN), and a classification head. The LQN employs text

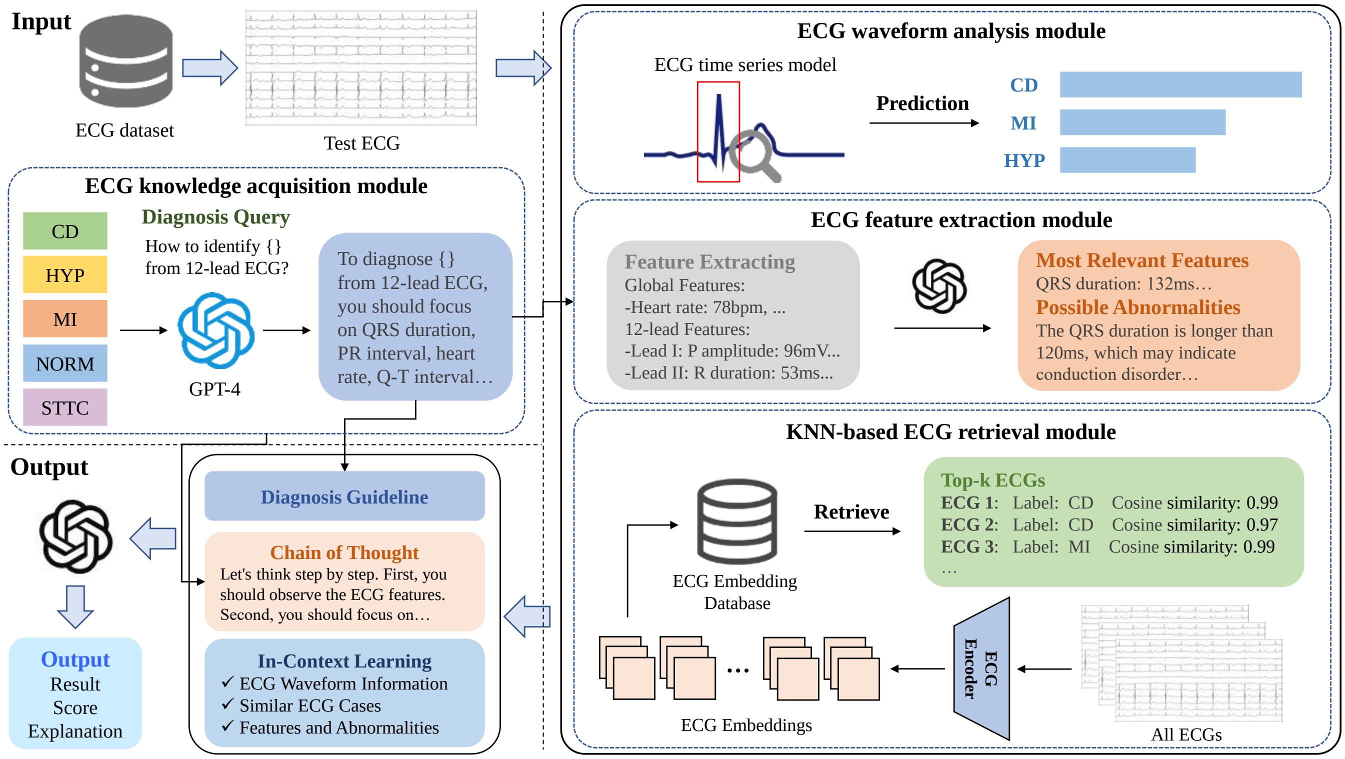

Figure 2: The overall framework of ECG-Doctor, which consists of four modules: ECG knowledge acquisition module, ECG feature extraction module, ECG waveform analysis module and KNN-based ECG retrieval module.

label encodings as queries and uses either ECG encodings or report encodings as keys and values to estimate label probabilities.

KED is pretrained with augmented signal-text-label contrastive learning (AugCL) to incorporate domain knowledge. Structured diagnostic labels are extracted from ECG reports and enriched with LLM-generated explanations, then aligned with enhanced reports and ECG signals through multimodal contrastive learning. Pretraining on MIMIC-IV-ECG [14] jointly optimizes AugCL and binary cross-entropy loss.

At inference, ECG signals and labels are separately encoded and passed through the LQN to compute category probabilities. Labels are assigned when probabilities exceed the thresholds that maximize the Matthews Correlation Coefficient (MCC). For each test ECG, the waveform analysis module predicts the diagnostic categories along with the corresponding probabilities.

3.5 KNN-based ECG Retrieval Module

The KNN-based ECG retrieval module plays a crucial role in enhancing diagnostic accuracy by introducing case-based analogical reasoning into ECG-Doctor. Its primary function is to identify the top-k ECG signals most similar to a given test ECG and to leverage both their diagnostic labels and cosine similarity scores as contextual information. This contextual data is provided to the LLM to guide ICL [5], thereby enabling the model to reason by analogy and closely emulating how expert electrocardiologists consult historical cases, especially in rare diagnostic scenarios.

To construct the retrieval system, we employ the xresnet1d_101 ECG encoder [19] in the KED model [59], to transform all ECG signals in the PTB-XL dataset into fixed-dimensional vector embeddings. These embeddings, which capture both morphological and temporal characteristics of the signals, are pre-computed and stored in an ECG embedding database for efficient retrieval. During inference, the same encoder is applied to generate an embedding for the test ECG, and cosine similarities are computed between this embedding and all other entries in the database, excluding the test sample itself. The top-k most similar ECGs are retrieved based on similarity scores. The diagnostic labels associated with these retrieved ECGs, along with their similarity values, are then formatted into a structured prompt and appended to the LLM's input context.

This process not only introduces clinically relevant prior examples but also encourages the LLM to perform pattern matching and diagnostic generalization grounded in empirical data. By explicitly incorporating analogical reasoning into the diagnostic process, the KNN-based module effectively leverages information from similar ECG cases and significantly enhances the model's generalization capability across diverse ECG morphologies. The full retrieval pipeline is detailed in Algorithm 1.

3.6 Prompts Preparation

Prompts play a critical role in eliciting the capabilities of large language models (LLMs), particularly in zero-shot and few-shot

Table 1: Description of various prompts in ECG-Doctor.

Description of Prompts
• **System Prompt** To further enhance task alignment, a system-level prompt such as "You are a professional electrocardiologist." is typically employed to explicitly assign an expert role to the LLM, thereby improving its problem-solving effectiveness in the ECG diagnostic domain.
• **Diagnosis Guideline** This component provides domain-specific ECG knowledge obtained from GPT-4, which has been reviewed and validated by certified electrocardiologists. Such knowledge can be retrieved by prompting GPT-4 with queries like: "You are a professional electrocardiologist. I need you to teach me how to identify conduction disturbance (CD) in 12-lead ECGs. What leads or features should I focus on? Please limit your response to under 100 words."
• **Feature Filtering Prompt** This prompt guides the LLM to identify and retain the most diagnostically relevant features from the input, based on the provided diagnosis guideline, and to assess potential abnormalities.
• **Feature Prompt** Feature Prompt. Serving as the output of the ECG feature extraction module, this prompt encapsulates the retained features along with the abnormalities identified by the LLM, as filtered through the feature filtering prompt.
• **Diagnosis Prompt** This prompt corresponds to the output of the ECG waveform analysis module. It includes the predicted diagnostic categories of the input ECG, as determined by the KED model, along with the associated probabilities.
• **KNN Prompt** As the output of the KNN-based ECG retrieval module, this prompt presents diagnostic labels and cosine similarities of the top-k retrieved ECGs. These contextual examples guide the LLM through ICL.
• **CoT Prompt** This prompt leverages CoT reasoning, typically initiated with the phrase "Let's think step by step." It guides the model to first focus on ECG waveform information, then on extracted features and potential abnormalities, and finally on similar ECG cases to support diagnosis. The model is instructed to integrate these sources of evidence into a coherent reasoning process, rather than treating them in isolation. In this way, the prompt emulates the diagnostic reasoning of electrocardiologists and enhances the LLM's analytical capability and diagnostic accuracy.
• **Output Prompt** This component instructs the LLM to classify the ECG into one or more of the following categories: CD, HYP, MI, NORM, or STTC. It also standardizes the output format as a JSON object with three fields: result, score, and explanation. Specifically, result denotes the predicted diagnostic categories; score denotes the certainty level of the LLM's diagnostic conclusion (ranging from 0 to 100), and explanation provides a rationale for the classification decision.

Algorithm 1: k-NN Few-Shot ECG Retrieval

Input: Number of neighbors k, Encoder function Encoder$(\cdot)$, dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, test ECG list $\mathcal{T} = \{(x_t, y_t)\}_{t=1}^{T}$
Output: Final result $\mathcal{R}$ (including labels and cosine similarities of the k closest ECGs) for each test ECG.
Initialize empty embedding map E;
for (x_i, y_i) *in dataset* $\mathcal{D}$ **do**
  $E_i \leftarrow$ Encoder(x_i) `// Precompute embeddings`
end
Initialize result list $\mathcal{R} \leftarrow \emptyset$;
for *test ECG* (x_t, y_t) *in* $\mathcal{T}$ **do**
  $C \leftarrow \emptyset$;
  Find k closest embeddings E_{t_i} to E_t, i.e. the unique indices $t_1, \ldots, t_k$ that maximize: $\sum_{i=1}^{k} CosineSimilarity(E_t, E_{t_i})$; such that $t_i \neq t$, and $c_{t_i} = CosineSimilarity(E_t, E_{t_i}) = \frac{E_t \cdot E_{t_i}}{\|E_t\| \|E_{t_i}\|}$;
  $C \leftarrow \{(c_{t_i}, y_{t_i})\}_{i=1}^{k}$;
  $\mathcal{R} \leftarrow \mathcal{R} \cup C$;
end
return $\mathcal{R}$

settings [49]. In the ECG-Doctor framework, prompts are systematically organized into seven components: diagnosis guideline, feature filtering prompt, feature prompt, diagnosis prompt, KNN prompt, CoT prompt, and output prompt. A detailed description of each prompt is provided in Table 1.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate the proposed ECG-Doctor framework on PTB-XL, a widely used benchmark dataset for ECG diagnosis. PTB-XL contains 21,837 clinical 12-lead ECG recordings, each 10 seconds long, collected from 18,885 patients. The recordings are available at two sampling rates: 100Hz and 500Hz. The dataset is gender-balanced (52% male, 48% female) and spans a wide age range (0–95 years; median: 62; interquartile range: 22). Each ECG record is annotated by cardiologists with one or more of 71 ECG statements, following the SCP-ECG standard. These annotations encompass form, rhythm, and diagnostic statements. For diagnostic evaluation, PTB-XL provides a hierarchical label structure comprising 5 diagnostic superclasses and 24 subclasses. The five superclasses are: Conduction Disturbance (CD), Hypertrophy (HYP), Myocardial Infarction (MI), Normal ECG (NORM), and ST/T Changes (STTC). Our evaluation focuses on these five diagnostic superclasses at both 100Hz and 500Hz sampling rates.

Implementation Details. We systematically evaluate the performance of ECG-Doctor on the ECG classification task using the PTB-XL dataset. To assess its robustness and portability, we conduct experiments under two sampling rates (100 Hz and 500 Hz), simulating variations in signal quality and device specifications commonly encountered in real-world scenarios. Furthermore, to demonstrate the versatility and scalability of ECG-Doctor across different model architectures, we integrate and compare a range of large language models (LLMs) with parameter sizes from 7B to 14B, including Qwen2.5-7B-Instruct [48], Llama-3.1-8B-Instruct [15],

Table 2: ECG diagnostic performance of different methods at 100Hz and 500Hz sampling rates on PTB-XL. The best results are bolded, with underline indicating the second highest.

Method	Model	100 Hz				500 Hz			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
SSL	SimCLR	0.010	0.135	0.534	0.213	0.051	0.408	0.396	0.291
	BYOL	0.011	0.152	0.603	0.241	0.054	0.431	0.419	0.308
	MoCo-v3	0.012	0.155	0.616	0.246	0.056	0.448	0.434	0.320
	CLOCS	0.011	0.146	0.580	0.232	0.054	0.428	0.416	0.306
	ASTCL	0.011	0.154	0.61	0.244	0.056	0.452	0.438	0.323
	CRT	0.011	0.148	0.586	0.234	0.057	0.457	0.444	0.327
	ST-MEM	0.010	0.131	0.518	0.207	0.049	0.391	0.379	0.279
	MERL	0.013	0.175	0.693	0.277	0.063	<u>0.504</u>	0.489	0.360
	KED	0.323	0.466	0.732	0.562	0.122	0.346	<u>0.530</u>	0.335
ECG-Doctor (ours)	Qwen2.5-7B-Instruct	0.380	0.526	0.713	0.577	0.279	0.403	0.486	0.401
	Llama-3.1-8B-Instruct	<u>0.430</u>	<u>0.546</u>	0.669	0.573	<u>0.313</u>	0.421	0.465	0.407
	gemma-2-9B-it	0.319	0.475	**0.804**	0.573	0.193	0.366	**0.574**	**0.436**
	Llama-3.2-11B-Vision-Instruct	0.342	0.523	<u>0.746</u>	<u>0.594</u>	0.264	0.414	0.510	<u>0.425</u>
	Qwen2.5-14B-Instruct	**0.440**	**0.562**	0.675	**0.599**	**0.428**	**0.507**	0.379	0.376

Gemma-2-9B-it [58], Llama-3.2-11B-Vision-Instruct[1], and Qwen2.5-14B-Instruct [48]. In all experiments, the number of retrieved similar ECGs k is set to 5. ECG features are directly obtained from the PTB-XL+ dataset, an extension of PTB-XL that leverages commercial and open-source toolkits to extract structured features such as base points, median beats, fiducial points, durations, and amplitudes from the original ECG signals. We adopt the official train-test split [61] and perform zero-shot inference on the test set without using any training data for fine-tuning.

Baselines. We compare our approach against three categories of baseline methods: 1) Comparative ECG self-supervised learning (eSSL) methods, which leverage temporal and spatial invariances or employ adversarial learning strategies during pretraining. This category includes SimCLR [6], BYOL [16], MoCo-v3 [8], CLOCS [27], and ASTCL [62]; 2) Generative eSSL methods, which learn ECG representations through masked segment reconstruction tasks, such as CRT [78] and ST-MEM [43]; 3) Multimodal self-supervised learning (SSL) methods, which enhance performance by aligning ECG signals with other modalities (e.g., textual descriptions), including MERL [35] and KED [59]. The detailed architecture and mechanism of the KED model are described in Section 3.4.

Metrics. We evaluate the classification performance of ECG-Doctor on the PTB-XL dataset using accuracy, macro precision, macro recall, and macro F1-score. In the context of multi-label classification, accuracy is a stringent metric because a prediction is considered correct only when the entire set of predicted diagnostic labels exactly matches the ground truth for a given ECG instance.

4.2 Main Results

We evaluate the classification performance of ECG-Doctor on the PTB-XL dataset at two different sampling rates: 100 Hz and 500 Hz. The detailed experimental results are presented in Table 2.

At a sampling rate of 100 Hz, ECG-Doctor consistently outperforms all baseline methods in terms of accuracy, precision, and macro-F1 across all LLM configurations, with the sole exception of gemma-2-9B-it, where the accuracy is marginally lower than that of the KED baseline. Notably, when using Qwen2.5-14B-Instruct, ECG-Doctor achieves the highest macro-F1 score of 0.599, outperforming the best-performing baseline (KED) at this sampling rate, which attains a macro-F1 of 0.562. This represents a relative improvement of 6.58%. At the higher sampling rate of 500 Hz, ECG-Doctor demonstrates superior performance across all LLMs in terms of both accuracy and macro-F1. In particular, with gemma-2-9B-it, it achieves a macro-F1 of 0.436, significantly surpassing the strongest baseline (MERL), which reaches 0.360, corresponding to a 21.11% relative improvement. Furthermore, at both sampling rates, ECG-Doctor consistently outperforms all baselines in accuracy, irrespective of the LLM used. Except for MERL, which ranks second in precision at 500 Hz, and KED, which ranks second in recall at the same sampling rate, ECG-Doctor achieves either the highest or second-highest scores across all other evaluation metrics, underscoring its comprehensive diagnostic performance.

These results demonstrate the superior diagnostic performance of ECG-Doctor compared to all baselines. Furthermore, when the KED model is incorporated into the ECG-Doctor framework, its diagnostic performance is substantially enhanced.

4.3 Versatility and Robustness

To evaluate the versatility of ECG-Doctor, we conduct experiments using LLMs from diverse model families (e.g., Qwen2.5 [48], Llama3 [15] and gemma2 series [58]), varying in scale (e.g., Llama-3.1-8B-Instruct, Qwen2.5-14B-Instruct) and modality (text-only vs. multimodal models such as gemma-2-9B-it and Llama-3.2-11B-Vision-Instruct[2]). As shown in Table 2, ECG-Doctor consistently outperforms all baseline methods across both 100 Hz and 500 Hz sampling rates, regardless of the underlying LLM employed. These results

[1]https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

[2]https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

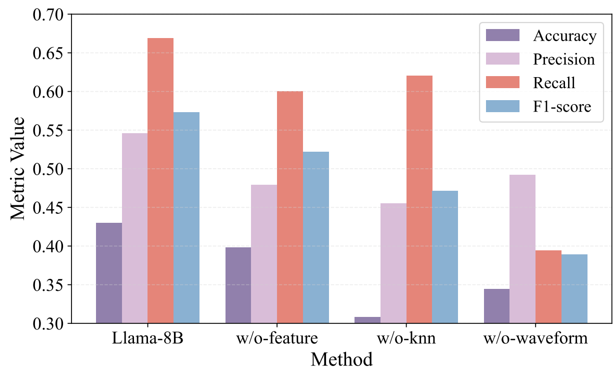

Figure 3: Ablation study on different modules of ECG-Doctor. Llama-8B denotes Llama-3.1-8B-Instruct. w/o-feature, w/o-waveform, and w/o-knn indicate the exclusion of the ECG feature extraction module, ECG waveform analysis module, and KNN-based ECG retrieval module, respectively.

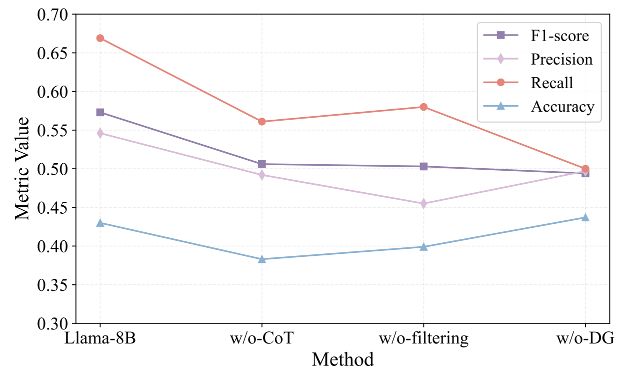

Figure 4: Ablation study on different prompts in ECG-Doctor. Llama-8B refers to Llama-3.1-8B-Instruct. w/o-CoT, w/o-filtering, and w/o-DG denote the exclusion of the CoT, feature filtering prompt, and diagnosis guideline, respectively.

highlight the strong versatility and adaptability of ECG-Doctor across diverse model architectures and diagnostic scenarios.

To examine the robustness of ECG-Doctor under varying signal conditions, we conducted experiments on the PTB-XL dataset using both 100 Hz and 500 Hz sampling rates, simulating the impact of signal quality variations or device discrepancies commonly encountered in real-world clinical settings. As shown in Table 2, ECG-Doctor consistently outperforms baseline methods in terms of accuracy and macro-F1 across both sampling rates and regardless of the underlying LLM employed. These results demonstrate the strong robustness of ECG-Doctor and highlight its great potential for deployment in complex and dynamic medical environments.

4.4 Scaling law

A detailed analysis of the experimental results reveals that LLMs with larger parameter scales generally exhibit superior ECG diagnostic performance within our framework, as shown in Table 2. For instance, at a sampling rate of 100 Hz, Qwen2.5-14B-Instruct and Llama-3.2-11B-Vision-Instruct outperform smaller models in terms of macro F1-score. Similarly, at 500 Hz, gemma-2-9B-it and Llama-3.2-11B-Vision-Instruct achieve higher macro F1-scores than Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. This performance gain may be attributed to the fact that larger models are typically pre-trained on more extensive corpora and possess stronger natural language understanding and reasoning capabilities, which are critical for complex diagnostic tasks.

4.5 Ablation Studies

Different Modules. We ablate the ECG feature extraction, waveform analysis, and KNN-based retrieval modules to evaluate their individual contributions. Experiments are conducted on the PTB-XL test set using Llama-3.1-8B-Instruct at 100 Hz. Results in Figure 3 show that removing any module significantly degrades accuracy, precision, recall, and F1-score. The ECG waveform analysis module proves to be the most critical component, as removing it results in a substantial macro-F1 drop from 0.573 to 0.389, underscoring the diagnostic importance of rhythm and morphological information in the ECG waveform. Eliminating the KNN-based ECG retrieval module leads to a decrease in macro-F1 to 0.471, highlighting the utility of analogical reasoning based on similar cases. Additionally, when the ECG feature extraction module is removed, the macro-F1 drops to 0.522, indicating the significance of structured features such as QTc, which typically require precise calculation and are not easily inferred from raw waveforms.

Different Prompts. The prompts in ECG-Doctor are primarily composed of seven components designed to support comprehensive ECG interpretation. Ablation of the diagnosis, KNN, and feature prompts corresponds to removing the ECG feature extraction, waveform analysis, and retrieval modules, respectively, and the results have already been presented in the previous section. The output prompt, which controls format for downstream tasks, is excluded from ablation. Thus, we focus on the diagnosis guideline, feature filtering, and CoT prompts, with ablation experiments conducted on the PTB-XL test set using Llama-3.1-8B-Instruct at 100 Hz, as shown in Figure 4. Removing any of these three prompts significantly reduces precision, recall, and F1-score, confirming their importance. The largest drop occurs when the diagnosis guideline is removed (F1 from 0.573 to 0.494), highlighting the necessity of external domain knowledge, particularly given the limited ECG expertise of sub-14B parameter LLMs. Excluding the feature filtering prompt lowers F1 to 0.503, likely due to noise from irrelevant features and input truncation. Removing the CoT prompt results in an F1 of 0.506, suggesting that guided reasoning is essential for integrating diverse diagnostic signals into accurate predictions.

4.6 Case Study

Interpretability is a key advantage of ECG-Doctor over traditional ECG time-series models. To evaluate the interpretability of ECG-Doctor, we conduct a case study involving three representative samples from the test dataset. These cases span all five diagnostic

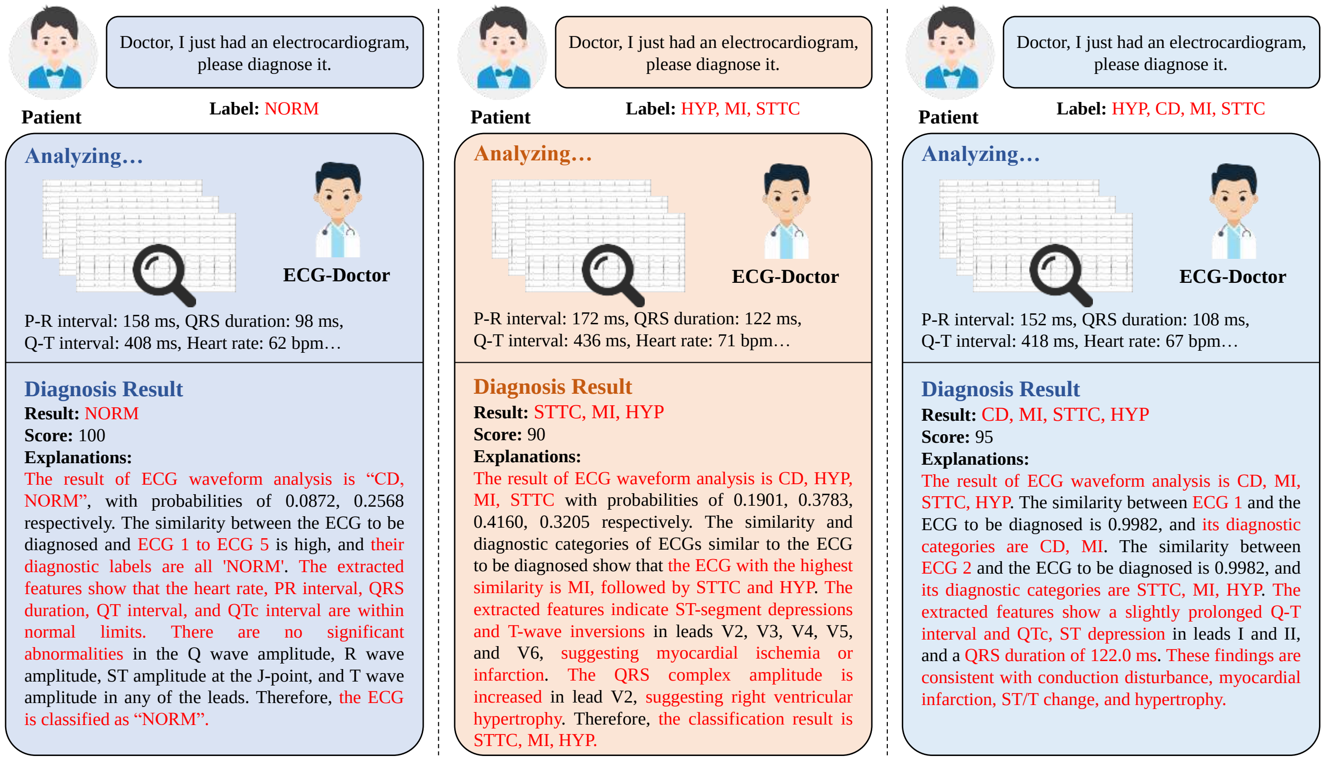

Figure 5: Case studies on the interpretability of ECG-Doctor across diverse ECG diagnostic scenarios.

superclasses defined in PTB-XL and include two particularly challenging multi-label classification instances. A certified electrocardiologist reviews and assesses the model's diagnostic rationale and explanations. The ground-truth labels, diagnostic results (including classification result, certainty score, and explanation) produced by ECG-Doctor for each case are presented in Figure 5.

Across all three cases, ECG-Doctor produces correct diagnoses with high certainty scores. Notably, in the complex multi-label scenarios represented by Case 2 and Case 3, the model maintains strong diagnostic accuracy and interpretability, demonstrating the robustness of our framework. An analysis of ECG-Doctor's interpretations shows that it identifies abnormalities, highlights disease risks, summarizes waveform analysis, and references similar cases, demonstrating a comprehensive diagnostic process that integrates diverse information sources.

Importantly, our analysis reveals that the modules within ECG-Doctor interact dynamically throughout the diagnostic process. Instead of aggregating results from different modules in a naive manner, ECG-Doctor performs integrated reasoning across the ECG feature extraction, waveform analysis, and KNN-based retrieval modules. For instance, in Cases 1 and 2, the waveform analysis initially suggests the presence of "CD", but this is later ruled out through cross-verification with feature-based and retrieval-based evidence, leading to a correct final diagnosis. In Case 3, although the KNN module fails to retrieve examples supporting "STTC" and "HYP", these conditions are accurately identified through feature and waveform analysis. After evaluation, the professional electrocardiologist believes that the interpretability provided by ECG-Doctor is clinically convincing. While achieving superior diagnostic performance compared to baseline methods, ECG-Doctor also offers strong interpretability by providing detailed, transparent explanations of its diagnostic decisions. This capability enhances clinical trust among both physicians and patients.

5 Conclusion

We present ECG-Doctor, an interpretable multimodal ECG diagnosis framework powered by LLMs. It integrates domain knowledge, structured ECG features, waveform analysis, and case retrieval to address LLMs' limitations in standardized diagnosis. By combining heterogeneous ECG data, ECG-Doctor emulates the cognitive reasoning of experienced electrocardiologists. Experiments on PTB-XL demonstrate superior performance over baselines, with enhanced robustness, versatility, and interpretability, highlighting its potential for real-world clinical use.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (U23A20319, 62441239, 2406303), Anhui Provincial Natural Science Foundation (No. 2308085QF229), Anhui Province Science and Technology Innovation Project (202423k09020010), the Fundamental Research Funds for the Central Universities (No. WK2150110034).

GenAI Usage Disclosure

We use ChatGPT-4o for grammar checking and error correction. Our proposed method uses some LLMs, see Section 3 for details.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Paul S Addison. 2005. Wavelet transforms and the ECG: a review. *Physiological measurement* 26, 5 (2005), R155.

[3] Ulas Baran Baloglu, Muhammed Talo, Ozal Yildirim, Ru San Tan, and U Rajendra Acharya. 2019. Classification of myocardial infarction with multi-lead ECG signals and deep CNN. *Pattern recognition letters* 122 (2019), 23–30.

[4] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M Bilginer Gulmezoglu. 2018. A survey on ECG analysis. *Biomedical Signal Processing and Control* 43 (2018), 216–235.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

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

[7] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2023. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. *arXiv preprint arXiv:2312.08078* (2023).

[8] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9640–9649.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[10] Yen-hsiu Chou, Shenda Hong, Yuxi Zhou, Junyuan Shang, Moxian Song, and Hongyan Li. 2020. Knowledge-shot learning: An interpretable deep model for classifying imbalanced electrocardiography data. *Neurocomputing* 417 (2020), 64–73.

[11] Ming Dong, Yujing Chen, Miao Zhang, Hao Sun, and Tingting He. 2024. Rich semantic knowledge enhanced large language models for few-shot chinese spell checking. *arXiv preprint arXiv:2403.08492* (2024).

[12] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. 2020. A review on deep learning methods for ECG arrhythmia classification. *Expert Systems with Applications: X* 7 (2020), 100033.

[13] Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. 2025. Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2504.14891* (2025).

[14] 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. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101, 23 (2000), e215–e220.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[16] 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. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[17] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. 2024. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379* (2024).

[18] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* 25, 1 (2019), 65–69.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34, 6 (2020), 1936–1962.

[22] Rahul Jain, Robin Singh, Sundermurthy Yamini, and Mithilesh K. Das. 2014. Fragmented ECG as a risk marker in cardiovascular diseases. *Current Cardiology Reviews* 10, 3 (2014), 277–286.

[23] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[25] Aofan Jiang, Chaoqin Huang, Qing Cao, Yuchen Xu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang. 2024. Self-supervised Anomaly Detection Pretraining Enhances Long-tail ECG Diagnosis. *arXiv preprint arXiv:2408.17154* (2024).

[26] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866* (2024).

[27] Dani Kiyasseh, Tingting Zhu, and David A Clifton. 2021. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*. PMLR, 5606–5615.

[28] Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. 2007. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. *Circulation* 115, 10 (2007), 1306–1324.

[29] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. 2023. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. *Nature Communications* 14, 1 (2023), 3741.

[30] Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. 2023. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research* (2023).

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[33] Jiahao Li, Quan Wang, Zhendong Mao, Junbo Guo, Yanyan Yang, and Yongdong Zhang. 2022. Improving chinese spelling check by character pronunciation prediction: The effects of adaptivity and granularity. *arXiv preprint arXiv:2210.10996* (2022).

[34] Chunyu Liu, Yongpei Ma, Kavitha Kothur, Armin Nikpour, and Omid Kavehei. 2023. BioSignal Copilot: Leveraging the power of LLMs in drafting reports for biomedical signals. *medRxiv* (2023), 2023–06.

[35] Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659* (2024).

[36] Changchun Liu, Kai Zhang, Junzhe Jiang, Zirui Liu, Hanqing Tao, Min Gao, and Enhong Chen. 2024. ARM: An Alignment-and-Replacement Module for Chinese Spelling Check Based on LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10156–10168.

[37] Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lv, Liya Ma, et al. 2024. Towards Automatic Evaluation for LLMs' Clinical Capabilities: Metric, Data, and Algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5466–5475.

[38] Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. 2021. Deep learning in ECG diagnosis: A review. *Knowledge-Based Systems* 227 (2021), 107187.

[39] Zengding Liu, Chen Chen, Jiannong Cao, Minglei Pan, Jikui Liu, Nan Li, Fen Miao, and Ye Li. 2024. Large language models for cuffless blood pressure measurement from wearable biosignals. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 1–11.

[40] Temesgen Mehari and Nils Strodthoff. 2022. Self-supervised representation learning from 12-lead ECG data. *Computers in biology and medicine* 141 (2022), 105114.

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

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

[43] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2024. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450* (2024).

[44] Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. 2022. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*. PMLR, 338–353.

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

[46] Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. 2025. Q-Heart: ECG Question Answering via Knowledge-Informed Multimodal LLMs. *arXiv preprint arXiv:2505.06296* (2025).

[47] Boris Pyakillya, Natasha Kazachenko, and Nikolay Mikhailovsky. 2017. Deep learning for ECG classification. In *Journal of physics: conference series*, Vol. 913. IOP Publishing, 012004.

[48] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[50] Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, and Dacheng Tao. 2024. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408* (2024).

[51] Giovanna Sannino and Giuseppe De Pietro. 2018. A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems* 86 (2018), 446–455.

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

[53] Shinnosuke Sawano, Satoshi Kodera, Hirotoshi Takeuchi, Issei Sukeda, Susumu Katsushika, and Issei Komuro. 2022. Masked Autoencoder-Based Self-Supervised Learning for Electrocardiograms to Detect Left Ventricular Systolic Dysfunction.. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*.

[54] Shraddha Singh, Saroj Kumar Pandey, Urja Pawar, and Rekh Ram Janghel. 2018. Classification of ECG arrhythmia using recurrent neural networks. *Procedia computer science* 132 (2018), 1290–1297.

[55] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine* (2025), 1–8.

[56] Jialu Tang, Tong Xia, Yuan Lu, Cecilia Mascolo, and Aaqib Saeed. 2025. 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)*. IEEE, 1–5.

[57] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[58] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[59] Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. 2024. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine* 5, 12 (2024).

[60] Rajesh Kumar Tripathy, Abhijit Bhattacharyya, and Ram Bilas Pachori. 2019. A novel approach for detection of myocardial infarction from ECG signals of multiple electrodes. *IEEE Sensors Journal* 19, 12 (2019), 4509–4517.

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

[62] Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. 2023. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[63] Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*. IEEE, 1578–1585.

[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

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

[66] Xue Xu, Sohyun Jeong, and Jianqiang Li. 2020. Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM. *Ieee Access* 8 (2020), 125380–125388.

[67] Jiaxian Yan, Jintao Zhu, Yuhang Yang, Qi Liu, Kai Zhang, Zaixi Zhang, Xukai Liu, Boyan Zhang, Kaiyuan Gao, Jinchuan Xiao, et al. 2025. BioMiner: A Multi-modal System for Automated Mining of Protein-Ligand Bioactivity Data from Literature. *bioRxiv* (2025), 2025–04.

[68] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*. Springer, 102–120.

[69] Han Yu, Peikun Guo, and Akane Sano. 2023. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In *Machine learning for health (ML4H)*. PMLR, 650–663.

[70] Han Yu, Peikun Guo, and Akane Sano. 2024. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text. *arXiv preprint arXiv:2405.19366* (2024).

[71] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).

[72] Yu Yuan, Shizhao Sun, Qi Liu, and Jiang Bian. 2025. CAD-Editor: A Locate-then-Infill Framework with Automated Training Data Synthesis for Text-Based CAD Editing. *arXiv preprint arXiv:2502.03997* (2025).

[73] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

[74] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18123–18133.

[75] Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. 2022. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement* 72 (2022), 1–15.

[76] Jing Zhang, Deng Liang, Aiping Liu, Min Gao, Xiang Chen, Xu Zhang, and Xun Chen. 2021. MLBF-Net: A multi-lead-branch fusion network for multi-class arrhythmia classification using 12-lead ECG. *IEEE journal of translational engineering in health and medicine* 9 (2021), 1–11.

[77] Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. 2021. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (2021), 377–389.

[78] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. 2023. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[79] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7368–7376.

[80] Zhaowei Zhu, Han Wang, Tingting Zhao, Yangming Guo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Xiang Lan, Xingzhi Sun, and Mengling Feng. 2020. Classification of cardiac abnormalities from ECG signals using SE-ResNet. In *2020 Computing in Cardiology*. IEEE, 1–4.