Benchmarking Multi-Modal Cardiological Diagnostics within the LLM-as-Agent Paradigm

Anonymous ACL submission

Abstract

Large language models (LLMs) have revolutionized cardiological diagnostics through agentic design. However, a significant challenge remains: the misalignment between real-world clinical reports used in hospitals and the publicly available datasets used to fine-tune LLMs. This discrepancy limits the reliability of LLMs in cardiological practices. In this work, we address this gap from two key perspectives. First, we introduce Z-BENCH, a benchmark derived from in-hospital cardiological reports, where patient records comprise multimodal electrocardiograms (ECGs) enriched with cardiological metrics. Second, we propose ZODIAC, an LLM-powered agentic framework designed to enhance cardiological diagnostics. ZODIAC operates by systematically extracting clinically relevant characteristics, detecting significant arrhythmias, and generating preliminary diagnostic reports, which are then reviewed and refined by cardiologists. Experimental results demonstrate that ZODIAC surpasses industry-leading LLMs from OpenAI, Meta, Google, and DeepSeek, as well as medical-specialist models such as Microsoft's BioGPT. Our findings highlight the transformative potential of specialized LLMs in healthcare, showcasing their ability to deliver medical solutions that meet the rigorous demands of cardiological guidelines.

1 Introduction

As technology continues to transform healthcare, large language models (LLMs) have become a pivotal component of *digital health* (FDA, 2020). With their human-like conversational abilities and extensive pre-trained knowledge, LLMs are increasingly being adopted as clinical agents (Boonstra et al., 2024; Gala and Makaryus, 2023; Xu et al., 2024). This shift has led to the development of various medical-specialist applications (Chen et al., 2023a, 2024b; ContactDoctor, 2024; Wang et al., 2024c; Luo et al., 2022; Chen et al., 2023b).

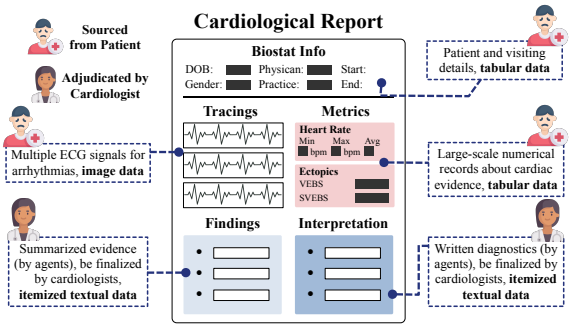

Figure 1: An example layout of cardiological report.

Despite these advancements, the integration of LLMs into real-world cardiac practice is hindered by two critical challenges. First, available public datasets, such as PTB-XL (Wagner et al., 2020), MIMIC-IV-ECG (Gow et al., 2023), and CODE (Ribeiro et al., 2020), exhibit misalignment between the available evidence and the clinical reports used in hospitals. As illustrated in Figure 1, clinical reports are typically synthesized from multimodal evidence (Kline et al., 2022; Cicerone et al., 2000). Training LLMs on misaligned cardiological evidence consequently disrupts their alignment with the standardized diagnostic process, potentially compromising their clinical reliability (Davenport and Kalakota, 2019; Asan et al., 2020).

Second, current LLM-based clinical agents often struggle to achieve cardiologist-level proficiency. While these models may be trained on a broad range of clinical tasks (Peng et al., 2023; Chen et al., 2024a), they lack the specialized alignment for medical decision-making (Khan et al., 2023; Wang et al., 2021; Kerasidou et al., 2022). Those gap underscores the need to enhance the reliability of LLMs in specialized medical fields.

Our Work. We address these gaps through two complementary layers: enhancing data-driven proficiency by benchmarking diagnostic scenarios and advancing technique-driven capabilities within the LLM paradigm:

I) Data Proficiency: We introduce Z-BENCH, a benchmark derived from in-hospital patient records, incorporating cardiologist-adjudicated texts and aligning with clinical guidelines. Z-BENCH ensures cardiologist-level proficiency in two aspects. First, it captures real-world cardiological characteristics, such as arrhythmias and their contributing factors, to accurately reflect clinical realities. Second, the direct involvement of human experts (cardiologists) ensures that the dataset encapsulates expert-level performance, while adherence to clinical guidelines mitigates potential biases and errors, thereby enhancing diagnostic accuracy and safety.

II) LLM-as-Cardiac-Agent: Next, we propose ZODIAC, a cardiologist-level diagnostic agent using **multi-agent** framework to analyze **multimodal** patient records. ZODIAC outperforms single-agent design while improving the identification of key characteristics and interpreting clinically significant arrhythmias (details in Section 5.3). Furthermore, we integrate **instruction tuning** and **in-context learning** into ZODIAC, wherein instruction tuning embeds data proficiency from Z-BENCH into the LLMs, while in-context learning provides professional demonstrations to further reinforce ZODIAC's diagnostics. Finally, we incorporate **fact-checking** against established cardiological guidelines (Goff Jr et al., 2014) to ensure the system generates accurate, expert-verified diagnostics.

Through extensive evaluations on Z-BENCH, we demonstrate that industry-leading LLMs, including OpenAI's ChatGPT, Google's Gemini, and Meta's Llama, fall short in performing cardiological diagnoses based on in-hospital clinical reports. In contrast, ZODIAC not only excels in numerical analysis but also generates expert-level diagnostic narratives and structured reports approbatory by cardiologists. Furthermore, we show that ZODIAC is generalizable in analyzing other ECG datasets beyond Z-BENCH, highlighting its adaptiveness to meet diverse organizational needs.

In summary, this work makes the following contributions:

- We introduce Z-BENCH, a benchmark aligned with in-hospital cardiological reports, enhancing data-driven proficiency in AI research and model development.
- We develop ZODIAC, which serves as a blueprint for constructing clinical-grade LLM agents while providing a scalable framework applicable across various clinical domains.
- Through evaluations, we demonstrate the practical applications of Z-BENCH in integrating human insights through the AI development and validate the effectiveness of ZODIAC in advancing clinical AI development with high reliability.

For anonymization, we temporarily release the benchmark and code at `https://anonymous.4open.science/r/Zbench-Zodiac-8A2A`.

2 Related Work

LLMs in Clinical Diagnostics. LLMs have shown considerable progress in processing and interpreting vast amounts of unstructured medical data, such as patient records, medical literature, and diagnostic reports. For example, Han et al. (2024) introduced a system that automatically summarizes clinical notes during interactions between patients and clinicians, while Ahsan et al. (2023) explored the role of LLMs in retrieving key evidence from electronic health records (EHRs). Despite these successes, concerns persist regarding LLMs' domain-specific expertise and professional performance in high-stakes, life-critical clinical settings (Nashwan and AbuJaber, 2023; Jahan et al., 2024; Wang et al., 2024a; Li et al., 2024). This work addresses these concerns by designing and validating ZODIAC through our design and experiments specifically for cardiological diagnostics.

Cardiological Diagnostic Systems. Current cardiological diagnostic systems primarily depend on rule-based algorithms or single-agent approaches for identifying cardiovascular risk factors or predicting cardiac events (Goff Jr et al., 2014; Sud et al., 2022; Olesen et al., 2012). In recent years, deep learning models have been introduced into cardiology (Hannun et al., 2019; Acharya et al., 2019). However, there remains a significant gap in incorporating recent LLMs into cardiological diagnostics—a gap that this work addresses significantly.

Multi-Agent Frameworks. Multi-agent frameworks have been extensively studied to enhance LLM capabilities in handling complex tasks and managing distributed processes (Wang et al., 2024b; Hong et al., 2023; Du et al., 2023; Chan et al., 2023). In healthcare, where collaboration across different expertise is essential, multi-agent frameworks have shown their potential in optimizing patient management, coordinating care between various agents (e.g., doctors, nurses, administra-

tive systems), and supporting decision-making processes (Furmankiewicz et al., 2014; Jemal et al., 2014; Shakshuki and Reid, 2015). Recent studies have also focused on leveraging multi-LLM agents to reduce manual tasks in healthcare workflows. For instance, Chen et al. (2024a) employed ChatGPT in distinct roles within a coordinated workflow, to automate tasks like database mining and drug repurposing, while ensuring quality control through role-based collaboration.

3 Z-BENCH: A Cardiac Benchmark Aligned with In-Hospital Diagnostics

This section presents Z-BENCH and its alignment with in-hospital diagnostics. We begin by introducing diagnostic tasks (Section 3.1), followed by the construction of Z-BENCH (Section 3.2).

3.1 Components of Cardiological Data

This paper focuses on diagnosing clinically significant arrhythmias using patient data. We classify the key components into two categories: patient records and diagnostic outputs.

Patient Data is comprised of three sections: *(1) Biostatistical information ($\mathcal{B}$)* provides details about the patient such as date of birth, gender, and age group. *(2) Metrics ($\mathcal{M}$)* summarize cardiological attributes and their numerical values presented in a tabular format, providing an overview of 24-hour monitored statistics for a patient. For example, *AF Burden: 12%* indicates that the patient experienced atrial fibrillation for 12% of the whole monitoring period. *(3) Tracings ($\mathcal{T}$)* includes ECG images depicting clinically significant arrhythmias such as AFib/Flutter (Atrial Fibrillation / Atrial Flutter), Pause, VT (Ventricular Tachycardia), SVT (Supraventricular Tachycardia), and AV Block (Atrioventricular Block). $\mathcal{T}$ presents a concise but representative segment of the 24-hour monitoring, such as a 10-second strip highlighting the highest degree of AV block.

Diagnostic Outputs is comprised of two elements: *Clinical Findings ($\mathcal{F}$)* and *Interpretation ($\mathcal{I}$)*, both presented as expert-crafted natural language statements by cardiologists. $\mathcal{F}$ outlines key observations from clinically relevant characteristics, while $\mathcal{I}$ offers the final diagnostics, interpreting these findings. For example, the finding *PR Interval is 210 milliseconds in the ECG tracings* leads to the interpretation: *The PR interval is slightly prolonged, suggesting a first-degree AV block.*

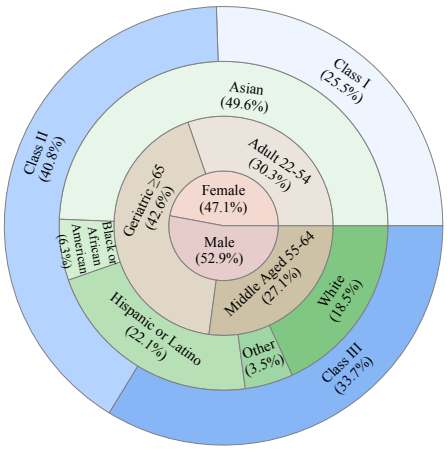

Figure 2: Statistics of Z-BENCH, subgrouped by gender, age, race, and arrhythmia classes – Class I: normal arrhythmias. Class II: clinically significant arrhythmias. Class III: life-threatening arrhythmias. Detailed clinical implications are provided in Appendix C.

Once $\mathcal{F}$ and $\mathcal{I}$ are completed by cardiologists (or by ZODIAC), a clinical end-of-study report is generated for the patient, including $(\mathcal{B}, \mathcal{M}, \mathcal{T}, \mathcal{F}, \mathcal{I})$, as illustrated in Figure 1.

3.2 Data Collection and Cardiologist-Incorporated Curation

Z-BENCH is characterized as *real*, *representative*, and *cardiologist-incorporated*.

Real-World Patient Data. Instead of relying on existing third-party or synthetic data—which often raise concerns about trustworthiness or misalignment with clinical practice (Chouffani El Fassi et al., 2024; Fehr et al., 2024; Youssef et al., 2024)—we collect ECG data sourced from our collaborating healthcare institutions under an IRB-approved protocol, with de-identified patient to ensure privacy protection. The raw data consists of 270+ metrics $\mathcal{M}$ and 5 ECG tracing $\mathcal{T}$ per patient. To ensure the clinical relevance, we engaged five cardiologists to review the data, resulting in a final dataset of 5,400+ samples. Of these, 1,500 were used for evaluation (Section 5), while the remainder were used for fine-tuning (Section 4.2).

Representative Groups. Z-BENCH adheres to the FDA's guidelines (Food et al., 2021) to ensure a representative population, encompassing comprehensive arrhythmia types and diverse racial groups while maintaining balanced age and gender demographics, as detailed in Figure 2.

Incorporating Cardiologist-Level Expertise. When reviewing the raw data, cardiologists are

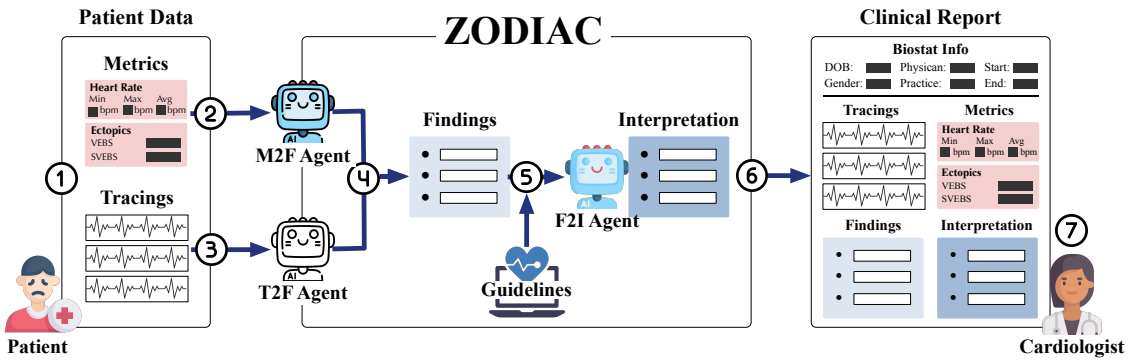

Figure 3: ZODIAC aligns with cardiological practice through a multi-agent framework that integrates patient data across various modalities: ① Patient data is collected in two modalities: tabular metrics and ECG tracings (images). ② A metrics-to-findings LLM agent processes the tabular metrics and generates text-based clinical findings. ③ An tracings-to-findings LLM agent analyzes the ECG tracings to produce additional text-based clinical findings. ④ The clinical findings from both agents are then combined. ⑤ A findings-to-interpretation LLM agent synthesizes these findings with clinical guidelines into comprehensive diagnostic interpretation. ⑥ ZODIAC generates a patient-specific report by integrating the metrics, tracings, clinical findings, and diagnostic interpretation. ⑦ A cardiologist validates the quality of the generated findings and interpretations (details in Section 5). For simplicity, we omit the biostatistics ($\mathcal{B}$) in this figure, which is considered in steps ①②③ by default.

asked to write professional findings ($\mathcal{F}$) and interpretation ($\mathcal{I}$) in accordance with established clinical guidelines ($\mathcal{G}$) (Association et al., 2023; Krumholz et al., 2020). This process facilitates the follow-up fine-tuning, embedding cardiologist-level expertise into LLMs. To save the cardiologists' time,Additionally, each cardiologist randomly audits at least 50% of their peers' drafts to identify and rectify issues such as incompleteness, inconsistencies, or diagnostic inaccuracies. This peer-review process helps standardize wording and structure across cardiologists, ensuring consistency and reliability in the reports.

4 ZODIAC: LLM-as-Cardiac-Agent

4.1 Formulation of Cardiological Diagnostics

Real-world cardiological diagnostics begins by reviewing the patient's data $(\mathcal{B}, \mathcal{M}, \mathcal{T})$ to identify clinically relevant characteristics, such as the *PR interval*, which are key for diagnosing arrhythmias. These identified characteristics are then summarized into natural language statements, referred to as findings $\mathcal{F}$, which integrate insights from both tabular metrics $\mathcal{M}$ and image-based ECG tracings $\mathcal{T}$. For example, the *PR interval* is derived from $\mathcal{T}$, while the *AF burden* is obtained from $\mathcal{M}$. Finally, cardiologists synthesize the findings $\mathcal{F}$ with their clinical expertise and the established guidelines $\mathcal{G}$ to form the final interpretation $\mathcal{I}$.

As illustrated in Figure 3, we follow established cardiological practices to develop ZODIAC. Rather than compressing multimodal data into a single LLM, ZODIAC employs a multi-agent collaboration framework, where each agent is responsible for a specific stage of the diagnostic process. This design enhances the LLM's focus on diagnostic behavior within each modality. ZODIAC comprises three specialized agents:

1. Metrics-to-Findings Agent ($\theta_{\texttt{M2F}}$): A table-to-text LLM that extracts key characteristics from metrics $\mathcal{M}$, while incorporating patient biostatistics $\mathcal{B}$ to generate clinical findings.

2. Tracings-to-Findings Agent ($\theta_{\texttt{T2F}}$): An image-to-text LLM that identifies key factors from ECG tracings $\mathcal{T}$, integrates relevant information from $\mathcal{B}$, and produces clinical findings.

3. Findings-to-Interpretation Agent ($\theta_{\texttt{F2I}}$): A text-based LLM that synthesizes findings $\mathcal{F}$ from both the $\theta_{\texttt{M2F}}$ and $\theta_{\texttt{T2F}}$, applies clinical guidelines $\mathcal{G}$, and generates interpretation $\mathcal{I}$.

ZODIAC formulates the diagnostic process as:

$$\begin{aligned} \mathcal{I} &\leftarrow \theta_{\texttt{F2I}}(\mathcal{F}, \mathcal{G}) \quad s.t. \\ \mathcal{F} &\leftarrow \theta_{\texttt{M2F}}(\mathcal{M}, \mathcal{B}) \cup \theta_{\texttt{T2F}}(\mathcal{T}, \mathcal{B}) \end{aligned} \tag{1}$$

wherein $\theta_{\texttt{M2F}}$ and $\theta_{\texttt{T2F}}$ independently generate clinical findings based on $\mathcal{M}$ and $\mathcal{T}$, respectively, which are then combined to form $\mathcal{F}$. This approach adheres to cardiological diagnostics as each finding in $\mathcal{F}$ corresponds to evidence derived from a specific modality – either metrics or ECG tracings.

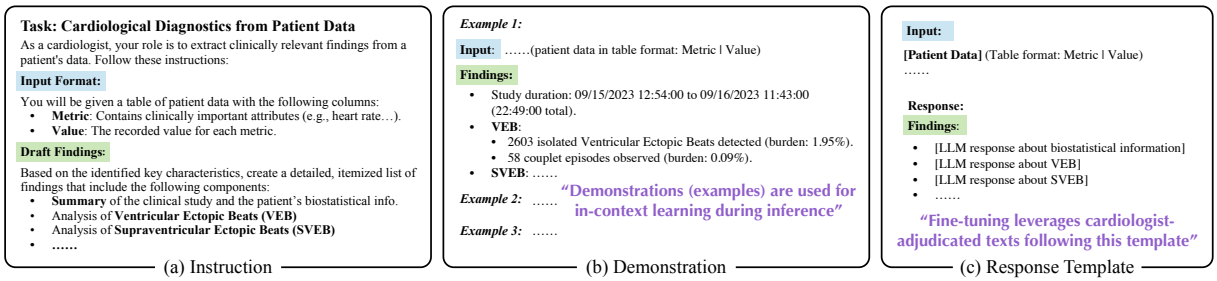

Figure 4: The prompts used for $\theta_{\texttt{M2F}}$ (prompts for $\theta_{\texttt{T2F}}$ and $\theta_{\texttt{F2I}}$ are in Figure 8): (a) the instruction (or "system prompt") used for both fine-tuning and inference; (b) the demos used for ICL during inference; and (c) the LLM response structure. During fine-tuning, (c) is filled with cardiologist-adjudicated texts, whereas during inference, (c) retains the format presented above to specify the response format.

4.2 Instruction Fine-Tuning

Instruction fine-tuning embeds cardiologist-level expertise from Z-BENCH into $\theta_{\texttt{M2F}}$, $\theta_{\texttt{T2F}}$, and $\theta_{\texttt{F2I}}$. We use Llama-3.2-3B as the base model for $\theta_{\texttt{M2F}}$ and $\theta_{\texttt{F2I}}$, and LLaVA-7B for $\theta_{\texttt{T2F}}$. Each model is fine-tuned individually on relevant subsets of Z-BENCH, tailored to its specific task. For example, as shown in Figure 4, $\theta_{\texttt{M2F}}$ is fine-tuned using system prompts as exemplified in (a) and cardiologist-adjudicated texts in the format of (c), aligning with its metric-to-findings task.

Let $\theta_{\texttt{Agent}}$ denote the trainable parameters of any LLM agent, with X and Y representing the instructional input and expected response from Z-BENCH, $\mathcal{D}$. The fine-tuning process is formulated as:

$$\theta^*_{\texttt{Agent}} = \arg\min_{\theta_{\texttt{Agent}}} \mathbb{E}_{(X,Y)\in\mathcal{D}} \mathcal{L}(\theta_{\texttt{Agent}}(X), Y) \quad (2)$$

The goal is to minimize the average of the summed loss $\mathbb{E}(\mathcal{L}(\cdot,\cdot))$ given each pair of (X, Y) within $\mathcal{D}$. Specifically, when $\theta_{\texttt{Agent}}$ is $\theta_{\texttt{M2F}}$, we have $X = (\mathcal{M}, \mathcal{B})$ and $Y = \mathcal{F}$. For $\theta_{\texttt{T2F}}$, $X = (\mathcal{T}, \mathcal{B})$ and $Y = \mathcal{F}$. Lastly, for $\theta_{\texttt{F2I}}$, $X = (\mathcal{F}, \mathcal{G})$ and $Y = \mathcal{I}$.

4.3 Inference

As outlined Figure 3, ZODIAC's inference involves a multi-agent collaboration. First, $\theta_{\texttt{M2F}}$ processes patient metrics $\mathcal{M}$ and $\theta_{\texttt{T2F}}$ handles ECG tracings $\mathcal{T}$, together generating findings $\mathcal{F}$. These findings are then interpreted by $\theta_{\texttt{F2I}}$ as the diagnostic interpretation ($\mathcal{I}$). Each agent leverages in-context learning to enhance diagnostic accuracy, with fact-checking applied afterward for self-correction.

In-Context Learning (ICL). For each fine-tuned LLM agent, we implement ICL using a set of demonstrations (or "demos", as shown in Figure 4-(b)) containing cardiologist-adjudicated $\mathcal{F}$ and $\mathcal{I}$. The content of each demo is tailored to the specific agent. To ensure relevance to the target patient's case, we categorize patient data into subgroups as shown in Figure 2. We then select three demos from the training dataset that match the patient's group (age, gender, race, and arrhythmia) and integrate them into the prompt for inference.

Algorithm 1: Fact Checking

Input: $\mathcal{F}$ – originally generated findings;
$\mathcal{I}$ – originally generated interpretation;
$\mathcal{G}$ – cardiological guidelines;
Output: Updated $\mathcal{F}$ and $\mathcal{I}$;

foreach *interpretation item* $i \in \mathcal{I}$ **do**
  Extract correlated findings $f_i \in \mathcal{F}$;
  while (f_i, i) *misaligned with* $\mathcal{G}$ **do**
    $g \leftarrow$ violated guidelines from $\mathcal{G}$;
    // $\theta \in \{\theta_{\texttt{M2F}}, \theta_{\texttt{T2F}}\}$
    $\theta \leftarrow$ agent that generates f_i;
    $p_f, p_i \leftarrow$ prompts about (f_i, i)'s misalignment with g ;
    $f_i^* \leftarrow \theta(p_f)$; $i^* \leftarrow \theta_{\texttt{F2I}}(p_i)$;
    break after 3 iterations;
  end
  Update $\mathcal{F}, \mathcal{I}$ by f_i^*, i^* ;
end
return $\mathcal{F}, \mathcal{I}$;

Fact-Checking. After $\theta_{\texttt{F2I}}$ generates the interpretation $\mathcal{I}$, ZODIAC applies cardiological guidelines $\mathcal{G}$ to verify whether the findings $\mathcal{F}$ correctly support $\mathcal{I}$. As detailed in Algorithm 1, since $\mathcal{F}$, $\mathcal{I}$, and $\mathcal{G}$ are structured as itemized lists, each item $i \in \mathcal{I}$ is evaluated against its corresponding finding $f_i \in \mathcal{F}$ to check for violations of $\mathcal{G}$.

If discrepancies $g \in \mathcal{G}$ are detected, ZODIAC automatically prompts its agents with (f_i, i, g) for regeneration, continuing until an aligned (f_i^*, i^*) set is produced or the maximum iteration (3 by

Model	Z-BENCH: Finding ($\mathcal{F}$)					Z-BENCH: Interpretation ($\mathcal{I}$)				
	P	R	F1	IoU	BS	P	R	F1	IoU	BS
GPT-4o	0.8032	0.8429	0.8226	0.6986	0.8112	0.7932	0.8561	0.8235	0.6999	<u>0.8563</u>
Gemini-Pro	0.7892	0.8604	0.8233	0.6996	<u>0.8441</u>	0.7625	0.8151	0.7879	0.6501	0.7925
Llama-3.2-90B	0.7415	0.8279	0.7823	0.6425	0.7785	0.6385	0.7684	0.6975	0.5355	0.8076
Mixtral-8x22B (+V)	<u>0.8203</u>	<u>0.8765</u>	<u>0.8475</u>	<u>0.7353</u>	0.8228	<u>0.8125</u>	<u>0.8864</u>	<u>0.8478</u>	<u>0.7359</u>	0.8247
DeepSeek-Janus-7B	0.6548	0.6221	0.6380	0.4685	0.7604	0.6109	0.7242	0.6627	0.4956	0.7785
LLaVA-13B	0.7356	0.7934	0.7634	0.6173	0.8025	0.7072	0.8169	0.7581	0.6104	0.8290
BioGPT-Large (+V)	0.2214	0.1206	0.1561	0.0847	0.5562	0.1275	0.0894	0.1051	0.0555	0.5168
Meditron-70B (+V)	0.6336	0.7127	0.6708	0.5047	0.7582	0.5782	0.6450	0.6098	0.4386	0.7864
Med42-70B (+V)	0.5209	0.4317	0.4721	0.3090	0.6337	0.4078	0.3710	0.3885	0.2411	0.5872
ZODIAC	**0.9902**	**0.9710**	**0.9805**	**0.9618**	**0.8831**	**0.9442**	**0.9683**	**0.9561**	**0.9159**	**0.9012**

Table 1: Results on Z-BENCH. P: precision, R: Recall, BS: BERT Score. For text-based LLMs, we integrate LLaVA-13B to process ECG, labeling as (+V). **Boldface** and <u>Underline</u> highlight the best and second-best results.

default) is reached. Due to space constraints, examples of $\mathcal{G}$ and further fact-checking details are provided in Appendix D.

5 Experiments

5.1 Experimental Setting

Our experiments aim to address the following research questions:

RQ$_1$: How effective are LLMs on Z-BENCH?
RQ$_2$: What are the influential factors to ZODIAC?
RQ$_3$: Is ZODIAC generalizable to other datasets?
RQ$_4$: Is ZODIAC helpful for cardiologists?

Baseline. We evaluate three groups of baselines: *(1) Industry-Leading LLMs:* GPT-4o, Gemini, Llama-3.2, Mixtral, DeepSeek, and LLaVA. *(2) Clinical-Specialist LLMs:* BioGPT-Large (Luo et al., 2022), Meditron (Chen et al., 2023b), and Med42 (Christophe et al., 2024), all derived from Llama. *(3) Ablations:* This includes a single-agent version of ZODIAC, a dual-agent variant, and an ablated version of ZODIAC with key components removed, as detailed in Section 5.3. For text-based baselines (e.g., Mixtral and BioGPT), we integrate the vision-capable LLM, LLaVA-13B, to assist in analyzing ECG tracings.

Dataset. We utilize two groups of datasets for evaluation: *(1)* Z-BENCH*:* By default, we assess 1,500 patient records from Z-BENCH while using the remaining data to fine-tune ZODIAC. Evaluating on Z-BENCH can test the varying proficiency of LLMs in in-hospital cardiological diagnostics, as it reflects real-world, representative, and cardiologist-incorporated diagnostics (3.2). *(2) Generalization Assessment:* To demonstrate ZODIAC's capability beyond Z-BENCH, we also evaluate it on PTB-XL (Wagner et al., 2020), MIMIC-IV-ECG (Gow et al., 2023), and CODE (Ribeiro et al., 2020) (in 5.4).

Metrics. We separately evaluate findings ($\mathcal{F}$) and interpretation ($\mathcal{I}$) using the following metrics, leveraging their itemized nature: *(i) Precision*– The ratio of items in the generated $\mathcal{F}$ and $\mathcal{I}$ that match those in Z-BENCH. *(ii) Recall*– The ratio of items in Z-BENCH that are successfully generated. *(iii) F1 Score*– The harmonic mean of precision and recall. *(iv) IoU (Intersection over Union)*– Also known as Jaccard similarity, measuring the overlap between generated outputs and items in Z-BENCH. *(v) BERT Score*– A text similarity measure using BERT to compare the generated $\mathcal{F}$ and $\mathcal{I}$ against their corresponding items in Z-BENCH.

To assess precision and recall, we use GPT-01 for binary classification, labeling each item in $\mathcal{F}$ and $\mathcal{I}$ as either "Matched" or "Not Matched".

5.2 Diagnostic Effectiveness (RQ$_1$)

Table 1 presents evaluation results on Z-BENCH. Notably, there is a significant performance gap between ZODIAC and other LLMs. For instance, ZODIAC surpasses GPT-4o by approximately 19% in precision for diagnosing findings and 15% in generating interpretations. Importantly, the 1,500 test reports from Z-BENCH were never seen by ZODIAC during development, and ZODIAC operates with only 13B parameters (3B+7B+3B). This highlights the necessity of specialized customization in redeveloping LLMs—particularly in life-critical fields like healthcare—rather than relying solely on general-purpose LLMs, despite their broad success across multiple domains.

Interestingly, medical-specialist LLMs performed worse than generic LLMs. While the small scale of BioGPT-Large (1.5B parameters) understandably limits its diagnostic capabilities, a more critical issue is that the data used for fine-

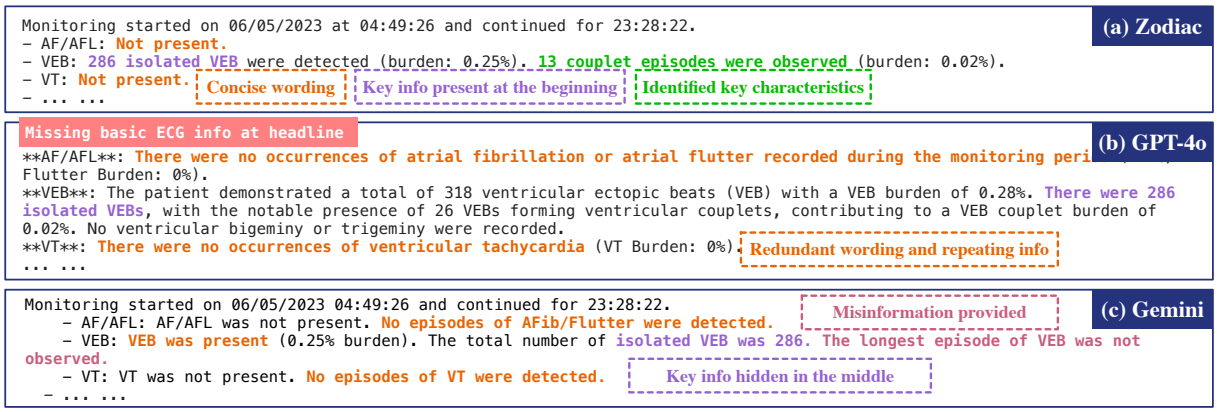

Figure 5: Examples of interpretation generated by ZODIAC, GPT-4o, and Gemini-Pro.

Model	Z-BENCH: Finding ($\mathcal{F}$)					Z-BENCH: Interpretation ($\mathcal{I}$)				
	P	R	F1	IoU	BS	P	R	F1	IoU	BS
Single Agent (T2F only)	0.8762	0.8671	0.8716	0.7725	0.8459	0.8292	0.8782	0.8530	0.7437	0.8341
Dual Agent (M2F←F2I)	0.8825	0.8741	0.8783	0.7830	0.8686	0.8758	0.9153	0.8951	0.8101	0.8849
Dual Agent (T2F←M2F)	0.9362	0.9174	0.9267	0.8634	0.8589	0.9031	0.9527	0.9272	0.8643	0.8652
Dual Agent (T2F←F2I)	0.9675	0.9481	0.9577	0.9188	0.8652	0.9122	0.9528	0.9321	0.8728	0.8467
w/o Fine-Tuning	0.7462	0.6859	0.7148	0.5562	0.7658	0.7069	0.7833	0.7431	0.5913	0.7496
w/o ICL	0.9636	0.9627	0.9631	0.9289	0.8816	0.9204	0.9450	0.9325	0.8736	0.8864
w/o Fact-Checking	0.9374	0.9317	0.9345	0.8771	0.8637	0.9078	0.9192	0.9135	0.8407	0.8872

Table 2: Ablation study results, where "Dual Agent" reuses one agent to perform another's function. For example, "M2F←F2I" indicates the removal of $\theta_{\texttt{F2I}}$, with $\theta_{\texttt{M2F}}$ performing its tasks.

tuning models like Meditron-70B appear to be misaligned with real-world clinical practice. Even when aided by in-context learning demos, these specialist LLMs struggle to meet the specific requirements and security demands of clinical tasks.

Case Study. Figure 5 compares the interpretations generated by ZODIAC, GPT-4o, and Gemini. ZODIAC produces concise, well-structured statements that allow cardiologists to efficiently extract key information. In contrast, other LLMs exhibit several limitations: GPT-4o omits critical details (e.g., missing the diagnostic headline), Gemini introduces inaccuracies (e.g., erroneous numerical summaries), and both models tend to generate redundant wording, making their outputs harder for cardiologists to rely on with confidence.

5.3 Ablation Study ($\mathbf{RQ_2}$)

We conduct an ablation study on two levels: (1) reducing the number of agents and (2) removing key components from ZODIAC.

Single and Dual-Agent. Table 2 examines the impact of agentic design variations on ZODIAC 's performance, revealing clear limitations in diagnostic accuracy. For example, removing $\theta_{\texttt{F2I}}$ results in an 11% decrease in F1 score for summarizing findings. This suggests that a single LLM struggles to effectively handle multiple diagnostic stages (e.g., both M2F and F2I), as each stage requires a distinct focus—M2F emphasizes information retrieval and summarization, while F2I integrates cardiological expertise. These results highlight the necessity of a collaborative multi-agent approach to distribute tasks efficiently and enhance diagnostic precision.

Ablative Component. Table 2 also presents results from removing fine-tuning, ICL, and fact-checking. Notably, fine-tuning has the most significant impact on diagnostic performance, demonstrating its critical role in embedding domain expertise directly into the LLMs' parameters. Additionally, ICL and fact-checking further refine the model's proficiency, emphasizing the importance of integrating these techniques to enhance diagnostic accuracy and reliability.

5.4 Generalization ($\mathbf{RQ_3}$)

While ZODIAC is developed on Z-BENCH, we evaluate its generalization capability using additional ECG datasets: PTB-XL (Wagner et al., 2020) for diagnosing 71 clinical statements, and MIMIC-IV-ECG (Gow et al., 2023) and CODE (Ribeiro et al., 2020) for identifying arrhythmia types. For each

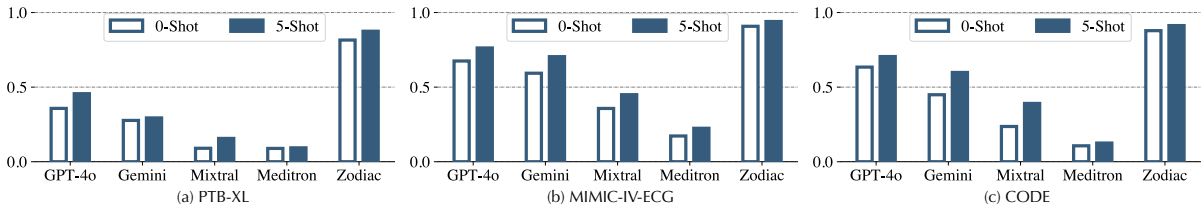

Figure 6: Diagnostic accuracy in the generalization evaluation using three additional ECG datasets.

Val Metric	Rubric (Items in $\mathcal{F}$ and $\mathcal{I}$ should ...)
Accuracy (ACC)	have correct statements, aligning with $\mathcal{G}$.
Completeness (CPL)	contain complete items from patient data.
Accessibility (ACS)	be brief and easy-to-follow the key info.
Professionalism (PRO)	use professional wording or abbreviations.

Table 3: Human validation metrics and descriptions of ideal cases. Each metric is rated on a scale from 1 to 5, where: 1 — Not at all; 2 — Below acceptable; 3 — Acceptable; 4 — Above acceptable; 5 — Excellent.

Model	Human Val Metric			
	ACC	CPL	ACS	PRO
GPT-4o	3.7 (±1.1)	4.1 (±1.1)	4.0 (±1.0)	3.2 (±1.4)
Gemini-Pro	3.7 (±1.1)	4.1 (±1.1)	3.9 (±1.0)	2.8 (±1.7)
Mixtral +(V)	3.6 (±1.0)	4.2 (±1.0)	4.1 (±0.7)	3.1 (±1.2)
Meditron +(V)	3.3 (±1.1)	3.3 (±1.2)	3.6 (±1.1)	2.3 (±1.2)
ZODIAC	**4.7 (±0.2)**	**4.8 (0.1)**	**4.7 (±0.4)**	**4.6 (±0.3)**

Table 4: LLM diagnostic performance across human validation metrics. Each cell presents "mean (±std)" among ratings from all cardiologists across all test data (same as Table 1). **Boldface** highlight the best results.

dataset, we utilize the provided ECG signals as tracings ($\mathcal{T}$) and incorporate available patient information as metrics ($\mathcal{M}$). Notably, the amount of $\mathcal{M}$ in these datasets is significantly lower than in Z-BENCH, which includes over 270 metrics.

Figure 6 presents diagnostic accuracy across these datasets, where we compare ZODIAC against the best medical-specialist agent (Meditron-70B) and the top three industry-leading LLMs (GPT-4o, Gemini-Pro, and Mixtral) from Table 1. We further evaluate two inference settings: (1) 0-shot (using only the patient's $\mathcal{T}$ and $\mathcal{M}$ as instructions) and (2) 5-shot (incorporating five demonstrations from patients). Notably, even in the 0-shot setting, ZODIAC achieves strong accuracy (>80%), surpassing other baselines. Incorporating 5-shot demonstrations further enhances its performance, highlighting ZODIAC 's adaptability for in-hospital applications. This demonstrates that ZODIAC can be easily customized with task-specific demonstrations to meet organizational needs.

5.5 Human Validation (RQ_4)

Involving human experts in validation is essential for enhancing the credibility and acceptance of advanced techniques (Tierney et al., 2024; Sallam et al., 2024). To this end, we engaged five cardiologists to evaluate ZODIAC using four metrics, as detailed in Table 3. To streamline the assessment process, we developed a structured questionnaire that begins with patient data, followed by generated findings and interpretations, and concludes with rating options on a 1–5 scale. **Notably, we anonymized the model names to prevent cardiologists from assigning biased scores based on their familiarity with or perceived reputation of specific models.**

As baselines, we include the same choice of best subset as in Section 5.4. The results show that ZODIAC not only achieves the highest performance across all human validation but also delivers more stable diagnostics, as evidenced by its lower standard deviation (e.g., ±0.1 CPL). These findings underscore the importance of incorporating refined technical strategies to improve consensus among cardiologists and enhance real-world applicability.

6 Conclusion

We introduce Z-BENCH, a cardiologist-adjudicated dataset comprising real-world, representative patient reports. Additionally, we develop ZODIAC, an LLM-powered multi-agent framework designed to achieve cardiologist-level diagnostics. Together, Z-BENCH and ZODIAC bridge the gap between clinicians and LLMs in cardiology. Through clinical validation, we demonstrate that ZODIAC outperforms other LLMs while exhibiting strong generalizability and practical utility. In conclusion, ZODIAC marks a significant step toward developing clinically viable LLM-based diagnostic tools.

Limitations

Data Scale. Clinical benchmarks are typically extensive, as exemplified by PTB-XL (Wagner et al., 2020), which contains approximately 19K patient records. While we have demonstrated that Z-BENCH is sufficient for developing cardiologist-level agents and conducting robust clinical evaluations, expanding the dataset remains a key long-term objective. As we establish collaborations with more institutions, we aim to continuously enrich Z-BENCH with diverse, high-quality clinical data, further enhancing its representativeness and utility in real-world cardiological diagnostics.

Development with Trustworthiness. As emphasized by FDA's guiding principles (FDA, 2024), securing the development and deployment of LLMs is as important as achieving functional effectiveness. While our current evaluation addresses professionalism, the next phase will prioritize further development of security measures to enhance trust. This will involve investigating third-party adversarial influences in data, identifying inherent weaknesses in LLMs that could lead to vulnerabilities (e.g., backdoors), proposing defensive strategies to safeguard ZODIAC in life-critical diagnostic applications, and promoting transparency to foster human understanding and effective collaboration in human-machine intelligence.

References

U Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, and Ru San Tan. 2019. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ecg signals. *Applied Intelligence*, 49:16–27.

Hiba Ahsan, Denis Jered McInerney, Jisoo Kim, Christopher Potter, Geoffrey Young, Silvio Amir, and Byron C Wallace. 2023. Retrieving evidence from ehrs with llms: Possibilities and challenges. *arXiv preprint arXiv:2309.04550*.

Onur Asan, Alparslan Emrah Bayrak, Avishek Choudhury, et al. 2020. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, 22(6):e15154.

American Heart Association, American College of Cardiology, et al. 2023. 2023 aha/acc/accp/aspc/nla/pcna guideline for the management of patients with chronic coronary disease. *Circulation*, 148(12):e213–e249.

Machteld J Boonstra, Davy Weissenbacher, Jason H Moore, Graciela Gonzalez-Hernandez, and Folkert W Asselbergs. 2024. Artificial intelligence: revolutionizing cardiology with large language models. *European Heart Journal*, 45(5):332–345.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Haoran Chen, Shengxiao Zhang, Lizhong Zhang, Jie Geng, Jinqi Lu, Chuandong Hou, Peifeng He, and Xuechun Lu. 2024a. Multi role chatgpt framework for transforming medical data analysis. *Scientific Reports*, 14(1):13930.

Kezhen Chen, Rahul Thapa, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2024b. Dragonfly: Multi-resolution zoom supercharges large visual-language model. *arXiv preprint arXiv:2406.00977*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023a. Meditron-70b: Scaling medical pretraining for large language models. In *ArXiv e-prints*.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco

Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Sammy Chouffani El Fassi, Adonis Abdullah, Ying Fang, Sarabesh Natarajan, Awab Bin Masroor, Naya Kayali, Simran Prakash, and Gail E Henderson. 2024. Not all ai health tools with regulatory authorization are clinically validated. *Nature Medicine*, pages 1–3.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Keith D Cicerone, Cynthia Dahlberg, Kathleen Kalmar, Donna M Langenbahn, James F Malec, Thomas F Bergquist, Thomas Felicetti, Joseph T Giacino, J Preston Harley, Douglas E Harrington, et al. 2000. Evidence-based cognitive rehabilitation: recommendations for clinical practice. *Archives of physical medicine and rehabilitation*, 81(12):1596–1615.

ContactDoctor. 2024. Bio-medical-multimodal-llama-3-8b-v1: A high-performance biomedical multimodal llm. https://huggingface.co/ContactDoctor/Bio-Medical-MultiModal-Llama-3-8B-V1.

Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94–98.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Joel Eapen and VS Adhithyan. 2023. Personalization and customization of llm responses. *International Journal of Research Publication and Reviews*, 4(12):2617–2627.

FDA. 2020. What is digital health?

FDA. 2024. Transparency for machine learning-enabled medical devices: Guiding principles.

Jana Fehr, Brian Citro, Rohit Malpani, Christoph Lippert, and Vince I Madai. 2024. A trustworthy ai reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6:1267290.

U.S. Food, Drug Administration (FDA), Health Canada, United Kingdom's Medicines, and Healthcare products Regulatory Agency (MHRA). 2021. Good machine learning practice for medical device development: guiding principles. *FDA*.

Apache Software Foundation. 2004. Apache license, version 2.0.

Małgorzata Furmankiewicz, Anna Sołtysik-Piorunkiewicz, and Piotr Ziuziański. 2014. Artificial intelligence and multi-agent software for e-health knowledge management system. *Informatyka Ekonomiczna/Uniwersytet Ekonomiczny we Wrocławiu*, (2 (32)):51–63.

Dhir Gala and Amgad N Makaryus. 2023. The utility of language models in cardiology: a narrative review of the benefits and concerns of chatgpt-4. *International Journal of Environmental Research and Public Health*, 20(15):6438.

David C Goff Jr, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B D'agostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J O'donnell, et al. 2014. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Circulation*, 129(25_suppl_2):S49–S73.

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

Jiyeon Han, Jimin Park, Jinyoung Huh, Uran Oh, Jaeyoung Do, and Daehee Kim. 2024. Ascleai: A llm-based clinical note management system for enhancing clinician productivity. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

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):65–69.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 171:108189.

Hanen Jemal, Zied Kechaou, and Mounir Ben Ayed. 2014. Swarm intelligence and multi agent system in healthcare. In *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 423–427. IEEE.

Charalampia Xaroula Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. 2022. Before and beyond trust: reliance in medical ai. *Journal of medical ethics*, 48(11):852–856.

Bangul Khan, Hajira Fatima, Ayatullah Qureshi, Sanjay Kumar, Abdul Hanan, Jawad Hussain, and Saad Abdullah. 2023. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials & Devices*, 1(2):731–738.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171.

Harlan M. Krumholz, Michael S. Lauer, et al. 2020. *Clinical Cardiology: Current Practice Guidelines*. Oxford University Press.

Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yonfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. 2024. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Inc. Meta Platforms. 2024. Meta llama 3 community license.

Abdulqadir J Nashwan and Ahmad A AbuJaber. 2023. Harnessing the power of large language models (llms) for electronic health records (ehrs) optimization. *Cureus*, 15(7).

Massachusetts Institute of Technology. 1988. The mit license (mit).

Jonas Bjerring Olesen, Christian Torp-Pedersen, Morten Lock Hansen, and Gregory YH Lip. 2012. The value of the cha2ds2-vasc score for refining stroke risk stratification in patients with atrial fibrillation with a chads2 score 0–1: a nationwide cohort study. *Thrombosis and haemostasis*, 107(06):1172–1179.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of International Conference on Learning Representations (ICLR)*.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. 2020. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760.

Malik Sallam, Muna Barakat, Mohammed Sallam, et al. 2024. A preliminary checklist (metrics) to standardize the design and reporting of studies on generative artificial intelligence–based models in health care education and practice: Development study involving a literature review. *Interactive Journal of Medical Research*, 13(1):e54704.

Elhadi Shakshuki and Malcolm Reid. 2015. Multi-agent system applications in healthcare: current technology and future roadmap. *Procedia Computer Science*, 52:252–261.

Maneesh Sud, Atul Sivaswamy, Anna Chu, Peter C Austin, Todd J Anderson, David MJ Naimark, Michael E Farkouh, Douglas S Lee, Idan Roifman, George Thanassoulis, et al. 2022. Population-based recalibration of the framingham risk score and pooled cohort equations. *Journal of the American College of Cardiology*, 80(14):1330–1342.

Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3):CAT–23.

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):1–15.

Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "brilliant ai doctor" in rural clinics: Challenges in ai-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–18.

Jinge Wang, Qing Ye, Li Liu, Nancy Lan Guo, and Gangqing Hu. 2024a. Scientific figures interpreted by chatgpt: strengths in plot recognition and limits in color perception. *NPJ Precision Oncology*, 8(1):84.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024b. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024c. Llama3-8b-chinese-chat (revision 6622a23).

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Alaa T Youssef, David Fronk, John Nicholas Grimes, Lina Cheuy, and David B Larson. 2024. Beyond the black box: Avenues to transparency in regulating radiological ai/ml-enabled samd via the fda 510 (k) pathway. *medRxiv*, pages 2024–07.

A A Real-world Cardiological Report

Figure 7 presents a real-world report on patient data and diagnostics (including findings and interpretation), with identifying information (such as patient name, date of birth, physician name, and company name) anonymized. The report layout is identical to that shown in Figure 3.

B Additional Prompts

Corresponding to Figure 4-(a)(b)(c), we provide the prompts used for agents $\theta_{\texttt{T2F}}$ and $\theta_{\texttt{F2I}}$ in Figure 8.

C Details about Arrhythmia Classes

In this work, we categorize arrhythmias into three subgroups:

Class I — Normal Arrhythmias: Also known as benign or physiological arrhythmias, these irregular heart rhythms can occur in healthy individuals and typically do not lead to serious health issues. They are generally considered harmless and may not require treatment. In our patient data, Class I arrhythmias include Sinus Bradycardia, Sinus Tachycardia, and Sinus Arrhythmia.

Class II — Clinically Significant Arrhythmias: These arrhythmias involve abnormal heart rhythms that can cause symptoms, lead to complications, or require medical intervention. They may disrupt the heart's ability to pump blood effectively, increasing the risk of serious events such as stroke, heart failure, or sudden cardiac death. In our patient data, Class II arrhythmias include Pause ($<$3s), Ventricular Premature Beat (PVC), and Atrial Fibrillation (AF).

Class III — Life-Threatening Arrhythmias: These abnormal heart rhythms can result in severe consequences, such as cardiac arrest, stroke, or sudden cardiac death, requiring immediate medical attention and often emergency intervention. In our patient data, Class III arrhythmias include Ventricular Flutter (VF), Complete Heart Block (Third-Degree AV Block), Atrial Fibrillation (AFib) with Rapid Ventricular Response, Prolonged Pause, Atrial Flutter (AFL), Ventricular Tachycardia (VT), and Supraventricular Tachycardia (SVT).

In our experiments, we use these arrhythmia classes (I, II, III) for subgroup analysis rather than specific arrhythmias to avoid the limitations of small patient sample sizes for individual conditions. Subgroup analysis based on arrhythmia classes provides a comprehensive view of the LLMs' diagnos-

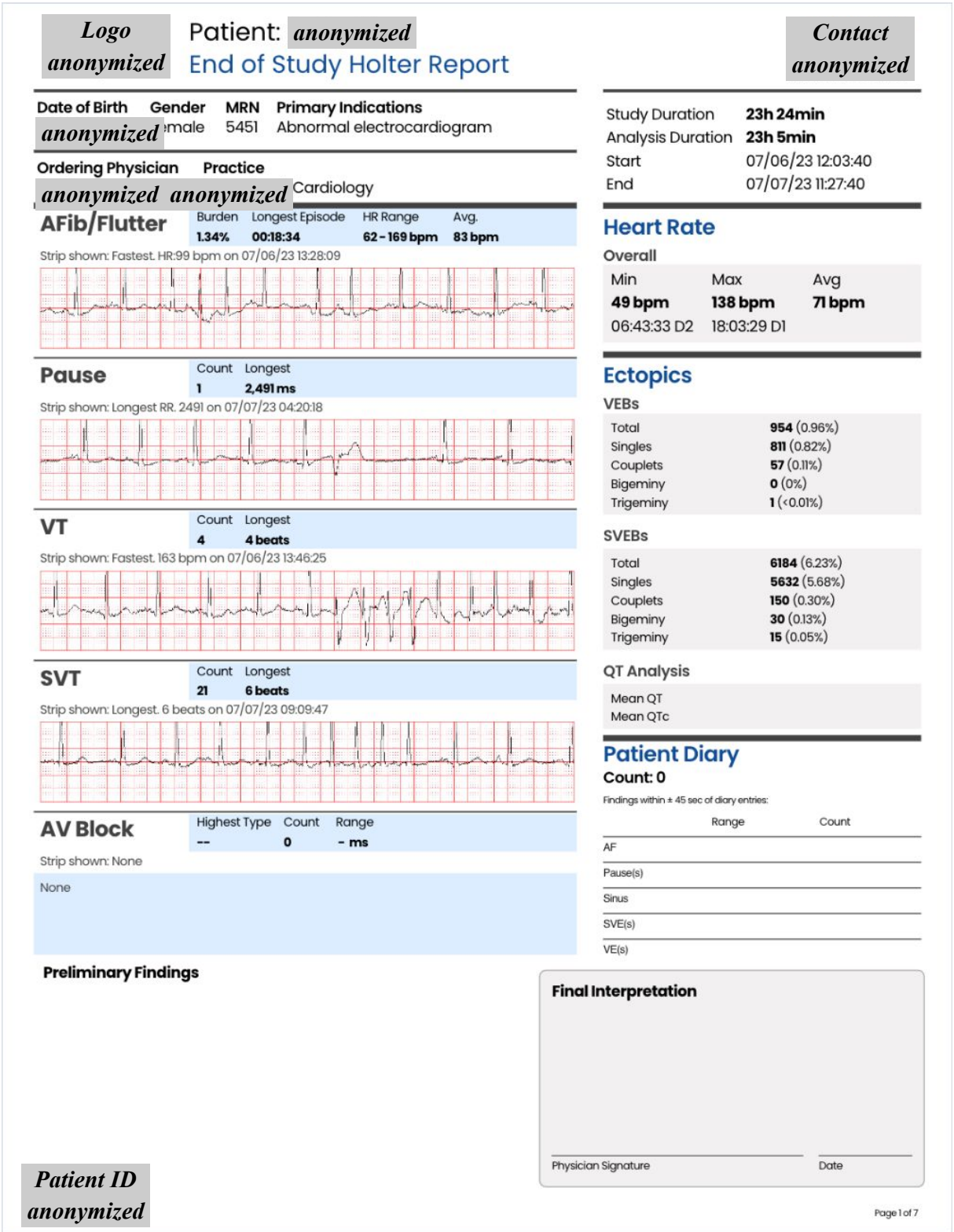

Figure 7: A real-world cardiological report, with identify-related information anonymized.

tic capabilities across different levels of urgency, offering valuable insights for data collection and performance improvement toward more balanced diagnostics.

D Fact Checking Using Clinical Guideline

Clinical guidelines are systematically developed statements designed to assist healthcare providers and patients in making decisions about appropriate health care for specific clinical circumstances.

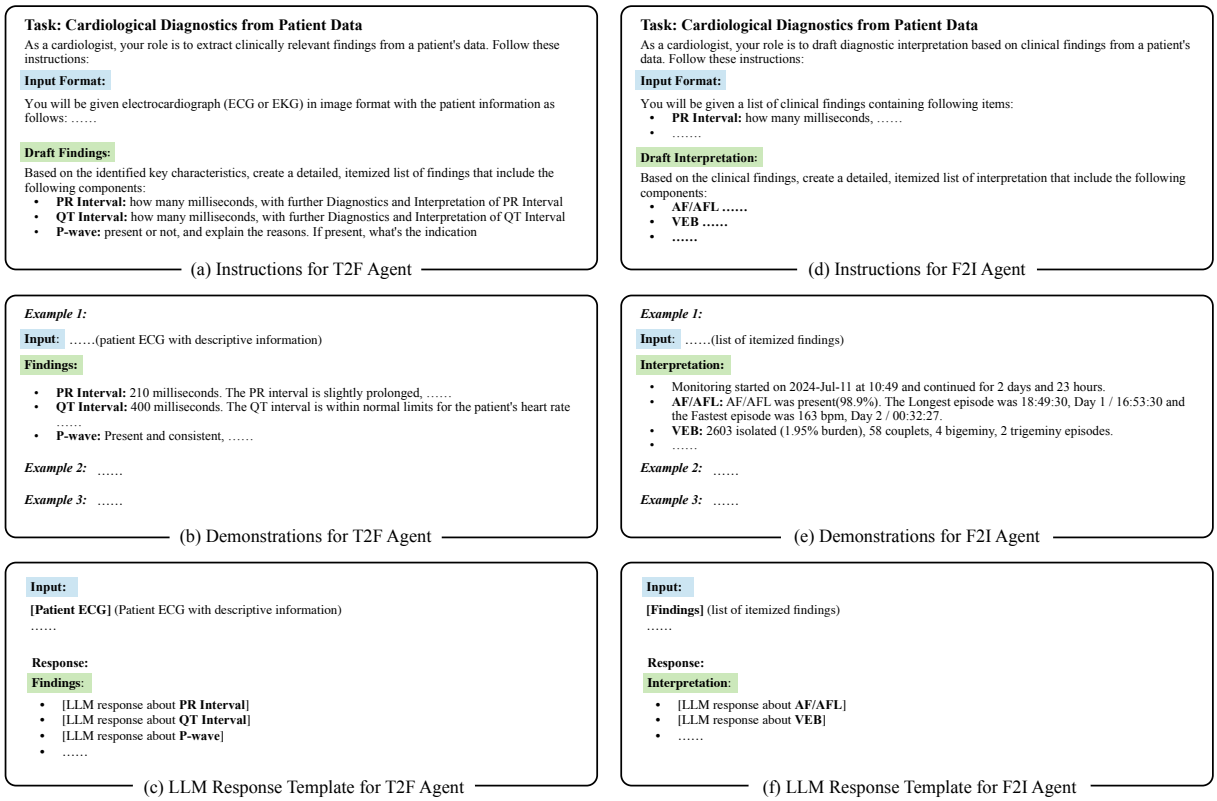

Figure 8: Prompts used for $\theta_{\texttt{T2F}}$ and $\theta_{\texttt{F2I}}$: (a)(d) – instructions or "system prompt"; (b)(e) –demonstrations used during in-context learning; (c)(f) – LLM response template.

These guidelines are based on the best available evidence and aim to standardize care, improve the quality of treatment, and ensure patient safety. For example, a section of clinical guidelines about PR Interval is provided in Figure 9.

Fact-Checking using Guidelines. We perform fact-checking by enumerating every itemized finding and corresponding interpretation to identify any misalignment with established guidelines. For example, if the PR interval exceeds 200 milliseconds, the interpretation should include a diagnosis of "a prolonged PR interval, which may indicate a first-degree AV block or the potential for a more advanced block". Failure to include such a diagnosis would signal an inaccurate assessment by ZODIAC. In response, we prompt the relevant LLM agents ($\theta_{\texttt{T2F}}$ and $\theta_{\texttt{F2I}}$ in this case) to re-examine the patient data, verify the accuracy of the findings, and update the interpretation accordingly.

E Experimental Configurations

We conducted our experiments using a set of NVIDIA RTX A6000 ADA GPUs, each equipped with 48GB of memory and running CUDA version 12.3. Table 5 provides a detailed overview of the default hyper-parameters and experimental settings.

Moreover, our experiments use a fixed set of hyperparameters as commonly used among other works (Qi et al., 2023; Yang et al., 2023) without hyperparameter search.

Models and Fine-Tuning (Customization)	
Training Data (& Statistics)	Z-BENCH (3,907 samples)
Test Data (& Statistics)	Z-BENCH (1,500 samples) PTB-XL (2,000 samples) MIMIC-IV-ECG (2,000 samples) CODE (2,000 samples)
LLMs	Llama-3.2-3B (x2) LLaVA-7B
Max sequence length	2,048 (Train); 4,096 (Test)
Batch size	16
Training epochs	50
Learning rate	1e-5
Optimizer	AdamW
Fine-Tuning Method	LoRA
GPU Hours	1.62
Inference	Temperature: 1.0 top-p 0.95

Table 5: Implementation and evaluation details of ZODIAC.

PR Interval:

1. Definition of PR Interval

The PR interval measures the period from the onset of atrial depolarization (beginning of the P wave) to the onset of ventricular depolarization (beginning of the QRS complex). It reflects the time taken for the electrical impulse to travel from the sinus node through the atria, AV node, His bundle, bundle branches, and Purkinje fibers to reach the ventricular myocardium.

2. Range

- **Normal: 120-200 milliseconds**
- **Prolonged: >200 milliseconds, indicating first-degree AV block**
- **Shortened: <120 milliseconds, may suggest pre-excitation syndromes like Wolff-Parkinson-White syndrome**

3. Clinical Relevance

- **Normal PR Interval**
  - **Finding: PR interval within 120-200 milliseconds.**
  - **Interpretation: Indicates normal atrioventricular (AV) conduction. The electrical signal travels from the atria to the ventricles through the AV node and His-Purkinje system within the expected time frame, suggesting healthy cardiac electrical function.**

- **Prolonged PR Interval**
  - **Finding: PR interval longer than 200 milliseconds.**
  - **Interpretation:**
    - **First-Degree AV Block: The prolongation is uniform across all heartbeats. This is often benign but can be associated with increased vagal tone, intrinsic AV nodal disease, or effects of certain medications (like beta-blockers, calcium channel blockers, or digoxin).**
    - **Higher degree AV block predisposition: Indicates potential for progression to higher degree AV block, especially in the setting of structural heart disease or acute myocardial infarction.**

- **Short PR Interval**
  - **Finding: PR interval less than 120 milliseconds.**
  - **Interpretation:**
    - **Pre-excitation Syndromes: Such as Wolff-Parkinson-White (WPW) syndrome where there is an accessory pathway (like the bundle of Kent) allowing premature ventricular activation.**
    - **Junctional Rhythms: If associated with an abnormal P wave morphology or positioning, may indicate that the impulse originates near or within the AV node rather than the atria.**

- **Variable PR Interval**
  - **Finding: Fluctuating PR intervals across different heartbeats.**
  - **Interpretation:**
    - **Second-Degree AV Block Type I (Wenckebach): Progressive lengthening of the PR interval until a P wave is not followed by a QRS complex.**
    - **Atrial Fibrillation with Variable Conduction: If associated with an irregularly irregular rhythm, indicates atrial fibrillation where AV nodal conduction is unpredictably variable.**

- **PR Interval with Grouped Beating**
  - **Finding: Groups of beats with a consistent PR interval followed by a longer pause.**
  - **Interpretation:**
    - **Second-Degree AV Block Type II: Typically associated with fixed PR intervals on conducted beats, interspersed with non-conducted P waves without prior change in the PR duration.**
    - **Mobitz Type II or Advanced Block: Often a precursor to complete heart block, requiring immediate assessment and potentially pacing intervention.**

- **Alternating PR Interval**
  - **Finding: Alternation in the length of the PR interval from beat to beat.**
  - **Interpretation:**
    - **Electrophysiological Variability: May be due to alternating dominance of different AV nodal pathways, a rare phenomenon or related to autonomic tone fluctuations.**
    - **Underlying Heart Disease: Consider evaluation for ischemic heart disease or infiltrative cardiac conditions that may intermittently affect AV nodal conduction.**

Figure 9: Part of clinical guidelines.

F Responsible Checklist

This section elaborates on the checklist for ARR submission:

F.1 Potential Risks

This work address the LLM-as-Agent contribution on cardiological domain. Even though the research lies on healthcare, which is life-sensitive, our contributed benchmark and agent help to advance the

LLM development in this domain. Moreover, this work doesn't introduce any harmful or sensitive message, so the contributed benchmark and model are safe to use in research and development purpose.

As to the data, we de-identified all patient data before releasing. So no identifiable message could be recognized. Thus we protect potential privacy leakage.

F.2 Use of Artifacts

This work utilizes public LLMs and datasets.

Model Licenses. The licenses for the LLMs used in this work vary depending on the model. Llama-3.2-3B follows the "Llama 3" license, which permits public use of its open-source model in accordance with Llama's user policy (Meta Platforms, 2024). LLaVA-7B is released under the Apache 2.0 license (Foundation, 2004), which allows free model download, modification, distribution, and even commercialization.

Data and Other Licenses. The datasets employed in this study are sourced from public GitHub repositories and our IRB-approved data. The use of public data adhere to the MIT license (of Technology, 1988). This license permits free use, modification, and distribution of the data and code made available in these repositories. Our contributed benchmark is with IRB: No. Pro00065572.

Artifact Use Consistent With Intended Use. Given the permissions granted by the model, data, and other licenses involved in this work, our development aligns with the intended use of these artifacts, ensuring compliance with their original licensing terms.

Offensive Content. There is no offensive content in this paper.

Personally Identifiable Information. This work does not involve any personally identifiable information. All used data is de-identified.

F.3 Descriptive Statistics

We have elaborated on the statistics of metrics in (1) Section 5.1 regarding the descriptions of metrics and (2) Section 5.5 regarding additional descriptions of evaluation metrics (mean and std).

F.4 Use of Packages

Our evaluations largely rely on Hugging Face TRANSFORMER packages and TORCH, other packages are regular Python libraries such as NUMPY and MATPLOTLIB, which can be seen at our released codes: https://anonymous.4open.science/r/Zbench-Zodiac-8A2A

F.5 Use of AI

This work primarily uses AI as an evaluator to compute whether two items are match and BERT score. This approach aligns with prior studies (Eapen and Adhithyan, 2023; Qi et al., 2023; Yang et al., 2023; Chen et al., 2021).

F.6 Instructions Given to Participants

Introduction Thank you for participating in this study. Your expertise as a cardiologist is crucial in annotating clinical reports and evaluating the performance of our AI-based model. These instructions will guide you through the annotation and evaluation process.

Objective The goal of this task is to:

1. Annotate clinical reports by writing structured **findings** and **interpretation**.
2. Evaluate the performance of an AI model by comparing its generated reports against expert annotations.

Annotation Guidelines Each clinical report will be provided with raw data, including imaging results, ECG readings, and physician notes. Your task is to:

1. **Findings**: Summarize the key observations in a structured manner.
2. **Interpretation**: Provide a concise clinical interpretation of the findings.

Formatting Rules

- Use complete sentences and precise medical terminology.
- Follow a structured format for each report.
- Avoid subjective language; remain clinically objective.

Model Performance Evaluation You will be provided with AI-generated reports alongside your annotated reports. Please evaluate the AI output using the following criteria:

Evaluation Criteria

1. **Accuracy**: How well the AI-generated findings and interpretation match expert observations and align with guidelines.

2. **Completeness**: Whether the AI-generated findings and interpretation obtain comprehensive information from patient's report.

3. **Accessibility**: Whether AI maintains brief statements across reports, with highlighted information easy to be caught by cardiologists.

4. **Professionalism**: Whether AI use professional wording and jarjons among experts.

Scoring System For each criterion, use the following rating scale:

- 1 - Not at all
- 2 - Below acceptable
- 3 - Acceptable
- 4 - Above acceptable
- 5 - Excellent

Submission Instructions

- Submit your annotations and evaluations via the provided portal.
- Use the provided spreadsheet template for scoring.
- Ensure all reports are reviewed within the given timeframe.

Confidentiality and Ethical Considerations

- Maintain patient confidentiality at all times.
- Do not share data outside the scope of this study.
- Follow HIPAA and institutional guidelines for handling clinical information.

F.7 Payment Policy

There is no payment assigned to our collaborative cardiologists.

F.8 Discussion on Consent and Ethical Considerations (IRB Protocol No. Pro00065572)

This study is conducted under the ethical oversight of the Institutional Review Board (IRB) at *"Anonymized Name"* under Protocol No. **Pro00065572**. The research involves the use of multi-modal clinical data, including cardiological diagnostic reports, for benchmarking in an LLM-as-Agent framework. Given the sensitivity of clinical data, the study strictly adheres to regulatory and ethical guidelines for data handling, consent, and privacy protection.

The dataset used in this study originates from *"Anonymized Collaborative Institutions"*. The data includes structured and unstructured cardiological diagnostic reports, imaging results, and corresponding clinical notes. If the dataset is publicly available, it has been previously de-identified and made accessible for research use, eliminating the need for direct patient consent. If institutionally collected, the data was obtained through *"Anonymized Collaborative Institutions"*, following IRB-approved data access protocols.

To ensure privacy protection, all personally identifiable information (PII) has been removed or anonymized before data processing. Anonymization steps include removing patient names, medical record numbers, and geographic identifiers, de-identifying text-based clinical reports using automated and manual review methods, and encrypting and restricting access to dataset files, allowing only approved personnel to interact with the data. The research follows HIPAA and GDPR compliance standards to prevent re-identification risks and ensure data security.

This dataset is used strictly for benchmarking multi-modal cardiological diagnostics and will not be re-shared outside the scope of the IRB-approved study. Any further data use beyond the LLM-as-Agent benchmarking framework would require additional IRB review and approval.

This study aligns with ethical principles outlined in the Belmont Report, ensuring respect for persons, beneficence, and justice. The IRB approval under Protocol Pro00065572 ensures compliance with institutional and federal regulations, including the Common Rule (45 CFR 46) for human subject research, the Health Insurance Portability and Accountability Act (HIPAA) for medical data privacy, and the General Data Protection Regulation (GDPR) for data processing involving European patients (if applicable).