LEAP-LLM: PRESERVING LEAD-TIME STRUCTURE FOR QUESTION-CONDITIONED ECG INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The electrocardiogram (ECG) records cardiac electrical activity across multiple leads and is used for detecting cardiac abnormalities. Recent ECG-language models enable large language models (LLMs) to interpret cardiac signals and answer clinically relevant questions. However, existing ECG-language models typically encode the ECG independently of the clinical question, forcing the ECG representation to retain broad and potentially irrelevant signal information. Moreover, these models often aggregate information across leads which can obscure lead-specific details needed for a particular question. To tackle these limitations, we introduce LEAP-LLM, an ECG-LLM architecture for **L**ead-aware **E**CG **A**lignment with **P**rompt conditioning. LEAP-LLM uses the clinical question to guide ECG feature extraction while preserving explicit lead and temporal structure. Task-trained instruction representations distill information from the question and guide two complementary branches. The Lead branch captures question-relevant temporal patterns within each ECG lead, while the Temporal branch summarizes relevant information across leads at each temporal position. Together, they provide the LLM with compact, question-relevant ECG information while preserving lead and temporal structure. LEAP-LLM reduces modality-token count by 89.0% and prefill FLOPs by 7.2× on average relative to comparable ECG-LLMs. Across four ECG benchmarks, LEAP-LLM achieves state-of-the-art performance, including a 3.4-percentage-point improvement in *G12EC* over the strongest prior ECG-specific method. We further show that the model adaptively emphasizes the clinically relevant anatomical lead groups for different cardiac abnormalities, with lead importance patterns consistent with established diagnostic knowledge, providing interpretable evidence that the predictions are grounded in clinically meaningful ECG structure rather than black-box representations.

1 INTRODUCTION

The electrocardiogram (ECG) is a widely used, non-invasive tool for detecting cardiac abnormalities such as arrhythmias, ischemia, and infarction (Namjouyan et al., 2025). It captures cardiac electrical activity across multiple leads over time, with clinically relevant information distributed across both lead and temporal dimensions (Berkaya et al., 2018; Hannun et al., 2019). Traditional deep learning methods have achieved strong performance on ECG analysis, but are typically developed for fixed prediction tasks such as abnormality classification (Hong et al., 2020; Zhu et al., 2021; Kiyasseh et al., 2021; Lan et al., 2022; Yang et al., 2023; Zhao et al., 2024; Li et al., 2024; Lan et al., 2024). Clinical ECG interpretation, however, is inherently more flexible meaning the same recording may be queried for rhythm, morphology, conduction abnormalities, etc. (Ruoqi et al., 2024).

Large Language Models (LLMs) provide a natural interface for extending ECG models beyond fixed predictions toward question-conditioned interpretation. Early ECG-LLM models explored ECG-text alignment for diagnosis and report generation (Cai et al., 2023; Zhao et al., 2025). More recent models such as PULSE (Ruoqi et al., 2024) encodes rendered ECG images and projects their features into visual tokens, while GEM (Lan et al., 2025) projects signal and image features before combining them with the textual query. These interfaces provide question-independent ECG tokens, leaving the LLM to identify the information relevant to each request during answer generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, the evidence needed from an ECG depends on the question being asked. Different questions may need to refer to different leads and temporal regions within the same recording. For example, myocardial infarction (MI) is assessed through abnormalities in anatomically related lead groups, whereas rhythm assessment relies on temporal relationships between successive beats (Wagner et al., 2009; Varma et al., 2021). A question-independent representation may therefore retain information irrelevant to the current question while discarding fine-grained details needed to answer it. This motivates using the question to guide evidence selection and aggregation when constructing the ECG tokens supplied to the LLM.

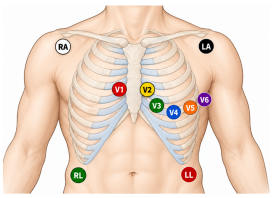


Figure 1: **Electrode placement for 12-lead ECG acquisition.** Ten electrodes generate the 12 ECG leads.

Several challenges remain in developing a question-guided ECG LLM. Firstly, global representation learning poses challenges for retaining fine-grained details and lead identities. Existing ECG encoders do not explicitly preserve lead-specific waveform information in the representations provided to the LLM. Question-guided extraction requires representations that preserve these details and explicitly identify their source leads and temporal locations, allowing relevant evidence to be selected while retaining its structural context. Secondly, question-ECG interaction poses challenges for efficient information aggregation and structural preservation. Relevant evidence may involve temporal patterns within individual leads or comparisons across leads at a given time. Simply flattening all ECG patches for attention does not specify how the aggregated output should retain these structures. The challenge is to produce compact ECG tokens that preserve both lead-specific and time-indexed information. And Lastly, variation in question length introduces both computational and semantic challenges. Question length does not necessarily reflect the amount of information relevant for ECG analysis. Clinical requests may be interspersed with redundant context or instructions. Conditioning ECG features on the entire prompt at each interaction introduces unnecessary computation and can make signal aggregation sensitive to linguistically irrelevant tokens.

To address these challenges, we introduce LEAP-LLM, an ECG-LLM framework for ***L**ead-aware **E**CG **A**lignment with **P**rompt conditioning*. We first use an encoder that maintains separate lead-specific features, with structural embeddings that make their lead identities and temporal locations explicit. Building on these representations, we introduce question-conditioned dual-attention alignment to aggregate question-relevant evidence without collapsing structural axes. The *Lead Branch* aggregates temporal evidence within each lead, whereas the *Temporal Branch* aggregates across leads at each temporal position, yielding lead-specific and time-indexed representations, respectively. Both branches are guided by task-trained instruction representations that distill variable-length questions into a fixed set of task-oriented queries. Together, the this design provides the LLM with compact, question-conditioned ECG tokens that retain complementary lead and temporal organization.

We evaluate LEAP-LLM on PTB-XL, CODE-15%, CSN, and G12EC, spanning multi-label abnormality detection and multiple-choice ECG interpretation. LEAP-LLM out performances across all benchmarks, improving over prior ECG-specific methods by 0.7 percentage points on CODE-15% AUC, 3.4 points on G12EC accuracy, and 0.4 points on CSN accuracy. At the same time, LEAP-LLM reduces the ECG modality-token count by 89.0% and achieves a 7.2$\times$ average fewer prefill FLOPs relative to comparable ECG-LLMs.

We further show that LEAP-LLM uses clinically meaningful information preserved through explicit lead and temporal structure. specifically, Lead-time attention analysis reveals selective focus on clinically relevant temporal regions and anatomical lead groups, while targeted lead-token interventions demonstrate the LLM's selective reliance on anatomically relevant lead representations for diagnosis. Our main contributions are:

- We introduce LEAP-LLM, a query-conditioned ECG-LLM framework that preserves explicit lead and temporal structure while extracting query-relevant signal information.
- We introduce complementary lead and temporal branches, guided by question-conditioned instruction tokens, to produce compact structured ECG representations for the LLM.

- We achieve state-of-the-art performance on four ECG benchmarks while reducing the ECG modality-token count by 89.0% and prefill FLOPs by 7.2× on average relative to comparable ECG-LLMs.
- We establish the clinical relevance of the learned representations through attention and intervention analyses, showing temporally selective attention around salient cardiac events and selective reliance on anatomically relevant ECG leads.

2 Related Work

2.1 Language Models for ECG Understanding

Early work explored connecting ECG representations with language for diagnosis and report generation. Yu et al. (2023) introduced retrieval-augmented approach for zero-shot ECG diagnosis, while JoLT (Cai et al., 2023) and MERL (Liu et al., 2024a) learned aligned ECG-text representations for language-guided prediction. ECG-CoCa and ECG-Chat (Zhao et al., 2025) further adapted vision-language architecture to directly process ECG time series input. Wan et al. (2025) formulate ECG report generation as instruction tuning, transforming ECG-report pairs into conversational examples.

More recent ECG-specific multimodal language models substantially expand this paradigm. PULSE (Ruoqi et al., 2024) converts ECG signals into rendered images and instruction-tunes a multimodal LLM. GEM (Lan et al., 2025) jointly incorporates raw ECG time series and 12-lead ECG images through separate encoders and cross-modal alignment, together with ECG-Grounding supervision. Other approaches operate directly on the ECG signal by converting it into discrete tokens. ECG-Byte (Han et al., 2024) adapts byte-pair encoding to tokenize ECG waveforms, while HeartLLM (Yang et al., 2026) uses lead-wise encoding and quantization to construct ECG tokens for language-model processing. Beyond representation design, ECG-R1 (Jin et al., 2026) combines signal and image inputs with reinforcement learning for grounded ECG reasoning.

These approaches primarily advance ECG-language modeling through representation alignment, multimodal integration, or reasoning supervision. In contrast, LEAP-LLM focuses on the organization of the ECG representation exposed to the LLM. Rather than representing the ECG as an image, a globally mixed feature sequence, or a single stream of discrete signal tokens, LEAP-LLM explicitly retains the physical lead and temporal axes, allowing the clinical question to selectively access lead-specific and temporally resolved evidence.

2.2 Time-Series-Language Alignment

Beyond ECG, recent work has increasingly explored interactions between time series and language models. Several approaches incorporate textual information to improve conventional time-series prediction or classification tasks (Jin et al., 2024a;b; Wang et al., 2025a; 2024b), while others use language models to generate descriptions or identify patterns from temporal signals (Liu et al., 2024b). These methods are typically designed for generic time series, where channels are treated primarily as parallel signal dimensions. This formulation does not directly transfer to ECG, where each lead has a fixed physical identity and clinically meaningful relationship to the other leads.

More recent work has moved toward treating time series as a native modality for language-based reasoning and question answering. OpenTSLM (Langer et al., 2026) evaluates natural-language reasoning across multivariate sensor domains, while ITFormer (Wang et al., 2025b) introduces question-conditioned instruction tokens that guide channel and temporal aggregation for multimodal QA. LEAP-LLM builds on this broader direction but targets the structured nature of multi-lead ECGs. Rather than reducing ECG leads into a shared temporal representation, LEAP-LLM preserves complementary lead- and time-indexed representations, allowing the LLM to access question-relevant information while retaining explicit correspondence to physical ECG leads and temporal locations.

3 Method

We formulate ECG-language alignment as question-conditioned extraction over a structured lead-time representation. We design LEAP-LLM to preserve this structure while selecting question-relevant information before it enters the LLM. Given a multi-lead ECG X and a clinical question

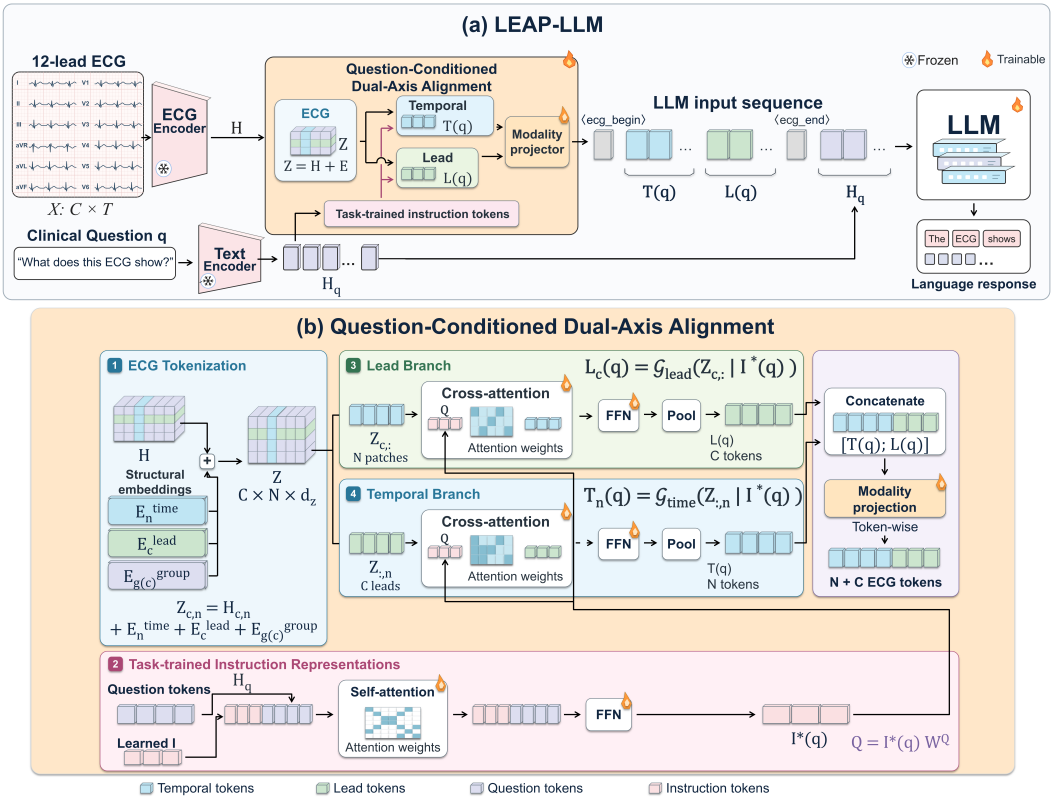


Figure 2: **Overview of LEAP-LLM. (a)** A multi-lead ECG and clinical query are encoded separately. Task-trained instruction tokens condition a dual-axis alignment module to produce compact structured tokens for the LLM. **(b)** The ECG is represented as lead-specific temporal tokens augmented with structural embeddings. Question-derived instruction tokens guide two complementary cross-attention branches: the *Lead Branch*, which summarizes temporal evidence within each lead, and the *Temporal Branch*, which summarizes cross-lead evidence at each temporal location. The resulting tokens are concatenated and projected into the LLM embedding space.

q, our encoder first extracts lead-specific temporal features and incorporates temporal, lead-identity, and lead-category embeddings. Task-trained instruction representations then distill the question into a fixed set of queries for two complementary attention branches. The Lead Branch retains lead-specific evidence, while the Temporal Branch retains time-indexed evidence. Together, they form a compact, question-conditioned ECG representation for answer generation, as illustrated in Figure 2.

3.1 ECG Encoding with Structural Embeddings

Question-guided extraction requires lead-specific temporal features to remain separately accessible. Our ECG encoder therefore maintains both axes throughout signal encoding. We represent the ECG signal as $X \in \mathbb{R}^{C \times T}$, where C is the number of leads and T is the number of samples per lead. To preserve local temporal continuity, each lead is partitioned into overlapping patches of length P and stride S, yielding $N = \lfloor (T - P)/S \rfloor + 1$ temporal positions. We construct an ECG encoder that processes each lead independently and produces a feature tensor $H \in \mathbb{R}^{C \times N \times d_z}$ based on PatchTST (Nie et al., 2022). Each feature $H_{c,n}$ represents a temporal patch from a specific lead, allowing subsequent question-conditioned attention to access temporal patterns within individual leads and compare features across leads.

Structural embeddings. To make each feature's temporal location and lead information explicitly available to the alignment module, we introduce three complementary embeddings. A *temporal-position embedding* E_n^{time} marks each patch's location within the recording, allowing the alignment module to distinguish similar waveforms at different times. A *lead-identity embedding* E_c^{lead} iden-

tifies the source lead (Na et al., 2024; Kligfield et al., 2007), since the lead's electrical view affects how its waveform is interpreted. And finally to represent shared aspects of the lead measurement systems, we further introduce a *lead-category embedding* $E^{\text{group}}_{g(c)}$.

The standard 12-lead ECG comprises three established lead families, standard limb leads $\{\text{I}, \text{II}, \text{III}\}$, augmented limb leads $\{\text{aVR}, \text{aVL}, \text{aVF}\}$, and precordial leads $\{\text{V1}, \ldots, \text{V6}\}$. These families differ in their electrode combinations and reference construction (Kligfield et al., 2007). Our category embedding follows this clinical classification, providing a shared component within each lead family, while the lead-identity embedding distinguishes individual leads. The resulting representation is

$$Z_{c,n} = H_{c,n} + E^{\text{time}}_{n} + E^{\text{lead}}_{c} + E^{\text{group}}_{g(c)}, \tag{1}$$

where $g(c)$ maps each lead to its category and $Z \in \mathbb{R}^{C\times N\times d_z}$ retains separate lead and temporal indices for subsequent aggregation.

3.2 Task-Trained Instruction Representations

Querying the ECG with all question tokens can introduce redundant linguistic context and makes cross-modal attention cost dependent on the variable or padded length of the question sequence. We therefore learn fixed-length instruction representations to distill task-relevant semantics and provide a uniform query budget for both branches.

Let $H_q \in \mathbb{R}^{L_q\times d_I}$ denote the question embeddings. We contextualize M learnable instruction tokens $I \in \mathbb{R}^{M\times d_I}$, with the question, through a self-attention and feed-forward module F_ϕ:

$$I^*(q) = \big[F_\phi\big([I; H_q]\big)\big]_{1:M} \in \mathbb{R}^{M\times d_I}. \tag{2}$$

The instruction positions can attend to the entire question, and their contextualized outputs form the queries used by both ECG branches. Both I and F_ϕ are optimized through the answer-generation objective in Section 3.4. Thus, the representation is trained for extracting signal evidence that supports the answer, rather than simply summarizing the question. This task supervision encourages the contextualizer to emphasize relevant semantics and downweight peripheral linguistic content.

Fixing M independently of L_q provides a uniform query length for subsequent cross-modal attention. The full question is provided directly to the LLM, allowing the instruction representations to focus on ECG evidence extraction while preserving task semantics and response requirements.

3.3 Question-Conditioned Dual-Attention Alignment

The structured feature grid and task-trained queries provide the two inputs needed for question-conditioned evidence extraction. To extract question-relevant evidence while preserving lead and temporal structure, we use parallel Lead and Temporal branches, each aggregating over one axis while retaining the other in its output.

In each branch, the instruction representations guide the aggregation of ECG features through multi-head cross-attention, with $I^*(q)$ as queries and an ECG slice S as keys and values. For branch $b \in \{\text{lead}, \text{time}\}$, we express the resulting aggregation as

$$\mathcal{G}_b(S \mid I^*(q)) = \frac{1}{M}\sum_{m=1}^{M}\big[\mathcal{F}_b\big(\text{CA}_b\big(I^*(q), S, S\big)\big)\big]_{m,:}, \tag{3}$$

where CA_b and $\mathcal{F}_b$ denote branch-specific cross-attention and feed-forward transformations. The branches share the instruction representations but have independent parameters. Mean pooling is applied only over the M instruction outputs, yielding one d_a-dimensional representation for each ECG slice. Residual connections and normalization are left implicit for clarity.

Lead Branch: temporal evidence within each lead. For lead c, the branch accesses its complete encoded temporal sequence $Z_{c,:} \in \mathbb{R}^{N\times d_z}$ and produces $L_c(q) = \mathcal{G}_{\text{lead}}(Z_{c,:} \mid I^*(q))$. Each attention head assigns question-dependent weights to the temporal patches through an $M \times N$ attention map. The resulting representation summarizes relevant temporal evidence within that lead without combining it with other leads. Applying this operation separately to all leads produces $\mathbf{L}(q) \in \mathbb{R}^{C\times d_a}$, preserving one question-conditioned representation per lead.

Temporal Branch: cross-lead evidence at each temporal position. To retain temporal organization alongside the lead-specific summaries, the Temporal Branch accesses all leads at each encoded patch position, $Z_{:,n} \in \mathbb{R}^{C \times d_z}$, and produces $T_n(q) = \mathcal{G}_{\text{time}}(Z_{:,n} \mid I^*(q))$. Its $M \times C$ attention map weighs the leads according to the question, aggregating cross-lead evidence while retaining the temporal index n. Repeating this operation across the recording produces $\mathbf{T}(q) \in \mathbb{R}^{N \times d_a}$, a time-indexed sequence that complements the lead-specific representations.

Structured ECG tokens for the LLM. We concatenate the temporal and lead representations along the token dimension and apply a learned projector to map each token into the LLM embedding space. This tokenwise mapping preserves the correspondence to individual leads and temporal positions without further pooling. The combined sequence thus provides complementary views of the ECG: lead-specific summaries of temporal evidence and time-indexed summaries of cross-lead evidence. These question-conditioned ECG tokens are supplied to the LLM together with the original question.

3.4 Training Procedure

Stage 0: ECG encoder pretraining. We first pretrain the ECG encoder through self-supervised masked signal reconstruction. Each lead is encoded independently, and a lightweight reconstruction head maps the encoded features back to the waveform space. Reconstruction is supervised by the original waveform patches using mean squared error evaluated only at masked positions. This provides waveform-level supervision for the lead-specific representations before multimodal alignment. After pretraining, we discard the reconstruction head and keep the ECG encoder frozen throughout Stages I and II. Pretraining details are provided in Appendix B.2.

Stage I: ECG-language alignment. We freeze the pretrained ECG encoder and the LLM and train the ECG-LLM interface, including the structural embeddings, the instruction tokens and contextualizer, the dual-attention branches, and the modality projector. Let θ denote the interface parameters and ψ the LLM parameters. Given training examples $(X, q, a) \sim \mathcal{D}$, we optimize the answer-generation objective

$$\mathcal{L}_{\text{ans}}(\theta, \psi) = -\mathbb{E}_{(X,q,a)\sim\mathcal{D}} \sum_{r=1}^{|a|} \log p_{\theta,\psi}\big(a_r \mid a_{<r}, X, q\big). \tag{4}$$

With ψ fixed, the answer loss updates the entire interface, jointly training the instruction representations and ECG aggregation modules along with the projector. This stage learns to extract question-relevant ECG evidence and present it for answer generation by the pretrained LLM.

Stage II: ECG-language fine-tuning. Starting from the Stage-I checkpoint, we jointly fine-tune all LLM parameters and all interface components, while keeping the ECG encoder frozen. We retain the same objective in Equation equation 4. Unlike the preceding stage, which adapts the interface to a fixed language model, this full-parameter fine-tuning allows the LLM to adapt to the complementary lead-specific and time-indexed ECG representations while continuing to refine question-conditioned evidence extraction. Additional training details are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. We train LEAP-LLM on ECGInstruct (Ruoqi et al., 2024) and ECG-Grounding (Lan et al., 2025), comprising 1.16M and 30K instruction-response pairs, respectively. ECGInstruct spans ECG feature recognition, rhythm analysis and pathology identification across multiple question formats, while ECG-Grounding provides supervision linking diagnostic interpretations to fine-grained ECG evidence. We evaluate on four ECG-Bench (Ruoqi et al., 2024) datasets, PTB-XL and CODE-15% for multi-label abnormality detection, and G12EC and CSN for multiple-choice abnormality detection. Following ECG-Bench, we report macro AUC, F1, and Hamming Loss (HL) for multi-label tasks, and accuracy for multiple-choice tasks. All downstream validation and test records are excluded from both ECG encoder pretraining and ECG-language training. We additionally evaluate the diagnostic information retained in the multimodal representations through linear probing on PTB-XL and CODE-15 provided in Appendix C.

Table 1: ECG-Bench abnormality detection results. Best results are shown in **bold**, and second-best results are underlined.

Model	PTB-XL			CODE-15%			G12EC	CSN
	F1 ↑	AUC ↑	HL ↓	F1 ↑	AUC ↑	HL ↓	Acc. ↑	Acc. ↑
Random	33.2	50.3	50.1	15.0	48.8	32.1	12.1	11.6
General-purpose MLLMs								
InternVL2-8B	14.3	50.6	27.8	16.1	55.8	17.7	37.5	47.7
Qwen2-VL-7B	22.4	51.3	30.8	24.8	60.7	20.5	32.9	25.5
Qwen2.5-VL-7B Instruct	16.8	51.5	27.6	26.2	57.9	19.5	13.8	12.2
Proprietary MLLMs								
GPT-4o	28.3	55.6	26.2	24.9	59.9	15.7	49.2	57.5
Claude-3.5-Sonnet	27.5	54.0	29.6	20.3	58.3	17.8	51.4	51.5
ECG-specialized Methods								
ECG-GPT	53.9	69.5	20.1	40.1	68.9	17.4	N/A	N/A
HeartLLM	46.1	65.2	19.5	65.1	79.0	8.0	11.4	13.3
PULSE	74.8	82.4	11.0	85.4	90.7	5.0	78.2	85.2
GEM	75.8	83.4	11.0	86.4	91.5	4.7	80.5	86.2
ECG-R1	73.7	81.7	11.4	86.7	91.4	4.6	84.5	90.4
LEAP-LLM$_{\text{LLaVA}}$	**76.2**	**83.6**	**10.5**	87.2	**92.2**	4.5	85.2	90.6
LEAP-LLM$_{\text{Qwen3}}$	75.9	83.5	10.7	**87.3**	91.7	**4.5**	**87.9**	**90.8**

Baselines. We compare LEAP-LLM against both general-purpose and ECG-specialized language models. General-purpose baselines include InternVL (Chen et al., 2024), Qwen-VL variants (Wang et al., 2024a; Bai et al., 2025), GPT-4o (Hurst et al., 2024), and Claude-3.5-Sonnet (Anthropic, 2024). ECG-specific baselines include ECG-GPT (Khunte et al., 2024), HeartLLM (Yang et al., 2026), PULSE (Ruoqi et al., 2024), GEM (Lan et al., 2025), and ECG-R1 (Jin et al., 2026).

Implementation Details. ECGs are resampled to 500 Hz and represented as 10-second, 12-lead signals. We use a frozen PatchTST-style ECG encoder with hidden dimension 512, producing 99 temporal patches per lead, and 128 question-conditioned instruction tokens. The proposed interface compresses each ECG into 111 modality tokens before passing them to the language model. We evaluate a LLaVA based (Vicuna-7B) and a Qwen3-VL-8B backbone. Training follows the two-stage procedure in Section 3. Stage I optimizes only the ECG-LLM interface, while Stage II jointly fine-tunes the interface and language model with the ECG encoder frozen. Models are trained on four NVIDIA L40S GPUs. Full implementation and pretraining details are provided in Appendix B.

4.2 Results

Predictive Performance. Table 1 summarizes performance across the four ECG-Bench datasets. LEAP-LLM consistently improves over prior ECG-specific models across both PTB-XL and CODE-15% as well as the out-of-domain G12EC and CSN benchmarks. With the same Vicuna-7B backbone as GEM, the LLaVA-based variant achieves stronger results on both PTB-XL and CODE-15%, indicating that the gains arise from the structured lead-time ECG interface. Importantly, these improvements are not confined to a single evaluation metric. LEAP-LLM improves F1 and AUC while also reducing HL on both multi-label benchmarks. The improvements extend to unseen datasets, where the Qwen3-based variant surpasses ECG-R1 on both G12EC and CSN, including a 3.4-percentage-point gain on G12EC. Both LEAP-LLM variants outperform the strongest prior ECG-specific results on these out-of-domain benchmarks, suggesting that the advantage is not specific to a particular backbone. Together, these results show that preserving lead-time structure provides consistent benefits across datasets, evaluation metrics, and language-model backbones.

Inference Efficiency. Table 2 shows that LEAP-LLM is substantially more efficient than existing ECG-LLM approaches. Its structured interface represents each ECG with only 111 modality tokens, reducing the amount of ECG information processed by the LLM while preserving strong down-

Table 2: Inference efficiency comparison with ECG-LLM baselines.

Method	LLM Backbone	ECG Tokens	Prefill FLOPs (T)	Prefill Latency (ms)	TTFT (ms)
HeartLLM	LLaMA-3.2-3B	313	**2.74**†	108.0	138.0
GEM	Vicuna-7B	676	10.73	65.8	90.7
PULSE	Vicuna-7B	2,879	43.66	242.7	270.9
ECG-R1	Qwen3-8B	862	13.52	145.9	149.1
LEAP-LLM	Vicuna-7B	**111**	3.13	**33.1**	**43.6**

†Lower FLOPs mainly reflect the smaller backbone, despite using 2.8× more modality tokens than LEAP-LLM. 4-bit NF4 inference adds dequantization overhead not captured by FLOPs, increasing latency.

stream performance. This compact representation translates directly into lower inference cost, with 3.4-14.0× fewer prefill FLOPs and 2.1-6.2× lower time to first token than the compared baselines. LEAP-LLM also achieves the lowest wall-clock prefill latency, and notably attains a lower TTFT than HeartLLM despite using a substantially larger language-model backbone. These results show that preserving lead-time structure enables a more compact and computationally efficient ECG-LLM interface without sacrificing interpretation quality. Full profiling details are provided in Appendix D.

4.3 Ablation

We ablate each branch of the dual-axis alignment module in the LEAP-LLM+LLaVA variant to isolate its contribution. The *Lead-only* variant removes the Temporal Branch, retaining question-conditioned representations for each physical lead but discarding explicit cross-lead representations at each temporal location. The *Temporal-only* variant removes the Lead Branch, retaining cross-lead representations over time but discarding separate lead-specific representations. Both ablations substantially reduce performance across the four benchmarks (Table 3). Retaining both branches achieves 76.20 F1 on PTB-XL, 87.20 F1 on CODE-15, 90.60% accuracy on CSN, and 85.16% on G12EC. Together, these results show that the two branches capture complementary information and neither is sufficient in isolation.

Table 3: Architectural ablation of LEAP-LLM LlaVA across ECG-Bench datasets.

Dataset	Metric	Lead Only	Temporal Only	Full
PTB-XL	F1	<u>51.68</u>	42.43	**76.20**
CODE-15	F1	65.08	<u>77.59</u>	**87.20**
CSN	Acc.	74.30	<u>82.25</u>	**90.60**
G12EC	Acc.	56.01	<u>64.52</u>	**85.16**

5 Structural and Functional Analysis

We test whether LEAP-LLM meaningfully preserves and uses lead-time structure through three complementary analyses. First, we evaluate whether different myocardial infarction territories show greater attention to their clinically associated lead groups. Second, we evaluate temporal selectivity by aligning attention to detected R peaks and measuring how attention varies across cardiac phase. Finally, we test functional dependence of the tokens by replacing selected patient-specific lead tokens with their dataset-mean representations and measuring the resulting change in the likelihood of the correct answer. Together, these analyses assess whether the learned representations are temporally localized, anatomically structured, and functionally used by the LLM.

5.1 Lead-Temporal Selectivity.

Lead attention exhibits clear anatomical selectivity (Figure 3). Inferior MI patients receive significantly greater attention than anterior MI patients over the inferior leads (II, III, and aVF), whereas anterior MI patients receive greater attention than inferior MI patients over the anterior leads (V1-

V4). Thus, the lead-attention pattern changes with infarction territory and emphasizes the lead group clinically associated with the corresponding diagnosis (Zimetbaum & Josephson, 2003). This selectivity is also localized in time. Both groups show their strongest attention immediately before the QRS complex, with a secondary peak in the ST region. These regions correspond to waveform features used in myocardial infarction diagnosis and localization (Wagner et al., 2009). Together, these patterns indicate that LEAP-LLM extracts diagnostic evidence from the appropriate anatomical leads at clinically relevant locations within the cardiac cycle. Details of R-peak alignment and attention calculation are provided in Appendix E.

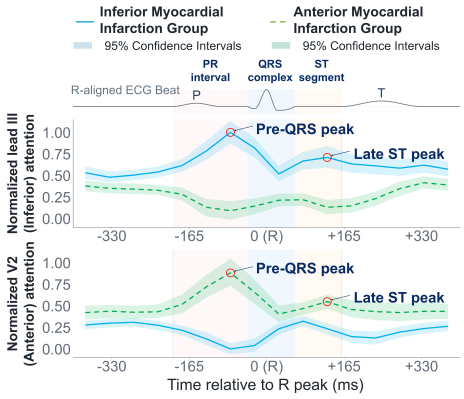


Figure 3: **LEAP-LLM attention patterns in myocardial infarction groups.** Inferior MI shows higher attention in inferior leads, while anterior MI shows higher attention in anterior leads. Attention peaks near clinically relevant QRS and ST-segment regions.

5.2 Functional Dependence on Lead-Specific Representations

Figure 4 provides functional evidence that the LLM selectively relies on lead representations associated with the relevant infarction territory. We report the reduction in the likelihood of the correct response, where larger values indicate a stronger effect of replacing the corresponding lead tokens with their dataset-mean representations. For inferior MI samples, replacing the inferior leads reduces likelihood by 50.5%, whereas replacing the anterior leads produces only a 6.8% reduction. The pattern reverses for anterior MI, replacing the anterior leads causes a 55.0% reduction, compared with only 5.5% when the inferior leads are replaced. This crossed interaction provides functional evidence that the LLM selectively relies on anatomically relevant lead representations

Predictions for inferior and anterior MI are disproportionately affected when the lead group associated with the corresponding infarction territory is disrupted. Thus, preserving lead identity does not merely retain structural information, but determines which ECG evidence is used by the downstream language model.Further details on the intervention procedure are provided in Appendix F.

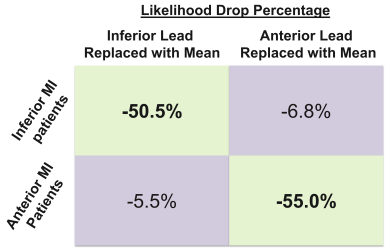


Figure 4: Reduction in the likelihood of the correct answer for different patients after replacing lead-specific tokens with their dataset-mean representations.

6 Conclusion

We presented LEAP-LLM, a question-conditioned ECG-LLM interface designed to preserve the native lead and temporal organization of multi-lead ECG signals when translating them into language-model inputs. LEAP-LLM maintains complementary lead-specific and temporally resolved tokens while using the clinical question to selectively extract relevant signal evidence. Across multiple ECG benchmarks, this structured interface achieves strong performance while requiring substantially fewer modality tokens and lower inference cost than existing ECG-language approaches. Our structural analyses further show that the learned representations exhibit clinically meaningful organization across both lead and temporal dimensions, while targeted interventions demonstrate that the LLM selectively relies on anatomically relevant lead representations when forming its responses. Together, these findings suggest that preserving the intrinsic structure of physiological signals provides an effective and efficient interface for language-based reasoning over ECGs.

AI DISCLOSURE

In this work, we used generative AI tools to assist with manuscript writing and editing, including grammar correction, sentence restructuring, improving clarity and conciseness, and refining phrasing. The tools were not used to develop the research methodology, design experiments, generate or interpret experimental results, or formulate the scientific conclusions of this work. We reviewed all AI-assisted text and revised it as necessary. We take responsibility for the final content of this work, including all text, claims, and other material.

REFERENCES

AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3(6), 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Günal, and Mehmet Bilginer Gülmezoglu. A survey on ecg analysis. *Biomed. Signal Process. Control.*, 43: 216–235, 2018. URL https://api.semanticscholar.org/CorpusID:21690511.

Yifu Cai, Mononito Goswami, Arjun Choudhry, Arvind Srinivasan, and Artur Dubrawski. Jolt: Jointly learned representations of language and time-series. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

William Han, Chaojing Duan, Michael A Rosenberg, Emerson Liu, and Ding Zhao. Ecg-byte: a tokenizer for end-to-end generative electrocardiogram language modeling. *arXiv preprint arXiv:2412.14373*, 2024.

E William Hancock, Barbara J Deal, David M Mirvis, Peter Okin, Paul Kligfield, and Leonard S Gettes. Aha/accf/hrs recommendations for the standardization and interpretation of the electrocardiogram: part v: electrocardiogram changes associated with cardiac chamber hypertrophy: 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*, 119(10):e251–e261, 2009.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in biology and medicine*, 122:103801, 2020.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiarui Jin, Haoyu Wang, Xingliang Wu, Xiaocheng Fang, Xiang Lan, Zihan Wang, Deyun Zhang, Bo Liu, Yingying Zhang, Xian Wu, et al. Ecg-r1: protocol-guided and modality-agnostic mllm for reliable ecg interpretation. *arXiv preprint arXiv:2602.04279*, 2026.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=Unb5CVPtae.

Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: What can large language models tell us about time series analysis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan

Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22260–22276. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/jin24i.html.

Akshay Khunte, Veer Sangha, Evangelos K Oikonomou, Lovedeep S Dhingra, Arya Aminorroaya, Andreas Coppi, S Vasisht Shankar, Bobak J Mortazavi, Deepak L Bhatt, Harlan M Krumholz, et al. Automated diagnostic reports from images of electrocardiograms at the point-of-care. *Medrxiv: the Preprint Server for Health Sciences*, 2024.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International conference on machine learning*, pp. 5606–5615. PmLR, 2021.

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, Olle Pahlm, Pentti Rautaharju, and Galen S. Wagner. Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology. *Journal of the American College of Cardiology*, 49(10):1109–1127, 2007. doi: 10.1016/j.jacc.2007.01.024. URL https://www.jacc.org/doi/10.1016/j.jacc.2007.01.024.

Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4532–4540, 2022.

Xiang Lan, Hanshu Yan, Shenda Hong, and Mengling Feng. Towards enhancing time series contrastive learning: A dynamic bad pair mining approach. In *International conference on learning representations*, volume 2024, pp. 10359–10379, 2024.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. Gem: Empowering mllm for grounded ecg understanding with time series and images. *arXiv preprint arXiv:2503.06073*, 2025.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott R. Ceresnak, Aydin Zahedivash, Alexander T Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Oppers Aalami, Filipe Barata, and Paul Schmiedmayer. OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FHDRzhKm0f.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pp. 402–415. PMLR, 2024.

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

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024b.

George Moody, Tom Pollard, and Benjamin Moody. WFDB Software Package. *PhysioNet*, June 2022. doi: 10.13026/gjvw-1m31. URL https://doi.org/10.13026/gjvw-1m31. Version 10.7.0.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WcOohbsF4H.

Kamran Namjouyan, Ervin Sejdić, Mark S Link, Antonio Pelliccia, Benjamin Glicksberg, Natalia Trayanova, and Chayakrit Krittanawong. Precision medicine for electrocardiogram interpretation: Clinical relevance, challenges, and advances. *Reviews in Cardiovascular Medicine*, 26(12):47007, 2025.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

Liu Ruoqi, Bai Yuelin, Yue Xiang, and Ping Zhang. Teach multimodal llms to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

Niraj Varma, Iwona Cygankiewicz, Mintu P. Turakhia, et al. 2021 ISHNE/HRS/EHRA/APHRS collaborative statement on mhealth in arrhythmia management: Digital medical tools for heart rhythm professionals. *European Heart Journal - Digital Health*, 2(1):7–48, January 2021. doi: 10.1093/ehjdh/ztab001.

Galen S Wagner, Peter Macfarlane, Hein Wellens, Mark Josephson, Anton Gorgels, David M Mirvis, Olle Pahlm, Borys Surawicz, Paul Kligfield, Rory Childers, et al. Aha/accf/hrs recommendations for the standardization and interpretation of the electrocardiogram: part vi: acute ischemia/infarction: 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*, 119(10):e262–e270, 2009.

Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Jing Xiong, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. Meit: Multimodal electrocardiogram instruction tuning on large language models for report generation. In *Findings of the association for computational linguistics: ACL 2025*, pp. 14510–14527, 2025.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12694–12702, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153, 2024b.

Yilin Wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li, and Zhongyu Wei. Itformer: Bridging time series and natural language for multi-modal qa with large-scale multitask dataset. *arXiv preprint arXiv:2506.20093*, 2025b.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Jinning Yang, Wenjie Sun, and Wen Shi. Heartllm: Discretized ecg tokenization for llm-based diagnostic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34250–34258, 2026.

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

Qinghao Zhao, Shijia Geng, Boya Wang, Yutong Sun, Wenchang Nie, Baochen Bai, Chao Yu, Feng Zhang, Gongzheng Tang, Deyun Zhang, et al. Deep learning in heart sound analysis: From techniques to clinical applications. *Health Data Science*, 4:0182, 2024.

Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Zhaowei Zhu, Xiang Lan, Tingting Zhao, Yangming Guo, Pipin Kojodjojo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Han Wang, Xingzhi Sun, et al. Identification of 27 abnormalities from multi-lead ecg signals: an ensembled se_resnet framework with sign loss function. *Physiological Measurement*, 42(6):065008, 2021.

Peter J Zimetbaum and Mark E Josephson. Use of the electrocardiogram in acute myocardial infarction. *New England Journal of Medicine*, 348(10):933–940, 2003.

A ELECTROCARDIOGRAM (ECG) BACKGROUND

ECG Aquisition A standard 12-lead electrocardiogram (ECG) records cardiac electrical activity from multiple spatial viewpoints using ten physical electrodes, four limb electrodes and six precordial electrodes. The limb electrodes generate six limb leads, with the augmented leads derived from the standard limb leads as

$$\mathrm{aVR} = -\frac{\mathrm{I} + \mathrm{II}}{2}, \qquad \mathrm{aVL} = \frac{\mathrm{I} - \mathrm{III}}{2}, \qquad \mathrm{aVF} = \frac{\mathrm{II} + \mathrm{III}}{2}.$$

The six precordial electrodes provide leads $\{\mathrm{V1}, \mathrm{V2}, \mathrm{V3}, \mathrm{V4}, \mathrm{V5}, \mathrm{V6}\}$, resulting in 12 ECG leads from 10 physical electrodes.

ECG Temporal Structure and Diagnostic Information Many cardiac abnormalities are identified by presence of a waveform change at different temporal locations.. The P wave reflects atrial depolarization, the QRS complex primarily reflects ventricular depolarization, and the T wave reflects ventricular repolarization. Consequently, different abnormalities manifest through changes in distinct temporal regions or relationships between them. For example, rhythm abnormalities may be identified from beat-to-beat regularity, R-R intervals, and P-wave–QRS relationships, whereas conduction and morphological abnormalities may appear through changes in QRS duration, waveform shape, or ST-T segments (Berkaya et al., 2018; Varma et al., 2021). Thus, preserving temporal location is important because diagnostically relevant evidence is distributed non-uniformly across the cardiac cycle.

Lead-dependent diagnostic information. Many cardiac abnormalities are identified not only by the presence of a waveform change, but also by the leads in which that change occurs. This is particularly important for myocardial infarction (MI), where lead groups correspond approximately to different anatomical regions of the myocardium. Inferior MI is primarily associated with abnormalities in leads II, III, and aVF, whereas anterior infarction is reflected predominantly in the precordial leads V1-V4 (Wagner et al., 2009; Zimetbaum & Josephson, 2003).

These properties motivate treating a multi-lead ECG as a structured lead-time signal rather than as an unordered collection of features. The temporal axis describes how cardiac electrical activity evolves within and across beats, while the lead axis identifies the spatial viewpoint from which that activity is measured. In LEAP-LLM, these two axes are explicitly preserved so that a clinical question can selectively retrieve evidence from the leads and temporal regions relevant to the requested interpretation.

B IMPLEMENTATION AND TRAINING DETAILS

B.1 ECG PREPROCESSING

All ECG recordings are resampled to 500 Hz and represented as 10-second, 12-lead signals. Each input therefore contains $C = 12$ leads and $T = 5000$ temporal samples. Signals are provided to the ECG encoder in their original voltage scale. We do not apply per-record standardization because absolute waveform amplitude contains diagnostically relevant information. (Hancock et al., 2009)

Each lead is partitioned into overlapping patches of length $P = 100$ samples with stride $S = 50$, corresponding to a patch duration of 200 ms and a displacement of 100 ms between consecutive patches. The resulting number of patches per lead is

$$N = \left\lfloor \frac{T - P}{S} \right\rfloor + 1 = \left\lfloor \frac{5000 - 100}{50} \right\rfloor + 1 = 99. \tag{5}$$

Thus, the ECG encoder receives 99 temporal patches from each of the 12 leads.

B.2 ECG ENCODER PRETRAINING

We pretrain the PatchTST encoder on PTB-XL and MIMIC using a self-supervised masked signal reconstruction objective. The temporal patches are embedded into a 512-dimensional latent space. During pretraining, 40% of the input patches are randomly masked, and the encoder is trained to reconstruct the corresponding raw ECG signal from the remaining visible context. A lightweight

Table 4: Optimization settings for ECG encoder pretraining and the two ECG-language training stages.

Setting	**Stage 0** *Encoder Pretraining*	**Stage I** *Alignment*	**Stage II** *Fine-tuning*
ECG encoder	Trainable	Frozen	Frozen
ECG-LLM interface	–	Trainable	Trainable
Language model	–	Frozen	Trainable
Training objective	Masked reconstruction	Answer generation	Answer generation
Optimizer	AdamW	AdamW	AdamW
Peak learning rate	1×10^{-3}	5×10^{-4}	1×10^{-5}
Scheduler	Cosine decay	Cosine decay	Cosine decay
Warm-up	5%	5%	5%
Epochs	10	1	1
Batch size	48	48	48
Gradient accumulation	2	8	8
Precision	bf16	bf16	bf16

reconstruction head maps the encoded representations back to the signal space, and the model is optimized using mean squared error computed over the masked regions:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 ,$$

where $\mathcal{M}$ denotes the set of masked patches, $\mathbf{x}_i$ is the original ECG patch, and $\hat{\mathbf{x}}_i$ is its reconstruction. This objective encourages the encoder to capture local waveform morphology as well as dependencies across longer temporal contexts without relying on diagnostic labels. After pretraining, the reconstruction head is discarded and the pretrained PatchTST encoder is used to initialize the ECG encoder in our multimodal framework.

B.3 Model Configuration

We use a 6-layer PatchTST-style encoder with a hidden dimension of $d_{\text{enc}} = 512$ and 8 self-attention heads. The encoder processes each lead independently.

Our architecture contains 128 learnable instruction tokens. These tokens are contextualized by the clinical question and shared between two independently parameterized cross-attention branches. The Lead Branch produces one representation for each of the 12 physical leads, whereas the Temporal Branch produces one representation for each of the 99 temporal patches. Their outputs are concatenated to form

$$N_{\text{ECG}} = C + N = 12 + 99 = 111 \tag{6}$$

structured ECG tokens. A learned modality projector maps these tokens from the interface dimension into the embedding dimension of the language model without altering the token axis.

We instantiate the proposed interface with the two language-model backbones reported in Table 1: a LLaVA-based backbone for the first model variant and Qwen3-VL-8B for the second variant. More details about the training is presented in 4

C Diagnostic Information in Multimodal Representations

C.1 Linear Probing

We use linear probing to evaluate how much diagnostically discriminative information remains accessible in the multimodal representations supplied to the language model. We probe the output of each model's alignment immediately before it is consumed by the LLM, rather than probing the underlying ECG or vision encoder. This isolates the quality of the modality representation exposed to the language model.

Table 5: Per-class linear probing ROC-AUC on PTB-XL and CODE-15%. The best result for each class is shown in bold.

Dataset	Class	LEAP-LLM$_{\text{LlaVA}}$	GEM	PULSE
PTB-XL	NORM	**0.8914**	0.8199	0.7614
	MI	**0.9290**	0.8313	0.6415
	STTC	**0.9377**	0.8735	0.8404
	CD	**0.8657**	0.8095	0.7955
	HYP	**0.8515**	0.7701	0.7677
	Macro	**0.8951**	0.8208	0.7613
CODE-15%	Normal ECG	**0.9918**	0.9108	0.8756
	1dAVb	**0.9663**	0.9316	0.8415
	RBBB	**0.9877**	0.9824	0.9347
	LBBB	**0.9981**	**0.9981**	0.9892
	SB	**0.9935**	0.9668	0.9711
	ST	**0.9952**	0.9813	0.9621
	AF	**1.0000**	0.9659	0.9326
	Macro	**0.9904**	0.9624	0.9295

Representation extraction. For each model, we extract the modality representation immediately before it is provided to the LLM. For LEAP-LLM, this corresponds to the concatenated lead and temporal branch outputs after projection into the LLM embedding space, yielding 111 tokens: 12 lead tokens and 99 temporal tokens. For GEM, we concatenate the projected time-series and image tokens. For PULSE, we extract the projected visual tokens produced from the rendered ECG image.

All three models expose 4096-dimensional representations at the LLM interface. We use the LLaVA-based variant of LEAP-LLM for this comparison so that LEAP-LLM, GEM, and PULSE share the same decoder family and interface dimensionality. All multimodal model parameters remain frozen during representation extraction and probe training.

Evaluation cohorts. We conduct probing experiments on the official ECGBench PTB-XL and CODE-15% evaluation cohorts. The corresponding native ECG waveforms are used for LEAP-LLM, while the official ECGBench-rendered images are used for PULSE. GEM receives both representations of the same examination. The evaluated records are held out from the instruction- tuning data used to train the compared models.

The PTB-XL cohort consists of 2,082 records from the official test split. PTB-XL defines five diagnostic superclasses: normal ECG (NORM), myocardial infarction (MI), ST/T change (STTC), conduction disturbance (CD), and hypertrophy (HYP). Because these superclasses are not mutually exclusive, we retain only examinations assigned to exactly one superclass, providing an unambiguous target for this probing analysis.

For CODE-15%, we evaluate seven abnormalities: normal ECG, first-degree atrioventricular block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), sinus tachycardia (ST), and atrial fibrillation (AF). We retain the multi-label formulation and independently probe each abnormality using a one-vs-rest target.

Probe training. The extracted representations are frozen before probe fitting. We train an independent one-vs-rest logistic regression classifier for every diagnostic category. This gives five classifiers per model for PTB-XL and seven classifiers per model for CODE-15%.

For each dataset, examinations are randomly divided into 80% training and 20% testing partitions using seed 0. For PTB-XL, this results in 1,274 training and 319 testing examinations. The identical examination-level partition is used for LEAP-LLM, GEM, and PULSE, ensuring that differences cannot be attributed to the probe split.

We use scikit-learn's `LogisticRegression` implementation with the `lbfgs` optimizer and `max_iter=1000`. All other hyperparameters use the default scikit-learn configuration. Performance is measured using the area under the receiver operating characteristic curve (ROC-AUC) for each one-vs-rest classifier.

Results. As shown in Table 5, LEAP-LLM produces the strongest linearly separable representations on both datasets. On PTB-XL, LEAP-LLM achieves an AUC of 0.8951, improving upon GEM and PULSE-7B by 0.0743 and 0.1338, respectively. The improvement is observed across all five diagnostic superclasses, including morphology-oriented categories such as MI, STTC, and HYP.

On CODE-15%, LEAP-LLM achieves a macro-AUC of 0.9904, compared with 0.9624 for GEM and 0.9295 for PULSE. LEAP-LLM obtains the highest or tied-highest AUC for all seven abnormalities. These results indicate that our model consistently preserves diagnostically relevant information across diverse abnormalities, demonstrating the benefit of maintaining explicit lead and temporal structure at the ECG-LLM interface.

D Inference Efficiency Measurement

We evaluate the inference efficiency of LEAP-LLM and the ECG-LLM baselines under a unified profiling protocol. All methods are evaluated using the same input example, prompt, hardware configuration, batch size, and decoding length. We report modality-token count, prefill FLOPs, prefill latency, time to first token (TTFT), end-to-end generation latency, and peak GPU memory.

D.1 Evaluation Protocol

All models are evaluated with batch size 1 on a single NVIDIA A100-SXM4-40GB GPU. We use the same ECGBench PTB-XL test example (`ptb-xl-14149`) and the same benchmark question for all methods. Image-based methods receive the rendered ECG image provided by ECGBench, whereas signal-based methods receive a 12×5000 input corresponding to a 10-second, 12-lead ECG sampled at 500 Hz.

For latency measurements, we perform three warm-up iterations followed by ten timed iterations and report the median. End-to-end generation uses one warm-up and three timed iterations because of its substantially higher cost. CUDA synchronization is performed within each timed region to ensure that asynchronous GPU execution is included in the measurement.

To make autoregressive generation directly comparable, all models use greedy decoding with `min_new_tokens=max_new_tokens=256`. Fixing the generated sequence length prevents differences in early stopping behavior from being interpreted as computational differences.

We report the following quantities:

- **Modality tokens**: the number of non-text tokens inserted into the LLM input to represent the ECG.
- **Prefill FLOPs**: analytical LLM forward-pass FLOPs required to process the complete input prompt before autoregressive decoding.
- **Prefill latency**: wall-clock latency of the LLM forward pass over the input prompt.
- **Time to first token (TTFT)**: latency of a complete generation call until the first generated token, including modality encoding, multimodal alignment, LLM prefill, and the first decoding step.
- **End-to-end latency**: wall-clock time required to generate all 256 output tokens.

TTFT and prefill latency measure different portions of the inference path. For LEAP-LLM, GEM, and PULSE, prefill latency is measured after modality embeddings have been computed and therefore isolates LLM prompt processing. TTFT instead includes the complete model path and is consequently the primary cross-model wall-clock measure of input-processing latency. For ECG-R1, the public model implementation does not expose an equivalent interface for independently timing the LLM prefill, so its reported prefill latency additionally includes its modality encoders.

D.2 FLOP Accounting

We analytically compute the forward-pass FLOPs of the language model using the convention that one multiply-accumulate operation corresponds to two FLOPs. Let p_{body} denote the number of parameters in the transformer blocks excluding the embedding table and language-model head, L

Table 6: Detailed inference-cost comparison. All latency measurements use batch size 1 on an NVIDIA A100-SXM4-40GB. FLOPs correspond to LLM computation; end-to-end measurements generate exactly 256 tokens.

Method	Non-LLM Params (M)	Modality Tokens	Prefill Tokens	Prefill FLOPs (T)	Decode FLOPs (T)	Total FLOPs (T)	Prefill (ms)	TTFT (ms)	E2E (ms)	Peak VRAM (MB)
HeartLLM†	**0.03**†	313	457	**2.74**†	**1.75**†	**4.49**†	108.0	138.0	21,896	**2,468**
GEM	411	676	802	10.73	3.51	14.23	65.8	90.7	8,678	14,787
PULSE	324	2,879	3,005	43.66	3.80	47.46	242.7	270.9	8,972	16,993
ECG-R1	682	862	936	13.52	4.04	17.56	145.9	149.1	14,200	17,389
LEAP-LLM	235	**111**	**239**	3.13	3.43	6.56	**33.1**	**43.6**	**8,344**	13,585

†HeartLLM uses a 3.3B LLM quantized to 4-bit NF4 with unmerged LoRA adapters, whereas the remaining models use 6.7–8.2B LLMs in FP16/BF16. Its FLOPs and VRAM therefore reflect a different model size and numeric execution path and should not be interpreted as a like-for-like arithmetic comparison.

the number of transformer layers, d the hidden dimension, V the vocabulary size, T the number of input tokens, and n the number of generated tokens. The prefill cost is

$$\mathrm{FLOPs}_{\mathrm{prefill}} = 2p_{\mathrm{body}}T + 4LdT^2 + 2dV, \tag{7}$$

where the first term accounts for parameterized transformer operations and the second accounts for the attention matrix products.

For autoregressive decoding with a KV cache, the corresponding cost for n generated tokens is (Hoffmann et al., 2022; Pope et al., 2023)

$$\mathrm{FLOPs}_{\mathrm{decode}} = 2p_{\mathrm{body}}n + 4Ld\left(nT + \frac{n(n-1)}{2}\right) + 2dVn. \tag{8}$$

The $nT + n(n-1)/2$ term accounts for the increasing number of cached key positions attended to across successive decoding steps. Embedding lookup and element-wise operations such as normalization, rotary position encoding, nonlinear activations, and softmax are excluded. We also exclude the modality encoders from the FLOP calculation so that the reported quantity isolates the computational consequence of the representation supplied to the LLM. Their parameter counts and their contribution to wall-clock TTFT are measured separately.

D.3 Modality Tokenization and Timing Protocol

Image tokenization differs across the evaluated models. GEM resizes the input to 336×336 and therefore uses a fixed visual-token count. PULSE employs an AnyRes input strategy; for the evaluated PTB-XL images, the selected 672×672 configuration produces four high-resolution crops in addition to the base image. ECG-R1 uses dynamic visual resolution with its released evaluation configuration capped at 768 visual tokens. In contrast, LEAP-LLM directly consumes the ECG waveform, and its 111 modality tokens are fixed by the dimensions of the lead-temporal interface.

Finally, each latency result corresponds to the median of repeated timed iterations for one benchmark example. Modality-token counts and analytical FLOPs are deterministic for the evaluated input configuration, while the observed prefill and TTFT timing ranges were below 2% across repeated iterations.

D.4 Detailed Efficiency Results

Table 6 provides the full breakdown of inference cost across the evaluated ECG–LLMs. The main differences arise during input processing: LEAP-LLM uses the fewest modality tokens and correspondingly has the lowest prefill latency and TTFT among the full-precision models. In contrast, the end-to-end latency of LEAP-LLM, GEM, and PULSE is relatively similar despite large differences in modality-token count, since the fixed 256-token output introduces a substantial autoregressive decoding cost that is shared across these similarly sized language models.

HeartLLM exhibits a different efficiency profile. Its quantized 3.3B backbone results in substantially lower peak memory and analytical FLOPs, but its measured latency is considerably higher.

This illustrates that analytical FLOPs are most informative when comparing models with similar model sizes and numerical execution paths, whereas wall-clock latency can also depend strongly on quantization and implementation details.

E LEAD-TIME STRUCTURAL ANALYSIS

We analyze the attention maps produced by the Temporal Branch to examine whether the learned ECG-LLM interface organizes information jointly according to lead identity and cardiac phase. All analyses use the PTB-XL test split from ECG-Bench and the same fixed classification prompt, such that only the ECG input varies across examinations. We analyze the raw post-softmax attention weights without additional normalization or post-hoc z-scoring.

Because attention weights indicate where the interface concentrates information but do not establish causal importance, we interpret this analysis as evidence of model selectivity. Functional dependence on the resulting lead representations is evaluated separately through the lead-token interventions in Appendix F.

Myocardial infarction cohorts. We construct myocardial infarction cohorts from the PTB-XL SCP diagnostic codes using an any-code inclusion criterion. Inferior myocardial infarction includes {IMI, ILMI, IPMI, IPLMI}, with the clinically associated leads II, III, and aVF, while anterior myocardial infarction includes {AMI, ASMI}, associated with leads V1-V4 (Zimetbaum & Josephson, 2003). This yields 310 inferior-MI examinations, 264 anterior-MI examinations, and 925 NORM examinations. We use the NORM cohort as a common reference to determine whether the lead-time attention patterns observed in each infarction group differ from those present in normal ECGs.

R-peak detection and cardiac-phase alignment. To align attention patterns across ECGs with different heart rates, we define temporal position relative to the cardiac cycle. R peaks are detected from lead II using the WFDB XQRS (Moody et al., 2022) algorithm. Each ECG patch is assigned a temporal offset according to the distance between its center and the nearest detected R peak. We retain ECGs with at least three detected R peaks and at least one RR interval within the physiologically plausible range of 0.3–2.0 s (30–200 bpm), which helps exclude records affected by obvious detection errors. Using these criteria, 1,997 of the 2,082 PTB-XL test ECGs (95.9%) are successfully aligned for the phase-resolved analysis.

Phase-resolved lead attention. The Temporal Branch operates independently at each temporal patch, with the question-conditioned instruction tokens attending across the 12 ECG leads. Thus, every temporal position is associated with a distribution of attention across leads. Retaining these weights over all temporal patches produces a two-dimensional lead–time attention map for each ECG.

Patch centers are grouped into 50 ms phase bins according to their temporal offset from the nearest R peak. Within each phase bin, we average the Temporal Branch attention assigned to each of the 12 leads, yielding a phase-resolved lead-attention map for every ECG. We denote the resulting attention assigned to lead c in cardiac-phase bin τ for ECG x as $A_{c,\tau}(x)$.

To isolate disease-associated changes from global lead preferences learned by the model, we compare each infarction cohort against the NORM reference rather than interpreting absolute attention values. For each lead and cardiac-phase bin, we compute

$$\Delta A_{c,\tau} = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{MI}}}[A_{c,\tau}(x)] - \mathbb{E}_{x\sim\mathcal{D}^{w}_{\mathrm{NORM}}}[A_{c,\tau}(x)], \tag{9}$$

where $\mathcal{D}_{\mathrm{MI}}$ denotes the corresponding inferior- or anterior-MI cohort and $\mathcal{D}^{w}_{\mathrm{NORM}}$ denotes its heart-rate-matched NORM reference cohort.

Uncertainty estimation. We estimate 95% confidence intervals using 2,000 record-level bootstrap resamples. Infarction and NORM examinations are independently resampled within each comparison, and the heart-rate-matching weights and lead-by-phase contrasts are recomputed for each bootstrap sample. The shaded regions in Figure 3 denote the resulting percentile confidence intervals.

F Functional Dependence on Lead-Specific Representations

Lead-token intervention. The Lead Branch preserves one representation for each physical ECG lead. This allows individual lead tokens to be manipulated immediately before they are supplied to the LLM. For an intervened lead set S, we replace each patient-specific lead token with the mean representation of the corresponding lead computed over the reference cohort:

$$\widetilde{\boldsymbol{L}}_c = \begin{cases} \overline{\boldsymbol{L}}_c, & c \in S, \\ \boldsymbol{L}_c, & c \notin S, \end{cases} \tag{10}$$

where $\boldsymbol{L}_c$ is the original representation for lead c and $\overline{\boldsymbol{L}}_c$ is its dataset-level mean representation. Replacement is performed separately for each lead, thereby removing examination-specific information while preserving lead identity and avoiding the out-of-distribution perturbation introduced by zeroing a token. The temporal tokens and all non-intervened lead tokens remain unchanged.

We evaluate three primary interventions: replacing the inferior leads $\{\mathrm{II}, \mathrm{III}, \mathrm{aVF}\}$, replacing the anterior leads $\{\mathrm{V1}, \ldots, \mathrm{V4}\}$, and replacing all 12 lead tokens.

Likelihood-based evaluation. We quantify the effect of each intervention using the likelihood of the ground-truth answer. For examination i with target answer $y_i = (y_{i,1}, \ldots, y_{i,m_i})$, we first compute the mean token log-likelihood

$$\ell_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \log p\left(y_{i,j} \mid y_{i,<j},\, \mathrm{ECG}_i,\, q_i\right). \tag{11}$$

For diagnostic group G and intervention S, we first compute the geometric-mean likelihood ratio between the intervened and intact representations,

$$R(G,S) = \exp\left(\frac{1}{|G|} \sum_{i \in G} \left[\ell_i^S - \ell_i^{\mathrm{intact}}\right]\right). \tag{12}$$

We report the corresponding reduction in gold-answer likelihood,

$$D(G,S) = 1 - R(G,S), \tag{13}$$

where ℓ_i^S and ℓ_i^{intact} denote the mean token log-likelihoods under the intervened and original representations, respectively. A value of $D = 0$ indicates no change in the model's support for the correct answer, while larger positive values indicate a greater reduction in likelihood and therefore stronger dependence on the replaced representations. For ease of interpretation, we report $D(G,S)$ as the percentage reduction in gold-answer likelihood.

Control interventions. We include two controls to distinguish territory-specific dependence from generic sensitivity to token perturbation. First, we replace three randomly selected lead tokens, matching the number of tokens in the inferior-lead intervention. Secondly, we replace 12 randomly selected temporal tokens, matching the number of tokens removed by the all-lead intervention.

Uncertainty estimation. We estimate 95% confidence intervals (Tibshirani & Efron, 1993) using 2,000 record-level bootstrap resamples within each diagnostic group. For each resample, we recompute the mean difference between the intervened and intact log-likelihoods, convert it to the likelihood ratio R, and report the corresponding likelihood reduction $D = 1 - R$. Confidence intervals are obtained from the bootstrap distribution of D.

Additional results. The control interventions further support the territory specificity of the main findings. Replacing three randomly selected lead tokens reduces the gold-answer likelihood by 17.2% for inferior MI and 17.6% for anterior MI, substantially smaller than the 50.5% and 55.0% reductions produced by replacing the anatomically matched lead groups, respectively. In contrast, replacing 12 randomly selected temporal tokens has essentially no effect, with $D = -0.2\%$ for both infarction groups.

G Limitations and Future Work

Limitations. Our evaluation is limited to standardized 10-second, 12-lead ECG recordings sampled at 500 Hz, and therefore does not establish how LEAP-LLM behaves under different recording

Table 7: Full lead-token intervention results. Values report the reduction in gold-answer likelihood, $D = 1 - R$, with 95% bootstrap confidence intervals. Higher values indicate greater dependence on the replaced representations. Bold values denote the anatomically matched territory-specific interventions.

Tokens replaced	**Inferior MI**	**Anterior MI**
All leads	73.5% [71.3, 75.5]	80.7% [78.3, 82.7]
Inferior leads	50.5% [46.1, 54.4]	5.5% [1.6, 9.4]
Anterior leads	6.8% [3.2, 11.0]	55.0% [49.9, 59.5]
Random three leads	17.2% [13.6, 20.8]	17.6% [11.8, 23.4]
Random 12 temporal tokens	−0.2% [−0.5, 0.1]	−0.2% [−0.6, 0.1]

durations, sampling rates, or reduced-lead configurations. In addition, our current experiments use a limited set of LLM backbones, we do not systematically evaluate LEAP-LLM across a broader range of LLMs , so the generality of the proposed interface across architectures remains to be further established. Finally, our interpretability analysis is limited to inferior and anterior myocardial infarction, and therefore does not yet establish whether the same lead-selective behavior holds across other cardiac abnormalities.

Future Work. Future work will extend LEAP-LLM to more diverse ECG acquisition settings, including variable recording durations, sampling rates, and reduced- or missing-lead configurations. We also plan to evaluate the proposed interface across a broader range of LLM backbones to better characterize its architectural generality. Finally, we will expand the interpretability analysis beyond inferior and anterior myocardial infarction to additional rhythms, conduction abnormalities, and morphological findings, providing a more comprehensive view of how LEAP-LLM uses lead-specific and temporal information across diagnostic settings.