PhysioBench-ECG: Confidence-Aware Benchmarking of Clinical Time-Series Reasoning

Anonymous ACL submission

Abstract

We introduce PHYSIOBENCH-ECG, a large-scale confidence-aware benchmark for evaluating clinical time-series reasoning over raw 12-lead ECGs. While existing ECG benchmarks largely evaluate high-level diagnosis-, report-, or clinical-decision-oriented QA, they do not directly test the lower-level time-series interpretation skills that make ECG reasoning possible: identifying fiducials, measuring intervals, recognizing rhythm and morphology patterns, and comparing waveform structure across beats and leads. PhysioBench-ECG fills this gap by converting raw 12-lead ECGs into auditable evidence-to-answer QA pairs that target intermediate waveform interpretation rather than final diagnosis prediction. From 800,035 MIMIC-IV-ECG records, we construct 19,996,104 QA pairs across 31 clinically motivated question types. Each answer is grounded through multi-method waveform alignment, including fiducials, intervals, rhythm statistics, morphology features, and cross-lead measurements, and is derived using deterministic ECG guideline-based rules. Evaluation under 3 input conditions shows that models perform best when rule-relevant evidence is explicitly provided, but struggle when they must recover the same evidence from raw waveform or visual ECG representations. PhysioBench-ECG therefore enables diagnostic evaluation of model accuracy, evidence reliability, and failure modes in physiological signal reasoning.

1 Introduction

Current medical AI evaluations are largely confined to text-based, "textbook" scenarios. While large language models (LLMs) achieve high performance on biomedical reading comprehension and exam-style reasoning benchmarks such as PubMedQA and MedQA (Jin et al., 2019, 2021), real-world clinical reasoning requires grounding these linguistic rules in raw, multi-channel physiological signals. Electrocardiograms (ECGs), electroencephalograms, and bedside waveforms are complex time series where interpretation depends on precise timing, morphology, and spatial relationships across sensors. A model that easily retrieves the definition of atrial fibrillation may still fail catastrophically when forced to execute the complex multimodal reasoning required to identify an irregular beat, measure QRS width, or determine if an ST deviation is localized to specific leads.

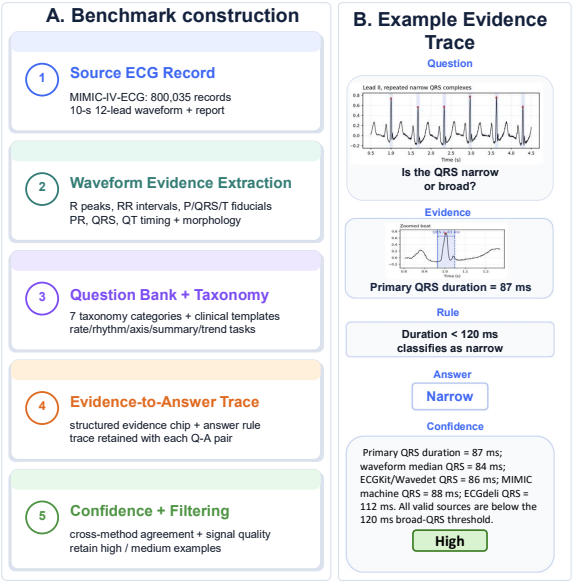

Figure 1: PhysioBench-ECG construction and evidence-to-answer trace. (A) Each MIMIC-IV-ECG record is converted into auditable QA pairs by extracting waveform evidence, applying clinical templates and deterministic ECG rules, and assigning confidence labels using source agreement and signal-quality guardrails. (B) Example QRS-width trace showing how structured evidence is mapped through a rule to an answer and confidence label.

Existing ECG resources do not adequately isolate this reasoning capability. Large-scale datasets like PTB-XL and MIMIC-IV-ECG provide 12-lead ECGs paired with diagnostic statements and clinical metadata (Wagner et al., 2020; Gow et al.,

2023), while ECG-QA formulates interpretation using expert-validated templates (Oh et al., 2023). However, these approaches compress the continuous waveform into final clinical diagnoses. Crucially for NLP evaluation, because open-access clinical datasets and their derived labels may appear in public training or evaluation corpora, diagnosis-only evaluation risks measuring memorized associations rather than active waveform reasoning. Furthermore, generic time-series benchmarks, which test structural features like seasonality and anomalies (Fons et al., 2024; Merrill et al., 2024), are detached from domain-specific measurement standards and clinical consequences.

To bridge this gap, we introduce PHYSIOBENCH-ECG, a benchmark designed to evaluate complex multimodal instruction following and physiological grounding over raw 12-lead ECG time series. Instead of testing whether a model can output a memorized diagnosis, PhysioBench-ECG evaluates whether it can recover the intermediate signal observations that support interpretation, such as P-wave organization, PR measurability, and beat-to-beat variability. These intermediate temporal and spatial reasoning tasks force active, in-context signal grounding that cannot simply be retrieved from textual discharge summaries.

Because generating a clinical gold-standard at scale requires extreme rigor, PhysioBench-ECG relies on an auditable, calibrated consensus engine. Each answer is derived from waveform measurements and deterministic interpretation rules, producing a QA row that includes the final answer, a natural-language rule trace, structured numerical evidence, and a construction-time confidence label based on cross-method agreement and signal-quality guardrails.

Our contributions are:

- **Waveform-Grounded Multimodal Benchmark.** We introduce PhysioBench-ECG, a large-scale QA dataset built from MIMIC-IV-ECG (19.9M QA pairs, 800K ECGs) that bypasses textual data contamination by isolating intermediate physiological reasoning.

- **Auditable Evidence-to-Answer Engine.** We replace opaque labels with auditable traces. Each answer is derived from waveform fiducials, rhythm statistics, and cross-lead measurements, paired with an explicit rule trace and a rigorously assigned confidence label.

- **Diagnostic evaluation of physiological signal grounding.** We benchmark LLMs and VLMs under serialized-waveform, question-specific-evidence, and ECG-plot settings to identify where models fail. We find that models perform best when rule-relevant evidence is explicitly provided, but struggle when they must extract intervals, morphology cues, rhythm statistics, and cross-lead relationships from raw waveform or visual ECG representations.

2 Related Work

2.1 Medical and Clinical QA Benchmarks

Medical QA benchmarks have been central to evaluating whether language models encode biomedical knowledge and clinical reasoning. Early resources such as PubMedQA, MedQA, and MedMCQA evaluate biomedical reading comprehension and exam-style medical reasoning (Jin et al., 2019, 2021; Pal et al., 2022), while MultiMedQA and Med-PaLM broaden evaluation to professional, research, and consumer medical questions (Singhal et al., 2023, 2025). More recent clinical benchmarks move beyond exam questions toward realistic care settings. EHRNoteQA grounds patient-specific questions in MIMIC-IV discharge summaries and clinician review (Kweon et al., 2024); ClinicBench evaluates LLMs across clinical language understanding, generation, and open-ended decision-making tasks (Liu et al., 2024); and MedHELM proposes a clinician-validated taxonomy covering broad medical workflows and tasks (Bedi et al., 2026). These benchmarks show the importance of clinically grounded evaluation, but their evidence is primarily textual. PhysioBench instead evaluates whether models can answer questions whose evidence is contained in physiological waveforms.

2.2 ECG Datasets, ECG QA, and ECG-Language Models

Large ECG datasets provide the substrate for waveform-grounded evaluation. PTB-XL offers public 12-lead ECGs with expert diagnostic statements (Wagner et al., 2020), and MIMIC-IV-ECG scales public diagnostic ECG research to hundreds of thousands of ECG records linked to the MIMIC clinical ecosystem (Gow et al., 2023). ECG-QA is the closest prior benchmark to our setting: it formulates ECG interpretation as question answer-

ing using expert-validated templates over ECG records (Oh et al., 2023). Recent ECG-language and ECG-multimodal work further shows growing interest in interactive ECG interpretation, including MEIT for multimodal ECG report generation (Wan et al., 2025), ECG-LM and ECG-Chat for ECG-language modeling (Yang et al., 2025; Zhao et al., 2025), ECG-Expert-QA for multimodal diagnostic ECG QA (Wang et al., 2025), Q-HEART for knowledge-informed ECG QA (Pham et al., 2025), and PULSE/ECGBench for ECG image instruction tuning and benchmark evaluation (Liu et al., 2026). These efforts mainly evaluate diagnostic QA, report generation, or ECG image interpretation. PhysioBench is complementary: it targets intermediate waveform-grounded observations such as fiducial timing, interval measurability, rhythm regularity, and lead-localized deviations, with each answer linked to explicit numerical evidence and construction rules.

2.3 Language Models for Time-Series Reasoning

A parallel line of work studies whether language models can process or reason about time series. LLMTime represents numerical sequences as text for zero-shot forecasting (Gruver et al., 2023), Time-LLM adapts language models for forecasting through reprogramming (Jin et al., 2024), and TimesFM trains a decoder-only foundation model for time-series forecasting (Das et al., 2024). Evaluation work shows that forecasting ability does not imply robust time-series reasoning. Fons et al. propose an EMNLP benchmark and taxonomy for LLM understanding of univariate and multivariate time-series features, including trend, seasonality, anomalies, volatility, structural breaks, stationarity, correlation, and cross-correlation (Fons et al., 2024). Merrill et al. show that language models still struggle with zero-shot time-series reasoning, including etiological reasoning, QA, and context-aided forecasting (Merrill et al., 2024). PhysioBench builds on this feature-centered view, but adds clinical grounding: ECG time-series categories are tied to standardized physiological measurements, waveform fiducials, lead relationships, and rule-based clinical interpretation.

Together, these lines of work motivate PhysioBench: medical QA benchmarks are clinically motivated but mostly text-based; ECG-language benchmarks are ECG-specific but often diagnosis- or report-oriented; and time-series reasoning benchmarks are structurally rigorous but clinically generic. PhysioBench connects these directions by evaluating waveform-derived, clinically meaningful time-series observations through auditable QA.

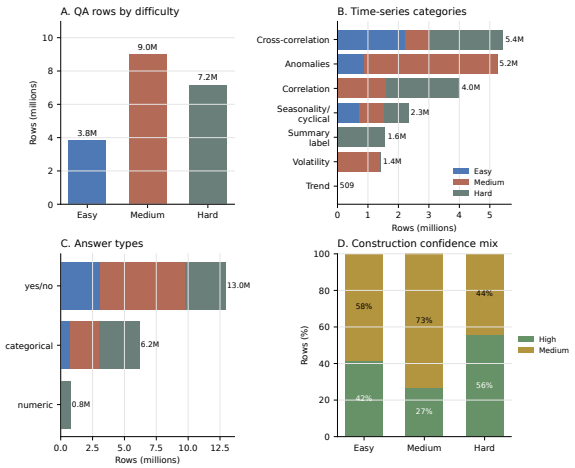

Figure 2: PhysioBench ECG composition: (A) QA row counts for the final easy, medium, and hard question groups. (B) QA rows by generic time-series category, stacked by difficulty. (C) QA rows by answer type, stacked by difficulty. (D) Construction-confidence mix within each difficulty group; construction confidence reflects source agreement, repeated waveform evidence, clinical-rule applicability, and guardrail checks, whereas difficulty is assigned from waveform-evidence text-model Canonical Accuracy.

3 PhysioBench ECG Construction

To construct PhysioBench-ECG, we started with raw ECG input signal from MIMIC-IV-ECG (Gow et al., 2023), as shown in Figure 3. After we ground waveform evidence through multi-method alignment of signal features, we convert this evidence into answers using deterministic ECG guideline-based rules. Last, Multi-method alignment and extracted signal pattern consistency sis used to assign confidence score. Each released QA row contains the ECG identifier, question, canonical answer, structured evidence, rule trace, time-series category, construction-time confidence label, and model-facing difficulty label. Figure 1 shows an example PhysioBench-ECG data entry.

3.1 Question Construction

The ECG question set follows a standard twelve-lead interpretation workflow, but targets intermediate waveform observations rather than diagnosis labels alone. We derive 31 questions per ECG from

Question group	Signal evidence	Time-series category
Rate and rhythm	Estimate heart rate and rhythm class from R-peak timing and RR-interval regularity.	Seasonality/Cyclical; Volatility
P-wave organization	Detect P-wave presence, regularity, duration, and temporal organization relative to QRS complexes.	Anomalies; Volatility; Cross-Correlation
Axis and lead polarity	Compare QRS polarity across limb leads, especially Lead I and aVF, to infer axis-related patterns.	Correlation
PR interval and AV conduction	Measure atrial-to-ventricular timing and screen for normal/prolonged/shortened PR, fixed PR prolongation, Wenckebach pattern, dropped QRS, AV dissociation, and AV block.	Cross-Correlation; Trend; Anomalies
QRS morphology and conduction	Measure QRS duration and screen for narrow/broad QRS, RSR′ morphology, and notched lateral R-wave patterns.	Anomalies
ST/T/QT and infarction evidence	Assess pathologic Q-wave evidence, ST-segment deviation, T-wave abnormality, QT thresholding, and QTc prolongation.	Anomalies; Cross-Correlation
Composite clinical summaries	Aggregate supported rhythm, conduction, morphology, repolarization, and ischemia-related findings into summary-level ECG impressions.	Summary label

Table 1: PhysioBench-ECG question design for 31 question types. Each group follows a clinically recognizable step in 12-lead ECG interpretation, while the signal-operation column summarizes the waveform evidence used to construct answers. The final column maps each group to the time-series categories (Fons et al., 2024).

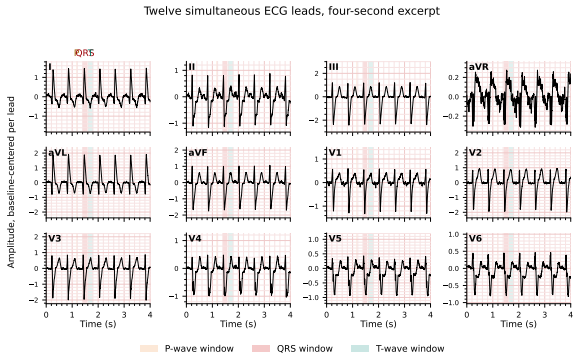

Figure 3: Example four-second excerpt from a de-identified MIMIC-IV-ECG record. The highlighted P, QRS, and T windows illustrate the temporal anchors used by PhysioBench; the twelve simultaneous leads provide different projections of the same cardiac events.

Statistic	Value
ECG records	800,035
Unique patients	161,352
Number of unique question types	31
Final QA pairs	19,996,104
Easy set	7 IDs; 3,827,023 rows
Medium set	13 IDs; 8,997,454 rows
Hard set	11 IDs; 7,171,627 rows
High-confidence rows	7,999,786 (40.0%)
Medium-confidence rows	11,996,318 (60.0%)

Table 2: PhysioBench ECG statistics. Difficulty sets follow the final question-level assignment in Table 6; evaluated labels use waveform-evidence mean text-model Canonical Accuracy with easy $\geq$80%, medium 50–79.9%, and hard $<$50%.

established ECG reading and measurement criteria, covering rate and rhythm, axis, P-wave assessment, PR interval and AV conduction, QRS morphology, ST/T/QT repolarization, and ischemia-related waveform findings (Sattar and Chhabra, 2023; Noble et al., 1990; Kligfield et al., 2007; Surawicz et al., 2009; Rautaharju et al., 2009; Wagner et al., 2009; Kusumoto et al., 2019; Thygesen et al., 2007).

Table 1 maps each question group to its signal operation and corresponding time-series category, following the taxonomy of Fons et al. (2024), including cyclical repetition, volatility, anomalies, correlation, cross-correlation, trend, and summary labels. Signal-quality checks are used as confidence guardrails during construction rather than as a separate question family.

3.2 Answer Construction

Answer construction has two stages. First, PhysioBench grounds ECG waves, fiducial points, and intervals from the raw twelve-lead waveform. Second, it converts these grounded measurements into question-specific answers using deterministic rules derived from ECG interpretation standards.

Stage 1: Fiducial and interval grounding. For each MIMIC-IV-ECG record, the pipeline filters the raw twelve-lead waveform, detects beat anchors, extracts candidate P/QRS/T fiducials, and stores interval, rhythm, morphology, and cross-lead

measurements. MIMIC-IV-ECG includes machine-generated measurements and report text, which are useful references, but we do not treat them as gold labels. Machine-derived intervals may be unreliable when waveform components are absent, low-amplitude, merged, displaced, or difficult to identify, as can occur in atrial fibrillation/flutter, conduction disease, ectopy, paced rhythms, noisy recordings, and abnormal P-wave morphology. We therefore use MIMIC machine measurements as weak comparators and guardrails rather than as the default source of benchmark labels.

To calibrate waveform grounding, we use LUDB, an external twelve-lead ECG database with cardiologist annotations of P-wave, QRS-complex, and T-wave boundaries and peaks (Kalyakulina et al., 2020, 2021). We evaluate multiple delineation and interval-estimation methods against LUDB annotations, including ECGKit/Wavedet (Martínez et al., 2004; Demski and Llamedo Soria, 2016), ECGdeli (Pilia et al., 2021), WFDB-style QRS/RR detection, local waveform fallbacks, U-Net3+ point detectors, and a LUDB-calibrated P-wave candidate filter. For LUDB validation, patient/record-level splits prevent beats from the same ECG appearing in both training and evaluation; interval errors are computed from record-level median intervals against cardiologist annotations. The goal is not to select one global detector for all ECG properties. Instead, we select primary sources separately for each target and use cross-method agreement, repeated-beat evidence, clinical-rule applicability, and threshold margin to assign confidence.

For target type t, such as RR, PR, QRS duration, QT interval, or P-wave presence, let A_t indicate that the relevant candidate methods agree within a target-specific tolerance, and let $B_{t,k}$ indicate that at least k analyzable beats, intervals, or repeated observations are available. We estimate the reliability of each construction rule on LUDB as

$$\hat{p}_{t,k} = P(C_t = 1 \mid A_t = 1, B_{t,k} = 1),$$

where C_t denotes correctness relative to LUDB reference annotations. High-confidence labels require sufficient repeated evidence and agreement among the relevant methods; rows are downgraded when methods disagree, evidence is sparse, the answer is near a clinical threshold, or the physiology makes the question fragile. The full probabilistic motivation for cross-method agreement and repeated-beat aggregation is given in Appendix B.

Table 7 summarizes the source-selection decisions used by the QA generator. For example, ECGKit/Wavedet is used as the primary QRS-duration source because it achieves the lowest QRS duration error in our LUDB interval evaluation, while local waveform estimates and MIMIC machine measurements serve as guardrails. PR and P-wave questions use LUDB-calibrated P fiducials and ECGdeli PR as interval evidence, but are marked not applicable when organized P waves or stable P-to-QRS relationships are absent. QT uses a multi-method median across valid ECGKit, ECGdeli, local waveform, and MIMIC machine candidates because T-wave endpoint placement is more method-sensitive.

Stage 2: Rule-based answer generation. After waveform measurements are grounded, PhysioBench generates answers using deterministic rules based on standard ECG interpretation criteria. Clinical thresholds and terminology follow ECG standardization and interpretation guidelines, including AHA/ACCF/HRS recommendations for ECG measurement, QRS and conduction abnormalities, ST/T/QT interpretation, bradycardia and AV-conduction concepts, and infarction-related Q-wave evidence (Kligfield et al., 2007; Surawicz et al., 2009; Rautaharju et al., 2009; Wagner et al., 2009; Kusumoto et al., 2019; Thygesen et al., 2007). For example, the QRS-width question reads `primary_qrs_ms`; values below 120 ms yield *Narrow*, values at or above 120 ms yield *Broad*, and missing or clinically invalid evidence yields a non-answerable label. Similarly, PR questions are answered only when organized P waves and stable P-to-QRS relationships are present; QT questions use calibrated QT evidence and repolarization guardrails; and ST/T/Q morphology questions are downgraded when wide QRS, pacing, hypertrophy, bundle-branch morphology, or baseline instability makes the rule screening-like rather than definitive.

Each released QA row contains the final answer and its construction trace: the natural-language question, answer, answer type, confidence label, question family, time-series category, derivation method, and evidence JSON. The `method` field describes the rule used to derive the answer, while the `evidence` field stores the numerical values, source methods, agreement status, and guardrail flags used by that rule. This makes the benchmark auditable: model predictions are compared with answers whose physiological basis can be traced

Class	Core rule	Validation criteria
high	Clinically applicable answer with sufficient repeated evidence and no major conflict among sources.	**Cross-method:** primary source agrees with at least one independent source within target tolerance τ_t; threshold labels are far from the decision boundary. **Cross-wave / pattern:** repeated beats show consistent fiducials, intervals, or morphology; questions usually require $\geq$5 analyzable beats or intervals, and hard time-series questions require $\geq$6 repeated observations.
medium	Answerable label, but one grounding condition is weakened.	**Cross-method:** valid sources are available but show moderate disagreement, require median/consensus aggregation, or trigger a guardrail. **Cross-wave / pattern:** fewer observations, partial wave recovery, borderline threshold margin, or screening-like morphology rule.
low	Answer can be emitted, but evidence is weak or physiologically fragile.	**Cross-method:** limited source agreement or only one usable source. **Cross-wave / pattern:** sparse analyzable beats, unstable P/QRS/T candidates, ectopy, noisy baseline, or rhythm context that makes the rule unreliable.
not assessable/not applicable	The ECG lacks the physiological condition or reliable waveform evidence required to answer that question.	**Cross-method:** no reliable method produces the required target. **Cross-wave / pattern:** insufficient fiducials, too few beats, missing waveform evidence, or unusable signal quality.

Table 3: Confidence-label criteria used in PhysioBench. Confidence is assigned at construction time. Each label combines clinical-rule applicability, cross-method validation among ECG measurement sources, and cross-wave or repeated-pattern validation within the ECG. Low confidence and not assessable/not applicable QA pairs are removed from the final release to keep the published benchmark clean.

back to waveform-derived measurements.

3.3 Confidence Assignment

PhysioBench uses confidence-aware label construction because waveform-derived ECG measurements vary in reliability across signal quality, rhythm context, and interval type. We assign a `confidence` field for each answer that represents how strongly a generated label is supported by waveform evidence, clinical-rule applicability, signal quality, and cross-method agreement. We use `high`, `medium`, and `low` for answerable rows. Rows are marked `not_applicable` when the question is physiologically invalid for that ECG, such as PR follow-up questions when organized sinus P waves are absent. Rows are marked `not_assessable` when the question is clinically applicable in principle but the required waveform evidence is unavailable or unreliable.

Confidence is assigned in two steps. First, the pipeline sets a base level from evidence quantity: most questions require at least five analyzable beats or intervals for high confidence and three to four for medium confidence, while pattern-level questions generally require at least six repeated observations for high confidence and at least three for medium confidence. Second, target-specific guardrails downgrade confidence when candidate algorithms disagree, the signal is noisy or incomplete, the value is close to a clinical threshold, or the clinical rule is only screening-like in the observed rhythm or morphology context. Appendix B gives the calibration rationale, and Table 3 summarizes how the public confidence labels should be used during evaluation.

4 Experiments

4.1 Benchmark Performance

Table 4 reports Canonical Accuracy on a randomly selected 1,500-row PhysioBench-ECG QA set, evalauted across representative open-source, biomedical, and API-protected models, including Gemma 3 (Gemma Team, 2025), Llama 3/3.2 (Dubey et al., 2024; Meta AI, 2024), Mistral models (Jiang et al., 2023; Mistral AI, 2024, 2025a,b), MedGemma (Sellergren et al., 2025), Med42 (Christophe et al., 2024; M42 Health AI Team, 2024), BioMistral-Zephyr-SLERP (Labrak et al., 2024; Tunstall et al., 2023; Goddard et al., 2024; Kabster, 2024), Gemini 2.5 Flash (Gemini Team, Google, 2025), and GPT-4.1-mini (OpenAI, 2025b,a). Responses are scored by canonical answer normalization and exact-match accuracy. The main takeaway is that models perform best when the rule-relevant waveform evidence is explicitly provided: QUESTION-SPECIFIC EVIDENCE + RULE gives the strongest Overall performance, whereas performance drops substantially when models must recover evidence from raw serial-

Model	Overall	Season.	Vol.	Anom.	Corr.	X-Corr.	Trend	Summ.
FULL WAVEFORM + RULE								
Gemma-3-4B-IT	38.2	32.2	49.0	40.9	21.6	52.1	100.0	2.0
Gemma-3-12B-IT	37.4	44.3	12.2	38.9	33.2	61.8	7.4	3.3
Llama-3.2-3B-Instruct	43.8	36.9	50.0	35.3	50.0	61.5	100.0	0.0
Llama-3.1-8B-Instruct	50.6	46.3	42.9	73.8	22.4	67.5	38.9	2.7
Mistral-7B-Instruct-v0.3	45.8	35.6	100.0	58.4	34.8	52.1	7.4	0.0
Mistral-Small-3.1-24B-Instruct	46.8	30.9	44.9	67.1	45.6	53.3	0.0	6.7
MedGemma-4B-IT	54.8	51.7	100.0	56.2	34.4	72.9	100.0	0.0
Llama3-Med42-8B (sel. leads)	41.6	45.6	38.8	39.1	49.6	61.8	0.0	0.7
BioMistral-7B-Zephyr-SLERP	6.6	20.1	16.3	5.1	11.2	0.6	0.0	0.0
Gemini-2.5-Flash (API)	50.2	30.9	21.4	63.8	68.4	58.1	46.3	0.0
GPT-4.1-mini (API)	62.5	57.1	100.0	68.7	43.0	77.2	100.0	0.0
QUESTION-SPECIFIC EVIDENCE + RULE								
Gemma-3-4B-IT	56.9	67.1	68.4	72.7	26.0	70.9	70.4	6.0
Gemma-3-12B-IT	65.9	71.8	65.3	65.1	63.6	90.0	57.4	13.3
Llama-3.2-3B-Instruct	46.4	49.0	22.4	31.5	67.2	74.1	20.4	14.0
Llama-3.1-8B-Instruct	59.1	69.1	51.0	66.0	34.8	84.3	55.6	16.7
Mistral-7B-Instruct-v0.3	55.5	59.1	0.0	79.0	49.2	65.8	59.3	3.3
Mistral-Small-3.1-24B-Instruct	77.6	73.8	86.7	88.8	85.2	88.0	57.4	12.0
MedGemma-4B-IT	59.2	64.4	59.2	78.5	24.4	79.8	59.3	6.0
Llama3-Med42-8B	58.9	72.5	6.1	64.7	67.2	78.9	37.0	10.0
BioMistral-7B-Zephyr-SLERP	65.0	64.4	78.6	64.9	76.0	72.9	90.7	11.3
Gemini-2.5-Flash (API)	81.7	87.2	84.7	94.9	86.8	94.9	57.4	4.7
GPT-4.1-mini (API)	84.5	87.2	78.6	94.9	97.6	98.0	57.4	11.3
VLM 12-LEAD PLOT + RULE								
Gemini-2.5-Flash (API)	60.5	45.0	69.4	73.8	68.0	62.1	0.0	36.0
GPT-4.1-mini (API)	56.8	46.3	72.4	65.1	71.2	65.8	7.4	5.3

Table 4: Canonical Accuracy on the PhysioBench-ECG evaluation set. Rows are grouped by input setting. FULL WAVEFORM + RULE provides serialized raw ECG waveform values plus the question rule, without extracted evidence fields. QUESTION-SPECIFIC EVIDENCE + RULE provides the extracted evidence fields used by the construction rule for the target question. VLM 12-LEAD PLOT + RULE provides a rendered 12-lead ECG image plus the question rule. **Overall** reports accuracy after canonical answer normalization over all evaluated QA rows. Category columns report accuracy within each generic time-series category: Seasonality/Cyclical, Volatility, Anomalies, Correlation, Cross-Correlation, Trend, and Summary label. Green cells mark the best value within each input setting and column; ties are highlighted jointly.

ized waveform representations, with an average decrease of 21.1 percentage points. The image-based VLM 12-LEAD PLOT + RULE setting performs between structured evidence and serialized waveform input, indicating that rendered ECG plots provide useful visual signal context but still do not reliably expose precise intervals, morphology cues, rhythm statistics, and cross-lead relationships. Category-level results show strong performance on anomaly, correlation, and cross-correlation questions under evidence-guided prompting, while summary-label questions remain difficult across models, highlighting the challenge of synthesizing multiple waveform findings into a clinical impression.

4.2 Confidence-Aware Analysis

Because PhysioBench-ECG is designed as a confidence-aware benchmark, we further examine whether construction-time confidence is reflected in model behavior. Across both prompting settings, average accuracy is higher on high-confidence rows than on medium-confidence rows, indicating that construction confidence captures meaningful differences in waveform-evidence reliability and rule applicability. This trend is es-

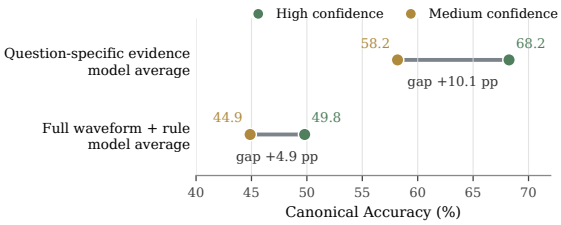

Figure 4: Model-average confidence gap. Dots show Canonical Accuracy for high- and medium-confidence rows under Question-Specific Evidence and Full Waveform+Rule input settings. Gap labels report the high-minus-medium difference in percentage points.

Prompt setting	Accuracy
Text evidence / rule settings	
Rule only	39.0
Evidence only	64.0
Evidence + rule	**74.0**
Raw waveform settings	
Waveform only	42.0
Waveform + rule	46.0
ECG plot settings	
ECG plot only	51.0
ECG plot + evidence	69.0
ECG plot + rule	62.0

Table 5: GPT ablation results on the 1,500-row PhysioBench-ECG subset. Accuracy is Canonical Accuracy after answer normalization. Evidence denotes extracted question-specific waveform evidence fields, rule denotes the deterministic ECG interpretation rule, waveform denotes serialized raw ECG waveform values, and ECG plot denotes the rendered 12-lead ECG image. The best setting is bolded.

pecially important because confidence is assigned before model evaluation, using source agreement, repeated waveform evidence, threshold margins, and clinical guardrails, rather than post hoc model performance.

4.3 Input Representation Ablation

Table 5 further isolates the role of evidence, rules, waveforms, and images in model performance. Paired McNemar tests confirm that extracted evidence is more important than rule text alone: the model performs significantly better when question-specific waveform evidence is provided, and performs best when that evidence is paired with the corresponding deterministic rule. In contrast, serialized waveform prompting remains much weaker, suggesting that the model struggles to recover clinically relevant intervals, morphology cues, rhythm patterns, and cross-lead relationships directly from raw waveform values. ECG plot prompting provides useful visual context and outperforms serialized waveform input in this ablation, but still remains below evidence-based prompting. Overall, these results indicate that current models can often apply ECG interpretation rules once the relevant signal evidence is exposed, but reliable extraction of that evidence from raw or visual ECG representations remains a major bottleneck.

4.4 Discussion

PhysioBench-ECG evaluates whether models can reason over physiological time-series evidence rather than only answer text-based clinical questions or predict final ECG labels. By converting raw 12-lead ECGs into auditable QA pairs with evidence fields, rule traces, and confidence labels, the benchmark targets intermediate observations used in ECG interpretation, including intervals, fiducial timing, rhythm regularity, morphology, and cross-lead relationships.

Our results show that input representation strongly determines model performance. Models perform best when question-specific waveform evidence is explicitly provided with the rule. Our ablation further supports this interpretation: evidence-only prompting substantially outperforms rule-only prompting, and combining evidence with the rule gives the best performance. Thus, the main bottleneck is not only medical knowledge, but access to reliable, structured signal evidence.

The category-level results reveal a second challenge: models handle many localized rule-based questions better than summary-level reasoning. Anomaly, correlation, and cross-correlation questions are relatively strong under evidence-guided prompting, but summary-label questions remain difficult across models. This indicates that synthesizing multiple waveform findings into a clinical impression remains a major unsolved capability, even when individual evidence fields are available.

The confidence-aware analyses show that construction confidence and model-facing difficulty capture different properties. High-confidence rows generally yield higher model accuracy than medium-confidence rows, supporting the idea that construction confidence reflects evidence reliability and rule applicability.

5 Conclusion

PHYSIOBENCH-ECG provides a confidence-aware benchmark for evaluating clinical time-series summarization and interpretation. Compared with existing ECG benchmarks, PHYSIOBENCH-ECG shifts the evaluation target from high-level diagnostic outputs to intermediate waveform reasoning over fiducials, intervals, rhythm patterns, morphology, and cross-lead relationships. By making physiological evidence explicit and auditable, PHYSIOBENCH-ECG establishes a new evaluation paradigm for clinical multimodal NLP: models must not only know medical rules, but also ground them in raw biomedical signals.

Limitations

PhysioBench-ECG is intended for model evaluation and research, not for clinical diagnosis, triage, or patient-care decision-making.

PhysioBench-ECG uses algorithmically constructed labels rather than fully clinician-adjudicated ground truth. Although labels are grounded in ECG interpretation standards, waveform-derived measurements, LUDB validation, and cross-method agreement, the full benchmark has not yet undergone systematic cardiologist review. It should therefore be viewed as a large-scale, auditable evaluation resource for physiological time-series reasoning, not as a substitute for expert-confirmed ECG diagnosis.

Confidence labels reduce but do not eliminate label uncertainty. High-confidence answers have stronger support from repeated waveform evidence, applicable rules, and source agreement, but remain construction-time reliability estimates. Medium-confidence rows should be interpreted with this metadata, while low-confidence, not-assessable, and not-applicable rows are excluded from the main release.

Finally, PhysioBench-ECG is limited to diagnostic 12-lead ECGs from MIMIC-IV-ECG and does not yet cover other physiological signals or longitudinal clinical context. Future work should add clinician review, targeted adjudication of subtle cases, and extensions to EEG, photoplethysmography, arterial blood pressure, and bedside telemetry.

Ethical Considerations

Intended use. PHYSIOBENCH-ECG is intended as a research benchmark for evaluating clinical time-series reasoning, physiological signal grounding, and model robustness. It is not intended for clinical diagnosis, triage, patient-care decision-making, or replacement of clinician judgment. Model performance on PHYSIOBENCH-ECG should therefore not be interpreted as evidence that a system is safe for deployment in clinical practice.

Data privacy and patient protection. PHYSIOBENCH-ECG is derived from MIMIC-IV-ECG, a de-identified clinical ECG resource. We do not attempt to re-identify patients, link records to external data sources, or recover protected health information. All benchmark construction and evaluation are conducted on de-identified records, and released artifacts will exclude raw identifiable information.

Label uncertainty and clinical risk. The benchmark uses algorithmically constructed labels derived from waveform measurements, deterministic ECG interpretation rules, cross-method agreement, signal-quality guardrails, and confidence labels. Although this design improves scalability and auditability, the labels are not a substitute for systematic cardiologist adjudication. To reduce misuse, PHYSIOBENCH-ECG reports construction-time confidence metadata and excludes low-confidence, not-assessable, and not-applicable rows from the main release. Users should interpret results as measurements of model reasoning behavior rather than as validated clinical claims.

Bias and representativeness. Because PHYSIOBENCH-ECG is derived from MIMIC-IV-ECG, it may inherit limitations of the source population, data collection process, device characteristics, clinical workflows, and documentation practices. Performance measured on this benchmark may not generalize to all patient populations, institutions, ECG acquisition settings, or downstream clinical use cases. Future work should include targeted subgroup analysis, external validation, and clinician review to better characterize fairness, robustness, and generalizability.

Artifact licensing and access. PHYSIOBENCH-ECG is derived from MIMIC-IV-ECG and therefore follows the access and usage restrictions of PhysioNet/MIMIC. We will release benchmark construction code, question templates, evaluation scripts, and metadata under an open-source license where permitted. Any data-derived artifacts that require protected or credentialed access will be distributed only to users who have completed the required PhysioNet credentialing process and agreed to the applicable data use agreement.

References

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi Haredasht, Ivan Lopez, and 64 others. 2026. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine*, 32(3):943–951.

Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco A. F. Pimentel. 2024. Med42-v2: A suite of clinical LLMs. *Preprint*, arXiv:2408.06142.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR.

Andrés Julio Demski and Mariano Llamedo Soria. 2016. ecg-kit: A MATLAB toolbox for cardiovascular signal processing. *Journal of Open Research Software*, 4(1):e8.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. 2024. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21598–21634, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Gemma Team. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. *Preprint*, arXiv:2403.13257.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. 2023. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*. Version 1.0.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Kabster. 2024. BioMistral-Zephyr-Beta-SLERP: Model card. `https://huggingface.co/Kabster/BioMistral-Zephyr-Beta-SLERP`. Accessed: 2026-05-26.

Alena Kalyakulina, Igor Yusipov, Viktor Moskalenko, Alexander Nikolskiy, Konstantin Kosonogov, Nikolai Zolotykh, and Mikhail Ivanchenko. 2021. Lobachevsky university electrocardiography database. *PhysioNet*. Version 1.0.1.

Alena I. Kalyakulina, Igor I. Yusipov, Viktor A. Moskalenko, Alexander V. Nikolskiy, Konstantin A. Kosonogov, Grigory V. Osipov, Nikolai Yu. Zolotykh, and Mikhail V. Ivanchenko. 2020. Ludb: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8:186181–186190.

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. 2007. Recommendations for the standardization and interpretation of the electrocardiogram: Part i: The electrocardiogram and its technology. *Circulation*, 115(10):1306–1324.

Fred M. Kusumoto, Mark H. Schoenfeld, Coletta Barrett, James R. Edgerton, Kenneth A. Ellenbogen, Michael R. Gold, Nora F. Goldschlager, Robert M. Hamilton, Jose A. Joglar, Robert J. Kim, Richard Lee, Joseph E. Marine, Christopher J. McLeod, Keith R. Oken, Kristen K. Patton, Christine N. Pellegrini, Craig A. Selzman, Andrew Thompson, and Paul D. Varosy. 2019. 2018 ACC/AHA/HRS guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay. *Journal of the American College of Cardiology*, 74(7):e51–e156.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. EHRNoteQA: An LLM benchmark for real-world clinical practice using discharge summaries. In *Advances in Neural Information Processing Systems*, volume 37. Datasets and Benchmarks Track.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. 2026. Teaching multimodal LLMs to comprehend 12-lead electrocardiographic images. *npj Digital Medicine*, 9(1):349.

M42 Health AI Team. 2024. Llama3-Med42-8B: Model card. https://huggingface.co/m42-health/Llama3-Med42-8B. Accessed: 2026-05-26.

Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. 2004. A wavelet-based ECG delineator: Evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581.

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3512–3533, Miami, Florida, USA. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2: Model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/. Accessed: 2026-05-26.

Mistral AI. 2024. Mistral-7B-Instruct-v0.3: Model card. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed: 2026-05-26.

Mistral AI. 2025a. Mistral Small 3.1. https://mistral.ai/news/mistral-small-3-1. Accessed: 2026-05-26.

Mistral AI. 2025b. Mistral-Small-3.1-24B-Instruct-2503: Model card. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503. Accessed: 2026-05-26.

R. Joe Noble, J. Stanley Hillis, and Donald A. Rothbaum. 1990. Electrocardiography. In H. Kenneth Walker, W. Dallas Hall, and J. Willis Hurst, editors, *Clinical Methods: The History, Physical, and Laboratory Examinations*, 3 edition. Butterworths, Boston.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. 2023. ECG-QA: A comprehensive question answering dataset combined with electrocardiogram. In *Advances in Neural Information Processing Systems*, volume 36. Datasets and Benchmarks Track.

OpenAI. 2025a. GPT-4.1 mini: Model documentation. https://developers.openai.com/api/docs/models/gpt-4.1-mini. Accessed: 2026-05-26.

OpenAI. 2025b. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2026-05-26.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. 2025. Q-HEART: ECG question answering via knowledge-informed multimodal LLMs. In *Proceedings of the European Conference on Artificial Intelligence*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 4545–4552. IOS Press.

Nicolas Pilia, Claudia Nagel, Gustavo Lenis, Silvia Becker, Olaf Dössel, and Axel Loewe. 2021. ECGdeli: An open source ECG delineation toolbox for MATLAB. *SoftwareX*, 13:100639.

Pentti M. Rautaharju, Borys Surawicz, Leonard S. Gettes, James J. Bailey, Rory Childers, Barbara J. Deal, Anton Gorgels, E. William Hancock, Mark Josephson, Paul Kligfield, Jan A. Kors, Peter Macfarlane, Jay W. Mason, David M. Mirvis, Peter Okin, Olle Pahlm, Gerard van Herpen, Galen S. Wagner, and Hein Wellens. 2009. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part iv: The ST segment, T and U waves, and the QT interval. *Journal of the American College of Cardiology*, 53(11):982–991.

Yasar Sattar and Lovely Chhabra. 2023. Electrocardiogram. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island (FL). Updated June 5, 2023. PMID: 31747210.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau,

and 1 others. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Scharli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Borys Surawicz, Rory Childers, Barbara J. Deal, Leonard S. Gettes, James J. Bailey, Anton Gorgels, E. William Hancock, Mark Josephson, Paul Kligfield, Jan A. Kors, Peter Macfarlane, Jay W. Mason, David M. Mirvis, Peter Okin, Olle Pahlm, Pentti M. Rautaharju, Gerard van Herpen, Galen S. Wagner, and Hein Wellens. 2009. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part iii: Intraventricular conduction disturbances. *Journal of the American College of Cardiology*, 53(11):976–981.

Kristian Thygesen, Joseph S. Alpert, Harvey D. White, and Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction. 2007. Universal definition of myocardial infarction. *Journal of the American College of Cardiology*, 50(22):2173–2195.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *Preprint*, arXiv:2310.16944.

Galen S. Wagner, Peter Macfarlane, Hein Wellens, Mark Josephson, Anton Gorgels, David M. Mirvis, Olle Pahlm, Borys Surawicz, Paul Kligfield, Rory Childers, Leonard S. Gettes, James J. Bailey, Barbara J. Deal, E. William Hancock, Jan A. Kors, Jay W. Mason, Peter Okin, Pentti M. Rautaharju, and Gerard van Herpen. 2009. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part vi: Acute ischemia/infarction. *Journal of the American College of Cardiology*, 53(11):1003–1011.

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):154.

Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Jing Xiong, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. 2025. MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14510–14527, Vienna, Austria. Association for Computational Linguistics.

Xu Wang, Jiaju Kang, Puyu Han, Ruida Liu, Luqi Gong, and Fanda Fan. 2025. ECG-Expert-QA: A benchmark for evaluating medical large language models in ECG. In *2025 IEEE International Conference on Bioinformatics and Biomedicine*, pages 5718–5723. IEEE.

Kai Yang, Massimo Hong, Jiahuan Zhang, Yizhen Luo, Suyuan Zhao, Ou Zhang, Xiaomao Yu, Jiawen Zhou, Liuqing Yang, Ping Zhang, Mu Qiao, and Zaiqing Nie. 2025. ECG-LM: Understanding electrocardiogram with a large language model. *Health Data Science*, 5:0221.

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

A Appendix

Table 6: List of the 31 PhysioBench ECG questions, their waveform-evidence difficulty labels, and their mapping to generic time-series categories. Difficulty is assigned from mean text-model Canonical Accuracy using waveform-evidence prompts: easy $\geq$80%, medium 50–79.9%, and hard $<$50%.

Question	**Difficulty**	**Time-series category**
Analytic ECG interpretation questions		
What is the heart rate?	Hard	Seasonality and Cyclical Patterns
Is the heart rate normal, bradycardic, or tachycardic?	Medium	Seasonality and Cyclical Patterns
Is the rhythm regular or irregular?	Medium	Volatility
What is the rhythm?	Easy	Seasonality and Cyclical Patterns
Are P waves present?	Easy	Anomalies
Is the QRS narrow or broad?	Medium	Anomalies
Are the P waves regular?	Medium	Volatility
Do P waves precede the QRS complexes?	Easy	Cross-Correlation
What is the cardiac axis?	Hard	Correlation
Is the QRS positive in both Lead I and aVF?	Medium	Correlation
Is the QRS positive in Lead I but negative in aVF?	Hard	Correlation
Is the QRS negative in Lead I but positive in aVF?	Hard	Correlation
Are both QRS complexes negative in Lead I and aVF?	Medium	Correlation
Is the P-wave duration 120 ms or less?	Medium	Anomalies
Is the PR interval between 120 and 200 ms?	Hard	Cross-Correlation
Is the PR interval normal, prolonged, or shortened?	Hard	Cross-Correlation
Is there a fixed PR interval greater than 200 ms?	Easy	Cross-Correlation
Does the PR interval get longer and longer until a beat drops?	Hard	Trend
Is there a dropped QRS without PR prolongation?	Easy	Anomalies
Are the P-waves and QRS complexes beating independently?	Hard	Cross-Correlation
Is any AV block present?	Medium	Cross-Correlation
Is the QRS duration less than 120 ms?	Medium	Anomalies
Is there an RSR′ M-shape pattern in V1 and V2?	Medium	Anomalies
Is there a notched R-wave in V5, V6, or Lead I?	Medium	Anomalies
Is there a Q-wave greater than 0.1 mV deep or greater than 40 ms wide?	Medium	Anomalies
Is the ST-segment elevated/depressed by at least 0.1 mV?	Medium	Anomalies
Are T-wave abnormalities present?	Easy	Anomalies
Is the QT interval less than 440 ms?	Easy	Cross-Correlation
Is the QTc prolonged?	Hard	Cross-Correlation
Summary rows		
What is the overall ECG impression?	Hard	Summary label
What is the most clinically significant finding?	Hard	Summary label

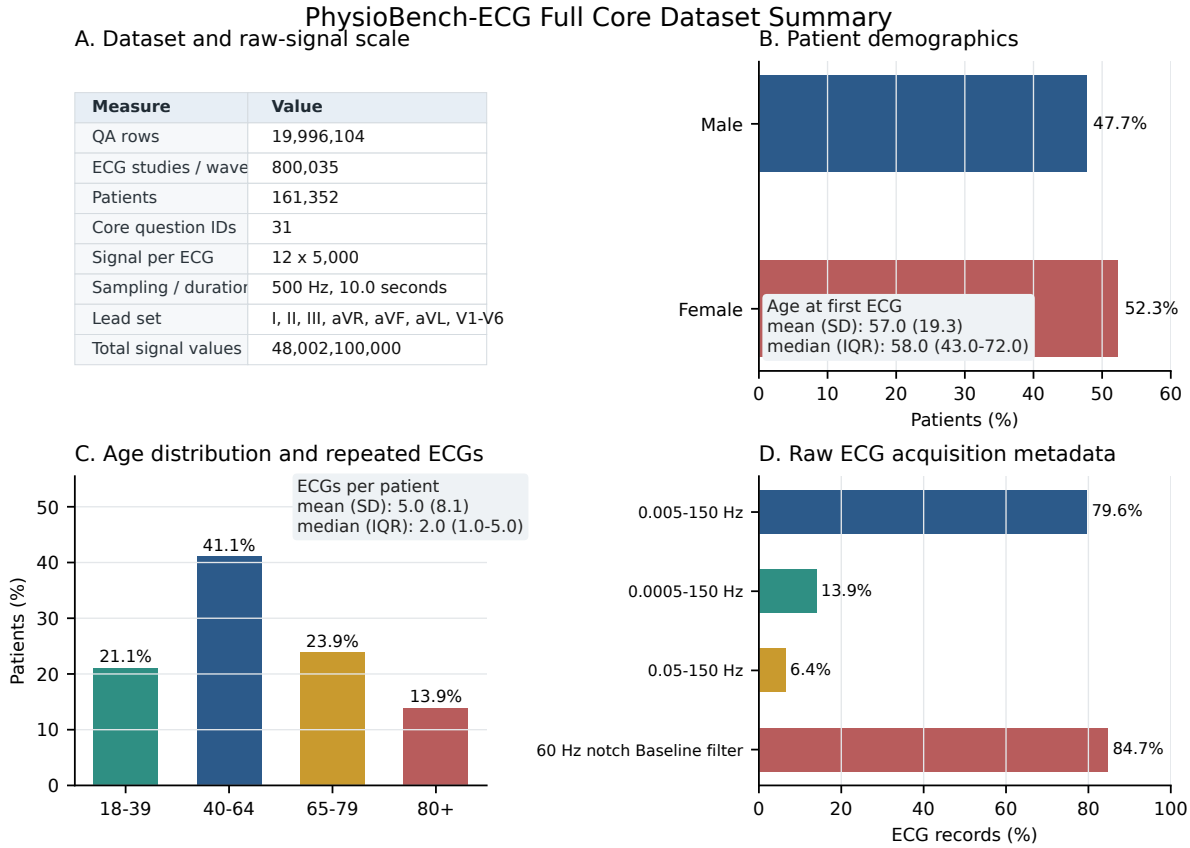

Figure 5: PhysioBench-ECG full core dataset summary, including cohort scale, patient demographics, repeated ECG coverage, and raw ECG signal properties.

Target	Primary source	Agreement	Downstream QA use
RR and heart rate	ECGdeli RR; waveform fallback when unavailable	WFDB/wqrs-j is retained as a reference comparator; confidence is downgraded when RR sources disagree or too few R peaks are detected.	Heart rate, rhythm regularity, RR periodicity, RR variability, temporal outlier beats.
P waves and PR interval	LUDB-calibrated P fiducials; ECGdeli PR as interval guardrail	Require organized P-wave evidence and stable P-to-QRS relation. Mark PR follow-up questions not applicable when organized P waves are absent or PR is physiologically undefined.	P-wave presence, P morphology, PR measurability, PR class, fixed PR $>$200 ms, P-to-QRS lag stability, AV-conduction screening.
QRS duration and morphology	ECGKit/Wavedet QRS duration	Use local waveform and MIMIC machine QRS as guardrails. Downgrade confidence when ECGKit conflicts with both secondary sources or when the value is near the 120 ms threshold.	QRS narrow/broad, QRS $<$120 ms, QRS-duration stability, bundle-branch morphology screening, axis rules.
QT and T-wave timing	Median of valid QT candidates from ECGKit, ECGdeli, local waveform, and MIMIC machine measurements	Use multi-method dispersion as a confidence signal. Downgrade when QT candidates disagree or T-offset evidence is weak.	QT $<$440 ms, QTc evidence, T/R amplitude ratio, ST/T repolarization guardrails.
ST, T, and Q morphology	Fiducial-anchored lead-wise waveform measurements	Use beat-specific baselines and relevant lead groups. Downgrade when wide QRS, pacing, hypertrophy, bundle-branch morphology, or baseline instability confounds interpretation.	ST deviation, localized ST elevation, pathologic Q-wave evidence, T-wave abnormality, T/R amplitude ratio.

Table 7: Target-specific evidence sources used for PhysioBench answer construction. The table summarizes how waveform-derived measurements are selected, checked, and converted into QA evidence.

B Confidence Assignment and Cross-Method Agreement Rationale

PhysioBench confidence labels are construction-time estimates of label reliability, not model-predicted probabilities. They are based on four factors: external LUDB validation, cross-method agreement, the number of analyzable observations, and clinical-rule applicability.

Cross-method agreement. For a target type t, let $x_{t,m}$ denote the estimate produced by method m. Let $m_t^\star$ be the primary method selected for target t based on LUDB validation, and let $\mathcal{M}_t$ be the set of available secondary methods. We define cross-method agreement as

$$A_t = \mathbf{1}\left[\min_{m \in \mathcal{M}_t} \left|x_{t,m_t^\star} - x_{t,m}\right| \le \tau_t\right],$$

where τ_t is a target-specific tolerance. For numerical intervals, τ_t is measured in milliseconds. For categorical labels derived from intervals, such as QRS narrow/broad or PR normal/prolonged, agreement is evaluated after mapping each method's interval estimate to the same clinical class.

For threshold-derived categorical labels, we additionally require margin from the clinical decision boundary. If $\hat{x}_t$ is the selected interval estimate and c_t is the clinical cutoff, then high-confidence labels require

$$|\hat{x}_t - c_t| > q_{t,0.95},$$

where $q_{t,0.95}$ is the 95th-percentile absolute error for target t estimated from LUDB. This prevents borderline cases from being treated as equally reliable as values far from the cutoff.

Empirical calibration. Let C_t denote the event that the generated label for target t is correct relative to the LUDB reference. Let $B_{t,k}$ denote the event that at least k analyzable beats, intervals, or repeated observations are available. For each target and evidence rule, we estimate

$$\hat{p}_{t,k} = P(C_t = 1 \mid A_t = 1, B_{t,k} = 1)$$

on LUDB. High-confidence labels require high empirical correctness under this conditional rule. Medium-confidence labels remain answerable but fail at least one strict agreement, repeated-evidence, or margin criterion. Low-confidence labels have weak support and are reserved for robustness analysis. Rows are marked `not_assessable` when the required fiducials or physiology are unavailable, and `not_applicable` when the question is physiologically invalid for that ECG.

Why agreement improves reliability. If two methods have correctness probabilities p_1 and p_2 for the same binary clinical label, and their errors are independent, then the probability that an agreed label is correct is

$$P(\text{correct} \mid \text{agree}) = \frac{p_1 p_2}{p_1 p_2 + (1-p_1)(1-p_2)}.$$

This expression shows why agreement between two accurate methods can be more reliable than either method alone. For example, the LUDB-calibrated P-wave detector achieves macro P-point F1 of 96.82%, PPV of 96.89%, and MAE of 16.8 ms, while ECGdeli PR provides an interval-level guardrail with PR MAE of 10.1 ms. Under the independence approximation, agreement between two sources with correctness near these levels would imply an agreed-label reliability substantially higher than either individual source. However, ECG algorithms can make correlated errors on difficult waveforms. Therefore, PhysioBench does not treat this expression as a guarantee; it is used as motivation for the agreement rule, while the operative confidence labels are calibrated empirically through $\hat{p}_{t,k}$.

Why repeated beats improve reliability. The requirement for multiple analyzable beats has a similar motivation. If a beat-level detector has correctness probability p and a record-level decision is made by majority or median aggregation across k beats, then the probability that the aggregated decision is correct under independent beat-level errors is

$$P_{\text{maj}}(k,p) = \sum_{j=\lceil k/2 \rceil}^{k} \binom{k}{j} p^j (1-p)^{k-j}.$$

For $k = 5$, this majority-aggregation probability is above 99% when p is in the range observed for the LUDB-calibrated fiducial detectors in Table 8. This motivates the repeated-evidence requirement used for high-confidence QAs. As with method agreement, this calculation is an idealized rationale rather than a standalone proof, because beat-level errors may be correlated in noisy or pathological ECGs. The final confidence categories are therefore validated empirically using LUDB and applied with target-specific guardrails.

Target-specific use. The agreement rule is applied differently across targets. RR and heart-rate labels require sufficient R peaks or RR intervals and agreement among RR sources when

available. P-wave and PR labels require organized P-wave evidence, measurable P-to-QRS lag, and PR agreement or guardrails; when organized P waves are absent, downstream PR questions are marked not applicable. QRS-duration labels use ECGKit/Wavedet as the primary source and downgrade confidence when local waveform or MIMIC machine evidence conflicts. QT labels use a median over valid ECGKit, ECGdeli, local waveform, and MIMIC machine candidates, with downgrades when candidate dispersion is large. ST/T/Q morphology labels use fiducial-anchored lead-wise measurements and are downgraded when wide QRS, pacing, hypertrophy, bundle-branch morphology, or unstable baseline confounds interpretation.

C Implementation Details for the Main Benchmark

Evaluation scope. The main benchmark table evaluates a mix of local text LLMs and API-based language or vision-language models on a 1500 sample split of the dataset. Rows are scored with Canonical Accuracy: each model response is parsed as JSON, the `answer` field is regex parsed and normalized according to the answer type, then compared with the dataset answer. API-based Gemini-2.5-Flash and GPT-4.1-mini are evaluated separately under the same answer-normalization protocol.

Evaluated model set. The table includes nine local instruction-tuned language models: Gemma-3-4B-IT, Gemma-3-12B-IT, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Mistral-Small-3.1-24B-Instruct, MedGemma-4B-IT, Llama3-Med42-8B, and BioMistral-7B-Zephyr-SLERP.

Input setting: question-specific evidence plus rule. In the QUESTION-SPECIFIC EVIDENCE setting, each prompt contains the natural-language question, the answer type, allowed answers when applicable, the machine-extracted evidence fields needed for that question, and the corresponding `derivation_method` rule, see figure /6.

Compute environment. Each array task uses one H100 GPU, eight CPU cores, and 120 GB of memory.

Acknowledgments

AI Usage Acknowledgement. The authors used ChatGPT for language editing, wording suggestions, and LaTeX formatting assistance.

Wave	Method	Point	F1	Se	PPV	MAE (ms)
P	Amp./edge-gated P filter	P onset	96.83	96.76	96.91	17.4
P	Amp./edge-gated P filter	P peak	96.81	96.77	96.86	16.1
P	Amp./edge-gated P filter	P offset	96.81	96.73	96.90	17.0
P	Amp./edge-gated P filter	Macro	96.82	96.75	96.89	16.8
QRS	U-Net3+	QRS onset	95.85	98.76	93.11	5.6
QRS	U-Net3+	QRS peak	95.85	98.76	93.11	3.8
QRS	U-Net3+	QRS offset	95.85	98.76	93.11	6.2
QRS	U-Net3+	Macro	95.85	98.76	93.11	5.2
T	U-Net3+	T onset	94.89	96.69	93.16	13.2
T	U-Net3+	T peak	94.84	96.64	93.11	5.6
T	U-Net3+	T offset	94.74	96.54	93.01	10.4
T	U-Net3+	Macro	94.83	96.62	93.09	9.7

Table 8: Point-level fiducial detection performance on LUDB. F1, sensitivity (Se), positive predictive value (PPV), and mean absolute timing error (MAE) are reported for the main PhysioBench detectors; U-Net3+ rows are measured by record-held-out LUDB cross-validation.

Method	PR MAE (ms)	QRS MAE (ms)	QT MAE (ms)	RR MAE (ms)
ECGdeli	10.1	25.4	20.2	10.4
ECGKit/Wavedet	16.3	8.8	13.5	72.2
NeuroKit DWT	39.5	59.6	36.5	9.3
ECGdeli based mimic denoised	9.4	35.2	14.1	34.2
Waveform fallback	13.1	9.4	24.9	34.2
WFDB/wqrs-j	–	14.5	–	6.1

Table 9: Record-level interval evaluation on LUDB. Each cell reports mean absolute error against LUDB cardiologist-derived record medians; dashes indicate intervals not produced by that method. These scores guide the source selection summarized in Table 7.

Prompt A: Full Waveform + Rule.

```
System:
You are an expert in ECG interpretation.
You answer ECG time-series interpretation questions.
Use only the provided raw ECG signal and guideline.
Return only one valid JSON object with exactly one key, "answer".

User:
Question:
{question}

Guideline:
{guideline}

Read the ECG signal below:
{raw_time_series}

Use the guideline as the criterion for deciding the answer.
If the required feature cannot be determined from the raw ECG signal, answer
"Not assessable".
Return exactly JSON like {"answer": "..."}. Do not explain.
```

Prompt B: Question-Specific Evidence + Rule.

```
System:
You are an expert in ECG interpretation.
You answer ECG time-series interpretation questions from structured evidence.
Use only the provided evidence and guideline.
Return only one valid JSON object with exactly one key, "answer", without any explanation.

User:
Question:
{question}

Guideline:
{guideline}

Read the evidence below:
{evidence}

Use the guideline as the criterion for deciding the answer.
If allowed_answers are provided, choose exactly one allowed answer.
If the required evidence is missing or invalid, answer "Not assessable".
Return exactly JSON like {"answer": "..."}. Do not explain.
```

Prompt C: 12-Lead ECG Plot + Rule.

```
System:
You are an expert in ECG interpretation.
You answer ECG interpretation questions from a standard all-12-lead ECG image.
Use only the provided ECG image and guideline.
Return only one valid JSON object with exactly one key, "answer".

User:
Question:
{question}

Guideline:
{guideline}

Use the guideline as the criterion for deciding the answer from the ECG image.
If allowed_answers are provided, choose exactly one allowed answer.
If the required feature is not visible enough in the image, answer
"Not assessable".
Return exactly JSON like {"answer": "..."}. Do not explain.
```

Figure 6: Prompt templates used for PhysioBench-ECG evaluation. Prompt A tests whether models can recover relevant information from serialized raw ECG time series. Prompt B tests rule application when question-specific structured evidence is provided. Prompt C tests visual ECG interpretation from a rendered 12-lead ECG plot.