Peak-R1: Instruction-Tuned Large Language Models for Robust J-Peak Detection in Cardiomechanical Signals

Jiahui Li[1], **Yida Zhang**[1], **Zixuan Zeng**[1], **Jiayu Chen**[1],
Xiang Zhang[2], **Jin Lu**[1], **Wenzhan Song**[1], **Fei Dou**[1]

[1]University of Georgia
[2]University of North Carolina at Charlotte

{jl57095, Yida.Zhang, zxzeng, jiayu.chen, jin.lu, wsong, fei.dou}@uga.edu
xiang.zhang@charlotte.edu

Abstract

Accurate peak detection across diverse cardiacmechanical signals, including the Ballistocardiogram (BCG), and Bodyseismography (BSG), is fundamental for cardiovascular monitoring but is often hindered by artifacts and signal variability. Conventional algorithms are typically engineered with expert knowledge for a single signal modality, limiting their generalizability. Conversely, deep learning-based methods often lack interpretability, raising concerns about their clinical trustworthiness and hindering expert-computer interaction. To address these limitations, we introduce Peak-R1, a novel framework that leverages instruction-tuned Large Language Models (LLMs) for robust, cross-modal, and explainable peak detection. A core innovation of our framework is a "peak-representation" technique that transforms time-series data into a condensed format, preserving critical event information while significantly reducing signal length. This representation provides a crucial inductive bias, guiding the LLM to reason over physiologically meaningful events rather than raw, noisy data. The model is optimized through a two-stage process: supervised fine-tuning (SFT) followed by reinforcement learning (RL) with a multi-objective reward function. The model's self-explanation capabilities are cultivated by fine-tuning on a custom-built Peak-Explanation dataset. Across four modalities—BCG and BSG—spanning seven datasets (six public benchmarks plus one real-world cohort), Peak-R1 demonstrates consistently excellent performance, achieving best or tied-best detection under clinically relevant temporal tolerance. Beyond accuracy, the generated rationales surface failure modes and support human-in-the-loop annotation. Together, these results indicate a single, generalizable, and interpretable solution to the complex challenge of peak detection across multiple physiological signals.

1 Introduction

Continuous cardiovascular monitoring is fundamental to the paradigm of proactive and personalized healthcare, enabling the early detection and management of cardiac conditions in daily life. Unobtrusive sensing modalities, particularly Ballistocardiogram (BCG), and Bodyseismography (BSG), have become cornerstone technologies for the longitudinal assessment of an individual's health status [1, 2, 3, 4, 5]. The ability of these sensors to facilitate timely identification of cardiac

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Learning from Time Series for Health.

arrhythmias—such as Premature Atrial Contractions (PACs), Premature Ventricular Contractions (PVCs), and Atrial Fibrillation (AFib)—is critical for preventing adverse health outcomes. However, the efficacy of these monitoring systems hinges on the accurate and robust detection of key prominent points, or "peaks," within the collected physiological data.

The primary challenge in this domain lies in the inherent heterogeneity of these signal modalities. Each captures a different facet of the cardiac cycle through distinct physical principles, resulting in unique signal morphologies (Fig 3). The BCG provides a contactless method by measuring the body's subtle vibrations from cardiac ejections, characterized by a prominent J-peak. While promising for unobtrusive settings like beds or chairs, BCG signals are highly susceptible to motion artifacts, posing significant analytical challenges [6, 7, 8]. The BSG delivers richer information content by effectively containing features of both BCG (whole-body movement) and SCG (direct precordial vibrations) [3, 9, 10]. This diversity in signal origin and quality necessitates a versatile and resilient approach to peak detection [11, 12].

Existing peak detection methodologies, however, fall short of providing a universally effective solution. Traditional signal-processing approaches rely on modality-specific domain knowledge, employing handcrafted features and heuristics that are often brittle and require meticulous parameter tuning [13, 14, 15]. Consequently, these methods lack generalizability across different signal types. While deep learning models offer greater adaptability by learning features directly from data, they are often criticized as "black boxes." [16, 17, 18, 19]. This lack of transparency and interpretability creates a significant barrier to trust and adoption in critical clinical applications, where understanding the model's reasoning is as important as its output.

To address these limitations, we explore a new paradigm that leverages the advanced reasoning and human-style interaction capabilities of Large Language Models (LLMs) [20, 21]. While recent work has applied LLMs to conceptual tasks in physiology, such as report generation [22, 23, 24]. their application to precise numeric inference—like peak localization—is hindered by a fundamental challenge: performance degradation when processing long, continuous numerical sequences [25]. We posit that the key to unlocking LLM potential in this area is not to process the raw signal directly, but to first transform it. We introduce a novel Peak Representation, which converts lengthy, sparse physiological signals into condensed, information-rich symbolic sequences. This abstraction retains the essential temporal and morphological information necessary for peak detection while making the data more amenable to LLM processing.

Building on this representation, we introduce Peak-R1, an instruction-tuned LLM designed for accurate, cross-modal, and explainable peak detection. To achieve this, we first constructed a Peak-Explanation Dataset through a novel data generation pipeline, specifically to enhance the model's self-explanatory capabilities. We then trained Peak-R1 using a two-stage instruction-tuning strategy. The first stage, Supervised Fine-Tuning (SFT), establishes a reliable and correctly formatted output structure. The second stage employs Reinforcement Learning (RL) with Group Relative Policy Optimization (GRPO), optimizing the model against a multi-objective reward function designed to concurrently enhance format validity, heart-rate consistency, positional accuracy, and detection completeness.

In summary, this work makes the following contributions:

- We introduce a modality-agnostic **Peak Representation** that summarizes physiological signals as timestamped local extrema with signal value, enabling token-efficient, auditable reasoning. It effectively compresses BCG data by 89%, respectively, while maintaining high fidelity for signal reconstruction with a correlation coefficient of 0.97.

- We develop **Peak-R1**, A two-stage SFT→RL pipeline with a multi-objective reward jointly optimizes detection accuracy, temporal consistency, and concise, structured rationales, which is supported by a scalable data-generation pipeline that builds a self-explainable peak-detection dataset.

- Across BCG/BSG on three public datasets and one real-world cohort, Peak-R1 achieves best-or-tied-best detection under clinically relevant tolerance with competitive HR/HRV errors, while providing step-by-step rationales.

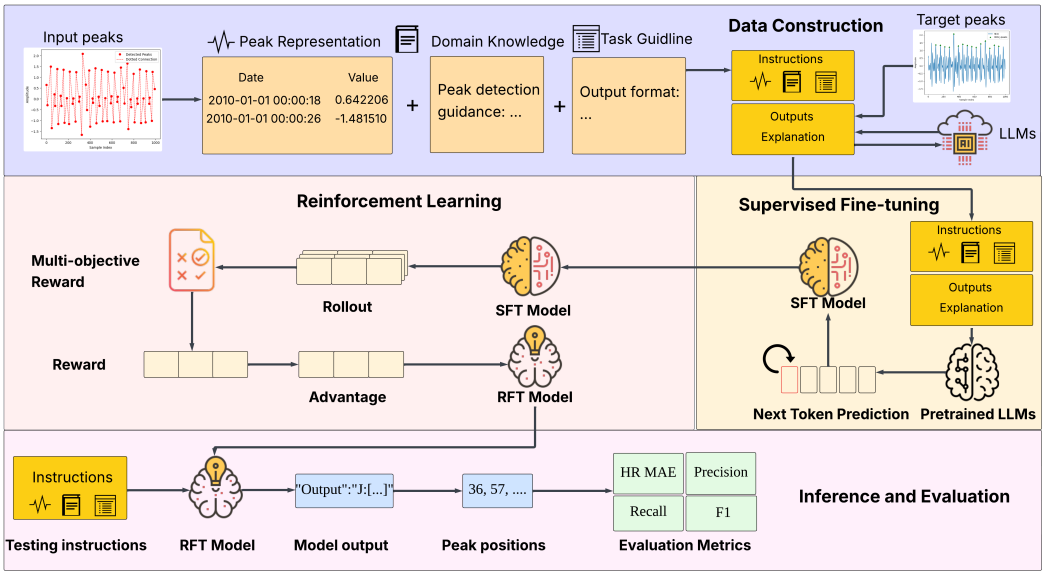


Figure 1: Peak-R1 Framework

2 Method

This section details the design and implementation of Peak-R1, our proposed framework for accurate and explainable cross-modal peak detection, the framework is shown in Fig 1. The methodology is presented in a structured progression. We begin by introducing the foundational component: The Peak Representation, which transforms dense, numerical physiological signals into a compact, text-based format amenable to LLM processing. To train a model capable of interpreting this representation, we then detail the construction of the Peak-Explanation Dataset, a novel corpus generated via a semi-automated pipeline that pairs signal data with machine-generated expert explanations. With the data and representation established, we describe the core Peak-R1 architecture. Finally, we elucidate our Two-Stage Instruction Tuning Strategy, a hybrid approach that leverages Supervised Fine-Tuning (SFT) to establish robust formatting and explanatory capabilities, followed by Reinforcement Learning (RL) to directly optimize for peak detection accuracy and physiological consistency.

2.1 Peak Representation: From Signal to Sequence

Our approach assumes that the most informative content in a physiological time-series is primarily encoded around **local extrema** that delineate each cardiac cycle. Intermediate samples can be reconstructed via interpolation between these key points, as illustrated in Fig. 2. This observation motivates transforming a dense numeric signal into a compact, structured sequence that an LLM can reason over. The proposed *Peak Representation* operationalizes this idea and serves as the input to *Peak-R1*.

Step 1: Inclusive Local-Extrema Extraction. We perform a **lenient local-extrema search** [26] on each preprocessed segment $x = \{x_t\}_{t=1}^{T}$ to construct a comprehensive candidate set of peaks $\mathcal{C} = \{c_i\}$ by collecting a high-recall set of both maxima and minima. Each candidate c_i stores its temporal index t_i and normalized amplitude a_i. This "inclusive-first" strategy ensures that all potential fiducial points are preserved for downstream reasoning.

Step 2: Timestamp Transformation. A key point of our representation is encoding each peak's temporal position as a human-readable timestamp rather than a raw numeric index. Motivated by evidence that LLMs excel at calendrical and timestamp reasoning compared to long numeric sequences [25], each sample index t_i is converted into a high-resolution **synthetic calendar timestamp**:

$$T_i(t_i) = \text{YYYY-MM-DD HH:MM:SS}$$

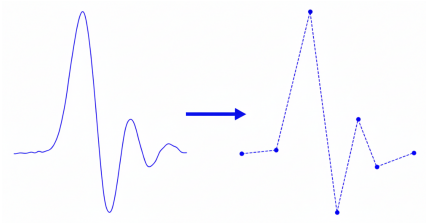


Figure 2: Signal approximation by interpolating peaks.

A fixed reference time $T_0 =$ `2020-01-01 00:00:00` is used, and absolute indices are computed across segments to prevent window-local ambiguities. The index t_i is transformed into elapsed seconds by dividing by the sampling frequency F_s (i.e., calendar-second $= t_i/F_s$), then formatted into the "HH:MM:SS" string. For example, at $F_s = 1$ Hz, a peak at $t_i = 97$ corresponds to 97 s $\rightarrow$ "00:01:37", which is appended to the reference date to yield the final timestamp. Timestamps are reset at the beginning of each segment to maintain bounded token lengths while allowing the model to reason about relative temporal intervals.

Step 3: Serialization into a Compact Sequence. Each detected peak is represented as a key–value pair and serialized into a structured textual sequence. All entries are arranged chronologically to preserve temporal coherence, forming a compact and information-dense representation well suited for LLM processing.

2.2 Peak-Explanation Dataset Construction

To train a model for both accurate peak detection and transparent reasoning, we constructed the Peak-Explanation Dataset. Each sample in this dataset is a structured triplet designed for instruction-tuning, comprising: (1) an Instruction providing signal context (e.g., modality, sampling frequency) and defining the dual task of peak identification and explanation; (2) the Input, a textual sequence of candidate peaks from our Peak Representation method; and (3) the Output, a ground-truth response containing the correct peak list and a human-readable explanation. For Details, see Appendix B.

3 Experiment Setup and Evaluation

In this section, we describe the datasets, evaluation metrics, hyperparameter selection, performance of Peak-R1 compared to baseline models. We also evaluate the application of Peak-R1 in arrhythmaia detection task.

3.1 Experimental Setup

Datasets and Preprocessing. We evaluated Peak-R1 on seven diverse physiological datasets covering BCG/BSG (Kansas [27], BCG Arrhythmia [28], and a self-collected ICU dataset). For datasets lacking ground truth, a semi-automatic annotation process was employed (see Appendix G). To ensure consistency, all signals were segmented into 1000-sample windows, processed via a fourth-order Butterworth bandpass filter (0.6–15,Hz) to remove noise, and z-score normalized.**Baselines and Metrics.**We benchmarked performance against nine state-of-the-art methods (detailed in Appendix H), including heuristic signal-processing algorithms (e.g., Pan-Tompkins [29] Elgendi [30], and Pino [31]) and deep learning models (CNN-SWT [32], 1D-UNet++ [33], FR-Net [34]). Evaluation metrics included Precision, Recall, and F1-score, using a strict $\pm$30,ms tolerance window for True Positives [35, 36]. Furthermore, to assess clinical utility, we calculated the Mean Absolute Error (MAE) for derived Heart Rate (HR) and Heart Rate Variability (HRV).

3.2 Results

Table 1 presents a comprehensive evaluation of Peak-R1 against established baselines across six physiological datasets. The results illustrate a clear hierarchy of performance: traditional signal-processing algorithms struggle with generalizability outside their target modalities, while deep learning models (e.g., FR-Net, 1D-UNet++) offer improved robustness but lack interpretability.

Table 1: Peak Detection Performance Comparison across Signal Modalities and Baselines. Lower values are better for MAE (HR(s), HRV(ms)), higher values are better for F1, Pre, Rec. Best performance in each metric for each dataset is highlighted in bold.

		Metric	Pan-T	Nabian	Elg	Bishop	Pino	Choi	CNN-SWT	1D-UNet++	FR-Net	Peak-R1
BCG	Kansas	HR MAE	29.12	7.79	24.74	153.17	3.77	6.08	4.21	5.23	4.04	**2.36**
		HRV MAE	161.73	71.28	128.19	94.22	70.84	83.02	80.98	91.42	75.92	**44.32**
		F1	0.0501	0.9033	0.7006	0.4073	0.8398	0.8273	0.9156	0.9103	0.9409	**0.9543**
		Pre	0.0403	0.8517	0.5779	0.2567	0.7856	0.7865	0.9064	0.9088	0.9594	0.9541
		Rec	0.0663	0.9617	0.8893	0.9865	0.9021	0.8725	0.9250	0.9119	0.9230	0.9544
	Arrhy	HR MAE	41.34	4.34	19.37	147.81	2.66	3.72	3.08	3.34	4.60	**1.40**
		HRV MAE	145.96	47.44	138.01	55.67	69.30	58.49	79.21	78.32	113.18	**37.05**
		F1	0.2092	0.8945	0.7986	0.4629	0.9250	0.8876	0.9435	0.9463	0.9418	**0.9476**
		Pre	0.1720	0.9209	0.7293	0.3073	0.9280	0.9030	0.9199	0.9699	0.9729	0.9408
		Rec	0.2671	0.8696	0.8824	0.9380	0.9221	0.8728	0.9682	0.9237	0.9120	0.9545
BSG	ICU	HR MAE	44.96	**6.74**	14.06	160.55	8.06	7.16	7.68	7.96	12.23	8.17
		HRV MAE	143.32	95.68	132.05	235.05	105.31	100.10	91.40	127.64	124.00	**83.67**
		F1	0.1241	0.8127	0.0908	0.4145	0.6688	0.6850	0.8005	0.7842	0.7910	**0.8332**
		Pre	0.0965	0.8054	0.0801	0.2657	0.6534	0.6773	0.7578	0.8866	0.8822	0.7835
		Rec	0.1739	0.8201	0.1047	0.9420	0.6850	0.6929	0.8482	0.7031	0.7169	0.8897

In contrast, Peak-R1 achieves highly competitive and often state-of-the-art (SOTA) performance across all evaluated domains. It secures achieves SOTA metrics on both BCG and BSG datasets, notably handling the complex artifacts inherent to Ballistocardiography. These results validate the efficacy of our LLM-based approach, demonstrating that transforming peak detection into a language-reasoning task yields superior accuracy and generalization compared to conventional methods.

4 Conclusion

In this paper, We introduced Peak-R1, a framework that redefines cardiac peak detection by transforming signal analysis into a language-based reasoning task. Through our novel Peak Representation and instruction-tuning strategy, the model achieves state-of-the-art performance and robust generalization across three datasets (BCG, and BSG). Unlike traditional "black-box" methods, Peak-R1 combines numerical precision with interpretable, text-based reasoning, establishing a new standard for transparent and trustworthy clinical AI.

References

[1] Jeremy W Gordon. Certain molar movements of the human body produced by the circulation of the blood. *Journal of anatomy and physiology*, 11(Pt 3):533, 1877.

[2] Chang-Sei Kim, Stephanie L Ober, M Sean McMurtry, Barry A Finegan, Omer T Inan, Ramakrishna Mukkamala, and Jin-Oh Hahn. Ballistocardiogram: Mechanism and potential for unobtrusive cardiovascular health monitoring. *Scientific reports*, 6(1):31297, 2016.

[3] Yingjian Song, Bingnan Li, Dan Luo, Glenna S Brewster Glasgow, Bradley G Phillips, Yuan Ke, and Wenzhan Song. Real-time continuous blood pressure estimation with contact-free bedseismogram. In *ICC 2024-IEEE International Conference on Communications*, pages 214–219. IEEE, 2024.

[4] Thomas Penzel, Jan W Kantelhardt, Ronny P Bartsch, Maik Riedl, Jan F Kraemer, Niels Wessel, Carmen Garcia, Martin Glos, Ingo Fietze, and Christoph Schöbel. Modulations of heart rate, ecg, and cardio-respiratory coupling observed in polysomnography. *Frontiers in physiology*, 7:460, 2016.

[5] Andriy Temko. Accurate heart rate monitoring during physical exercises using ppg. *IEEE Transactions on Biomedical Engineering*, 64(9):2016–2024, 2017.

[6] Yong-Xian Li, Jiong-Ling Huang, Xin-Yu Yao, Si-Qi Mu, Shou-Xin Zong, and Yan-Fei Shen. A ballistocardiogram dataset with reference sensor signals in long-term natural sleep environments. *Scientific Data*, 11(1):1091, 2024.

[7] Jianwei Su, Xuezhou Zhu, Xiaodong Zhang, Jintian Tang, and Lei Liu. Ballistocardiogram measurement system using three load-cell sensors platform in chair. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–4. IEEE, 2009.

[8] S Azhaginiyan, M Anish, Menon K Shivaranjan, and M Ganesan. Denoising of bcg signal using multi resolution analysis. In *2019 5th International conference on advanced computing & communication systems (ICACCS)*, pages 1005–1008. IEEE, 2019.

[9] Zaid Farooq Pitafi, Yingjian Song, Zaipeng Xie, Benjamin Brainard, and Wenzhan Song. Contactless vital signs monitoring for animals. *IEEE Internet of Things Journal*, 2025.

[10] Yingjian Song, Haotian Xiang, Zixuan Zeng, Jiayu Chen, Yida Zhang, Zaid Farooq Pitafi, He Yang, Qin Lu, Xiang Zhang, Bradley G Phillips, et al. Multi-granularity supervised contrastive learning with online adaptation for contactless in-bed posture classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(2):1–32, 2025.

[11] Guy JJ Warmerdam, Rik Vullings, Lars Schmitt, Judith OEH Van Laar, and Jan WM Bergmans. Hierarchical probabilistic framework for fetal r-peak detection, using ecg waveform and heart rate information. *IEEE Transactions on Signal Processing*, 66(16):4388–4397, 2018.

[12] Parry Fung, Guy Dumont, Craig Ries, Chris Mott, and Mark Ansermino. Continuous noninvasive blood pressure measurement by pulse transit time. In *The 26th annual international conference of the IEEE engineering in medicine and biology society*, volume 1, pages 738–741. IEEE, 2004.

[13] Abhishek Chakraborty, Deboleena Sadhukhan, and Madhuchhanda Mitra. A robust ppg onset and systolic peak detection algorithm based on hilbert transform. In *2020 IEEE Calcutta Conference (CALCON)*, pages 176–180. IEEE, 2020.

[14] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one*, 8(10):e76585, 2013.

[15] Srinivas Kuntamalla and L Ram Gopal Reddy. An efficient and automatic systolic peak detection algorithm for photoplethysmographic signals. *International Journal of Computer Applications*, 97(19), 2014.

[16] Kianoosh Kazemi, Juho Laitala, Iman Azimi, Pasi Liljeberg, and Amir M Rahmani. Robust ppg peak detection using dilated convolutional neural networks. *Sensors*, 22(16):6054, 2022.

[17] Pritam Sarkar and Ali Etemad. Cardiogan: Attentive generative adversarial network with dual discriminators for synthesis of ecg from ppg. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 488–496, 2021.

[18] Hanhui Xu and Kyle Michael James Shuttleworth. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intelligent Medicine*, 4(1):52–57, 2024.

[19] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[20] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[21] Zhenhua Wang, Guang Xu, and Ming Ren. Llm-generated natural language meets scaling laws: New explorations and data augmentation methods. *arXiv preprint arXiv:2407.00322*, 2024.

[22] Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, et al. Opentslm: Time-series language models for reasoning over multivariate medical text-and time-series data. *arXiv preprint arXiv:2510.02410*, 2025.

[23] Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. *arXiv preprint arXiv:2408.08849*, 2024.

[24] Jinning Yang and Wen Shi. Diagecg: An llm-driven framework for diagnostic reasoning via discretized ecg tokenization. *arXiv preprint arXiv:2508.15338*, 2025.

[25] Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. *arXiv preprint arXiv:2404.16563*, 2024.

[26] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[27] Charles Carlson, Vanessa-Rose Turpin, Ahmad Suliman, Carl Ade, Steve Warren, and David E Thompson. Bed-based ballistocardiography: Dataset and ability to track cardiovascular parameters. *Sensors*, 21(1):156, 2020.

[28] Jing Zhan, Zhengying Li, Xiaoyan Wu, Chao Zhang, Tao Zhao, Kewei Chen, and Zhibing Lu. A multi-pathology ballistocardiogram dataset for cardiac function monitoring and arrhythmia assessment. *Scientific Data*, 12(1):963, 2025.

[29] MAZ Fariha, Ryojun Ikeura, Soichiro Hayakawa, and Shigeyoshi Tsutsumi. Analysis of pan-tompkins algorithm performance with noisy ecg signals. In *Journal of Physics: Conference Series*, volume 1532, page 012022. IOP Publishing, 2020.

[30] Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.

[31] Esteban J Pino, Javier AP Chávez, and Pablo Aqueveque. Bcg algorithm for unobtrusive heart rate monitoring. In *2017 IEEE healthcare innovations and point of care technologies (HI-poct)*, pages 180–183. IEEE, 2017.

[32] Donghwan Yun, Hyung-Chul Lee, Chul-Woo Jung, Soonil Kwon, So-Ryoung Lee, Kwangsoo Kim, Yon Su Kim, and Seung Seok Han. Robust r-peak detection in an electrocardiogram with stationary wavelet transformation and separable convolution. *Scientific Reports*, 12(1):19638, 2022.

[33] Tengda Zhou, Shaoyang Men, Jingxian Liang, Baoxian Yu, Han Zhang, and Xiaomu Luo. 1d u-net++: an effective method for ballistocardiogram j-peak detection. *Journal of Mechanics in Medicine and Biology*, 21(10):2140058, 2021.

[34] Zhenqin Chen, Kaixiao Zheng, Junhua Shen, Yiwei Lin, Yan Feng, and Jinshan Xu. Sample point classification of abdominal ecg through cnn-transformer model enables efficient fetal heart rate detection. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2023.

[35] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[36] Chunsheng Zuo, Yu Zhao, and Juntao Ye. Tau: Modeling temporal consistency through temporal attentive u-net for ppg peak detection. *arXiv preprint arXiv:2503.10733*, 2025.

[37] C Brüser, Stefan Winter, and Steffen Leonhardt. Robust inter-beat interval estimation in cardiac vibration signals. *Physiological measurement*, 34(2):123, 2013.

[38] JH Shin, BH Choi, Yong Gyu Lim, DU Jeong, and KS Park. Automatic ballistocardiogram (bcg) beat detection using a template matching approach. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1144–1146. IEEE, 2008.

[39] Alaleh Alivar, Charles Carlson, Ahmad Suliman, Steve Warren, Punit Prakash, David E Thompson, and Balasubramaniam Natarajan. Motion artifact detection and reduction in bed-based ballistocardiogram. *Ieee Access*, 7:13693–13703, 2019.

[40] Yingjian Song, Bingnan Li, Dan Luo, Zaipeng Xie, Bradley G Phillips, Yuan Ke, and Wenzhan Song. Engagement-free and contactless bed occupancy and vital signs monitoring. *IEEE internet of things journal*, 11(5):7935–7947, 2023.

[41] Christoph Schranz, Christina Halmich, Sebastian Mayr, and Dominik PJ Heib. Surrogate modelling of heartbeat events for improved j-peak detection in bcg using deep learning. *Frontiers in Network Physiology*, 4:1425871, 2024.

[42] Zhenqin Chen, Mengying Wang, Meiyu Zhang, Wei Huang, Hanjie Gu, and Jinshan Xu. Post-processing refined ecg delineation based on 1d-unet. *Biomedical Signal Processing and Control*, 79:104106, 2023.

[43] Sricharan Vijayarangan, R Vignesh, Balamurali Murugesan, SP Preejith, Jayaraj Joseph, and Mohansankar Sivaprakasam. Rpnet: A deep learning approach for robust r peak detection in noisy ecg. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 345–348. IEEE, 2020.

[44] Kai Yang, Massimo Hong, Jiahuan Zhang, Yizhen Luo, Suyuan Zhao, Ou Zhang, Xiaomao Yu, Jiawen Zhou, Liuqing Yang, Ping Zhang, et al. Ecg-lm: Understanding electrocardiogram with a large language model. *Health Data Science*, 5:0221, 2025.

[45] Mikhail L Arbuzov, Alexey A Shvets, and Sisong Beir. Beyond exponential decay: Rethinking error accumulation in large language models. *arXiv preprint arXiv:2505.24187*, 2025.

[46] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

[47] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[48] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[49] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

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

[51] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[52] Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen S Quigley, Lisa F Barrett, and Sarah Ostadabbas. An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE journal of translational engineering in health and medicine*, 6:1–11, 2018.

[53] Steven M Bishop and Ari Ercole. Multi-scale peak and trough detection optimised for periodic and quasi-periodic neuroscience data. In *Intracranial Pressure & Neuromonitoring XVI*, pages 189–195. Springer, 2018.

[54] Byung Hun Choi, Gih Sung Chung, Jin-Seong Lee, Do-Un Jeong, and Kwang Suk Park. Slow-wave sleep estimation on a load-cell-installed bed: a non-constrained method. *Physiological measurement*, 30(11):1163, 2009.

A Related Work

A.1 Peak Detection in Cardiac Physiological Signals

The detection of fiducial points in cardiac physiological signals has been predominantly approached through two distinct paradigms: traditional signal-processing techniques and, more recently, deep learning models.

Signal-Processing Approaches. Classical methods for peak detection are characterized by their reliance on modality-specific domain knowledge and handcrafted heuristics. BCG J-peak detection presents unique challenges due to noise from motion artifacts. Consequently, methods often rely on template-matching or autocorrelation to identify a standard cardiac cycle. Bruser et al. estimate the cardiac interval length using a combination of autocorrelation, magnitude differences, and amplitude pairs [37], while Shin et al. identify IJK complexes by correlating the signal with a constructed template [38]. To mitigate noise, some methods explicitly detect and reconstruct motion-corrupted segments by modeling signal distributions [39]. For BSG signals, existing methods identify heart rate by locating the peak with maximum amplitude in the autocorrelation function (ACF) of the signal [40].

The principal limitation of these signal-processing approaches is their inherent specificity and lack of generalizability. They depend heavily on expert-driven parameter tuning and are often brittle, failing when signal quality degrades or when applied to a new modality without significant redesign. This fragility motivates our objective of developing a robust, cross-modal solution.

Deep Learning Approaches. To overcome the adaptability limitations of classical methods, deep learning models have been widely adopted. Architectures based on Convolutional Neural Networks (CNNs) have been successfully applied to peak detection across BCG and BSG signals [41, 16, 42]. These models typically reframe peak detection as a classification problem, where each time-step or signal segment is classified as either containing a peak or not. By learning salient features directly from data, models such as RPNet have demonstrated superior performance and noise robustness compared to their signal-processing counterparts [43]. However, despite their accuracy, these models suffer from a significant drawback: their "black-box" nature. This lack of transparency and interpretability poses a major barrier to clinical adoption, as physicians cannot verify the model's underlying reasoning. This critique establishes a clear need for our proposed explainable LLM.

A.2 Large Language Models in Physiological Time-Series Analysis

The application of Large Language Models (LLMs) to physiological time-series analysis is an emerging field, with current research primarily focused on high-level semantic interpretation. For example, OpenTSLM was proposed to perform diverse tasks such as time-series captioning and question-answering for sleep staging [22]. Similarly, ECG-LM was developed to diagnose arrhythmia types from a raw ECG signal [44]. While powerful, these models treat the physiological signal holistically and are not designed for precise numeric inference—that is, localizing specific events like an arrhythmia or a prominent peak at a precise time index.

This "numeric grounding" gap stems from a core architectural challenge: the input length limitation of auto-regressive models. As these models generate tokens sequentially, errors can accumulate, causing the model to "drift out of distribution" during prolonged generation tasks [45]. Research evaluating LLM performance on time-series feature understanding has confirmed this, finding that numeric retrieval accuracy consistently degrades as the input length increases [25]. This fundamental constraint inspired our approach to first distill the essential information within a physiological signal into a condensed Peak Representation, thereby shortening the input sequence while preserving its core informational content.

A.3 Explainable AI (XAI) for Clinical Decision Support

Explainable AI (XAI) has become a critical area of research in the medical domain, driven by ethical and practical imperatives. As patients have a right to an explanation for automated decisions affecting their care, clinicians require an understanding of *why* a model makes a certain prediction before they can confidently integrate it into their workflow [46]. Current XAI methods are often post-hoc, seeking to explain a model's decision by attributing importance to input features [47, 48]. Techniques like Gradient SHAP (GS), for instance, explain predictions by computing the gradients of outputs with respect to points along an interpolation path from a baseline reference to the input [49].

However, these post-hoc explanations can be complex to interpret and may not fully capture the model's internal logic. To bridge this gap, we propose to develop a model with inherent, self-explanatory capabilities. By leveraging a custom Peak-Explanation Dataset and instruction-tuning, our approach provides a more direct and intuitive solution to the explainability problem. This design

positions Peak-R1 not merely as an accurate tool, but as a trustworthy and transparent partner in clinical decision support.

B Details of Dataset Construction

This dataset was built using a semi-automated pipeline. First, raw signals are preprocessed and converted into the textual Input sequence as shown in section 2.1. To generate the corresponding Output, we leverage a powerful "teacher" LLM (e.g., GPT-4o). As illustrated in Figure 5, we provide this teacher model with the instruction, the candidate peaks, and the final list of ground-truth peaks. The model is then prompted to articulate a clear, logical rationale for filtering the candidates to arrive at the ground-truth set. This machine-generated explanation is then combined with the ground-truth peaks to form the complete, high-quality output label used for training.

To effectively train Peak-R1 for both robust performance and explainability across diverse physiological signals, we employ a sophisticated two-stage instruction tuning strategy. This approach synergistically combines Supervised Fine-Tuning (SFT) with Reinforcement Learning (RL) optimization, leveraging the strengths of each paradigm. We utilize a pre-trained open-source LLM (specifically `Qwen2.5` for our experiments) as the base model, fine-tuning it with our meticulously constructed Peak-Explanation Dataset. For Details, please see Appendix C.

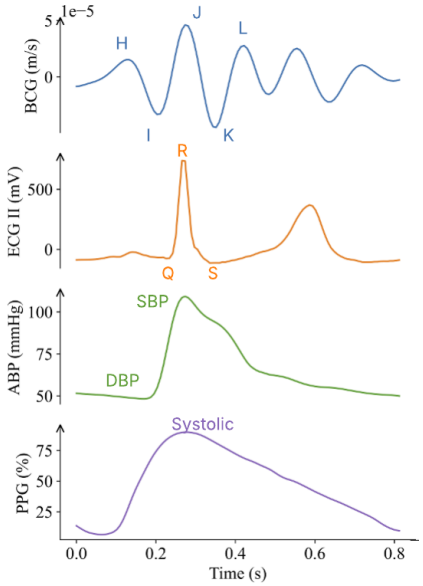


Figure 3: Physiology signals with labelled peaks

C Details of Two-Stage Instruction Tuning Strategy

Stage 1: Supervised Fine-Tuning, is foundational for adapting the base model to the specific domain of physiological peak detection and ensuring a consistent output format. A schematic of the SFT pipeline is shown in Figure 4a. During SFT, the model is trained on the instruction-response pairs from our Peak-Explanation Dataset. The LLM is trained in an auto-regressive fashion, performing next-token prediction [50] where, for a given context, it generates a probability distribution over its entire vocabulary for the subsequent token.

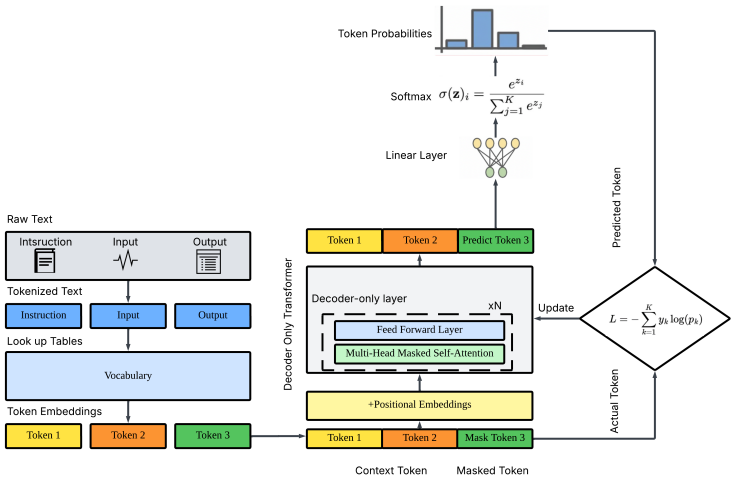
The training objective is to maximize the likelihood of generating the ground-truth output sequence by minimizing the token-level cross-entropy loss at each step. The loss for a single token prediction is defined as:

$$L = -\sum_{k=1}^{K} y_k \log(p_k) \tag{1}$$

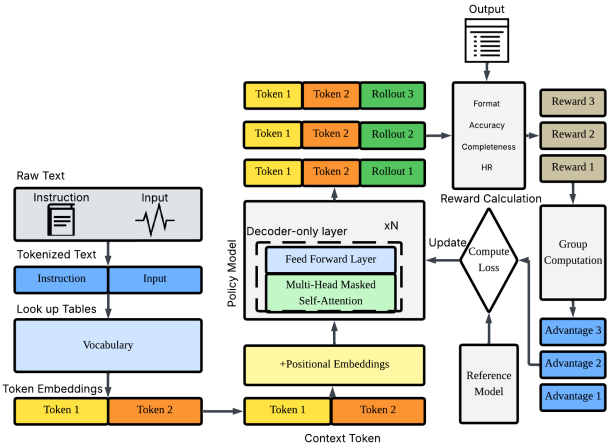
where K is the size of the vocabulary, y_k is a binary indicator (1 if the k-th token is the true next token, and 0 otherwise), and p_k is the model's predicted probability for the k-th token. By minimizing this loss function across the training corpus, the model's parameters are adjusted to assign higher probabilities to the ground-truth token sequences.

SFT serves three primary objectives: (1) it *stabilizes the learning dynamics* of the LLM, adapting its general language understanding to the specific domain of physiological signal analysis; (2) it *enforces a consistent and valid output format* (i.e., JSON structure with specified fields and explanatory text), which is critical for the downstream utility of Peak-R1; and (3) it *equips the model with foundational peak detection capabilities* by learning directly from the expert-annotated ground truth.

Stage 2: Reinforcement Learning Optimization. Following SFT, the model undergoes a second stage of optimization using Reinforcement Learning (RL) with Group Relative Policy Optimization



(a) SFT illustration



(b) GRPO illustration

Figure 4: Training strategies comparison

(GRPO) [51]. Unlike SFT, which optimizes at the token level, GRPO refines the model's policy at the *sequence level*, directly aligning the generated outputs with desired performance metrics. The core principle of GRPO is to sample multiple responses for the same prompt and compute an advantage score for each response based on its relative quality within that group. This allows the model to learn from comparative feedback, generating responses that better align with quantitative criteria, as illustrated in Fig 4b.

The objective of GRPO is to maximize the following function:

$$\begin{aligned}\mathcal{J}_{GRPO}(\theta) =& \mathbb{E}\left[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)\right] \\ & \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[r_t(\theta)\hat{A}_{i,t}, \text{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}\end{aligned} \tag{2}$$

where $r_t(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$ is the probability ratio between the new and old policies, $\hat{A}_{i,t}$ is the estimated advantage based on the relative rewards within the group, and ε is a clipping hyperparameter that constrains the policy update step. The objective function seeks to increase the probability of actions that yield a higher advantage, while the KL-divergence term, weighted by β, regularizes the policy to prevent it from deviating too far from a reference policy, thus stabilizing training.

In our implementation, the advantage $\hat{A}_{i,t}$ is derived from a custom, multi-objective reward function R_{total}, which guides the LLM towards high-fidelity peak detection and physiologically plausible outputs. This reward function is a weighted sum of four distinct components:

$$R_{\text{total}} = 0.1 \cdot R_{\text{format}} + 0.6 \cdot R_{\text{detection}} + 0.15 \cdot R_{\text{complete}} + 0.15 \cdot R_{\text{HR}} \tag{3}$$

The components are defined as follows: The Format Compliance Reward (R_{format}) is a binary reward ensuring the output adheres to the specified JSON syntax, defined as $R_{\text{format}} = \begin{cases} 1, & \text{if valid format} \\ 0, & \text{otherwise} \end{cases}$. The Detection Accuracy Reward ($R_{\text{detection}}$) is quantified by the F_1-score between the predicted and ground-truth peaks, using a strict 30 ms tolerance window. The Count Completeness Reward (R_{complete}) incentivizes the prediction of the correct total number of peaks, formulated as an exponential penalty for count errors, $R_{\text{complete}} = -|N_{\text{pred}} - N_{\text{gt}}|$. Finally, the Heart-Rate Consistency Reward (R_{HR}) assesses physiological plausibility by rewarding outputs where the derived heart rate closely matches the ground truth, defined as $R_{\text{HR}} = \exp\left(-2 \cdot \frac{|HR_{\text{pred}} - HR_{\text{gt}}|}{HR_{\text{gt}}}\right)$. By optimizing for this composite, sequence-level reward, the GRPO stage refines the model beyond the token-matching of SFT, enhancing its ability to generate responses that are quantitatively accurate and physiologically sound.

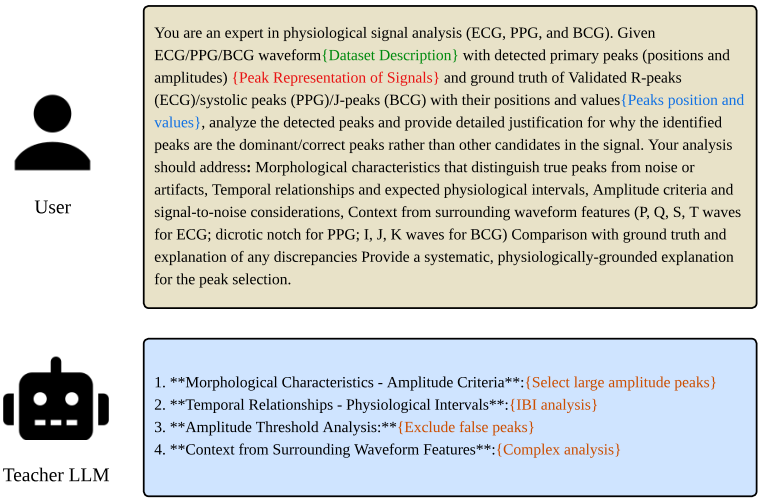


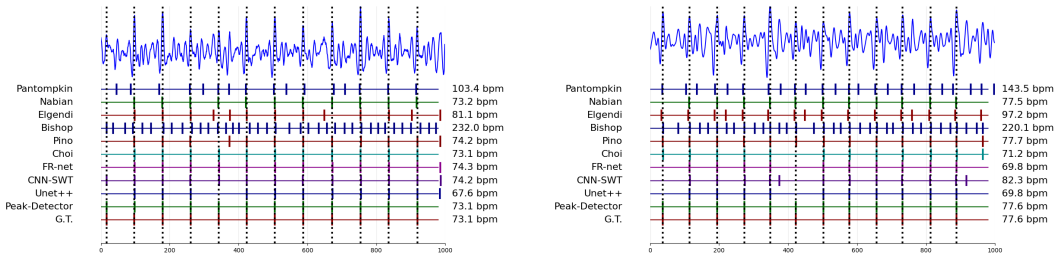
Figure 5: An example of the structured prompt provided to the teacher LLM (top) and the corresponding structured JSON output, including the step-by-step explanation (bottom), used to construct the Peak-Explanation Dataset.

D Interpretation and Explainability

Qualitative Analysis of Peak Detection. To provide a qualitative assessment of performance in challenging scenarios, Figure 6 illustrates the comparative detection results on representative segments of BCG, and BSG signals. This visual analysis highlights the practical failure modes of baseline algorithms and underscores the robustness of our proposed method.

Then, the BCG arrhythmia example (Figure 6a) highlights the difficulty of analyzing signals with irregular timing and morphology. Traditional algorithms, relying on fixed assumptions about rhythm and shape, struggle significantly; Pan-Tompkins and Elgendi exhibit considerable over-detection, while Bishop and Pino show incorrect localization. In contrast, data-driven approaches, and particularly Peak-R1, successfully identify the correct J-peak positions even amidst arrhythmia and reduced signal clarity. These qualitative observations across diverse modalities reinforce the quantitative findings, visually confirming that Peak-R1's contextual reasoning provides superior robustness against morphological anomalies, noise, and rhythm disturbances.

Figure 6: Qualitative visualization of peak detection performance across challenging segments of (a) BCG with arrhythmia (b) BSG in ICU.



(a) BCG J-Peak Detection

(b) BSG J-Peak Detection

Finally, the BSG ICU example (Figure 6b) illustrates performance on a highly challenging BSG signal, characterized by the elevated noise levels typical of an ICU environment. As observed, traditional algorithms are prone to over-detection, while deep learning models tend towards under-detection in this scenario. Peak-R1, however, correctly identifies all ground-truth peaks, demonstrating that its contextual reasoning is highly effective at distinguishing true cardiac events from artifacts, even under conditions of both high noise and arrhythmic variability.

E Explanation Evaluation

Evaluating the quality of machine-generated explanations is a nuanced challenge. To provide a holistic assessment, we designed a multi-faceted evaluation framework that assesses the generated explanations across five key dimensions. The results of this evaluation, summarized in the radar plot in Figure 7, demonstrate the high quality and reliability of Peak-R1's explanatory capabilities.

The evaluation framework is comprised of the following dimensions. **Faithfulness** measures whether the explanation's claims correspond to the signal's quantitative features, assessed by programmatically verifying assertions against the input data. **Robustness** evaluates the stability of the explanation's factual correctness under controlled input perturbations (clipping, noise addition, and masking). On both of these dimensions, Peak-R1 achieved a score of 98.1, indicating that its reasoning is not only strictly grounded in the provided data but also stable and reliable in the face of minor signal variations.

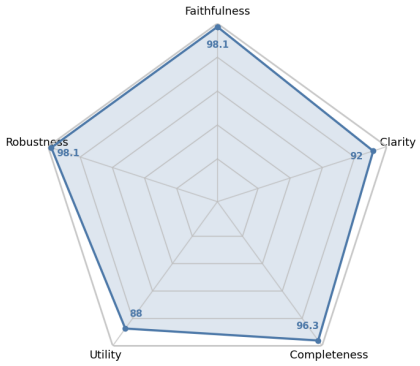


Figure 7: Radar Plot of Explanation Evaluation

Clarity and **Utility** were assessed through user studies to measure the intelligibility and practical helpfulness of the explanations, respectively. With high scores of 88 for both dimensions, the results confirm that the explanations are easily understood by both experts and non-experts and are practically useful for increasing an expert's confidence in the model's predictions. Finally, **Completeness** was quantified by checking for the presence of four key analytical dimensions (morphological, temporal, amplitude, and contextual). The model achieved a score of 96.3, highlighting its strong instruction-following capabilities and its ability to generate comprehensive, multi-faceted rationales. Collectively, these strong results across all five dimensions validate the high quality of Peak-R1's generated explanations.

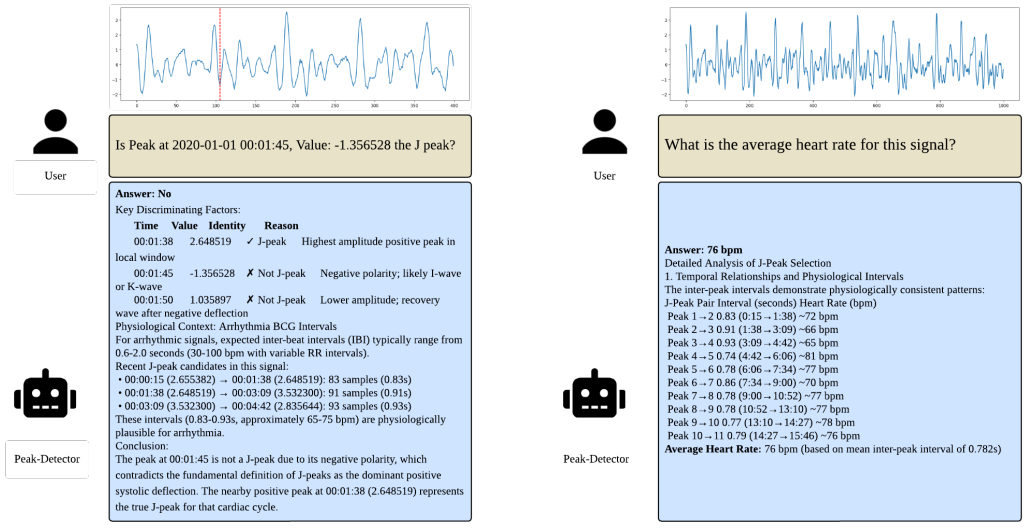
Analysis of Generated Explanations. Table 2 presents a side-by-side comparison between an explanation generated by the "teacher" LLM and one produced by our fine-tuned Peak-R1, revealing several key insights into our model's interpretability. The analysis confirms that Peak-R1 demonstrates a strong adherence to the predefined explanatory format, consistently integrating multiple analytical dimensions into its reasoning. These dimensions include morphological characteristics (peak shape), temporal relationships (physiological intervals), signal quality criteria (signal-to-noise ratio and

Table 2: Comparison of J-Peak Detection Results and Explanations from Teacher LLM and Fine-tuned Peak-R1

Peak-R1 Generated Output	**Teacher LLM Generated Output**
J: [2020-01-01 00:00:17, 2020-01-01 00:01:51, 2020-01-01 00:03:22, 2020-01-01 00:04:57, 2020-01-01 00:06:34, 2020-01-01 00:08:10, 2020-01-01 00:09:47, 2020-01-01 00:11:25, 2020-01-01 00:13:06, 2020-01-01 00:14:44, 2020-01-01 00:16:20]	J: [2020-01-01 00:00:17, 2020-01-01 00:01:51, 2020-01-01 00:03:22, 2020-01-01 00:04:57, 2020-01-01 00:06:34, 2020-01-01 00:08:10, 2020-01-01 00:09:47, 2020-01-01 00:11:25, 2020-01-01 00:13:06, 2020-01-01 00:14:44, 2020-01-01 00:16:20]
Explanation: **Step 1: Morphological Characteristics** The J-peak is defined as the **largest headward (positive) wave** occurring late in systole. Examining the detected peaks, I identify candidates with prominent positive amplitudes: • 2020-01-01 00:00:17: **2.915030** ✓ • 2020-01-01 00:01:51: **2.786782** ✓ • 2020-01-01 00:03:22: **2.842781** ✓ • 2020-01-01 00:04:57: **3.200422** ✓ (highest amplitude) • 2020-01-01 00:06:34: **2.703152** ✓ • 2020-01-01 00:08:10: **3.031469** ✓ • 2020-01-01 00:09:47: **2.775080** ✓ • 2020-01-01 00:11:25: **2.718626** ✓ • 2020-01-01 00:13:06: **2.302004** ✓ • 2020-01-01 00:14:44: **2.987384** ✓ • 2020-01-01 00:16:20: **3.164307** ✓ These peaks exhibit amplitudes >2.0, significantly higher than other positive deflections (<1.5), indicating they represent true cardiac events rather than noise. **Step 2: Temporal Relationships and Physiological Intervals** **Inter-beat intervals (RR-equivalents):** • 00:00:17 → 00:01:51: **94 seconds** (0.94s) • 00:01:51 → 00:03:22: **91 seconds** (0.91s) • 00:03:22 → 00:04:57: **95 seconds** (0.95s) • 00:04:57 → 00:06:34: **97 seconds** (0.97s) • 00:06:34 → 00:08:10: **96 seconds** (0.96s) • 00:08:10 → 00:09:47: **97 seconds** (0.97s) • 00:09:47 → 00:11:25: **98 seconds** (0.98s) • 00:11:25 → 00:13:06: **101 seconds** (1.01s) • 00:13:06 → 00:14:44: **98 seconds** (0.98s) • 00:14:44 → 00:16:20: **96 seconds** (0.96s) Mean interval: ∼0.97s (Heart rate: ∼62 bpm) This indicates bradycardia with arrhythmia, consistent with the signal description. The variability in intervals (91–101 seconds) confirms arrhythmic pattern. **Step 3: Signal-to-Noise Ratio and Amplitude Criteria** Amplitude threshold analysis: • True J-peaks: 2.915–3.200 (SNR excellent) • Other positive peaks: 0.096–1.950 (likely I-waves, K-waves, or noise) • Negative deflections: -0.191 to -2.100 (I-waves and K-waves) The 2:1 amplitude ratio between true J-peaks and other features provides clear discrimination. **Step 4: Waveform Context (I-J-K Complex)** Each identified J-peak follows the characteristic pattern: • **Preceding I-wave** (negative deflection): e.g., -1.943758 before 2.915030 • **J-peak** (dominant positive): The selected peaks • **Following K-wave** (negative deflection): e.g., -1.597640 after 2.786782 Example I-J-K complex (00:04:57 J-peak): • I-wave: 00:04:49 (-1.795014) • J-peak: 00:04:57 (3.200422) • K-wave: 00:05:06 (-1.629486) **Step 5: Exclusion of False Positives** Why smaller positive peaks are NOT J-peaks: • 00:00:49 (1.614529): Too small, occurs too soon after 00:00:17 (31s interval = 194 bpm, physiologically impossible) • 00:02:21 (1.366664): Low amplitude, inconsistent timing • 00:02:55 (1.562412): Only 54s after previous J-peak, too frequent • 00:07:52 (0.997150): Below amplitude threshold, irregular pattern Why negative peaks are NOT J-peaks: • All negative deflections represent I-waves or K-waves, part of the I-J-K complex	**Explanation:** **Step 1: Morphological Characteristics** The selected J-peaks represent the largest positive (headward) deflections in the BCG signal, consistent with the physiological definition of the J-wave as the most prominent systolic component: Peak Time: 00:00:17, Amplitude: 2.915030 ✓J-peak – Dominant positive peak in first cardiac cycle Peak Time: 00:01:51, Amplitude: 2.786782 ✓J-peak – Highest amplitude in surrounding 10s window Peak Time: 00:03:22, Amplitude: 2.842781 ✓J-peak – Clear dominant positive deflection Peak Time: 00:04:57, Amplitude: 3.200422 ✓J-peak – Maximum amplitude in entire signal Peak Time: 00:06:34, Amplitude: 2.703152 ✓J-peak – Prominent systolic wave Peak Time: 00:08:10, Amplitude: 3.031469 ✓J-peak – Second highest amplitude overall Peak Time: 00:09:47, Amplitude: 2.775080 ✓J-peak – Dominant in cycle Peak Time: 00:11:25, Amplitude: 2.718626 ✓J-peak – Clear J-wave morphology Peak Time: 00:13:06, Amplitude: 2.302004 ✓J-peak – Highest in irregular segment Peak Time: 00:14:44, Amplitude: 2.987384 ✓J-peak – Strong systolic component Peak Time: 00:16:20, Amplitude: 3.164307 ✓J-peak – Dominant terminal peak Rejected candidates (smaller positive peaks): 00:00:49 (1.614529), 00:01:22 (0.887839), 00:02:21 (1.366664) – All significantly smaller than nearby dominant peaks. These represent either I-waves, K-waves, or noise artifacts. **Step2: Temporal Relationships - Physiological RR Intervals** Analysis of inter-peak intervals reveals arrhythmic pattern consistent with the signal description: Interval Peak 1→2: Duration 94s, Heart Rate 64 bpm – Bradycardic/irregular Interval Peak 2→3: Duration 91s, Heart Rate 66 bpm – Irregular rhythm Interval Peak 3→4: Duration 95s, Heart Rate 63 bpm – Consistent with arrhythmia Interval Peak 4→5: Duration 97s, Heart Rate 62 bpm – Variable intervals Interval Peak 5→6: Duration 96s, Heart Rate 63 bpm – Arrhythmic pattern Interval Peak 6→7: Duration 97s, Heart Rate 62 bpm – Irregular Interval Peak 7→8: Duration 98s, Heart Rate 61 bpm – Bradyarrhythmia Interval Peak 8→9: Duration 100s, Heart Rate 60 bpm – Extended interval Interval Peak 9→10: Duration 98s, Heart Rate 61 bpm – Variable Interval Peak 10→11: Duration 96s, Heart Rate 63 bpm – Irregular rhythm Key observations: Mean interval: ∼96.2s (approximately 62 bpm) Coefficient of variation: Low, but intervals show arrhythmic irregularity No intervals <40s or >120s, ruling out extreme tachycardia/bradycardia Pattern consistent with sinus arrhythmia or atrial fibrillation with controlled ventricular response **Step3: Signal-to-Noise Ratio and Amplitude Criteria** Amplitude threshold analysis: Selected J-peaks: 2.30 to 3.20 (mean: 2.84) Rejected positive peaks: 0.06 to 1.64 (mean: 0.68) Clear bimodal distribution with >1.5 amplitude separation Signal-to-noise ratio: ∼4:1 for true J-peaks vs. artifacts Negative deflections (I-waves and K-waves): Range: -0.19 to -2.10 Properly excluded as J-peak is defined as positive/headward deflection These represent I-waves (preceding J) and K-waves (following J) **Step4: Contextual Waveform Features - IJK Complex** Each identified J-peak shows characteristic IJK morphology: Example cycle (00:04:57 J-peak): I-wave: 00:04:49 (-1.795014) – Negative deflection preceding J J-peak: 00:04:57 (3.200422) – Dominant positive peak ✓ K-wave: 00:05:06 (-1.629486) – Negative deflection following J Pattern verification across all cycles: 1. Each J-peak is preceded by negative I-wave (0.5–1.0s before) 2. Each J-peak is followed by negative K-wave (0.5–1.5s after) 3. Smaller positive deflections between cycles represent: L-waves (late diastolic oscillations) M-waves (atrial components) Respiratory artifacts

amplitude), and waveform context. Furthermore, the model's textual rationale is demonstrably faithful to the underlying signal data, as the provided justification directly supports the final peak selections, showing a strong alignment between its analysis and prediction. Finally, a notable trade-off between conciseness and verbosity is observed; the explanation from Peak-R1 is more succinct than that of the larger teacher model, while still effectively communicating the core logic for its decisions.

Emergent Analytical Capabilities. A significant advantage of leveraging a pre-trained LLM is its potential to generalize its learned knowledge beyond the explicit fine-tuning task. To explore this, we evaluated Peak-R1 on two zero-shot tasks not present in the training data: fine-grained peak classification and procedural heart rate calculation. The results, shown in Figure 8, demonstrate that our training cultivates a model with a foundational understanding of the underlying signal, rather than just a narrow peak detection function.



(a) Zero-Shot Peak Classification

(b) Zero-Shot HR Calculation

Figure 8: Demonstration of Peak-R1's emergent analytical capabilities on zero-shot tasks. (a) The model correctly classifies a non-J-peak and hypothesizes its identity. (b) The model executes a multi-step procedure to accurately calculate heart rate from the detected peaks.

As shown in Figure 8a, when queried to classify a specific point in the waveform, the model correctly identifies it as a non-J-peak. Critically, it goes further by hypothesizing that the point may correspond to a K-wave. This demonstrates a latent understanding of the complete IJK complex morphology, a concept it was never explicitly taught to classify. Furthermore, Figure 8b showcases the model's capacity for multi-step procedural reasoning. When tasked with calculating the heart rate, Peak-R1 spontaneously executes a correct analytical workflow: it first identifies all J-peaks, then calculates the individual peak-to-peak intervals, and finally averages these intervals to derive the mean heart rate.

These emergent capabilities highlight that our framework does not merely produce a static peak detector. Instead, it cultivates an interactive analytical tool with a deeper, more flexible understanding of cardiac signals, showcasing the broader potential for LLMs to serve as versatile partners in physiological data analysis.

F Analysis

In this section, we details the analysis on model ablations, comparison to LLMs predictions, and computational analysis.

Table 3: Ablation study results on the BCG Arrhythmia dataset, comparing precision, recall, F1-score, HR MAE (BPM), and HRV MAE under a 50 ms matching tolerance. Best performance per block is highlighted in bold. "-" indicates no valid result could be obtained.

	BCG Arrhythmia Dataset				
Method	Precision	Recall	F1-score	HR MAE	HRV MAE
Peak-R1 (Full Framework)	**0.9408**	**0.9545**	**0.9476**	**1.39**	**36.44**
Peak-R1 (SFT-only)	0.9408	0.9545	0.9476	1.40	37.05
Peak-R1 (w/o Peak Representation)	0.8073	0.8862	0.8449	7.87	129.22
Peak-R1 (w/o Explanation)	0.9525	0.9417	0.9471	1.85	50.86
Comparison with General-Purpose LLMs					
qwen2.5-3B-instruct (Base Model)	0.2662	0.1262	0.1712	32.22	893.27
Claude-Sonnet-4.5	0.7389	0.9579	0.8343	23.01	144.37
GPT-5	0.4981	0.9847	0.6615	66.21	125.05
Gemini-2.5-Pro	0.6833	0.9446	0.7929	29.09	154.98

F.1 Ablation Study

To dissect the individual contributions of each component within the Peak-R1 framework, we conducted a systematic ablation study on BCG Arrhythmia dataset, with the results detailed in Table 3. Our analysis begins with the baseline performance of a general-purpose, instruction-tuned LLM (`qwen2.5-3B-instruct`), which yields a very low F1-score of 0.1712, confirming that standard LLMs are ill-suited for this precise numerical inference task without specialized adaptation. The most significant performance gain is realized through Supervised Fine-Tuning (SFT) on our Peak-Explanation Dataset, which catapults the F1-score to 0.9476, demonstrating the effectiveness of SFT in teaching the model the fundamental task from our textual representation. The indispensable role of the Peak Representation is also starkly evident; removing this component causes the F1-score to plummet to 0.8449 and the HR MAE to increase dramatically to 7.87, underscoring the critical failure of LLMs to effectively process long, raw numerical sequences.

Further analysis reveals the more nuanced contributions of the remaining components. The inclusion of the explanation generation task during SFT provides a marginal yet meaningful improvement; removing it ('w/o Explanation') leads to a slight decrease in F1-score (0.9471) but a more noticeable degradation in physiological consistency, as the HR MAE increases from 1.40 to 1.85. This suggests that compelling the model to articulate its reasoning fosters the learning of more robust, physiologically grounded features. Similarly, the final RL optimization stage serves to refine the model's policy. While it does not change the F1-score, its primary contribution is enhancing physiological plausibility by reducing the HR MAE from 1.40 to 1.39, indicating that the reward function guides the model to select peak locations that better align with a stable cardiac rhythm. This specialized framework's profound advantage is highlighted when compared against several powerful, closed-source LLMs like Claude, GPT, and Gemini, which exhibit substantially lower F1-scores and orders-of-magnitude-higher HR MAEs. This result unequivocally demonstrates that our domain-specific approach—combining a novel signal representation with a targeted, two-stage tuning strategy—is essential for achieving expert-level performance in this challenging task.

F.2 Computational Complexity Analysis

To evaluate the computational feasibility and practical viability of our approach, we conducted a complexity analysis comparing Peak-R1 with all baseline models. Table 4 details key performance indicators including model parameter counts, training throughput, and inference throughput. Throughput is quantified in signal segments per second (SPS), with performance evaluated on input segments of 1000 data points each.

The analysis reveals a clear performance hierarchy. As expected, traditional signal-processing algorithms and specialized deep learning models exhibit the highest inference throughput, with speeds often exceeding thousands of SPS. This efficiency stems from their design as lightweight, highly optimized numerical processors. In contrast, general-purpose, closed-source LLMs operate at a significantly slower pace, with an inference throughput of approximately 0.07 SPS, rendering them impractical for all but offline, non-time-critical analysis.

Table 4: Computational complexity and throughput analysis for all evaluated models. Throughput is measured in segments per second (SPS). For traditional algorithms, which do not have a training phase, throughput reflects their processing speed during inference.

Model/Algorithm	Parameters	training(SPS)	Inference(SPS)
Large Language Models			
Peak-R1 (our model)	∼3 Billion	0.397	3.571
Claude-Sonnet-4.5	N/A	N/A	0.07
GPT-5	N/A	N/A	0.07
Gemini-2.5-pro	N/A	N/A	0.07
Deep Learning Models			
1D-UNet++	1,790,465	3,414	12,379
CNN-SWT	693,016	2,904	14,059
FR-Net	6,692,449	1,061	4,037
Signal Processing Algorithms			
Pino	N/A	N/A	2,110
Choi	N/A	N/A	5,278
Bishop	N/A	N/A	272
Elgendi	N/A	N/A	578
Nabian	N/A	N/A	624
Pan-Tompkins	N/A	N/A	1,550

Our proposed Peak-R1 occupies a crucial middle ground. With an inference throughput of 3.571 SPS, it is demonstrably slower than the numerical baselines. However, it is over 50 times faster than the general-purpose LLMs. This substantial speedup is directly attributable to our **Peak Representation**, which circumvents the primary computational bottleneck of LLMs by transforming long numerical sequences into condensed, information-rich text.

While Peak-R1 is not designed for beat-by-beat, low-latency streaming applications, its processing speed is well-suited for many practical clinical and remote monitoring scenarios. For instance, processing a standard 10-second segment can be completed in a timeframe suitable for asynchronous analysis, report generation, or batch processing of longer recordings. This capability for efficient, offline analysis—coupled with its high accuracy and unique explanatory power—positions Peak-R1 as a powerful and practical tool for real-world deployment.

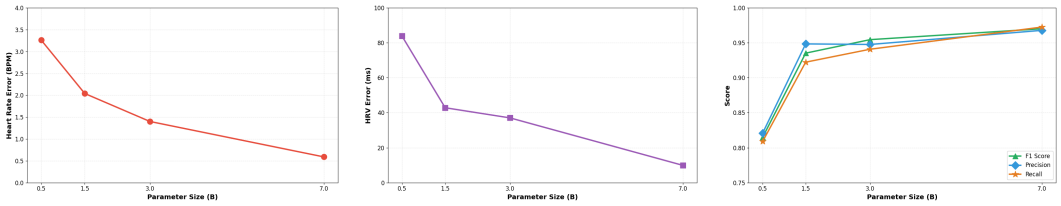
F.3 Impact of Model Scale

To investigate the relationship between model capacity and task performance, we conducted a scaling analysis by fine-tuning several variants of our base LLM architecture, with parameter counts ranging from 0.5 billion to 7 billion. The performance of each model variant was evaluated on the challenging BCG Arrhythmia Dataset.

As depicted in Figure 9, the results demonstrate a clear and consistent scaling law: performance across all metrics improves monotonically with the number of model parameters. Specifically, as the model size increased from 0.5B to 7B, the HR MAE decreased substantially from 3.26 to 0.59, and the HRV MAE dropped from 83.85 to 9.91. Concurrently, the F1-score rose from 0.8150 to 0.9701, driven by marked improvements in both precision (from 0.8210 to 0.9678) and recall (from 0.8091 to 0.9725).

This strong positive correlation suggests that larger models possess a greater capacity to learn the complex, non-linear patterns and subtle morphological cues inherent in arrhythmic physiological signals. The concurrent improvement in both detection accuracy (F1-score) and physiological consistency (HR/HRV MAE) indicates that increased model scale enhances not just pattern matching but also a deeper, more context-aware understanding of the underlying cardiac rhythm. This scaling trend suggests that the performance of the Peak-R1 framework is not yet saturated and could be further enhanced by leveraging even larger foundational models, affirming that the principles of scaling laws extend effectively to this specialized domain of time-series analysis.

Figure 9: The impact of model scaling on Peak-R1 performance, evaluated on the BCG Arrhythmia Dataset. As the base model size increases from 0.5B to 7B parameters, there is a consistent and significant improvement across all key metrics: (a) HR MAE and (b) HRV MAE decrease, while (c) F1-score, precision, and recall all increase. This demonstrates a strong positive scaling law for this task.



(a) HR MAE (b) HRV MAE (c) F1, Precision, and Recall

G Dataset Details

To thoroughly evaluate the cross-modal and explanatory capabilities of Peak-R1, we conducted experiments on a comprehensive suite of six publicly available physiological signal datasets. This collection includes two datasets each for BCG and BSG signals, ensuring broad coverage of signal modalities and varying data characteristics. A summary of these datasets is provided below:

- **Kansas Database** [27]: Developed by Kansas State University, this open-source dataset offers synchronized multimodal physiological signals, including BCG, ECG, PPG, and Arterial Blood Pressure (ABP) waveforms. BCG signals were captured using four electromechanical film (EMFi) sensors placed under the mattress and four load cells under the bed frame. Data were collected from 40 subjects (17 males, ages 18-65 years), with four subjects presenting cardiovascular-related conditions. The raw BCG signal is sampled at 1000 Hz; for our analysis, it was downsampled to 100 Hz to optimize for computational efficiency while retaining sufficient peak information within each segment. As with BIDMC, this dataset lacked pre-provided BCG J-peak labels, necessitating the adoption of a similar semi-automatic annotation procedure involving algorithmic detection and expert review.
- **BCG Arrhythmia Database** [28]: This dataset includes BCG recordings from 85 participants, encompassing individuals with sinus rhythm, heart failure (HF), and various cardiac arrhythmias such as Atrial Fibrillation (AF), Premature Ventricular Contractions (PVCs), and Premature Atrial Contractions (PACs). Signals are sampled at 100 Hz. BCG J-peaks in this database were labeled semi-automatically, combining Elgendi's algorithmic detection with subsequent manual verification by domain experts.
- **BSG ICU Database**: This clinical dataset was acquired from Yixing People's Hospital under an IRB-approved protocol with informed consent. It comprises 1,120 hours of continuous BSG and ECG recordings from 44 ICU patients. The cohort encompasses a wide demographic range, with patient ages spanning from 6 to 86 years (pediatric to geriatric). Both signals were sampled at a frequency of 100 Hz. The BCG J-peaks in this database were annotated using a rigorous semi-automatic protocol; this process involved initial algorithmic detection using Nabian's algorithm [52], followed by meticulous manual verification and correction by domain experts to ensure high-fidelity ground truth.

H Baseline Details

H.0.1 Signal-Processing Baselines

These methods leverage domain-specific heuristics and mathematical transformations tailored to individual signal modalities.

- **Pan-Tompkins** [29]: A widely adopted algorithm for ECG R-peak detection, it utilizes a series of signal processing steps including filtering, differentiation, squaring, and moving-

window integration to extract slope and energy information, followed by adaptive thresholding to identify R-peaks.

- **Nabian** [52]: This ECG R-peak detection method employs a sliding window technique. It identifies the maximum amplitude within each window, designating it as a potential R-peak, and subsequently refines these detections.
- **Elgendi** [30]: Designed for PPG systolic peak detection, this algorithm defines regions of interest by calculating two moving averages with distinct window sizes. Peaks are then identified as local maxima within these predefined regions.
- **Bishop** [53]: A multi-scale approach for PPG, it constructs a Local Maxima Scalogram (LMS) by analyzing the signal at various levels of smoothing. Peaks are then robustly identified by detecting common local maxima across these different scales.
- **Pino** [31]: This method focuses on BCG J-peak detection, employing techniques such as wavelet transformations, template matching, and signal envelopes to enhance and isolate the characteristic J-peak morphology within noisy BCG signals.
- **Choi** [54]: This BCG J-peak detection algorithm segments the signal based on an estimated mean heartbeat interval. It then identifies local maxima within each segment and eliminates false positives through analysis of peak-to-peak intervals, enhancing robustness to noise.

H.0.2 Deep Learning Baselines

These models represent advanced data-driven approaches, designed to learn complex features directly from the physiological signals.

- **CNN-SWT** [32]: Originally proposed for robust ECG R-peak detection, this model is a Convolutional Neural Network (CNN) architecture that leverages Stationary Wavelet Transform (SWT) coefficients as input, combined with separable convolutions to enhance feature extraction.
- **1D-UNet++** [33]: An extension of the U-Net architecture, this model utilizes nested dense skip connections between its encoder and decoder paths. Although initially designed for medical image segmentation, its 1D adaptation has demonstrated strong performance in time-series analysis, including BCG J-peak detection.
- **FR-Net** [34]: This is a specialized CNN-Transformer based network primarily developed for fetal R-peak detection in challenging ECG signals. Its architecture combines convolutional layers for local feature extraction with transformer blocks to capture long-range dependencies, making it suitable for complex peak detection tasks.

I Discussion

Implications of the Peak Representation Paradigm. The success of Peak-R1 is fundamentally rooted in the Peak Representation, which serves as a powerful abstraction layer between raw numerical data and the LLM. For peak detection, this transformation is intuitive, as the most salient information is concentrated at the peaks. However, its implications extend far beyond this specific task. We posit that this signal-to-sequence paradigm could be a transformative approach for a wide range of time-series analysis tasks where information is sparse or event-driven, such as anomaly detection, seizure detection in EEG, or sleep stage classification. By converting irregular time-series data into a symbolic, token-based format, this method naturally aligns with the architectural strengths of LLMs, potentially unlocking their sophisticated reasoning capabilities for problems previously dominated by specialized numerical models.

Interpretability and the Role of the Peak-Explanation Dataset. A key contribution of this work is the generation of explainable outputs, facilitated by our Peak-Explanation Dataset. By training on a corpus of machine-generated explanations, Peak-R1 learns to articulate a physiologically grounded rationale for its decisions. This represents a significant step towards building trustworthy AI systems for clinical applications. We acknowledge, however, that the fidelity of these explanations, while structurally sound, is currently benchmarked against a "teacher" LLM rather than a human expert. The nuances and deep clinical insights of a trained cardiologist may not be fully captured. This limitation also presents a clear direction for future work: the creation of a "gold-standard" explanation

dataset curated and validated by human experts. Such a dataset would not only enhance the model's explanatory power but also serve as a valuable resource for research into human-computer interaction in clinical decision support, with the LLM acting as an intelligent bridge.

Limitations and Future Directions. Despite its strong performance, Peak-R1 has several limitations that warrant discussion. First, as highlighted in our complexity analysis, its inference throughput, while vastly superior to general-purpose LLMs, remains lower than highly optimized numerical algorithms. This positions it as an ideal tool for offline or batch analysis rather than for real-time, beat-by-beat monitoring. Future work could explore model compression techniques such as quantization or knowledge distillation to mitigate this latency. Second, our framework's performance is inherently dependent on the recall of the preliminary peak extraction step; any true peaks missed at this initial stage cannot be recovered by the LLM. While our lenient thresholding minimizes this risk, investigating end-to-end detection remains a compelling, albeit challenging, research avenue. Finally, the potential applications of this framework are vast. Future research should explore its adaptation to other physiological signals (e.g., EEG, GSR) and its extension to more complex diagnostic tasks beyond peak detection, such as arrhythmia classification directly from the Peak Representation. make discussion more concise

J Detailed Experimental Protocol

Annotation Details. For datasets lacking ground-truth labels, we employed a rigorous, multi-stage semi-automatic annotation protocol to ensure high-fidelity annotations. The process began with an initial automated peak detection using Elgendi's algorithm [14]. These algorithmically generated labels were then independently reviewed and corrected by two domain experts. Following this, the two sets of human-corrected annotations were cross-validated to identify any discrepancies. All disagreements were subsequently resolved through a consensus discussion between the experts to produce a final, unified set of ground-truth labels. The guiding principle for this process was to ensure high recall, prioritizing the inclusion of all potential physiological peaks, even those with atypical morphology, to create a comprehensive and robust ground truth.

Supervised Fine-Tuning Configuration Peak-R1 was fine-tuned using Qwen2.5-3B-Instruct as the base model with a maximum sequence length of 2048 tokens and a dropout rate of 0.1. The model was trained for 5 epoachs using the AdamW optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate was set to 2×10^{-5} with a cosine decay schedule following 500 warmup steps. To prevent overfitting, we applied weight decay of 0.01 and gradient clipping with a maximum norm of 1.0. The training process utilized a per-device batch size of 32 with 4 gradient accumulation steps, yielding an effective batch size of 128. Mixed precision training with BF16 was employed to improve computational efficiency. The model was trained on 4$\times$ NVIDIA A6000 GPUs (80GB each).

GRPO Configuration The model was further optimized using Group Relative Policy Optimization (GRPO). The actor model was initialized from a supervised fine-tuned checkpoint with a learning rate of 1×10^{-6}. Maximum prompt and response lengths were set to 4000 and 2000 tokens, respectively. To control policy divergence, KL divergence loss was enabled with a coefficient of 0.001 using the low-variance formulation. The rollout generation utilized vLLM with tensor model parallelism and generated 8 samples per prompt. Training was conducted on 4 GPUs for a single epoch with a training batch size of 24, PPO mini-batch size of 12, and micro-batch size of 2 per GPU.

Table 5: Subject-level data splitting for training and testing.

Dataset	Total Subjects/IDs	Training Set	Test Set	Split Ratio (Train/Test)
MIT-BIH	48	38	10	79.2% / 20.8%
Incart	75	60	15	80.0% / 20.0%
BIDMC	50	40	10	80.0% / 20.0%
Capnobase	42	33	9	78.6% / 21.4%
Kansas	40	32	8	80.0% / 20.0%
BCG Arrhythmia	85	68	17	80.0% / 20.0%

Data Splitting. To ensure a rigorous evaluation and prevent data leakage, all datasets were partitioned into training and test sets at the subject level. This methodology guarantees that data from a given subject is confined to only one set (either training or testing), which is critical for assessing the model's ability to generalize to unseen individuals. We maintained an approximate 80/20 training/test

split ratio across all datasets. The specific subject-level distributions for each dataset are detailed in Table 5.

Code Availability. The source code for the Peak-R1 framework and the generated Peak-Explanation Dataset will be made publicly available in an online repository upon acceptance of this manuscript.

K Ethical Approval and Data Collection for BSG ICU dataset

The data collection for this study was conducted in the hospital ICU under an Institutional Review Board (IRB)-approved protocol. The patient consent process adhered strictly to ethical guidelines: data from patients, or their legal representatives, who declined consent—either during their ICU stay or after recovery—were excluded from the analysis.

The sensing devices were permanently installed beneath the hospital beds, seamlessly integrated as part of the bed infrastructure. The data acquisition process was entirely passive (i.e., involving no emission of active radiation) and contactless (i.e., without any physical contact with patients). As such, the system posed no risk to human subjects or interference with existing medical equipment. Importantly, the installation and operation of these devices did not disrupt standard ICU monitoring or treatment procedures, nor did they require any intervention from patients or clinical staff during routine care.

To minimize any potential disruption, device setup was meticulously planned and executed. Prior to installation, all devices were fully configured and validated for network connectivity and operational reliability. Installation was performed under the supervision of hospital personnel, ensuring a rapid, safe, and non-intrusive process. Once deployed, the devices operated autonomously, requiring no further human intervention. To further reduce maintenance demands within the ICU, devices were powered by a direct connection to the hospital mains, obviating the need for battery changes. In addition, the system software was engineered with robust fault-tolerance features, including automatic recovery from network or system errors, to maximize operational stability.

All data collection activities were performed solely by hospital staff. Data science researchers had access exclusively to de-identified data, as provided by the hospital. This strict separation ensured the protection of patient privacy and eliminated the risk of any information leakage.