Exploring RAG-driven Multimodal LLMs for Explainable ECG Interpretation

Minsol Kim
MIT Media Lab
Massachusetts Institute of Technology
75 Amherst St., Cambridge, MA 02139
minsol@mit.edu

Du Liu
MIT Media Lab
Massachusetts Institute of Technology
75 Amherst St., Cambridge, MA 02139
liudu@mit.edu

Ragasudha Botta
MMSCI
Harvard Medical School
25 Shattuck St., Boston, MA 02115
rbotta@hms.harvard.edu

Seok Gyu Han
Rush University Medical College
600 S Paulina St., Chicago, IL 60612
kevin_han@rush.edu

Man Jiang
Harvard T.H. Chan School of Public Health
677 Huntington Ave. Boston, MA 02115
mjiang1@hsph.harvard.edu

Jason Gusdorf
Clinical Fellow in Medicine
Beth Israel Deaconess Medical Center
330 Brookline Ave, Boston, MA 02215
jgusdorf@bidmc.harvard.edu

Pattie Maes
MIT Media Lab
Massachusetts Institute of Technology
75 Amherst St, Cambridge, MA 02139
pattie@media.mit.edu

Paul Pu Liang
MIT Media Lab
Massachusetts Institute of Technology
75 Amherst St, Cambridge, MA 02139
ppliang@mit.edu

Abstract

Foundation models offer a path toward safer, more generalizable biosignal analysis, yet their role in clinical Electrocardiograms (ECGs) interpretation remains unclear. We present a multimodal framework with Retrieval-Augmented Generation (RAG) that couples 12-lead ECG waveforms with structured patient metadata (age, sex) to guide large language models (LLMs) and a downstream ECG encoder. Using PTB-XL for benchmarking and MIMIC-IV-ECG for pretraining, we compare four LLMs within the same RAG pipeline—GPT-4, GPT-3.5, Claude 3.5 Sonnet, and Llama 3.2—for arrhythmia classification across five diagnostic superclasses. GPT-4 achieves the best overall performance (AUC = 0.940) and maintains stable accuracy when demographic features are ablated or added, suggesting reduced sensitivity to demographic shifts. Metadata provided the largest relative gains for models with lower performance in our benchmark, while qualitative analysis highlighted remaining failure modes such as hallucinations and retrieval mismatches. These results indicate that RAG-driven multimodal prompting can enhance ECG interpretation and fairness when paired with high-performing foundation models, and we outline practical levers—retrieval design, prompt structure, and metadata integration—for building equitable and scalable clinical ECG systems.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of 17.9 million deaths annually, accounting for 32% of global mortality [Organization, 2021]. ECGs are the gold standard for noninvasive diagnosis of cardiac conditions, such as arrhythmias and myocardial injuries [Goldberger et al., 2017], yet manual interpretation is labor-intensive and heavily dependent on scarce clinical expertise. Deep learning (DL) models have demonstrated success in tasks such as arrhythmia classification and sleep apnea detection [Mostafa et al., 2019]. However, their reliance on unidimensional signal data restricts their ability to assimilate broader clinical insights.

Integrating LLMs into the diagnostic process can address limitations by contextualizing ECG data with additional patient information and clinical metadata. Nonetheless, this integration poses challenges in embedding multimodal data, managing computational constraints, and handling limited datasets for quality assurance. Moreover, existing models often fail to generalize across diverse populations or unseen cardiac conditions, leading to potentially biased or incomplete predictions [Ström et al., 2023]. The *black-box* nature of LLMs also raises interpretability and trustworthiness concerns in clinical decision-making. There is an urgent need for multimodal systems that improve diagnostic accuracy while ensuring clarity and fairness for safe use in patient care. By combining multimodal data with the interpretive capabilities of LLMs, we aim to bridge the gap between raw ECG signals and clinically meaningful insights, using structured inputs and a Retrieval-Augmented Generation (RAG) system to improve the accuracy and usability of automated ECG interpretation. Furthermore, we compare and evaluate the performances of four open-source LLMs for classifying five major classes of arrhythmia conditions.

Previous studies have explored multiple ECG analysis methods. Śmigiel et al. [2021] introduced a 1D-Convolutional Neural Network model for arrhythmia classification that achieved an AUC of 0.91 across five arrhythmia classes but showed limited generalizability.Mehari and Strodthoff [2022] investigated self-supervised learning methods for 12-lead ECG sequences, highlighting the challenges such as distribution shifts, limited interpretability, and integrating multimodal inputs. Li et al. [2023] converted ECG signals into text for zero-shot learning for sinus rhythm detection but did not extend to more complex arrhythmias or multimodal integration. Recent work has shifted toward multimodal pipelines to improve diagnostic accuracy. LLaVa-Med and MEIT (Multi-Modal Electrocardiogram Instruction Tuning) fine-tune LLMs on biomedical corpora, improving ECG report generation and highlighting the value of multimodal data integration for benchmarking and handling signal variation.

Building on these directions, we investigate structured multimodal integration with LLMs to improve diagnostic accuracy. Although combining ECG records with patient metadata has shown promise, current models still struggle to generalize across diverse populations. To address this gap, we evaluate key factors—including retrieval-augmented generation (RAG), prompt design, output structure, and demographic diversity—that may shape model performance. Specifically, we explore: (1) Does incorporating multimodal data (e.g., ECG signals and patient information) improve interpretation? (2) Which factors—such as prompt design, retrieval-augmented generation, and demographic data—shape model performance? (3) How do different LLMs compare in ECG interpretation, and what features account for their differences? Ultimately, our work seeks to integrate high-performance AI models into clinical practice for greater trust and usability in real-world settings and extending applicability beyond ECGs to other biosignals.

2 Methodology

Datasets Our study draws on two complementary public ECG corpora. PTB-XL [Wagner et al., 2020] comprises 21,837 ten-second 12-lead recordings (100/500 Hz) annotated by cardiologists with 71 SCP-ECG diagnostic labels, accompanied by demographic and signal-quality metadata. This structured design provides a clean, standardized reference for model development and waveform evaluation. In contrast, MIMIC-IV-ECG [mim, 2024] contains approximately 800,000 12-lead recordings (500 Hz) collected across diverse clinical settings and automatically linked to the MIMIC-IV electronic health record database. This integration supports large-scale training under real-world variability and enables multimodal outcome analysis. Using PTB-XL for controlled benchmarking and MIMIC-IV-ECG for scale and clinical context allows us to build and validate models that generalize beyond curated data.

Considering Bias in Age and Gender Representation The PTB-XL dataset, while comprehensive, exhibits demographic imbalances that introduce critical biases. The gender representation is nearly balanced (52% male, 48% female), but the age distribution is biased toward older adults (median age 62, range 0–95). Pediatric and young adult cases are underrepresented, which can lead models to generalize poorly to younger populations, often defaulting to unrealistic pediatric predictions. Gender bias also emerges, as conditions such as myocardial infarction and atrial fibrillation are represented disproportionately by one gender. This leads to systematic misclassifications, particularly for underrepresented groups such as women with myocardial infarction. Inconsistencies in clinical annotation may further amplify these gender biases.

The clinical impact of these biases can be substantial. Misdiagnosis in pediatric cases can arise from incorrect age predictions and inappropriate treatment plans due to misclassifications of gender-specific presentations. To mitigate these issues, we propose a multipronged strategy. Dataset augmentation should include pediatric and young adult cases from external sources such as MIMIC-IV and employ data augmentation techniques to simulate pediatric clinical text. It is also vital to balance condition-specific cases across genders by adding underrepresented gender-condition pairs, such as myocardial infarction in women. Models should explicitly incorporate age and gender as structured features in the input pipeline to better distinguish demographic-specific patterns. As suggested in Perez Alday et al. [2022], the performance evaluation should be stratified by age groups (e.g., pediatric, adult, elderly) and gender categories (e.g., male, female) to address the disparities identified during these assessments. Finally, domain-specific fine-tuning of tools such as ClinicalBERT using data sets enriched with pediatric-specific clinical conditions and gender-balanced annotations could improve age range generalization and reduce condition-specific biases.

Multimodal ECG Interpretation Model For our main model for multimodal ECG interpretation, we adopted the ECG Semantic Integrator (ESI) [Yu et al., 2024]. It is a novel multimodal framework that leverages both ECG waveforms and textual descriptions to enhance LLM learning. The core components of the models are a signal encoder that captures both spatial and temporal features from 12-lead ECG signals, a text encoder that utilizes BioLinkBERT, a variant of the BERT architecture pre-trained on biomedical texts, and a cross-modality decoder that uses self-attention and cross-attention mechanisms to capture relationships between ECG signals and their textual annotations.

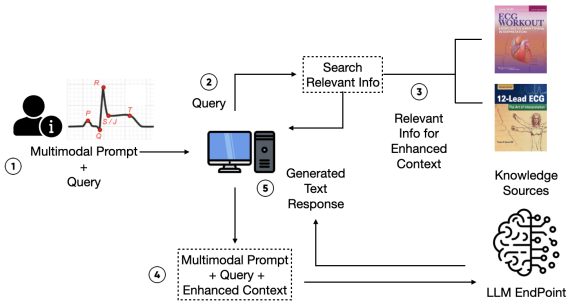

Figure 1: Pipeline of Retrieval-Augmented Generation (RAG) for multimodal ECG interpretation.

Among the components, our research focuses on evaluating the role of Retrieval-Augmented Generation (RAG), particularly in enriching ECG textual descriptions through domain-specific knowledge retrieval, as shown in Figure 1. Existing work has investigated the impact of top-k in vector search, the number of textbooks in RAG, and the impact of different LLM prompting strategies, but the impact of different LLMs has not been studied. Our goal is to assess the effectiveness of different Large Language Models (LLMs)—GPT-4, GPT-3.5, Claude 3.5 Sonnet, and Llama 3.2—in generating clinically meaningful textual prompts.

To understand the RAG structure of the model, the most essential component is the Cardio Query Assistant (CQA). It is a model that retrieves information from given domain expertise and generates descriptive prompts. These prompts are integrated into ESI as input for both pre-training and fine-tuning tasks. The pre-training task is self-supervised, with the training objective of a combination of dual-encoder contrastive loss and captioning loss. We define arrhythmia detection as the fine-tuning and evaluation task, with the training objective of cross-entropy loss of five diagnostic super-classes

of arrhythmia: STTC (ST segment and T wave changes), HYP (Hypertrophy), MI (Myocardial Infarction), CD (Cardiac Disorder), and Normal (NORM) [Strodthoff et al., 2020].

The CQA, employing a RAG pipeline, retrieves knowledge from two ECG textbooks ([Huff, 2006], [Garcia, 2015]). The knowledge forms a comprehensive vector database through the `amazon.titan-embed-text-v2:0` API. After constructing the knowledge base, prompts in the form of "How is [the ECG report in the dataset] reflected in ECG?" are input into different LLMs. We utilized Amazon Bedrock for accessing Claude and Llama models and OpenAI API for GPT models. The `llama_index` and `langchain` libraries were used for RAG construction, and multithreading was employed to speed up API queries for large datasets. LLMs retrieve and generate different descriptive contexts of specific waveform attributes. Finally, the enriched text of waveform attributes is combined with demographic information (sex and age), serving as input to ESI's text encoder.

LLMs Benchmark Evaluation This study employs the PTB-XL benchmark framework [Wagner et al., 2020] for standardized evaluation. Four LLMs—GPT-4, GPT-3.5, Claude 3.5 Sonnet, and Llama 3.2—were integrated into the RAG pipeline for generating textual prompts. Pretraining of the ESI model was conducted on the MIMIC-IV-ECG dataset. Fine-tuning was performed on the PTB-XL dataset. Performance was evaluated with the area under the receiver operating characteristic curve (AUC). Three sets of evaluations were conducted: (1) **Arrhythmia Detection**: Evaluating fine-tuned models on PTB-XL for diagnostic classification. AUC scores were computed for each diagnostic superclass; (2) **Demographic Data Impact:** Comparing model performance with and without demographic data (age and sex) to assess their influence.; (3) **Sex Bias Impact:** Comparing model performance when training and evaluating on male or female subsets to assess the model's bias. The evaluation framework included benchmarks from prior studies [Strodthoff et al., 2020], such as supervised and self-supervised models like ResNet50, LSTM, and ConvNeXt. These baselines provided a reference for understanding the added value of multimodal and RAG-enhanced approaches.

3 Evaluation Results and Discussion

The results are shown in Table 1. It demonstrates that GPT-4 achieved the highest AUC (0.940) for arrhythmia detection. Importantly, its performance remains consistent across patient metadata categories, including sex-only (0.940), age-only (0.939), and combined metadata (0.940). This indicates a robust model design that effectively integrates demographic information without introducing biases. Conversely, ESI-GPT3.5 shows a slight preference for age-specific data, achieving a higher AUC with age-only metadata (0.930) compared to sex-only (0.928) and combined metadata (0.929). ESI-Claude 3.5 Sonnet exhibits balanced performance with patient metadata, achieving 0.939 for females and 0.937 for males, suggesting a minor inclination toward female-specific data. ESI-Llama3.2, while uniform across categories, consistently underperforms with an overall AUC of 0.922, suggesting limited ability to fully leverage demographic information. These findings demonstrate ESI-GPT4's robustness and generalization ability, outperforming other models while minimizing potential demographic biases.

Qualitative analysis of the textual descriptions generated revealed that GPT-4 consistently produced accurate and coherent output. Claude 3.5 Sonnet demonstrated proficiency in logical reasoning. However, it occasionally struggled with retrieving interrelated context. Llama 3.2, while generally efficient, exhibited tendencies to 'hallucinate'; that is, the model could generate additional details or conclusions not supported by the retrieved content. For instance, when provided with a normal ECG report, Llama 3.2 might generate unsupported content such as: *"sinus rhythm with sinus arrest. The*

Table 1: Evaluation results (AUC across models and input data).

Method	AUC	AUC (sex)	AUC (age)	AUC (sex+age)
LSTM	0.927	-	-	-
ResNet50	0.930	-	-	-
Ensemble	0.934	-	-	-
ConvNeXt-Tiny	0.918	-	-	-
ESI-tiny	0.935	-	-	-
ESI-GPT4 (ours)	0.940	0.940	0.939	0.940
ESI-GPT3.5 (ours)	0.927	0.928	0.930	0.929
ESI-Claude3.5 Sonnet (ours)	0.935	0.939	0.938	0.938
ESI-Llama3.2 (ours)	0.922	0.922	0.923	0.924

sinus rhythm is normal, but there is a pause, causing the heart rate to slow down temporarily. The pause is followed by a return to the normal rhythm. The ST segment may be depressed, and T-waves may be inverted." This behavior indicates a potential challenge in using the model for reliable medical interpretations where precision is critical. Comparing the LLM models, we were also able to confirm that GPT-4 and Claude 3.5 Sonnet outperformed in arrhythmia classification tasks compared to other LLMs due to their ability to generate clinically meaningful textual descriptions with patient metadata.

Notably, top-performing models like GPT-4 maintained high fairness and reliability even when incorporating patient demographic data, such as sex and age, for the classification task. For instance, GPT-4 achieved nearly identical area under the curve (AUC) scores across male and female subgroups (0.942 and 0.940), showing its ability to mitigate subtle dataset biases. In contrast, GPT-3.5 and Llama 3.2 exhibited more pronounced demographic variations, highlighting the need for robust models when utilizing patient demographics. Notably, patient metadata played a critical role in improving the consistency and accuracy of less capable models, such as Llama 3.2, which performed poorly when excluding the data. Our findings suggest that integrating metadata can still ensure equitable and reliable performance, but bias considerations are needed for less advanced systems.

4 Conclusion

We investigated the potential of retrieval-augmented generation (RAG) and large language models (LLMs) for multimodal ECG interpretation. By integrating ECG signals and patient metadata, we evaluated whether multimodal data improve diagnostic performance. Our comparative evaluation of four LLMs identified GPT-4 as the most effective model for ECG interpretation tasks, while other models like Llama 3.2 relied more heavily on metadata for stability and robustness. Our exploration also uncovered system-level challenges within the RAG pipeline, including occasional hallucinations and retrieval mismatches. By benchmarking using AUC, sensitivity, and specificity, this study establishes an initial framework for selecting and optimizing LLMs in clinical applications. Moreover, integrating ECG signals with demographic data and textual prompts represents a step toward developing holistic diagnostic systems capable of navigating real-world clinical variability. However, our work also highlights that careful consideration of bias and fairness is essential in clinical contexts. While GPT-4 and Claude 3.5 Sonnet exhibited minimal demographic disparities, models such as GPT-3.5 and Llama 3.2 showed more pronounced variation across subgroups. These findings emphasize that multimodal integration alone does not guarantee equitable outcomes—robust retrieval design, bias-aware training, and systematic subgroup evaluation are necessary to ensure reliable and trustworthy deployment in clinical contexts.

5 Limitations and Future Work

While our contributions demonstrate the promise of RAG-driven multimodal LLMs for ECG interpretation, several aspects warrant careful consideration. The datasets used may not fully capture the diversity of cardiac conditions or patient populations observed in practice, and ethnicity information is unavailable. Additionally, the conversion of ECG signals into textual formats, while effective for LLM processing, may reduce the granularity of signal features, potentially impacting diagnostic precision. The computational demands of training and deploying LLMs on large-scale datasets also present a significant barrier to real-time clinical adoption.

Future work should consider: (i) **broader and more representative datasets**, including rare and underrepresented conditions, richer demographic attributes, and standardized subgroup or site/device splits for robust cross-hospital validation; (ii) **richer multimodal conditioning**, augmenting ECG with longitudinal context such as patient history, medications, clinical notes, and outcomes, leveraging unified vector encodings to align heterogeneous modalities, and aligning model outputs with clinical decision points; (iii) **fairness- and reliability-first evaluation**, including subgroup-aware reporting, shift-robustness and counterfactual tests, calibration and uncertainty estimation, and retrieval stress testing (e.g., hard negatives, off-topic passages, citation faithfulness); (iv) **cross-signal generalization and shared benchmarks**, extending to or combining with other biosignals (EEG, EMG, PPG, etc.), releasing open retrieval corpora and prompt templates, and adopting transparent, reproducible evaluation protocols; and most importantly (v) **human–AI collaboration and interpretability**, including prospective clinician-in-the-loop studies, waveform-grounded explanations, and workflows that integrate human feedback for continual refinement. We believe that advances along these directions could help translate foundation models from single-modality proofs-of-concept into reliable, equitable, and scalable systems for physiological signal analysis in clinical diagnosis and beyond.

Acknowledgments and Disclosure of Funding

We would like to thank Dr. Leo Anthony Celi and Jacques Kpodonu MD, FACC, for their guidance and support.

References

MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset v1.0. https://physionet.org/content/mimic-iv-ecg/1.0/, 2024. Accessed 2024-12-09.

Tomas B. Garcia. *12-Lead ECG: The Art of Interpretation*. Jones & Bartlett Learning, 2 edition, 2015. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=772374.

Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin. *Clinical Electrocardiography: A Simplified Approach*. Elsevier, 9 edition, 2017.

Jane Huff. *ECG Workout: Exercises in Arrhythmia Interpretation*. Lippincott Williams & Wilkins, illustrated edition, 2006.

Jie Li et al. Frozen language model helps ECG zero-shot learning. In *Proceedings of Machine Learning Research*, volume 227, pages 402–415, 2023.

Tewodros Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computer Methods and Programs in Biomedicine*, 141:105114, 2022.

Sherif S. Mostafa, Filipe Mendonça, Antonio G. Ravelo-García, F. Morgado-Dias, and Thomas Penzel. A systematic review of detecting sleep apnea using deep learning. *Sensors*, 19(4934):1–19, 2019. doi: 10.3390/s19224934.

World Health Organization. Cardiovascular diseases (cvds) fact sheet. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2021. Accessed 2024-12-10.

E. A. Perez Alday et al. Age, sex, and race bias in automated arrhythmia detectors. *Journal of Electrocardiology*, 74:5–9, 2022.

Robert Śmigiel et al. Challenges in automated ECG signal interpretation and classification: A review of machine learning and deep learning techniques. *Computer Methods and Programs in Biomedicine*, 207:106168, 2021. doi: 10.1016/j.cmpb.2021.106168.

Nils Strodthoff et al. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2020. doi: 10.1109/JBHI.2020.3022989.

Peter Ström et al. Generalization challenges in Deep Learning for Healthcare. *Patterns*, 4(4):100635, 2023. doi: 10.1016/j.patter.2023.100635.

Patrick Wagner et al. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

Han Yu, Peikun Guo, and Akane Sano. ECG semantic integrator (ESI): A foundation ECG model pretrained with LLM-enhanced cardiological text, 2024.