On the Potential of Large Language Models in ECG-based AFib and Sinus Rhythm Detection and Justification

Maria Slim[1], Chaymaa Abbas[1], Jad Assi[2], Hussein El Jebbawi[2], Alaaeddine El Ghazawi[2], Mariette Awad[*1], Fatme Charafeddine[*2], Marwan Refaat[*2], Fouad Zouein[2]

[1]Maroun Semaan Faculty of Engineering and Architecture, American University of Beirut
[2]Medical School, American University of Beirut
mas194@mail.aub.edu, cwa07@mail.aub.edu,jha24@mail.aub.edu, hne13@mail.aub.edu,
aae89@mail.aub.edu,mariette.awad@aub.edu.lb,fc24@aub.edu.lb, mr48@aub.edu.lb, fz15@aub.edu.lb

Abstract

Atrial fibrillation (AFib) is a common arrhythmia that is associated with increased stroke and mortality risk. It requires early and accurate detection for improved patient healthcare support. This study explores the application of vision-enabled large language models (LLMs)—specifically Llama-3.2-11B-Vision-Instruct and Qwen2-VL-7B-Instruct —for AFib and sinus rhythm detection using ECG images. We designed structured prompts to simulate clinical reasoning, evaluate rhythm features, and elicit model confidence. Models were tested on a curated PTB-XL subset under both full 12-lead and dual-lead (Lead II + V1) configurations. Results show that while Llama achieves higher diagnostic accuracy, especially with Chain-of-Thought prompting (up to 97% for AFib), both models struggle with consistent feature-level interpretation, particularly for sinus rhythm. Our findings underscore both the promise and current limitations of LLMs in ECG-based diagnosis. Bridging the gap between AI-generated outputs and clinical standards will require fine-tuning on ECG-specific data, robust prompting strategies, and hybrid approaches that integrate signal-level reasoning for improved interpretability and reliability in real-world settings.

1 Introduction

AFib is the most common and life-threatening heart rhythm disorder in the world. Today, it affects between 52 and 57 million people worldwide, nearly twice as many as in 1990 (Caffrey 2023; Tan et al. 2025). In the United States, a 2024 study found that approximately 10.5 million adults, or nearly 5% of the population, have been diagnosed with AFib, three times more than previous estimates (Noubiap et al. 2024). As the number of cases is projected to increase to 60% globally by 2050, there is an urgent need to better prevent, detect, and manage this life threatening health irregularity (Lippi, Sanchis-Gomar, and Cervellin 2021).

AFib is characterized by irregular rhythms and the absence of distinct P waves, among other characteristics. Early and accurate detection is essential, as it can greatly improve patient outcomes by allowing timely preventive interventions. As AI becomes increasingly integrated into clinical workflows, effective human-AI collaboration offers new opportunities to support clinicians to make faster and more accurate diagnoses, particularly in resource-limited settings where specialist interpretation of ECGs may be scarce.

This study investigates two such models—Llama-3.2-11B-Vision-Instruct and Qwen2-VL-7B-Instruct—for their ability to detect and justify diagnoses of AFib and sinus rhythm from ECG images. Structured prompts were crafted to guide the models to simulate cardiologist reasoning, classify rhythms, identify key waveform features, and report diagnostic confidence. We evaluated model performance across full 12-lead and reduced dual-lead (Lead II + V1) configurations, focusing on both diagnostic accuracy and feature-level justifications.

Our results reveal several important trends. Llama demonstrated better diagnostic accuracy, particularly for AFib detection, achieving up to 97% accuracy under Chain-of-Thought prompting in the dual-lead setting. However, its performance on sinus rhythm was modest, with just 33.67% accuracy on dual-lead ECGs. Moreover, despite strong AFib performance, Llama consistently struggled with PR interval interpretation (79.12%), indicating incomplete feature-level understanding. In contrast, Qwen performed poorly overall and frequently returned empty or incorrect responses, especially under sinus rhythm tasks. Even when non-empty outputs were isolated, Qwen's average performance improved marginally by 13%, and only in a few cases (e.g., Role Specification on sinus detection) did it slightly surpass Llama on Chain-of-Thought in the dual-lead setting. To further test interpretability, we conducted an electrophysiologist-guided re-prompting experiment focused on sinus rhythm morphology. Both models—especially Qwen—failed to reliably confirm even basic waveform features such as upright or biphasic P waves, underscoring limitations in their internal electrophysiological reasoning.

These findings highlight both the promise and critical gaps in using general-purpose vision-enabled LLMs for ECG interpretation. While prompting strategies can influence outcomes, they do not compensate for the lack of domain-specific internal knowledge. As such, we advocate for future efforts to include fine-tuning on clinically annotated ECG datasets and representational enhancements to support reliable feature-level reasoning. Our work empha-

[*]Corresponding Authors all sharing equal contributions

sizes the importance of human-AI collaboration, especially in high-stakes clinical environments.

The remainder of the paper is organized as follows: Section 2 provides background information, while Section 3 offers an expert overview of key ECG components. Sections 4 and 5 outline the methodology and present the results, respectively. Section 6 discusses the key insights derived from the findings, followed by the conclusion in Section 7. Appendices A and B include sample prompts used for AFib detection to illustrate the prompting approach and the tables of all results performed.

2 Background and Related Work

2.1 LLMs in Healthcare

LLMs' ability to comprehend and produce text that is human-like has quickly made them a focus of medical AI. Models such as GPT-3/GPT-4, PaLM, and others have been evaluated on medical knowledge benchmarks. With a score of roughly 67%, Google's Med-PaLM (based on PaLM) was the first to surpass the passing mark on USMLE-style questions. Its successor, Med-PaLM 2, also approached expert-level performance with a score of about 86% (Singhal et al. 2023).These models show that LLMs are capable of remembering and reasoning with an extensive amount of medical data, including clinical facts and guidelines. In some situations, they can even draft responses that doctors prefer. In addition to exams, LLMs have been used for tasks like creating clinical notes, summarizing patient interactions, and providing evidence-based recommendations to support decision-making. Nevertheless, the majority of these applications have been textual in nature, meaning that both the input and the output are text (medical questions, symptoms, etc.). A more recent development is the application of LLMs to multimodal inputs, like medical images. Introduced in late 2023, GPT-4 with vision demonstrated the ability to analyze images, from X-rays to dermatology photos, by incorporating them into its conversational context (Koga and Du 2025). GPT-4V was immediately probed by researchers using image tasks specific to a given domain. As mentioned, GPT-4V performed poorly in radiology when it came to diagnosing images, correctly responding to fewer than half of image-based questions. On visual inspection, the model frequently generated analyses that sounded plausible but were factually incorrect.(Medical Xpress 2024) These contradictory findings imply that general LLMs' visual comprehension is inferior to their textual comprehension, particularly when it comes to specialized medical imagery. However, the capability is still developing, and performance may be enhanced by domain specific fine-tuning. By assessing how well existing multimodal LLMs handle ECG images—which are different from natural images or standard radiographs and have their own complexities (they are essentially time-series plots of electrical signals)— our study advances this line of research.

2.2 Prompt Engineering and Reasoning

One major advantage of LLMs is their ability to change behavior through prompts, without any need to retrain or alter their underlying parameters. Researchers are actively developing efficient prompting strategies to improve performance on reasoning-intensive medical tasks. One such method is Chain-of-Thought prompting, which encourages the model to generate step-by-step explanations before arriving at a final answer. This approach loosely mirrors the way physicians reason through complex cases by breaking down analysis into sequential, manageable steps.Wu, Chen, and Chen (2023) extended Chain-of-Thought prompting to clinical diagnosis tasks and reported up to a 15% improvement in diagnostic accuracy using a "think-aloud" format.

Another emerging technique is role prompting, where the model is assigned a specific persona—such as a cardiologist or medical resident—to simulate expert-level reasoning. Early evidence suggests this can enhance factual consistency in medical Q&A by anchoring the model's output within the expected style and content of domain experts. In this study, we employ role prompting to ground the model's reasoning in a clinically relevant perspective.

While an LLM's self-reported confidence is not rigorously calibrated, the presence of a confidence estimate can be informative. Prior human factors research indicates that showing a confidence level can help users decide when to trust an AI and when to be skeptical (Turner et al. 2020). Having an AI flag for uncertainty could be very helpful for critical diagnoses like AFib, where false positives could result in unnecessary treatment and false negatives could miss a serious condition. A doctor would be able to do an additional review if the model indicated that it was uncertain. In our experiments, we employed a Confidence Assessment prompting strategy to elicit self-reported confidence levels for each diagnostic observation and evaluated how this approach influenced overall diagnostic and feature-level accuracy.

2.3 Interpretability and Trust

The ultimate objective of using LLMs for ECG interpretation is to improve interpretability and clinician confidence in AI, not just raw performance. In the field of medical artificial intelligence, there is a growing understanding that any algorithm used in practice needs to be explicable, or at the very least, provide evidence for its results. Even though they are still in their beginnings, traditional ECG AI systems have begun to incorporate some interpretable components (such as highlighting specific ECG segments or reporting which criteria were met). By producing free-text justifications, LLMs provide a more comprehensive explanation—basically, an interpretable story. However, one requires cautiousness because LLMs are also prone to factual hallucinations, which means they may boldly claim that an ECG has a particular feature when in fact it does not. If left unchecked, such behavior could be deceptive or even harmful. Thus, making sure the LLM's explanation is correct and consistent with the real ECG features is a crucial part of establishing trust. In this study, we assess not only whether the models accurately label rhythms but also whether their explanations align with clinical reality (for example, if the model asserts that "no P waves are seen" or "the rhythm is irregularly irregular", are those claims true based on the ground truth?). By dissecting

the LLMs' outputs in this way, we address the reliability of their interpretability. We also take into account more general ethical and societal issues: accountability is a concern because an LLM may make mistakes that a traditional algorithm wouldn't, and bias and fairness issues could arise if the model's performance varies across patient subgroups. Although an extensive ethical analysis is outside the purview of this article, we stress that this investigation is only the beginning and that any practical implementation of such technology would necessitate stringent validation, bias checks, and the establishment of explicit procedures for human supervision.

3 Clinical Criteria for AFib and Normal Sinus Rhythm

As the foundation for both human and model interpretations in this study, it is essential to define the distinguishing ECG features of AFib versus normal sinus rhythm before diving into methodology. The diagnostic criteria for normal sinus rhythm and AFib are different, and they can be summarized as follow:

- AFib: The ECG hallmark of AFib is an irregularly irregular rhythm – the R-R intervals (time between heartbeats) follow no repetitive pattern (Heidbuchel et al. 2016). In addition, P waves are absent; instead of the normal P wave before each QRS complex, AFib shows either chaotic fibrillatory waves or a flat baseline, reflecting the disorganized atrial activity (January et al. 2019). The ventricular response in AFib is typically rapid and highly variable; heart rates often range from 90 to 170 beats per minute, though slower or faster rates can occur. The QRS complexes in AFib usually remain narrow (duration less than 120 ms) since ventricular conduction is normal – unless a pre-existing bundle branch block or an accessory pathway is present, which would widen the QRS. Another distinguishing feature is the absence of "sawtooth" flutter waves; this helps differentiate AFib from atrial flutter, where flutter waves are present in a regular pattern.
- Normal Sinus Rhythm: Sinus rhythm is the standard cardiac rhythm originating from the sinus node. In ECG terms, it is characterized by a regular rhythm – the P-P intervals (and R-R intervals) are consistent, apart from slight natural variations with breathing (sinus arrhythmia). P waves are present in front of every QRS complex, and they have a uniform shape (all arising from the sinus node) and a normal direction on the ECG leads. The baseline between beats in normal sinus rhythm is stable and isoelectric (flat), indicating absence of pathological atrial activity. Heart rate in normal sinus rhythm typically falls between 60 and 100 bpm at rest for adults. Some other interval criteria can be present such as the PR interval (onset of P to onset of QRS) should be about 120–200 ms, reflecting normal AV conduction delay, and the QRS duration is less than 120 ms indicating normal intraventricular conduction in normal sinus rhythm.

4 Methodology

Figure 1 illustrates the proposed methodology for classifying AFib and sinus rhythm ECGs using LLMs. The process begins with a curated dataset containing ECG images labeled with ground truth diagnoses. These images are processed through various prompting techniques. Each technique is designed to guide the LLM toward clinically grounded and interpretable outputs. Initially, the methodology was applied to the full 12-lead ECGs. To further investigate the diagnostic relevance of individual leads, we subsequently retried the process using only Lead II and V1, which are commonly emphasized in rhythm analysis. The selected prompting strategy is passed to the LLM, which generates structured JSON outputs reflecting diagnostic interpretation. Finally, these outputs are evaluated through an automatic evaluation module that compares the model's predictions against ground truth labels to assess accuracy, consistency, and confidence calibration.

4.1 Dataset and Ground Truth Design

The data set used in this study was a subset of the PTB-XL data set (Wagner et al. 2022, 2020), consisting of 100 randomly chosen ECG images representing AFib and other 100 randomly chosen ECG images representing Sinus Rhythm. We define here the Sinus Rhythm as the standard rhythm originating from the sinus node. It may include cases of normal Sinus Rhythm, defined earlier as having normal PR interval, stable and isoelectric baseline activity, normal heart rate and normal QRS morphology. And it may also include cases of abnormal Sinus Rhythm, whereby only the P wave morphology defines the rhythm as sinus, while the other criteria of PR interval, baseline activity, heart rate, and QRS morphology are not normal. Therefore, for Sinus Rhythm, we define the ground truth as follows. In the first experiment, we simply required the presence of P waves based on the feedback of allied healthcare professionals. In the second experiment, three more accurate criteria had to be met: (1) P waves present; (2) P waves upright in leads II, III, and aVF (or lead II only in the dual-lead setting); and (3) P waves biphasic in lead V1. For AFib, we defined the ground truth as an irregularly irregular rhythm; absent P waves; and a PR interval that is not measurable. These standardized criteria were used as the rubric for evaluating model outputs.

4.2 Prompt Design for AFib and Sinus Rhythm

We employed the same three prompting strategies described in Section 2.2—Role Specification, Chain-of-Thought, and Confidence Assessment— with tailored templates for AFib and Sinus Rhythm. Example templates are provided in the Appendix.

4.3 Automation Workflow and Evaluation Metrics

To optimize efficiency and minimize human error, the entire experimental process was automated. The automated workflow consisted of several sequential steps:

API Integration: Python-based scripts automatically sent ECG images and prompts to the Llama and Qwen mod-

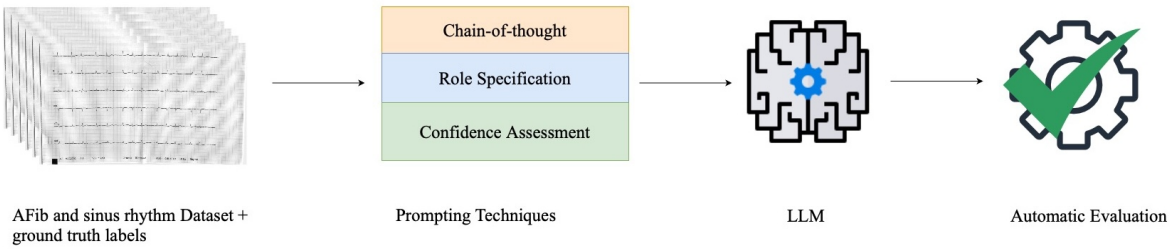

Figure 1: Methodology pipeline.

els via API calls. The responses of the models were automatically retrieved in structured JSON format and systematically stored for further processing. Llama-3.2-11B-Vision-Instruct was queried via the Hugging Face Inference API, using its underlying inference providers, while Qwen2-VL-7B-Instruct was served through a custom Hugging Face Inference endpoint and queried over HTTP

Automated Evaluation: To streamline our evaluation, we stored all Llama and Qwen outputs in a JSON file and used an LLM to map them in one standardized format. These unstructured LLM-based outputs were parsed using custom scripts designed to compare each predicted feature against the established ground truths. The scripts systematically computed multiple evaluation metrics for each diagnostic feature, including accuracy, error rates, at two granularities: individual ECG features (such as Rhythm Regularity, P waves, and PR Interval) and overall diagnostic classifications (AFib vs. Non-AFib, Sinus Rhythm vs. Non-Sinus Rhythm). A small number of API calls returned no response. In these cases, summary statistics were computed on the subset of calls that produced outputs, resulting in slightly smaller denominators for those metrics.

Visualization and Analysis: Prompt-level statistics were summarized in tables, facilitating a direct comparison between the effectiveness of Role Specification, Chain-of-Thought reasoning, and Confidence Assessment prompts. All experiments were conducted on Google Colab using an NVIDIA T4 GPU with 16 GB of memory; each API call to the LLMs took approximately one minute per ECG image, and all experiments were run only once.

5 Results

5.1 Performance on 12-Lead ECGs

We analyzed the performance of the Qwen and Llama models on one run of full 12-lead ECG data and focused on their ability to classify AFib and Sinus Rhythm cases and interpret relevant ECG features.

Qwen Model

AFib Cases The Qwen model demonstrated extremely limited diagnostic capability on AFib cases. It achieved a diagnosis accuracy of only 8.87%, with a recall of 0.09 and an F1-score of 0.16. Feature-level accuracy was similarly low: Rhythm Regularity was correctly identified in 11.26% of cases, P waves in 5.12%, and PR Interval interpretation in 9.22%.

Across prompting strategies, Role Specification yielded the highest diagnostic accuracy at 11.11%, while Chain-of-Thought and Confidence Assessment reached 8.16% and 7.29%, respectively. Under Role Specification, P waves peaked at 6.06%, while under Confidence Assessment, P wave detection fell to 5.21% and Rhythm regularity had an accuracy of 15.62%.

Importantly, only 15% of the correctly diagnosed AFib cases had full feature accuracy. The rest contained at least one error highlighting that Qwen was unable to justify its AFib classifications through consistent multi-feature reasoning. Additionally, Qwen returned empty responses in 56% of AFib cases, which were considered as incorrect.

Sinus Rhythm Cases The Qwen model's performance on Sinus Rhythm classification remained poor. It achieved a diagnosis accuracy of only 13.71%, with a recall of 0.14 and an F1-score of 0.24. Feature-level accuracy was also low: P waves was correctly identified in 23.41% of cases.

Prompt-wise, Chain-of-Thought provided the highest diagnostic accuracy at 16.00%, followed by Confidence Assessment at 13.00% and Role Specification at 12.12%. Under Chain-of-Thought, P wave accuracy rose to 28.00% , whereas under Role Specification this was 22.22% ; Additionally, Qwen returned empty responses in 56% of Sinus Rhythm cases, which were counted as incorrect.

Llama Model

AFib Cases The Llama model performed substantially better than Qwen, with a diagnosis accuracy of 84.85%, recall of 0.85, and F1-score of 0.92. In terms of feature–level performance: Rhythm Regularity was correctly identified in 88.22% of cases, P waves in 92.59%, and PR Interval in 79.12%.

Among the prompting strategies, Role Specification produced a diagnostic accuracy of 77.78% and relatively good feature extraction—85.86% for P waves and 74.75% for PR Interval. Chain–of–Thought improved further, achieving 93.94% diagnostic accuracy with P waves at 98.99% and PR

Interval at 84.85%. Confidence Assessment also performed well, with 82.83% diagnostic accuracy and feature accuracies of 92.93% for P waves and 77.78% for PR Interval.

84.85% of correctly diagnosed AFib cases had fully accurate supporting features; the remaining responses contained at least one error.

Sinus Rhythm Cases The Llama model's performance on Sinus Rhythm cases was modest, achieving a diagnosis accuracy of 25.51%, recall of 0.26, and F1-score of 0.41. P waves was correctly identified in 34.01% of the cases.

Prompt-wise, Chain-of-Thought delivered the best diagnostic accuracy at 35.71%, followed by Confidence Assessment at 22.45% and Role Specification at 18.37%. Under Chain-of-Thought, P wave extraction reached 44.90%; under Role Specification, P waves accuracy was only 24.49%.

5.2 Performance on Lead II and Lead V1

We then conducted a separate evaluation using only Lead II and V1 to examine how each model performs with minimal ECG input. These two leads were selected due to their relevance in rhythm interpretation.

Qwen Model

AFib Cases Using a reduced dual-lead configuration (Lead II and Lead V1), Qwen's diagnostic performance further deteriorated. Overall diagnosis accuracy fell to 7.46%, with recall of 0.07 and an F1-score of 0.14. Rhythm Regularity was correctly identified in 10.51% of cases, and feature extraction remained minimal: P waves in 6.44%, PR Interval in 7.46%.

Prompt-wise, Confidence Assessment achieved the highest diagnostic accuracy at 12.37%, followed by Role Specification at 6.06% and Chain-of-Thought at 4.04%. 13.6% of the correctly diagnosed cases provided complete feature justifications. Qwen returned empty responses in 52.67% of AFib cases, which were counted as incorrect.

Sinus Rhythm Cases Qwen's diagnostic performance on Sinus Rhythm improved modestly with the reduced dual-lead configuration (Lead II and Lead V1), achieving an overall diagnosis accuracy of 14.90%, with a recall of 0.15 and an F1-score of 0.26. Feature-level accuracy remained low: P waves was correctly identified in 24.04% of the cases.

Prompt-wise, Role Specification yielded the highest diagnostic accuracy at 18.84%, followed by Confidence Assessment at 13.24% and Chain-of-Thought at 12.68%. Under Role Specification, P waves reached 21.74%; under Chain-of-Thought, P waves peaked at 28.17%; Under Confidence Assessment, P waves reached 22.06%. Qwen returned empty responses in 48.83% of Sinus Rhythm cases, which were counted as incorrect.

Llama Model

AFib Cases Llama achieved strong diagnostic performance in this configuration: 84.33% accuracy, recall of 0.84, and F1-score of 0.92. Rhythm Regularity was correctly identified in 87.33% of cases, P waves in 92.67%, and PR Interval in 83.33% of the cases.

Among the prompting strategies, Chain-of-Thought produced the highest diagnostic accuracy at 97.00%, with P wave extraction of 99.00% and PR Interval accuracy of 92.00%. Role Specification followed, yielding 78.00% diagnostic accuracy and strong feature extraction for P waves (89.00%) and PR Interval (82.00%). Confidence Assessment also performed well, with 78.00% diagnosis accuracy, 90.00% for P waves, and 76.00% for PR Interval.

Out of all the correct AFib diagnoses, 91.7% achieved full feature-level accuracy, with the remaining cases containing at least one error.

Sinus Rhythm Cases The Llama model on the reduced dual-lead configuration achieved a diagnosis accuracy of 33.67%, with a recall of 0.34 and an F1-score of 0.50. Feature-level performance remained modest: P waves were correctly identified in 37.67% of the cases.

Prompt-wise, Chain-of-Thought produced the highest diagnostic accuracy at 44.00%, followed by Role Specification at 37.00% and Confidence Assessment at 20.00%. Under Chain-of-Thought, P waves reached 47.00%. Under Role Specification, P waves accuracy was 42.00%. Confidence Assessment lagged, with P waves at 24.00%.

We should note that even after discarding the empty files in Qwen, the results remained similar with only an average increase in accuracy of 13% in specific scenarios. More details are highlighted in Table 1, in the Appendix.

5.3 Electrophysiologist-Driven Re-Prompting Experiment

On the same ECG dataset, our initial Sinus Rhythm prompts—designed in consultation with allied healthcare professionals—relied on a broader feature whether P wave is present or not. In response, the electrophysiologists distilled the definition of Sinus Rhythm to three essential, unambiguous criteria: (1) P waves present, (2) P waves upright in leads II/III/aVF (or II only for dual-lead recordings), and (3) P waves biphasic in V1. These features formed the basis of our focused re-prompting experiment.

Sinus Rhythm ECGs tested previously in Section 5.2 were filtered to include only those image–prompt pairs in which the model had marked P waves as present. We then re-submitted those same ECGs and prompts to each LLM (Llama and Qwen) with a minimal JSON prompt requiring confirmation of only the three electrophysiologist criteria. Four conditions were tested for correctly diagnosed ECGs as Sinus Rhythm and for ECGs that were falsely not identified as Sinus Rhythm: Llama on full 12-lead ECGs and dual-lead ECGs, and Qwen on full 12-lead ECGs and dual-lead ECGs.

Results for Originally Correctly Diagnosed ECGs Under Llama on 12-lead ECGs, the overall diagnosis accuracy was 5.56%. P wave presence was confirmed in 27.78% of cases, upright-P morphology in 16.67%, and biphasic-P morphology in 8.33%. 2.78% of the responses satisfied all three electro-physiological criteria while retaining the Sinus Rhythm label, and another 2.78% of the responses retained the label despite missing at least one criterion.

In dual-lead ECGs under Llama, the accuracy of diagnosis was 7.29%, the presence of P waves accuracy was 31.25%,

upright P in lead II accuracy was 14.58% and biphasic P in V1 accuracy was 18.75%. 7.29% of the responses met all three criteria and retained the Sinus label, while 3.12% met some but not all.

Under Qwen and in 12-lead ECGs, the diagnosis accuracy was 7.89%, the presence of P waves was 13.16%, upright-P was none and biphasic-P was 2.63%. 7.89% of the responses retained the Sinus label despite missing criteria, but none satisfied all three. Qwen returned empty JSON outputs in 57.89% of these cases, which were counted incorrect.

In the dual-lead ECG with the Qwen model, none of the criteria was met. Qwen returned empty responses in 100% of the cases that were all treated as incorrect.

For Llama model on the 12-lead subset, the Role Specification prompt reached a 6.25% overall diagnosis accuracy while confirming P wave presence was 31.25%, upright-P morphology 18.75%, and biphasic-P morphology 12.50%. The Chain-of-Thought prompt achieved a 2.94% overall diagnosis accuracy with P wave presence at 26.47%, upright-P morphology at 17.65%, and biphasic-P morphology at 5.88%. The Confidence Assessment prompt scored a 9.09% overall diagnosis accuracy with P wave presence at 27.27%, upright-P morphology at 13.64%, and biphasic-P morphology at 9.09%.

On the dual-lead subset, Role Specification achieved a 2.94% overall diagnosis accuracy with P wave presence at 32.35%, upright-P morphology at 17.65%, and biphasic-P morphology at 20.59%. Chain-of-Thought reported a 13.95% overall diagnosis accuracy with P wave presence at 41.86%, upright-P morphology at 16.28%, and biphasic-P morphology at 23.26%. Confidence Assessment did not produce any correct diagnoses while confirming P wave presence, upright-P morphology, and biphasic-P morphology each resulted in 5.26%.

Role Specification achieved with the Qwen model on the 12-lead subset, an 8.33% overall diagnosis accuracy with P wave presence at 8.33% and there was no confirmations of upright-P or biphasic-P morphology. Chain-of-Thought achieved no correct diagnosis with P wave presence at 14.29% and biphasic-P morphology at 7.14%. Confidence Assessment achieved 16.67% overall accuracy with P wave presence at 16.67% and no confirmations of upright or biphasic morphology.

On the dual-lead subset, none of the prompts produced any correct diagnoses or confirmed any of the three electrophysiologist-defined criteria.

Results for Originally Misdiagnosed ECGs Among the Llama 12-lead cases with P waves present but falsely labeled Non-Sinus, diagnosis accuracy was 3.57%, P wave presence 17.86%, upright-P 10.71%, and biphasic-P none. Only 3.57% of the responses were correctly labeled as Sinus despite missing at least 1 criteria.

For Llama on dual-lead misdiagnoses, diagnosis accuracy was none, P wave presence 17.65%, upright-P none, and biphasic-P 11.76%;

For Qwen on 12-lead misdiagnoses, diagnosis accuracy was 3.12%, P wave presence 3.12%, upright-P and biphasic-P none. Qwen returned empty responses in 78.12% of these cases which were considered wrong.

For Qwen on the dual-lead misdiagnoses, none of the criteria were met and Qwen returned empty outputs in 100% of these cases which were considered wrong.

In the 12-lead misdiagnosed subset, Role Specification with Llama did not produce any correct diagnosis while confirming the presence of P waves in 25% and upright morphology in 25%, without biphasic-P confirmations. Chain-of-Thought produced 10% overall diagnosis accuracy with P wave presence at 20%, upright-P morphology at 10.00%, and no biphasic-P confirmations. Confidence Assessment produced no correct diagnosis with P wave presence at 10% and no confirmations of upright-P or biphasic-P morphology. In the dual-lead subset, no prompt elicited any correct diagnoses, but P wave presence was confirmed in 75% and biphasic-P morphology in 50% under Chain-of- Thought Prompting.

For Qwen on the misdiagnosed 12-lead cases, both the Role Specification and Confidence Assessment prompts failed to produce any correct diagnoses or confirm any ECG features. Chain-of-Thought produced an overall accuracy of 7.14% with diagnosis and presence of P waves, without biphasic-P and upright-P confirmations. On the dual-lead misdiagnosed subset, none of the prompts produced any correct diagnoses or confirmed P wave presence, upright-P morphology, or biphasic-P morphology.

6 Discussion and Insights

Based on our experimental setup, our results highlight several important trends:

First, model architecture plays an important role. Llama consistently outperformed Qwen across both ECG configurations for AFib cases, showing more robust diagnostic performance and greater consistency in feature extraction. This advantage is particularly evident for AFib detection, where Llama paired with Chain-of-Thought prompting achieved 93.94% accuracy on 12-lead ECGs and 97.00% on dual-lead (Lead II and V1) configurations. However, for Sinus Rhythm detection, while Llama generally performed better when including empty responses of Qwen, Qwen slightly outperformed Llama in certain specific scenarios when its empty responses were excluded. For instance, Qwen achieved 48.15% (Role Specification) vs. Llama's 44.00% (Chain of Thought) in diagnosis accuracy on dual-lead ECGs. Nonetheless, both models exhibited significant limitations in Sinus Rhythm diagnosis and P wave detection accuracy for Sinus Rhythm cases.

Second, prompting strategies had a substantial influence on model performance, with effectiveness varying by rhythm type and model. For Llama, Chain-of-Thought consistently provided the highest diagnostic accuracy across AFib cases.

Third, lead configuration impacted each model differently. For Llama, reducing leads from twelve to two (Lead II and V1) preserved its high AFib diagnostic accuracy (84.85% $\rightarrow$ 84.33%) and led to a noticeable improvement in Sinus Rhythm detection (25.51% $\rightarrow$ 33.67%). Conversely, Qwen experienced a slight decrease in AFib diagnostic accuracy (8.87% $\rightarrow$ 7.46%) under the dual-lead setup and only

minor improvement in Sinus Rhythm diagnostic accuracy (13.71% $\rightarrow$ 14.90%), indicating minimal benefit from lead reduction for this model.

Fourth, although Llama achieved strong AFib diagnostic performance (84.85%) in this experiment, the PR Interval was consistently its weakest feature-level interpretation (79.12% accuracy), clearly identifying this interval as the main source of residual errors.

Despite Llama's relative success compared to Qwen in our experimental setup, these results are preliminary and exploratory. Neither Llama nor Qwen demonstrated sufficient reliability to be trusted for clinical diagnosis or feature-level interpretation at this stage. The observed better performance of Llama should be viewed cautiously and not as evidence of readiness for clinical deployment. This study's main goal was to highlight critical weaknesses inherent in vision-enabled LLMs when applied to ECG analysis.

While prompt engineering can guide LLMs towards more structured reasoning, it does not fundamentally alter their underlying knowledge. Fine-tuning on domain-specific ECG datasets offers the potential to internalize clinically relevant patterns and terminology, improving both rhythm classification and feature-level interpretations such as P waves and PR Interval. This can lead to more consistent justifications, reduced hallucinations, and better generalization across lead configurations. Fine-tuning particularly enables models to learn temporal and morphological relationships intrinsic to ECG signals—such as timing between waveforms or rhythm regularity—that are often missed with prompt-only approaches. The poor performance of the models after our electrophysiologist-guided re-prompting experiment reinforces the notion that these models lack sufficient internal understanding of electro-physiological criteria, suggesting that structural or representational tuning is essential. Therefore, fine-tuning represents a critical next step toward enhancing the clinical reliability of LLM-based ECG analysis.

7 Conclusion

This study evaluated the capabilities of vision-LLMs, specifically Llama-3.2-11B-Vision-Instruct and Qwen2-VL-7B-Instruct, for diagnosing and explaining AFib and Sinus Rhythm from ECG images, using structured prompting strategies across one run of full 12-lead and dual-lead (Lead II + V1) configurations.

Our findings confirm that both model architectures and prompting strategies significantly affect diagnostic accuracy and feature-level reasoning. Llama consistently outperformed Qwen, achieving higher accuracy for AFib detection—especially under Chain-of-Thought prompting—and demonstrating strong performance on key features such as P waves. However, its performance on Sinus Rhythm interpretation remained modest, and neither model reliably produced full multi-feature explanations. Qwen performed substantially worse overall, with poor diagnostic and feature-level accuracy across configurations. However, in limited instances when empty responses were excluded, Qwen narrowly outperformed Llama for Sinus Rhythm detection, indicating some conditional utility under specific prompting strategies.

Lead configuration also influenced results: reducing to Lead II and V1 preserved Llama's strong AFib performance while noticeably improving Sinus Rhythm detection accuracy, though overall accuracy for Sinus Rhythm remained relatively low, highlighting continued challenges for both models. In contrast, Qwen saw minimal positive impact from lead reduction.

Our electrophysiologist-driven re-prompting experiment further revealed that LLMs frequently failed to confirm basic morphological criteria—such as upright P waves in inferior leads or biphasic-P morphology in lead V1—highlighting a lack of deep waveform understanding and limiting trustworthiness.

These results highlight the potential of LLMs in ECG-based diagnosis yet our work underscores the need for enhanced feature-level reasoning and generalization. Future work should explore dynamic prompt mixing, fine-tuning on clinically annotated ECG datasets to include detailed waveform features, testing across broader datasets and conducting multiple experimental runs. Promising directions include evaluating models explicitly trained for ECG image understanding, and conducting lead-level failure analyses to inform dynamic attention mechanisms and targeted model refinement.

Most importantly, our results highlighted the inherent weaknesses faced by LLMs,including both diagnostic classification and waveform-level feature identification, within our experimental pipeline. Although promising diagnostic accuracy was observed—particularly with Llama—these findings must not be interpreted as clinical validation or justification for clinical deployment. Particularly, when tasked beyond diagnosis, these LLMs are not able to provide reliable feature identification. Additionally, the text-to-text mapping step performed by LLMs introduced another source of potential hallucination and inaccuracies due to ambiguous interpretations or overconfidence in the GenAI produced JSON files. These limitations reinforce the essential role of human-AI collaboration in clinical ECG analysis. In their current maturity, AI systems should be deployed alongside human oversight to ensure reliability, accuracy, and clinical utility.

Acknowledgements

The authors would like to acknowledge the support of the University Research Board, the AI, Data Science and Computing Hub, the VIP and the Maroun Semaan Faculty of Engineering and Architecture at the American University of Beirut for funding this research and providing the essential infrastructure needed to conduct this study.

Appendix

Prompts used for AFib

Prompt 1: Role Specification

As an experienced cardiologist, analyze the provided ECG image to determine if it indicates an AFIB or NON-AFIB case. Respond strictly in JSON format.

```
{
  "Diagnosis": "[AFIB/NON-AFIB]",
  "Rhythm Regularity": "[Regular/
      Irregularly Irregular/Regularly
      Irregular]",
  "P Waves": "[Present/Absent]",
  "PR Interval": "[Normal/Variable/
      Prolonged/Short/Not Measurable]"
}
```

Prompt 2: Chain-of-Thought Reasoning

Analyze the ECG step by step. Respond in JSON.

```
{
  "Diagnosis": "[AFIB/NON-AFIB]",
  "Rhythm Regularity": "[Regular/
      Irregularly Irregular/Regularly
      Irregular]",
  "P Waves": "[Present/Absent]",
  "PR Interval": "[Normal/Variable/
      Prolonged/Short/Not Measurable]"
}
```

Prompt 3: Confidence Assessment

Evaluate the ECG. Respond in JSON with confidence levels.

```
{
  "Diagnosis":{
   "Observation":"[AFIB/NON-AFIB]",
   "Confidence Level": "[High/Medium/Low]"
  },
  "Rhythm Regularity": {
    "Observation": "[...]",
    "Confidence": "[High/Medium/Low]"
  },
  "P Waves": {
    "Observation": "[...]",
    "Confidence": "[High/Medium/Low]"
  },
  "PR Interval": {
    "Observation": "[...]",
    "Confidence": "[High/Medium/Low]"
  }
}
```

Qwen Non-Empty Files Results Summary

Type	Prompt	Feature	2-Lead (%)	12-Lead (%)
AFib	Role Spec.	Diagnosis	15.00	24.39
		Rhythm Reg.	22.50	21.95
		P Waves	10.00	12.20
		PR Interval	15.00	17.07
	Chain-of-Thought	Diagnosis	8.33	17.39
		Rhythm Reg.	10.42	17.39
		P Waves	8.33	8.70
		PR Interval	8.33	15.22
	Confidence	Diagnosis	18.00	18.42
		Rhythm Reg.	28.00	36.84
		P Waves	22.00	7.89
		PR Interval	24.00	31.58
Sinus	Role Spec.	Diagnosis	48.15	27.91
		P Waves	55.56	51.16
	Chain-of-Thought	Diagnosis	21.43	30.00
		P Waves	47.62	54.00
	Confidence	Diagnosis	25.71	26.83
		P Waves	37.14	43.90

Table 1: Accuracy (%) of Qwen in AFib/Sinus Detection by Prompt and Lead Type (non-empty only).

References

Caffrey, M. 2023. Global Burden of Atrial Fibrillation Rises Sharply Over 30 Years, Study Finds. *The American Journal of Managed Care*.

Heidbuchel, H.; Verhamme, P.; Alings, M.; Antz, M.; Diener, H.-C.; Hacke, W.; Oldgren, J.; Sinnaeve, P.; Camm, A. J.; Kirchhof, P.; and Group, E. S. D. 2016. Updated European Heart Rhythm Association practical guide on the use of non-vitamin-K antagonist anticoagulants in patients with non-valvular atrial fibrillation: Executive summary. *European Heart Journal*, 38(27): 2137–2149.

January, C. T.; Wann, L. S.; Calkins, H.; Chen, L. Y.; Cigarroa, J. E.; Cleveland, J. C.; Ellinor, P. T.; Ezekowitz, M. D.; Field, M. E.; Furie, K. L.; Heidenreich, P. A.; Murray, K. T.; Shea, J. B.; Tracy, C. M.; and Yancy, C. W. 2019. 2019 AHA/ACC/HRS Focused Update of the 2014 AHA/ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society in Collaboration With the Society of Thoracic Surgeons. *Circulation*, 140(2): e125–e151.

Koga, S.; and Du, W. 2025. From text to image: challenges in integrating vision into ChatGPT for medical image interpretation. *Neural Regeneration Research*, 20(2): 487–488. Epub 2024 Apr 3.

Lippi, G.; Sanchis-Gomar, F.; and Cervellin, G. 2021. Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. *International Journal of Stroke*, 16(2): 217–221. Epub 2020 Jan 19.

Medical Xpress. 2024. GPT-4 with vision shows poor accuracy in medical image interpretation. Accessed: 2025-03-29.

Noubiap, J. J.; Tang, J. J.; Teraoka, J. T.; Dewland, T. A.; and Marcus, G. M. 2024. Minimum National Prevalence of Diagnosed Atrial Fibrillation Inferred From California Acute Care Facilities. *JACC*, 84(16): 1501–1508.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617.

Tan, S.; Zhou, J.; Veang, T.; Lin, Q.; and Liu, Q. 2025. Global, regional, and national burden of atrial fibrillation and atrial flutter from 1990 to 2021: sex differences and global burden projections to 2046—a systematic analysis of the Global Burden of Disease Study 2021. *EP Europace*, 27(2): euaf027.

Turner, A.; Kaushik, M.; Huang, M.-T.; and Varanasi, S. 2020. Calibrating Trust in AI-Assisted Decision Making. Technical report, UC Berkeley School of Information.

Wagner, P.; Strodthoff, N.; Bousseljot, R.; Samek, W.; and Schaeffter, T. 2022. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). https://doi.org/10.13026/kfzx-aw45. PhysioNet.

Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Kreiseler, D.; Lunze, F. I.; Samek, W.; and Schaeffter, T. 2020. PTB-XL: A Large Publicly Available ECG Dataset. *Scientific Data*, 7: 154.

Wu, C.-K.; Chen, W.-L.; and Chen, H.-H. 2023. Large Language Models Perform Diagnostic Reasoning. arXiv:2307.08922.