Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Information Retrieval

Assignment 5: LLMs applications for a concrete medical problem

Authors:

José Antonio Ruiz Heredia

Joseph Tartivel

Álvaro Honrubia Genilloud

Teacher:

Victor Maojo

Date:

April 6, 2025

Contents

1 Introduction

AI is gradually finding its place in medicine, with large language models (LLMs) like ChatGPT even passing parts of the US Medical Licensing Exam (USMLE) without any specialized training[1], a challenge that requires years of study for human doctors. These models are beginning to assist with various medical tasks, such as diagnostics, writing medical documents, and supporting research, which can help make certain aspects of healthcare more efficient. LLMs are also being used to draft reports, develop training tools, and speed up research processes. However, despite these advancements, there are still significant concerns. Experts caution that LLMs need to be carefully evaluated for reliability, as they can sometimes provide incorrect or misleading information. While the potential for improving healthcare is clear, it's important not to overlook the risks, especially when it comes to patient safety and quality of care. [2]

Cardiovascular diseases remain a major global health issue, with millions of deaths each year. Despite advancements in treatment, cardiovascular diseases continues to be difficult to manage, especially due to the complex interactions between risk factors like high blood pressure, diabetes, and cholesterol. There is a big need for innovation to better understand these complexities and improve outcomes. LLMs hold promise in this area by enabling faster data analysis, identifying patterns, and uncovering potential risk factors. They can process large volumes of clinical, genetic, and trial data, potentially accelerating diagnostics and improving treatments.[3]

LLMs have various promising applications to cardiovascular disease. By integrating electronic health records, imaging data, and genomic information, these models could help identify patients at higher risk of cardiovascular events, allowing early treatment. Additionally, LLMs can assist in generating personalized treatment recommendations by analyzing datasets from clinical trials and real-world data. Despite their potential in diagnosing and treating cardiovascular disease, LLMs raise key concerns. They can generate incorrect or misleading conclusions, when misdiagnosis can have serious consequences. Bias in training data may also lead to inaccurate recommendations, affecting certain patient groups unfairly. Additionally, LLMs often function as "black boxes," making it difficult for doctors to understand or trust their decisions. Without transparency and validation, their real-world reliability remains uncertain.

In this work, we will explore the possible applications of LLMs to a concrete medical problem: a cardiovascular disease. To do so we will analyze and compare three LLMs. By studying their reasoning, limitations, and ethical concerns that result from it.

2 Cardiovascular disease choice

2.1 Introduction to cardiovascular diseases

Cardiovascular diseases (CVDs) are a group of problems that affect the heart and blood vessels. Some common types of CVDs are coronary artery disease, heart failure, arrhythmias, and stroke. Coronary artery disease happens when the blood vessels around the heart get blocked due to fat deposits. Heart failure occurs when the heart can't pump blood well enough, which can be caused by things like heart attacks or high blood pressure. Arrhythmias are problems with the heart's rhythm, making it beat too fast, too slow, or irregularly. Strokes happen when the blood supply to the brain is disrupted, either by blockage or rupture in the blood vessels. [4]

Even if there have been many advancements in understanding and treating cardiovascular diseases, some areas are still challenging. For example, treatments for coronary artery disease have improved[5], but it's still difficult to predict and treat heart failure and arrhythmias effectively. Researchers are still studying the best ways to manage arrhythmias and heart failure. Current treatments include medications and procedures like catheter ablation, but more research is needed to find more effective approaches.[6]

2.2 Focus on arrhythmias

We chose to focus on arrhythmias, irregular heart rhythms that can lead to death. The most common types are atrial fibrillation (AF), ventricular tachycardia, and bradycardia. AF is the most common and is often associated with high blood pressure, heart disease, and diabetes. It can lead to strokes and heart failure if not managed properly. Ventricular tachycardia is another serious arrhythmia that can lead to sudden cardiac arrest if not treated. Bradycardia, on the other hand, occurs when the heart beats too slowly, which can lead to dizziness, fainting, or heart failure in extreme cases. There are many causes of arrhythmias, including structural heart problems, electrolyte imbalances, genetic conditions, and lifestyle factors such as stress, alcohol consumption, and lack of physical activity. Data from electrocardiograms (ECGs), blood tests, and patient medical histories are the key to diagnosing and managing these conditions. However, early detection is often difficult due to the unpredictability and variability of arrhythmia patterns. It is essential to identify and monitor arrhythmias through continuous data analysis to prevent more serious complications.[7]

Because they are complex and unpredictable, heart conditions can be difficult to detect early. LLMs could help by analyzing large amounts of patient data, like ECG results and medical histories, to identify patterns and irregularities that may signal arrhythmias. They can assist in early detection, which is crucial for preventing severe complications, and help personalize treatment plans by suggesting therapies based on a patient's unique medical history. This ability to provide real-time analysis and tailor treatments can significantly improve decision-making and overall patient outcomes in arrhythmia management, making it a perfect choice for our exploratory work.

3 LLM choice

3.1 Introduction to large language models

Large Language Models (LLMs) are artificial intelligence systems designed to understand, generate, and interact with human language. They are trained on big amounts of text data to predict the next word in a sentence, which allows them to generate coherent responses and so simulate intelligence. In medicine, LLMs are increasingly being used to assist with tasks such as diagnosing diseases, drafting medical reports, and supporting research. For instance, they can analyze patient data to suggest possible diagnoses or treatment options, helping healthcare professionals make more informed decisions. However, while LLMs show promise in streamlining certain medical processes, they are not without flaws. They can sometimes provide incorrect or misleading information, and their lack of deep understanding can lead to errors in complex medical scenarios. Additionally, there are concerns about the reliability and ethical use of these models in healthcare settings.[8]

3.2 Choice justification

For our exploration of LLMs in the context of cardiovascular diseases, we have chosen to focus on ChatGPT by OpenAI, LeChat by Mistral, and DeepSeek. ChatGPT is selected for being the most popular model. It has demonstrated strong performance in various medical tasks, making it a valuable tool for our analysis. In a recent general benchmark study evaluating the performance of LLMs across different domains (MMLU)[9], ChatGPT 4o mini scored 82%[10], showcasing its ability to provide accurate and contextually relevant responses. Mistral 8x7b, a smaller model, is included for its efficiency and potential to reduce the environmental impact of AI, which is important considerations as the use of these models becomes more widespread. In the same benchmark, Mistral 8x7b scored 70% while using fewer computational resources, making it a sustainable choice for large-scale applications. DeepSeek R1, a "low-cost" reasoning model, that could be particularly beneficial for complex medical problems. In the benchmark, DeepSeek's R1 showed promise in handling complex medical queries, having better results than ChatGPT in raw performance metrics: 90%. For this work we will use the online platform from these company and so the precise model choice was impossible for LeChat because the platform choose for you what model to use, so we cannot assure that 8xb7 was used for the following work.

The choice of these models involves complex technological and geopolitical considerations. OpenAI, based in the US, has benefitted from large investment and industry support, enabling it to lead the AI market. However, recent events have led European countries to reconsider their dependence on US software and cloud services, highlighting concerns about reliance on American companies. DeepSeek, a Chinese company, offers a highly efficient model that competes with OpenAI, but raises significant concerns regarding privacy, ethics. Mistral AI, based in France, provides a European alternative, focusing on sustainability and ethical AI development. However, the question remains: is their model advanced enough for medical applications, or does it fall short compared to its US and Chinese counterparts? As European countries weigh their options, they face a dilemma. Should they continue depending on the US or China, or risk missing a significant opportunity to leverage AI for improving medical diagnosis and treatment? By examining these models, we aim to offer a view of LLM use in medicine, highlighting both the potential and challenges that must be addressed, from privacy and ethics to geopolitical concerns.

4 Reasoning capabilities

4.1 Definition

Reasoning in large language models refers to their ability to process information, draw logical conclusions, and make recommendations based on available data. For cardiovascular diseases like arrhythmias, this capability is essential as diagnosis often requires connecting complex patterns across various data points including ECG readings, patient histories, and clinical measurements. Models like ChatGPT, LeChat, and DeepSeek are three different types of LLMs and so, significant differences could emerged in their approach to complex cases. When evaluating LLM performance on medical reasoning tasks, several studies have shown varying capabilities across different models. A study found that larger models typically provide more comprehensive medical reasoning with better grounding in established literature, while smaller models often offer more concise analyses that may lack depth in complex cases.[11] Meanwhile, another demonstrated that certain specialized models show particular strength in connecting information across multiple sources, potentially identifying correlations not explicitly stated in any single reference.[12] These findings suggest that different LLMs may have distinct reasoning strengths and limitations when applied to cardiovascular disease analysis.

4.2 Case study experiment

4.2.1 Methodology

To evaluate the reasoning capabilities of the LLMs in our specific case, we developed a small experiment using two publicly available arrhythmia case studies from the website NetCe [13]. Here is a short summary of the case study by Chat GPT 4o mini: *The cases chosen covered a common arrhythmias (atrial fibrillation with rapid ventricular response) as well as a more challenging diagnoses (Atrial Fibrillation with a Slow Ventricular Response). Each case contained a clinical vignette, text-based ECG findings, and relevant patient history. The first case described a 68-year-old male with hypertension and coronary artery disease presents with shortness of breath and palpitations. His ECG shows an irregularly irregular rhythm, narrow QRS complex tachycardia, and fibrillatory waves, consistent with atrial fibrillation with rapid ventricular response. The second was a 65-year-old male with hypertension presents with palpitations. The ECG shows an irregularly irregular rhythm, no P waves, and a ventricular rate of 110-150 bpm, indicating atrial fibrillation.*[14] We presented the same cases to all three models using standardized prompts that requested: 1) a diagnosis, 2) an explanation of their reasoning process, 3) a discussion of differential diagnoses, and 4) suggestions for potential management approaches.

As non-medical students, we couldn't judge the answers of the models so we thought of an evaluation process based on another LLM to judge the medical accuracy and reasoning quality of the models' responses. Conscious that this method present important limitation, we stayed critic about both the answers and the evaluation of the LLMs. We tested the models' reasoning by asking Claude Sonnet 3.7, another LLM, to evaluate the quality of the answers of the three studied LLMs. Again with a standardized prompt.

4.2.2 Results

For the atrial fibrillation case mentioned above, this is how Claude summarized the evaluation of the different models: *ChatGPT 4.5 correctly identified the key diagnostic features—such as the irregularly irregular rhythm and absence of P waves—before discussing treatment options like rate control and the use of diltiazem, supported by clinical guidelines. Mistral 8x7b diagnosed the arrhythmia correctly but did not delve as deeply into the pathophysiology or specific treatment options. DeepSeek R1 recognized atrial fibrillation and acknowledged the patient's comorbidities, like hypertension and coronary artery disease, in relation to the arrhythmia's development, showcasing its ability to integrate clinical context.*[15]In this case, the different LLMs demonstrated varying levels of reasoning. ChatGPT provided a really good analysis of the situation. LeChat correctly diagnosed the arrhythmia but had a more simple justification. DeepSeek R1 gave a good analysis and showed its ability to connect clinical context to the diagnosis, this is certainly because it's a reasoning model.

In the second case, Claude summarized the evaluation of the different models as follows: *ChatGPT accurately identified the key diagnostic features of atrial fibrillation, such as the irregular rhythm and absence of P waves, and discussed treatment options, including rate control and rhythm control, referencing common risk scores like CHA_2DS_2-VASc. Mistral correctly diagnosed the arrhythmia but provided less detailed reasoning, particularly regarding treatment options and guidelines. DeepSeek also identified atrial fibrillation and linked the patient's hypertension to an increased risk of the condition, though its explanation of management strategies was less comprehensive.*[16] Theses results are again close to what we had in the first case, the only change is DeepSeek failing to provide a clear explanation (As judge by Claude).

4.2.3 Conclusion of the experiment

There are important limitations to our study. First, as non-medical students, we could not independently verify all aspects of the clinical accuracy of the models' outputs and verify the evaluation from Claude, even with serious online research. Second, the test cases were based only on text descriptions rather than actual ECG, leaving out the visual analysis that is often crucial in real-world diagnosis. Third, since the models could not interact with patients or request new tests, they were confined to the information provided in the prompts. The clinical diagnosis was not realist at all. These limitations point to the need for a better evaluation by medical experts using more complete clinical data, as well as new cases studies.

Despite these limitations, our experiment gives an idea of the reasoning capabilities and potential applications of the models. While we cannot fully confirm their clinical accuracy without asking an expert, the overall patterns in their reasoning suggest that each model has unique strengths. ChatGPT use a clear method and provide the best explanation in terms of clarity. This model is certainly the most versatile. DeepSeek is really good at connecting different concepts, this is certainly due to the fact that it can take more token and more time before giving his answers. This may be useful for exploring complex cases with multiple relation between factors. Mistral is good at giving simple diagnosis but may give limited explanation, this is certainly cause by the small size of the model (in terms of parameter).

5 Limitation

While Large Language Models show promise in the field of cardiovascular medicine, particularly in arrhythmia diagnosis and management, they face significant limitations.

5.1 Data access and interaction limitations

One of the most important limitation is the lack of access to real-time patient data. LLMs can only work with the information explicitly provided to them, making it impossible for them to request additional tests or examinations that might be crucial for accurate diagnosis. In real clinical settings, physician often need to order additional ECGs, blood tests, or imaging studies based on the first findings. The inability of LLMs to interact directly with patients or medical devices creates a gap between theoretical capabilities and practical utility.

Another critical aspect of this limitation is the inability of current LLMs to properly interpret visual data. ECGs, a fundamental tool in arrhythmia diagnosis, contain visual patterns that require specialized training to interpret correctly. While some models may be able to process text descriptions of ECG findings, they cannot directly analyze the visual ECG tracings themselves. Like some machine model can. This creates a significant gap in their diagnostic capabilities, as subtle visual abnormalities on an ECG can be crucial for accurate diagnosis but difficult to capture completely in text descriptions. This complexify the system and potentially create bias.

5.2 Training data and bias issues

A major limitation concerns the quality and representation of training data. LLMs like ChatGPT, LeChat, and DeepSeek learn from vast text that may not represent diverse patient populations or rare cardiovascular conditions. This can lead to significant biases in their recommendations, potentially resulting in less effective care for underrepresented groups. For example, arrhythmias may present differently in women compared to men, or in patients from different ethnic backgrounds. If these variations are not well-represented in training data, LLMs might miss important diagnostic cues or suggest inappropriate treatments for certain patient groups.[17]

Medical knowledge evolves rapidly, but LLMs have fixed training cutoff dates and most of them cannot automatically update their knowledge base with new research findings, updated guidelines, or new approved medications. For instance, if a new drug for atrial fibrillation is developed after a model's training period, the LLM would be unaware of these advancements unless specifically updated. This creates a significant risk of providing outdated recommendations.[18]

5.3 Transparency and reasoning limitations

The "black box" nature of these models presents a serious limitation in medical applications. Healthcare professionals need to understand the reasoning behind diagnostic and treatment recommendations to make informed decisions. However, even if LLMs provides textual justification for the diagnosis we are not sure that their justification was the real reasoning behind the results, because often they cannot explain the specific weights they assign to different factors or clearly articulate their decision-making process in a way that aligns with medical reasoning. This lack of transparency makes it difficult for clinicians to trust or verify the models' outputs, especially in complex arrhythmia cases where multiple factors need to be considered simultaneously.

LLMs also struggle with temporal reasoning and incorporating changes in patient status over time. Arrhythmias can change in response to treatments. But current models have limited ability to track and incorporate these temporal patterns into their reasoning, which is essential for proper management of conditions like atrial fibrillation. They work primarily with static information provided at a single point in time, missing the dynamic nature of cardiovascular disease progression.

5.4 Technical constraints

The limited context window of these models represents another significant constraint. Complex arrhythmia cases often involve extensive patient histories, multiple diagnostic tests, and detailed medication records. Even advanced models like DeepSeek R1, that have larger context windows than others, can still struggle to process and integrate all relevant information from a patient's complete medical record. This limitation forces users to carefully select which information to include in prompts, potentially omitting details that could be diagnostically significant.

Technical issues also limit practical implementation. LLMs require significant computational resources, making them expensive to run and potentially inaccessible in resource-limited healthcare settings. The environmental impact of running these models at scale also raises sustainability concerns, especially for more powerful models that require greater energy consumption. Issues of data privacy and security further complicate their integration into healthcare systems, as patient information used for prompts must be protected according to various regulations.

5.5 Practical implementation challenges

For these reasons, LLMs should be viewed as tools rather than replacements for trained medical staff. Their limitations necessitate a collaborative approach where human clinicians maintain final responsibility for diagnostic and treatment decisions, using LLM outputs as one of many information sources to consider. Future development should focus on addressing these limitations through improved transparency, better integration with clinical workflows, and rigorous validation in real-world cardiovascular care settings.

The integration of LLMs into existing healthcare systems presents practical challenges beyond the models themselves. Current electronic health record systems are not designed to work seamlessly with LLMs, creating workflow disruptions when trying to incorporate these tools. Additionally, healthcare providers would require specialized training to effectively prompt these models and critically evaluate their outputs, presenting an educational burden that many busy clinicians may not have time to overcome. These practical barriers must be addressed alongside the technical limitations to realize the potential benefits of LLMs in cardiovascular care.

6 Ethical issues

The application of Large Language Models in cardiovascular disease management, particularly for arrhythmias, raises significant ethical concerns that must be treated before implementation. These ethical issues extend beyond technical limitations and touch on fundamental questions about the role of AI in healthcare decision-making, patient autonomy, and equitable access to care.

6.1 Patient autonomy and consent

The first obvious concern is whether patients are informed about the use of LLMs in their care. Unlike traditional clinical decision support tools, LLMs the ones with studied operate with complex, opaque mechanisms that even healthcare providers may not fully understand. This raises important questions about informed consent: Can patients truly consent to have an LLM involved in their diagnosis or treatment planning if neither they nor their doctors fully understand how the model reaches its conclusions. The ”black box” nature of these models, as highlighted in our limitations section, directly impacts the trust of patient in their physicians.[19][20]

In the context of arrhythmia management, where quick decisions may be necessary in emergency situations, there may be additional complications regarding when and how to obtain informed consent. If an LLM suggests a particular intervention for a life-threatening arrhythmia, the time constraints may not allow for an explanation of the technology’s role to patient. This can creates tension between patient awareness and practical realities of emergency.

6.2 Responsibility

The question of who bears responsibility for errors made by LLMs in diagnosing or suggesting treatments for arrhythmias remains largely unresolved. If a model like ChatGPT misinterprets ECG findings described in a prompt and suggests an inappropriate treatment that leads to patient harm, determining who caused the error becomes complex. Is the AI company responsible? The physician who used the tool? The institution that implemented the system?

Our case studies showed varying levels of reasoning quality across different models, with none achieving perfect accuracy. In clinical practice, these inconsistencies could lead to serious consequences. Current legal and ethical frameworks are not fully equipped to address these questions of responsibility when AI systems participate in medical decision-making. This accountability gap is particularly concerning for cardiovascular conditions like arrhythmias, where incorrect decisions can have immediate and potentially fatal consequences.

6.3 Equity and access issues

The implementation of LLMs in healthcare risks to increase existing healthcare disparities. As noted in the limitations section, these models are trained on data that may underrepresent certain populations. For arrhythmias, this is particularly problematic as presentation can vary significantly based on factors like sex, age, and ethnicity. If LLMs perform better for well-represented groups and worse for underrepresented populations, they could inadvertently worsen health outcomes for already marginalized groups.

Additionally, access to LLM-enhanced care may be limited to well-resourced healthcare systems due to the computational requirements and implementation costs. This creates a scenario where the potential benefits of these technologies may be available only to privileged populations, widening the gap in care quality between high and low-resource populations. The geographical distribution of this technology could follow existing patterns of healthcare inequality, with rural and underserved areas having less access to these advanced diagnostic tools.

6.4 Dehumanization of care

The integration of LLMs into cardiovascular care can create dehumanization of the doctor-patient relationship. Arrhythmias can cause significant anxiety and distress for patients, who often need emotional support during medical intervention. Our analysis of different LLMs showed that while they can provide technical information about arrhythmias, they cannot replace the human judgment that considers a patient's specific circumstances, preferences, and quality of life considerations. There is a risk that overreliance on these models could reduce medicine to a purely technical exercise, losing the approach that considers the patient as a person rather than a collection of symptoms and data points.

6.5 Professional impact and medical authority

The introduction of LLMs into cardiovascular care may also affect the professional identity and authority of cardiologists and other specialists. If these models perform well on diagnostic tasks, as suggested by their performance on the USMLE and in our case studies, we still question the unique value that human specialists bring to arrhythmia management. This could lead to deskilling of medical professionals, with doctors potentially losing certain diagnostic abilities if they routinely defer to LLM suggestions.

There are also concerns about how LLMs might influence the power dynamics in medical decision-making. If a model suggests a particular treatment approach that differs from a doctor's judgment, patients might question their doctor's expertise, potentially hurting between those two. The authority of medical professionals could be challenged if patients or administrators place to much confidence in LLM outputs, despite the significant limitations we have identified.

6.6 Tension between innovation and safety

Perhaps the most fundamental ethical tension is between innovate and improve cardiovascular care and the imperative to protect patients from potential harms. Our analysis suggests that while these models show promise, they are far from perfect currently. The ethical question then becomes: at what point is it acceptable to implement these technologies in real clinical settings? What performance justifies their use when human lives are at stake?

For arrhythmias, where timely and accurate diagnosis can be life-saving, this tension is particularly rising. Waiting for perfect systems might delay the benefits of improved care for many patients, but implementing imperfect systems too quickly could lead to preventable harm. Balancing these considerations requires dialogue between technologists, physician, politics and patients themselves. A conversation that has only just begun.[21]

7 Discussion

7.1 Summary of findings

Our exploration of LLMs for arrhythmia management reveals both significant potential and serious limitations. The experiment we conducted, while simplified, showed that current models like ChatGPT, LeChat, and DeepSeek can correctly identify common arrhythmias from text descriptions and provide reasonably sound management suggestions. However, the varying quality of explanations across models highlights the uneven development of these technologies. ChatGPT provided more comprehensive analyses while Mistral's explanations were more limited, reflecting differences in model size and training approaches.

What stands out most clearly from our work is that LLMs should be viewed as tools to augment human clinicians rather than replace them. The black box nature of these models, their inability to directly interact with patients or analyze visual ECG data, and their potential biases all severely limit their standalone utility in real clinical settings. The ethical considerations we identified show the importance of keeping humans in the decision-making loop when using these technologies.

The choice between US, Chinese, or European models represents more than just a technical decision. It involves considerations of data sovereignty, privacy standards, and long-term technological independence. Our experiment suggests that the European model from Mistral is not as good as its American and Chinese competitors. Therefore there Europe as a choice to make. They can decide to use non-european model and face privacy risks and over-reliance on other countries. They can choose to use Mistral's model and then not provide the best tool to its healthcare providers. Or they can invest in AI and try to compete with OpenAI and DeepSeek. The choice is up to Europe's political leaders.

7.2 Addressing limitations and ethical concerns

Several approaches could help address the limitations and ethical concerns we identified. The black box nature of these models could be partially mitigated through the development of more transparent LLMs that provide clearer explanations of their reasoning process. Some research is already being conducted on "explainable AI" approaches that could make the decision-making process more transparent to both clinicians and patients, helping to build trust and ensure informed consent. [22][23]

The issue of responsibility remains particularly problematic. If an LLM misinterprets a case description and suggests inappropriate treatment, who is responsible for the negative outcomes? A potential solution would be the development of a frameworks that define roles and responsibilities when AI systems are used in clinical decision-making. These frameworks would need to balance innovation with patient protection and could include requirements for human oversight of all AI-generated recommendations.

The risk of increasing the unequal access to technologies should be addressed through efforts to train models on diverse datasets that represent various demographics. This can be included in the framework discussed just before. We can also prioritize the access to LLMs to underserved communities rather than well-resourced communities.

The static nature of these models' knowledge represents another significant limitation. Medicine evolves rapidly, with new treatments and guidelines emerging regularly. This could be addressed through the development of systems for regular model updates that incorporate new medical research and guidelines. Alternatively, models could be designed to acknowledge when their information might be outdated and prompt clinicians to check more current sources.

Technical integration with existing healthcare systems presents practical challenges that must be overcome for successful implementation. User-friendly interfaces that fit into existing clinical workflows rather than disrupting them would be essential for adoption. This might include developing specialized prompting templates for arrhythmia management that help clinicians efficiently provide the most relevant information to the models.

7.3 Potential applications in arrhythmia management

Despite these challenges, LLMs show promise for several applications in arrhythmia management. First, they could serve as educational tools for medical students and healthcare professionals, helping them understand complex arrhythmia patterns and treatment approaches through interactive case discussions. Our experiment demonstrated how these models can analyze case studies and suggest diagnostic considerations, which could be valuable for training purposes. The educational application carries lower risk than direct clinical use while still providing significant value.

In clinical practice, LLMs could function as decision support tools, helping physicians double-check their diagnostic reasoning by offering a "second opinion" based on text descriptions of patient presentations and ECG findings. This could be particularly helpful in resource-limited hospitals where specialist cardiologists might not be always available. The models' ability to process large amounts of information quickly could help highlight important factors in complex cases that human physicians might overlook due to time constraints or cognitive biases.

For research applications, LLMs could help analyze large volumes of cardiovascular literature, identifying patterns and correlations across studies that might inform new approaches to arrhythmia management. They could assist in protocol development for arrhythmia research, generating hypotheses worth exploring, or identifying potential gaps in current knowledge. The research application leverages the models' ability to process and synthesize large amounts of text-based information, which is one of their core strengths.

Patient education represents another promising application. Simplified versions of these models could help explain complex arrhythmia diagnoses and treatment plans in accessible language, potentially improving treatment adherence and patient satisfaction. Our experiment showed that models like ChatGPT can provide clear explanations of arrhythmias that could be adapted for patient education materials. Using LLMs to generate personalized educational content could help patients better understand their condition and treatment options.

LLMs might also help with medical documentation, drafting initial reports based on clinical findings which physicians could then review and modify. This could reduce the administrative burden on healthcare professionals, allowing them to spend more time on direct patient care. The documentation application takes advantage of the models' natural language generation capabilities without requiring the level of diagnostic precision needed for direct clinical decision-making.

In low-resource settings where specialized cardiac care is limited, these models could potentially help bridge knowledge gaps by providing guidance to general practitioners faced with complex arrhythmia cases. However, this application would require careful implementation and safeguards to ensure safe and appropriate use, perhaps through specialized versions of the models designed for this specific purpose.

7.4 Future directions

The tension between innovation and patient safety remains central to any discussion of LLMs in healthcare. Moving too quickly risks to harm patient and healthcare systems, while moving too slowly may delay potential benefits. Finding the right balance requires collaboration between AI researchers, medical professionals, ethicists, and regulatory bodies. It also requires rigorous real-world testing beyond the limited experiment we conducted.

The development of multimodal models that can directly analyze ECG tracings alongside textual information could address some current limitations. Greater transparency in model reasoning and specialized training on diverse cardiac datasets could also improve performance and trustworthiness. Integrating these models with electronic health record systems in thoughtful ways could help overcome some practical implementation barriers.

The varying strengths we observed across different models suggest that different types of LLMs might be suited for different applications in cardiovascular care. Larger models with more comprehensive reasoning capabilities might be better for complex diagnostic assistance, while smaller, more efficient models could be valuable for patient education or documentation support.

As this field continues to evolve rapidly, striking the right balance between innovation and patient safety will remain the central challenge. The decisions made today about how to develop, regulate, and implement these technologies will shape their impact on cardiovascular care for years to come, making thoughtful consideration of both their promise and their limitations essential.

8 Conclusion

LLMs show promising potential in supporting arrhythmia management through educational tools, clinical decision support, research assistance, and patient education. However, significant limitations remain, including the inability to directly analyze visual data, limited transparency in reasoning, potential biases in training data, and complex ethical concerns around responsibility and equity.

Our experiment, despite its limitations, demonstrated that current models can correctly identify common arrhythmias from text descriptions, with varying degrees of explanatory quality. However, the path to safe clinical implementation remains long and will require addressing numerous technical, ethical, and practical challenges.

The future of LLMs in cardiovascular care will depend on technological advancements that improve transparency and visual analysis capabilities, careful integration into clinical workflows, development of clear ethical and legal frameworks, and ongoing evaluation in real-world settings. Rather than viewing these models as replacements for human expertise, their greatest value likely lies in augmenting human clinicians, reducing administrative burdens, and improving education for both patients and healthcare providers.

References

[1] Dana Brin, Vera Sorin, Eli Konen,Girish Nadkarni, Benjamin S. Glicksberg, Eyal Klang (2024).
"How GPT models perform on the United States medical licensing examination: a systematic review" *Discov Appl Sci 6, 500*
Available: https://doi.org/10.1007/s42452-024-06194-5

[2] Xiangbin Meng and Xiangyu Yan and Kuo Zhang and Da Liu and Xiaojuan Cui and Yaodong Yang and Muhan Zhang and Chunxia Cao and Jingjia Wang and Xuliang Wang and Jun Gao and Yuan-Geng-Shuo Wang and Jia-ming Ji and Zifeng Qiu and Muzi Li and Cheng Qian and Tianze Guo and Shuangquan Ma and Zeying Wang and Zexuan Guo and Youlan Lei and Chunli Shao and Wenyao Wang and Haojun Fan and Yi-Da Tang. "The application of large language models in medicine: A scoping review." *iScience, 2024).* Available: https://doi.org/10.1016/j.isci.2024.109713

[3] Wikipedia contributors. "Cardiovascular disease" *Wikipedia, 2025).* Available: https://en.wikipedia.org/w/index.php?title=Cardiovascular_disease&oldid=1268136383

[4] Henry C. McGill, Jr, MD, C. Alex McMahan, PhD, and Samuel S. Gidding, MD. "Preventing Heart Disease in the 21st Century: Implications of the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Study" *Circulation, 2008).* Available: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.717033

[5] "Coronary Heart Disease Research" *NHLBI, 2025).* Available: https://www.nhlbi.nih.gov/research/coronary-heart-disease

[6] Stanley Nattel, Jason Andrade, Laurent Macle, Lena Rivard, Katia Dyrda, Blandine Mondesert, Paul Khairy. "New Directions in Cardiac Arrhythmia Management: Present Challenges and Future Solutions" *Canadian Journal of Cardiology, 2014).* Available: https://doi.org/10.1016/j.cjca.2014.09.027

[7] Wikipedia contributors. "Cardiovascular disease" *Wikipedia, 2025).* Available: https://en.wikipedia.org/wiki/Arrhythmia

[8] Wikipedia contributors. "Large language model" *Wikipedia, 2025).* Available: https://en.wikipedia.org/wiki/Large_language_model

[9] Hendrycks, Dan and Burns, Collin and Basart, Steven and Zou, Andy and Mazeika, Mantas and Song, Dawn and Steinhardt, Jacob. "Measuring Massive Multitask Language Understanding" *arXiv, 2020).* Available: https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

[10] OpenAI. "GPT-4o mini: advancing cost-efficient intelligence" *OpenAI, 2024.* Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[11] Zabir Al Nazi, Wei Peng. "Large Language Models in Healthcare and Medical Domain: A Review." *Informatics, 2024).* Available: https://doi.org/10.3390/informatics11030057

[12] Hanjie Chen, Zhouxiang Fang, Yash Singla, Mark Dredze. “Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions” *Rice University, Johns Hopkins University, 2024)* Available: https://doi.org/10.48550/arXiv.2402.18060

[13] NetCE. “Clinical Management of Atrial Fibrillation.” *NetCE, n.d.*. Available: https://www.netce.com/casestudies.php?courseid=2699

[14] OpenAI. “Summary of the case study by Chat GPT 4o mini” *OpenAI*, 2025. Available: https://openai.com

[15] Anthropic. “Summary of the evaluation of the first case study by Claude Sonnet 3.7” *Anthropic*, 2025. Available: https://Anthropic.com

[16] Anthropic. “Summary of the evaluation of the second case study by Claude Sonnet 3.7” *Anthropic*, 2025. Available: https://Anthropic.com

[17] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science, 2019.* Available: https://doi.org/10.1126/science.aax2342

[18] Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. “Using Recurrent Neural Network Models for Early Detection of Heart Failure Onset.” *Journal of the American Medical Informatics Association, 2016.* Available: https://doi.org/10.1093/jamia/ocw112

[19] Goodman, B., and Flaxman, S. “European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”.” *AI Magazine, 2017.* Available: https://doi.org/10.1609/aimag.v38i3.2741

[20] Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. “What Do We Need to Build Explainable AI Systems for the Medical Domain?” *arXiv, 2017.* Available: https://arxiv.org/abs/1712.09923

[21] Anuradha Mathrani, Teo Susnjak, Gomathy Ramaswami, and Andre Barczak. “Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics.” *Computers & Education: Artificial Intelligence*, 2021. Available: https://doi.org/10.1016/j.caeo.2021.100060

[22] Yujia Zhou, Yifan Peng, and Zhiyong Lu. “Explainable Biomedical Claim Verification with Large Language Models.” *arXiv, 2025.* Available: https://arxiv.org/abs/2502.21014

[23] Benjamin M. Gyori, et al. “Critique of Impure Reason: Unveiling the Reasoning Behaviour of Medical Large Language Models.” *arXiv*, 2024. Available: https://arxiv.org/abs/2412.15748