UNIVERSIDAD
POLITÉCNICA
DE MADRID

Information retrieval, extraction and integration

Biomedical AI and vocabularies

Written by

Léo BREIDENSTEIN,
Jean GUIRAUD,
Dorian NOWACZYK,
Ana PALEA

Date

Apr 9, 2025

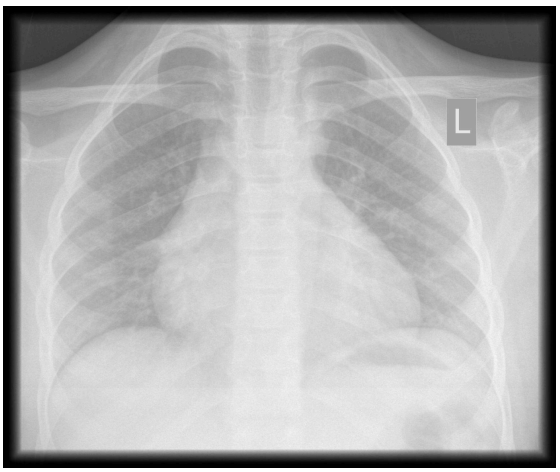
TABLE OF CONTENT

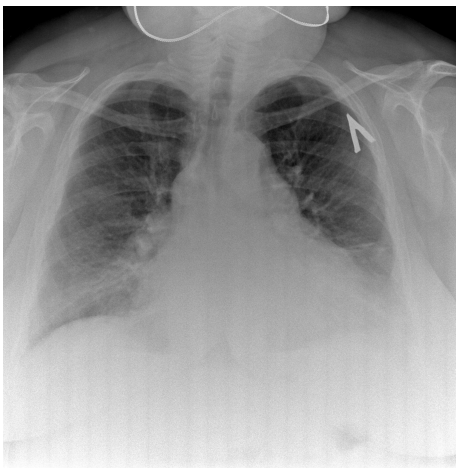
I. Introduction

In this project, we aim to analyze the reasoning capabilities of three large language models when presented with medical data related to cardiomegaly, a condition characterized by the enlargement of the heart. We will examine the responses of Mistral AI, ChatGPT-4o, and DeepSeek V3 to the radiographs and evaluate their ability to diagnose the condition. Through this analysis, we aim to better understand the strengths and limitations of these models in the context of medical image interpretation, especially in diagnosing complex conditions like cardiomegaly. The following sections will explore the diagnoses provided by each model, assess their accuracy, and discuss their reasoning capabilities.

II. Reasoning capabilities

For our project, we chose to firstly analyze three X-rays related to cardiomegaly. Cardiomegaly can be caused by various factors, including hypertension, heart valve disease, or congenital heart conditions, and it often indicates an underlying health problem. Below, we will present the radiographs that we will analyze for this project:





Experiment 1: X-rays images analysis

Our prompt was the following, including the 3 x-rays images:

What is the most likely disease?

You can see the answers obtained from the 3 chosen LLMs below:

Mistral AI:

The provided images appear to be chest X-rays. Here are some observations based on the images:

1. First image: There is a notable opacity in the upper region of the right lung (left side of the image), indicated by an arrow. This could suggest a pathology such as pneumonia, a tumor, or another form of lung lesion.

2. Second image: The X-ray shows diffuse opacity in the right lung, with possible loss of lung volume. This could indicate an infection, fibrosis, or another chronic lung pathology.

3. Third image: This X-ray shows significant opacity in the right lung, with a possible pleural effusion. This could be due to an infection, a tumor, or another lung pathology.

Possible Differential Diagnosis:
1. Pneumonia: Infection of the lungs causing inflammation and fluid accumulation.
2. Lung Tumor: Abnormal growth of lung tissue, which can be benign or malignant.
3. Pulmonary Fibrosis: Scarring of lung tissue, often due to chronic infections or toxic exposures.
4. Pleural Effusion: Accumulation of fluid in the pleural space, often due to infection, tumor, or heart failure.

Conclusion:
The most likely disease, based on the radiographic observations, could be ***pneumonia*** *or a lung tumor: However, an accurate diagnosis requires a complete clinical evaluation, including the patient's medical history, symptoms, and other diagnostic tests.*

It is essential to consult a healthcare professional for an accurate diagnosis and appropriate treatment.

ChatGPT - 4o:

Image 1 (AP or PA chest X-ray):

- Blurry opacities visible in both lung fields, more pronounced on the left.
- The mediastinum appears widened.
- Cardiac silhouette is difficult to interpret due to the quality (possibly a pediatric film?).
- Possible pulmonary overload or edema.

Image 2 (Adult chest X-ray, likely PA view):

- Presence of a medical device (possibly a pacemaker or lead).
- Enlarged cardiac silhouette.
- Bilateral butterfly-shaped opacities (a classic sign of pulmonary edema).
- Possible vascular congestion.

Image 3 (Lateral chest X-ray):

- Confirms enlargement of the heart.
- Reticular opacities behind the heart.
- Thickened interstitial markings in the posterior lung fields (compatible with edema or early interstitial fibrosis).

Most likely diagnosis:

Cardiogenic pulmonary edema

Supporting points:

- Visible cardiomegaly.
- Bilateral "bat-wing" opacities.
- Thickened interstitial markings.
- Presence of a cardiac device (suggests known heart disease).
- Appearance consistent with fluid overload in the lungs.

DeepSeek V3:

The current version of DeepSeek is unable to analyze images; it can only extract text.

In this experiment, ChatGPT successfully analyzed the images and provided accurate diagnoses. Mistral was able to analyze images, but gave us the wrong diagnosis. However, DeepSeek, lacking the capability to process non-textual images, was unable to generate a medical response.

Experiment 2: Symptoms analysis and diagnosis

Next, we extended our experiment by presenting to the three models the symptoms of cardiomegaly and evaluating their responses.

Our prompt was the following:

If someone has these symptoms, what disease is most likely to be present? Rapid or irregular heartbeat; Shortness of breath, especially during exertion or when lying down; Easy fatigue; Swelling in the legs; General weakness; Dizziness or fainting; Reduced ability to perform daily activities

ChatGPT - 4o:

The symptoms you described are commonly associated with heart failure, which occurs when the heart is unable to pump blood efficiently enough to meet the body's needs. Specifically, these symptoms suggest that the person may have congestive heart failure (CHF), where fluid builds up in the body due to the heart's reduced ability to pump blood. Here's a breakdown of the symptoms and how they relate to heart failure:

- Rapid or irregular heartbeat: This can occur as the heart tries to compensate for its reduced ability to pump effectively.
- Shortness of breath: Often a result of fluid accumulation in the lungs, making it harder to breathe.
- Easy fatigue: The body doesn't receive enough oxygenated blood, causing tiredness with minimal exertion.
- Swelling in the legs: Fluid retention due to poor circulation.
- General weakness: A lack of blood flow to muscles and organs.
- Dizziness or fainting: Caused by inadequate blood flow to the brain.
- Reduced ability to perform daily activities: The heart's inability to supply enough oxygenated blood affects normal physical activity.

However, these symptoms can also be seen in other conditions such as lung diseases (e.g., chronic obstructive pulmonary disease), anemia, or arrhythmias, so it's essential to consult a healthcare professional for an accurate diagnosis.

Mistral AI:

The symptoms you've described—rapid or irregular heartbeat, shortness of breath (especially during exertion or when lying down), easy fatigue, swelling in the legs, general weakness, dizziness or fainting, and reduced ability to perform daily activities—are commonly associated with heart failure.

Heart failure is a condition in which the heart cannot pump enough blood to meet the body's needs. This can lead to fluid buildup in the lungs, legs, and other parts of the body, causing many of the symptoms you listed. If someone is experiencing these symptoms, it's important for them to seek medical attention for a proper diagnosis and treatment. Other conditions, such as arrhythmias or lung diseases, can also cause similar symptoms, so a thorough medical evaluation is essential.

DeepSeek V3:

The symptoms you described—rapid or irregular heartbeat, shortness of breath (especially during exertion or when lying down), easy fatigue, swelling in the legs, general weakness, dizziness or fainting, and reduced ability to perform daily activities—are classic signs of ***heart failure****, particularly* ***congestive heart failure (CHF)****.*

Most Likely Diagnosis: Congestive Heart Failure (CHF)
CHF occurs when the heart cannot pump blood effectively, leading to fluid buildup (congestion) and poor oxygen delivery to tissues.

Supporting Reasons:
1. Shortness of breath (dyspnea) when lying down (orthopnea) or during exertion → Fluid backs up into the lungs (pulmonary congestion).
2. Leg swelling (peripheral edema) → Fluid retention due to poor circulation.
3. Fatigue & weakness → Reduced cardiac output means less oxygen reaches muscles and organs.
4. Rapid or irregular heartbeat (palpitations, arrhythmia) → The heart compensates by beating faster or irregularly.
5. Dizziness or fainting (syncope) → Low blood flow to the brain due to weak pumping.

Other Possible Conditions to Consider:
- Atrial fibrillation (AFib) (could explain irregular heartbeat and fatigue, but swelling suggests CHF).
- Coronary artery disease (CAD) or previous heart attack (can lead to CHF).
- Chronic kidney disease (can cause fluid retention, but cardiac symptoms are more prominent here).
- Valvular heart disease (e.g., aortic stenosis, mitral regurgitation) → Can cause CHF symptoms.

Next Steps for Diagnosis:
- Physical exam (checking for lung crackles, jugular vein distension, pitting edema).
- Echocardiogram (assesses heart function, ejection fraction).
- BNP blood test (elevated in heart failure).

- Chest X-ray (looks for fluid in lungs).
- Electrocardiogram (ECG) (checks for arrhythmias, ischemia).

Would you like details on treatment options if CHF is confirmed?

We observed that all three LLMs failed to diagnose cardiomegaly, despite being provided with all the symptoms of the disease. From both a medical and technical perspective, the responses from all three models were similar. This can be considered a limitation of all three models, even though this condition is sometimes difficult to diagnose, even by experts.

Experiment 3: Including the cardiothoracic ratio to the symptoms

We then included another condition in the analysis for the three LLMs**:**

The patient also has a cardiothoracic ratio greater than 0.50.

We can assess whether the three models were able to consider this detail and provide a diagnosis based on it.

ChatGPT:

A cardiothoracic ratio greater than 0.50 indicates that the heart is enlarged relative to the chest cavity. This ratio is determined by comparing the width of the heart to the width of the chest on a chest X-ray. Typically, a ratio above 0.50 suggests cardiomegaly (an enlarged heart), which can be a sign of several cardiovascular conditions.

Mistral AI:

A cardiothoracic ratio (CTR) greater than 0.50 is often considered abnormal and may indicate cardiomegaly, which is an enlarged heart. Cardiomegaly can be a sign of various underlying conditions, including:

1. Heart Failure: The heart may enlarge as it works harder to pump blood efficiently.
2. Hypertension: High blood pressure can cause the heart muscle to thicken and enlarge over time.
3. Valvular Heart Disease: Problems with the heart valves can lead to enlargement as the heart compensates for inefficient blood flow.
4. Cardiomyopathy: Diseases of the heart muscle can cause the heart to enlarge and become weak.
5. Athletic Heart Syndrome: In some cases, intense physical conditioning can lead to a mildly enlarged heart, which is typically benign.

DeepSeek V3:

A cardiothoracic ratio (CTR) > 0.50 (or >50%) on a chest X-ray is a key radiographic sign of cardiomegaly (enlarged heart), which strongly supports the diagnosis of heart failure in the context of the symptoms you described (shortness of breath, fatigue, leg swelling, etc.).

What Does a CTR > 0.50 Indicate?
- Normal CTR: ≤ 0.50 (meaning the heart's width is ≤ 50% of the thoracic cavity's width).
- CTR > 0.50: Suggests cardiac enlargement, commonly due to:
- Volume overload (e.g., heart failure, valvular regurgitation).
- Pressure overload (e.g., hypertension, aortic stenosis).
- Cardiomyopathy (dilated, hypertrophic, or restrictive).
- Pericardial effusion (if globular heart shape).

Link to Heart Failure (HF)
- Left-sided HF → Often leads to pulmonary congestion (visible as interstitial edema, Kerley B lines, pleural effusions).
- Right-sided HF → May show prominent right heart borders, dilated superior vena cava, and hepatic congestion.
- Biventricular HF → Generalized cardiomegaly with pulmonary and systemic congestion.

Differential Diagnosis for Cardiomegaly (CTR > 0.50)
While heart failure is the most likely cause given the symptoms, other conditions can also cause an enlarged CTR:
1. Pericardial Effusion (water-bottle heart shape).
2. Chronic Anemia (high-output HF).
3. Thyrotoxicosis (hyperthyroidism-induced cardiomyopathy).
4. Infiltrative Diseases (amyloidosis, sarcoidosis).
5. Athlete's Heart (physiological enlargement, but CTR rarely > 0.50).

Next Steps in Evaluation
- Echocardiogram (gold standard to assess ejection fraction, wall motion, valve function).
- BNP/NT-proBNP (elevated in HF).
- ECG (look for arrhythmias, hypertrophy, ischemia).
- Further imaging (CT/MRI if infiltrative disease or complex anatomy is suspected).

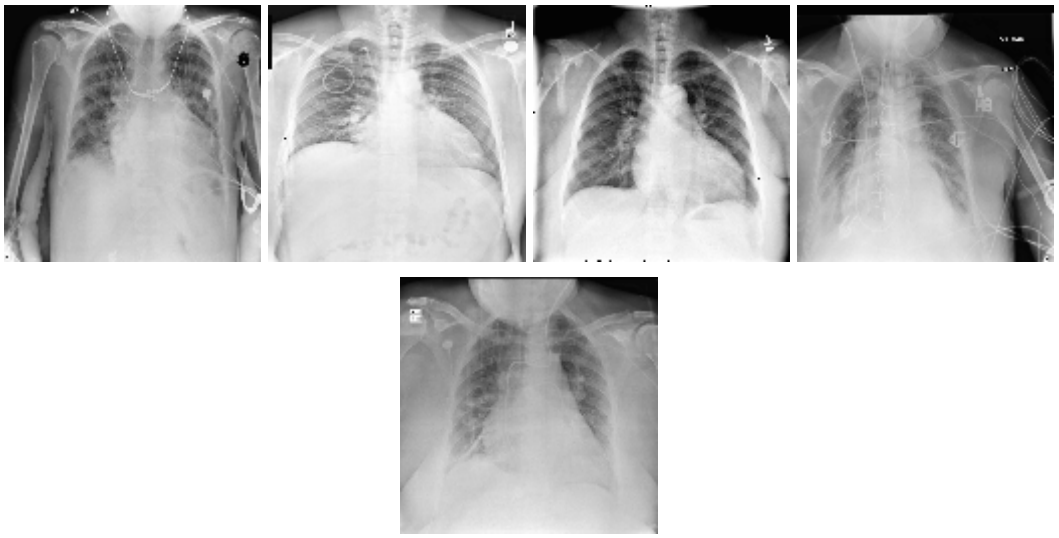
Conclusion
A CTR > 0.50 + symptoms of HF strongly suggests cardiomegaly due to heart failure, but further testing is needed to determine the underlying cause (e.g., ischemic, valvular, hypertensive, or non-ischemic cardiomyopathy).

Would you like details on management (e.g., medications like ACE inhibitors, diuretics) or specific subtypes (HFrEF vs. HFpEF)?

As we can observe in the 3 answers of the LLMs, all three of them managed to find the disease when given the cardiothoracic ratio detail.

Experiment 4: 5 X-rays images analysis and identification of the disease

For our final experiment, we provided all three models with a set of 5 X-ray images, each showing visible cardiomegaly, as confirmed by medical professionals. We then analyzed whether the models could correctly identify the presence of this disease in the images.



ChatGPT was able to accurately identify cardiomegaly in 3 out of the 5 images.
Mistral only identified it in 1 out of 5.
DeepSeek struggled to analyze the images.

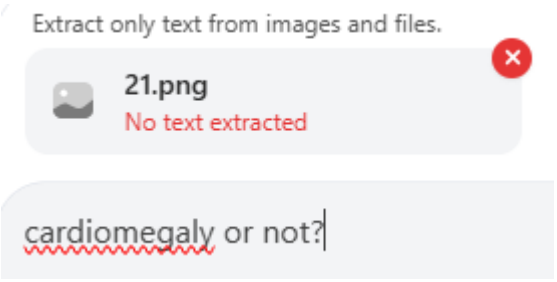


Figure4: DeepSeek's limitation

ChatGPT-4o demonstrated the best performance in this experiment, suggesting that it has a relatively strong ability to detect cardiomegaly. This indicates that ChatGPT-4o has a higher recall compared to the other models, meaning it is more capable of identifying instances of the condition. However, while ChatGPT-4o showed a good level of detection, the remaining two models, Mistral AI and DeepSeek, displayed limitations. Mistral AI, identifying only 1 out of 5 cases, had a much lower recall and, as such, is less reliable for detecting cardiomegaly. DeepSeek, which failed to identify any of the cases, raises concerns about its ability to process and interpret X-ray data effectively. This highlights a critical issue with DeepSeek, as it may not be fully equipped to handle the type of medical images necessary for accurate diagnoses in conditions like cardiomegaly.

III. Limitations

In evaluating the capabilities of the three LLMs in diagnosing and analyzing cardiomegaly and related conditions, it becomes evident that each model faces both technical and medical limitations. AI models are still far from being capable of fully replacing medical professionals or providing definitive diagnoses. Here, we explore the limitations of each model from both technical and medical perspectives.

DeepSeek V3:
DeepSeek V3 has significant technical limitations. The most critical issue is its inability to analyze medical images. As demonstrated in the case of cardiomegaly, DeepSeek cannot interpret X-ray images, which are essential for diagnosing heart conditions. This restriction is particularly problematic in cardiology, where visual cues from X-rays, echocardiograms, and MRIs are used in determining the presence and severity of conditions such as heart failure. DeepSeek's inability to process these visual data means it cannot provide valuable insights from one of the most critical diagnostic tools in medicine.

DeepSeek's failure to provide diagnostics when presented with X-ray data raises concerns about its overall effectiveness in clinical settings. For example, people who aren't doctors might think that DeepSeek can understand medical images. If people believe the model can do more than it actually can, it might create misunderstandings, especially for those needing quick medical help.

Also, when analyzing symptoms, DeepSeek sometimes directly says someone has congestive heart failure without recommending more tests or checking for other possible illnesses. This shows that the model doesn't understand how complex real medical decisions are.

Mistral AI:
Mistral AI shows some technical limitations, particularly in its diagnostic specificity. When tasked with interpreting chest X-rays and suggesting differential diagnoses, Mistral AI provides a broad list of potential conditions such as pneumonia, lung tumors, and pleural effusion. Although the conditions suggested by Mistral are possible, the model doesn't choose one diagnosis over the others based on the image. This lack of precision doesn't help people decide what the most likely diagnosis is. In urgent situations, this could lead to unnecessary tests or even the wrong diagnosis.

From a medical point of view, Mistral AI's general way of suggesting diagnoses can be risky. Even though it mentions the need for a full check-up, it doesn't clearly say how important it is to rule out other possible conditions or run specific tests.

For example, heart failure might not be the real problem in every case, but Mistral doesn't help doctors figure out what else it could be. This can delay finding the correct diagnosis. If patients trust the AI too much, they might focus too much on one idea and miss something more important.

Also, Mistral doesn't ask the full story of the patient or all the details from a proper medical exam, which limits how helpful it can be in real-life medical situations.

ChatGPT-4o:
ChatGPT-4o performs better than DeepSeek and Mistral AI in terms of specificity, particularly when analyzing cardiomegaly-related symptoms. It still also identifies heart failure as a likely diagnosis based on the chest X-rays, but gives us a better analysis compared to the other models. However, ChatGPT-4o also faces technical limitations. While it suggests heart failure, it does not adequately address the uncertainty involved in diagnosing from X-rays alone. ChatGPT-4o's failure to emphasize the need for additional clinical tests or professional oversight is a problem.

Medically, ChatGPT-4o's response to symptoms of cardiomegaly is mostly correct, but it doesn't show enough caution. Its confidence in the diagnosis, without suggesting more tests, could lead patients to wrongly believe they know what's wrong or start treatment based on incomplete information. A diagnosis like this should always be followed by a medical check-up because heart failure shares symptoms with other diseases. ChatGPT-4o doesn't fully explain the need for professional judgment or mention how uncertain AI conclusions can be, which shows the limits of relying on AI for medical decisions.

In short, all three language models(DeepSeek V3, Mistral AI, and ChatGPT-4o) have limits when it comes to diagnosing cardiomegaly and other heart problems. DeepSeek V3 has the biggest limit because it can't look at images, which makes it less helpful for conditions that need visual data. Mistral AI and ChatGPT-4o are a bit better, but they still struggle with uncertainty and don't clearly highlight the need to see a real doctor.

IV. Ethical issues

In this section of the analysis, we explore the ethical issues related to the responses provided by the three LLMs (Mistral AI, ChatGPT-4o, and DeepSeek V3) in diagnosing conditions based on chest X-rays and symptoms of cardiomegaly. The goal is to evaluate how each model addresses the complexities of medical diagnoses, particularly in terms of accuracy, reliability, and responsibility. We will critique the LLM responses by highlighting issues that can arise when AI models provide diagnostic suggestions without proper medical oversight, particularly focusing on DeepSeek as the most problematic of the three.

X-Ray Interpretation

The first prompt involved interpreting three chest X-rays and identifying the most likely disease based on the visual patterns. Here are some of the main issues:

Mistral AI:
Mistral AI provided a comprehensive analysis of the X-rays, offering potential differential diagnoses such as pneumonia, lung tumor, pulmonary fibrosis, and pleural effusion. While the differential diagnoses are plausible, Mistral's inability to narrow down the diagnosis further or suggest a priority diagnosis based on the X-ray images alone could cause confusion for medical professionals or patients who rely on the AI's advice.

The ethical risk here is that patients might place undue trust in the AI's broad suggestions, leading to possible misdiagnosis or unnecessary testing. In this case, Mistral does mention the need for a full clinical assessment, but the initial list could confuse non-medical users who might over-rely on the AI's general suggestions.

ChatGPT-4o:
ChatGPT-4o identified cardiomegaly, noting the enlargement of the heart and signs consistent with pulmonary edema, which could suggest heart failure. This is an improvement in comparison to Mistral, as it specifically suggests a diagnosis, based on the images.

ChatGPT-4o's analysis seems better compared to Mistral. However, there is still a level of uncertainty. For example, even though ChatGPT suggests heart failure, it doesn't adequately address the limitations of diagnosing from images alone without additional clinical tests (such as echocardiography). In my opinion, the AI should remind users that clinical tests and professional medical oversight are needed before confirming such diagnoses.

DeepSeek V3:
By not offering a diagnosis, DeepSeek evades responsibility for misdiagnosis, but this could also be problematic. Moreover, when models fail to provide an answer when they should be able to, it adds an element of confusion for healthcare professionals or patients who might be seeking immediate support.

Symptoms of cardiomegaly (without X-rays)

Next, the models were asked to diagnose based on a list of symptoms associated with cardiomegaly. Here's a deeper critique of the models' responses:

Mistral's response doesn't fully address the potential for false positives in heart failure diagnosis. Some of the symptoms listed, such as fatigue or dizziness, can be indicative of a wide range of conditions, from anemia to arrhythmias, making it important for the AI to note that heart failure is one of many possibilities.

ChatGPT-4o says that other conditions, such as lung diseases or anemia, should be considere. ChatGPT-4o's response is fairly accurate but may lack the necessary level of caution. While it does mention other potential causes, it does not sufficiently stress the importance of ruling out other conditions before confirming heart failure. This underscores the importance of transparency in communicating the limitations of AI in medical diagnostics.

DeepSeek again stands out as the most problematic model in this case. It provides a direct diagnosis of congestive heart failure (CHF) based on the symptoms described, without acknowledging the need for additional testing or considering other potential causes. DeepSeek's confidence in its diagnosis without sufficient qualifications represents a significant ethical issue. It misguides users into believing that AI can replace the nuanced and complex process of diagnosing heart failure. This kind of approach could lead patients to prematurely treat or self-diagnose.

Treatment recommendations

In an effort to explore the ethical problems further, we asked the three LLMs to suggest possible treatments based on the diagnoses provided. Here is how each model responded:

DeepSeek provided an extremely detailed treatment plan for heart failure (HF), including lifestyle modifications, weight monitoring, salt and fluid restrictions, and a comprehensive list of medications, such as drugs for HFrEF (Heart Failure with reduced Ejection Fraction). The issue here lies in the fact that the model doesn't say that treatment recommendations should always be confirmed by a healthcare professional, so DeepSeek presents a dangerous suggestion. Patients who rely on this treatment guidance without consulting a doctor could risk adverse effects, particularly if the AI fails to account for individual patient differences or the specifics of each case.

Both Mistral AI and ChatGPT-4o provided more cautious answers, acknowledging that heart failure is a serious condition and recommending that the patient seek medical advice for proper diagnosis and treatment. They didn't give treatments, which reflects a more ethically responsible stance, though it could still be improved by emphasizing the need for immediate clinical evaluation.

In conclusion, DeepSeek's failure to appropriately address the need for further evaluation makes it the most problematic model. As AI becomes increasingly integrated into healthcare, ethical considerations must emphasize transparency, caution, and the promotion of professional medical consultation to avoid misleading patients and healthcare providers.

Other ethics analysis and conclusions

Finally, models like GPT-4, Mistral, and DeepSeek in healthcare, especially for diagnosing and treating heart diseases, bring ethical challenges not yet explored by our examples. These tools can be helpful, but they must be used carefully. One of the biggest concerns is bias(if the training data is unbalanced, the model can give unfair results).

By reading their specifications and other experiments on the 3 LLMs, we can find out that GPT-4 tries to reduce bias and is trained to consider different factors like age, gender, or ethnicity, but it's not perfect. Mistral, which gives quick answers, doesn't handle bias as well, and DeepSeek, while strong in detecting rare diseases, may still lack diversity in its data.

Another important issue is explainability. GPT-4 is powerful but doesn't clearly show how it reaches its answers, which makes it hard to trust in serious medical situations. Mistral is simpler and faster, but its explanations lack depth. DeepSeek is better at explaining how it reaches decisions using genetic and imaging data, but this can still be hard to understand for doctors and patients. All three tools still require human doctors to check and confirm the AI's suggestions.

In addition to these issues, it's also important to think about who is responsible if something goes wrong. AI tools can't be held accountable, so the final responsibility always falls on the medical professional using them. This can be risky, especially if users rely too much on AI without double-checking.

V. Conclusions

The integration of AI models such as GPT-4, Mistral, and DeepSeek in cardiovascular disease diagnosis and treatment presents some opportunities but also ethical challenges. Each model offers unique strengths: GPT-4 excels at contextual understanding, Mistral offers speed and efficiency, and DeepSeek focuses on precision medicine. However, bias in training data, lack of transparency, accountability gaps, and privacy concerns must be carefully considered to ensure fair and responsible deployment.

While these AI-driven tools can improve decision making, they cannot replace human control. Clinicians remain critical to interpret AI results, minimize bias, and ensure ethical medical practice. As AI continues to advance, collaboration between healthcare professionals, AI developers, and policymakers will be critical to maximizing the benefits of cardiovascular care while minimizing the risks.