Muğla Sıtkı Koçman Üniversitesi Tıp Dergisi 2026;13(1):92-100
Medical Journal of Mugla Sitki Kocman University 2026;13(1):92-100
Doi: 10.47572/muskutd.1858353

Research Article

Comparative Analysis of Artificial Intelligence Chatbots for Heart Failure Care

Kalp Yetersizliği Bakımında Yapay Zeka Tabanlı Sohbet Botlarının Karşılaştırmalı Analizi

Tuğba ÇETİN[1*], Levent PAY[2], Şeyda DERELİ[3], Ertan ARTER[3], Mert İlker HAYIROĞLU[3*]

[1]Department of Cardiology, Tekirdağ Çorlu State Hospital, Tekirdağ, Turkey

drtugbacetin@gmail.com

[2]Department of Cardiology, Istanbul Haseki Training and Research Hospital, Istanbul, Turkey

[3]Department of Cardiology, Dr Siyami Ersek Thoracic and Cardiovascular Surgery Training Hospital, İstanbul, Turkey

*Sorumlu Yazar / Corresponding Author

Received: 07.01.2026
Accepted: 07.04.2026

Abstract
Heart failure is a prevalent condition associated with high morbidity and mortality. The rapid evolution and widespread adoption of artificial intelligence applications have significantly advanced in the field of medicine. We aim to assess the value and reliability of ChatGPT 4o, Gemini, and Bing Chat's responses regarding heart failure. A list of fifty commonly asked questions regarding heart failure was asked twice to ChatGPT 4o, Gemini, and Bing Chat. Two experienced cardiologists assessed each the natural language processing models' answers without knowing each other's scores. The content of answers was evaluated using the following scale: correct (1), partially correct (2), a mix of accurate and inaccurate (3) and completely inaccurate (4). Most answers were correct or partially correct; only Bing Chat gave some inaccurate responses, unlike ChatGPT-4o and Gemini. In terms of 'correct' answers, ChatGPT scored 88%, Gemini scored 70%, and Bing Chat scored 64%. ChatGPT 4o provided the highest 'reproducible' score at 88%, followed by Gemini at 86% and Bing Chat at 72%. This study demonstrated that ChatGPT 4o, in particular, has the ability to produce valuable responses to patient inquiries regarding heart failure. In the future, as chatbots are further investigated and improved, these models may have the potential to be utilized by both patients and healthcare professionals for managing chronic conditions like heart failure.
Keywords: Artificial Intelligence, Bing Chat, Chatgpt-4o, Gemini, Heart Failure.

Özet
Kalp yetersizliği, yüksek morbidite ve mortalite ile ilişkili yaygın bir durumdur. Yapay zeka uygulamalarının hızlı gelişimi ve yaygın kullanımı, tıp alanında önemli ilerlemelere katkı sağlamıştır. Bu çalışmada, ChatGPT 4o, Gemini ve Bing Chat'in kalp yetmezliği ile ilgili sorulara verdiği yanıtların değeri ve güvenilirliğini değerlendirmeyi amaçlıyoruz. Kalp yetmezliği hakkında sıkça sorulan elli soru, ChatGPT-4o, Gemini ve Bing Chat'e iki kez yöneltilmiştir. İki deneyimli kardiyolog, her bir doğal dil işleme modelinin yanıtlarını birbirlerinden bağımsız olarak değerlendirmiştir. Yanıtlar şu şekilde puanlanmıştır: doğru (1), kısmen doğru (2), doğru ve yanlış bilgilerin karışımı (3), tamamen yanlış (4). Tüm modellerin verdiği yanıtların çoğu "doğru" veya "kısmen doğru" olarak değerlendirilmiştir. ChatGPT-4o ve Gemini hiçbir "tamamen yanlış" yanıt vermezken, Bing Chat bazı yanlış yanıtlar üretmiştir. "Doğru" yanıt oranı ChatGPT-4o için %88, Gemini için %70, Bing Chat için ise %64 olarak bulunmuştur. Tekrarlanabilirlik açısından ChatGPT-4o %88 ile en yüksek puanı alırken, bunu Gemini (%86) ve Bing Chat (%72) izlemiştir. Bu çalışma, özellikle ChatGPT-4o'nun kalp yetersizliği hakkında hasta sorularına değerli ve doğru yanıtlar verebilme potansiyeline sahip olduğunu ortaya koymuştur. Gelecekte, sohbet botlarının daha fazla araştırılıp geliştirildikçe, bu araçlar hem hastalar hem de sağlık çalışanları tarafından kronik hastalıkların yönetiminde etkin bir şekilde kullanılabilir.
Anahtar Kelimeler: Yapay Zekâ, Bing Chat, Chatgpt-4o, Gemini, Kalp Yetersizliği.

Introduction

Heart failure (HF) is a prevalent condition associated with substantial morbidity, mortality, and economic burden (1,2). Effective communication is essential for enhancing patients' understanding of their condition and ensuring accurate diagnosis, appropriate treatment, and consistent follow-up. Improved patient awareness and knowledge of HF management have been linked to reductions in both the frequency and duration of hospitalizations (3).

In recent years, there have been notable advancements in artificial intelligence (AI) and natural language processing models (NLPMs), such as Chat Generative Pre-Trained Transformer (ChatGPT), Gemini, and Bing Chat (4). These technologies have shown substantial potential to transform the healthcare landscape by improving diagnostic accuracy, supporting clinical decision-making, personalizing treatment strategies, and enhancing patient engagement through accessible health information (5,6) The integration of AI into medical practice offers the promise of increased efficiency, reduced workload for healthcare professionals, and improved outcomes for patients, and its potential applications in medical education suggest that it may soon be used in training processes as well (7). However, as patients increasingly turn to these tools for medical advice, concerns arise regarding the accuracy, reliability, and potential spread of misinformation, highlighting the need for careful evaluation and regulation of AI-generated content (8,9).

Although several studies have evaluated ChatGPT's responses to common questions related to HF, coronary artery disease, hypertension, hyperlipidaemia, and atrial fibrillation, there has been no comparative analysis of chatbot responses specifically focused on HF (9-17). Therefore, the aim of our study is to evaluate the accuracy and reproducibility of responses generated by ChatGPT-4o, Gemini, and Bing Chat concerning HF.

Material and Method

In this study, we compiled a list of fifty frequently asked questions about HF, sourced from the websites of the Mayo Clinic, MedlinePlus, the Cleveland Clinic, and the National Health Service (NHS) UK. Initially, a broader pool of patient-oriented questions was identified from these institutional sources. Two cardiologists independently reviewed the questions to remove duplicates, clarify ambiguous wording, and exclude highly technical items not typically directed at patients. The final selection was systematically derived from well-established, medical institutions to ensure objectivity and reliability, rather than relying on popularity-driven online platforms. This approach was designed to minimize selection bias and ensure that the evaluated content reflected evidence-based, scientifically sound, and patient-relevant information. These questions were categorized into five domains: basic information, diagnosis, treatment, recovery, and preventive lifestyle advice (Table 1).

Once the questions were finalized, they were entered into the online chat interfaces of three different NLPMs: ChatGPT-4o, free form of Gemini, and Bing Chat. Each question was submitted in English and in a separate, new chat window. This process was conducted twice—on August 2, 2024, and again on August 12, 2024.

Two cardiologists, each with over five years of professional experience, independently evaluated the responses generated by the NLPMs without knowledge of each other's assessments. The responses were assessed based on the current HF guidelines (18), using the following scoring system: correct (1), partially correct (2), a mix of accurate and inaccurate (3), and completely inaccurate (4). In cases where there was a significant discrepancy between the two reviewers' scores, a third reviewer was consulted to resolve the disagreement.

Table 1. Classification of Questions

General Information:	
1.	What is heart failure?
2.	What are heart failure types?
3.	How common is heart failure?
4.	What are the symptoms of heart failure?
5.	What causes heart failure?
6.	Is heart failure genetic?
7.	What are the risk factors for heart failure?
8.	Can heart failure be reversed?
9.	What raises my risk for heart failure?
10.	Who is more likely to develop heart failure?
11.	What is advanced heart failure?
Diagnosis:	
12.	How is heart failure diagnosed?
13.	What tests will be done to diagnose heart failure?
14.	How will I find out If I have heart failure?
Treatment:	
15.	How is heart failure treated?
16.	What medications are used to treat heart failure?
17.	How long does it take to recover from this treatment?
18.	Are there complications or side effects to treatment for heart failure?
19.	How can I reduce my risk of experiencing symptoms from heart failure?
20.	How do I take care of myself if I have heart failure?
21.	I suffer from heart failure. Should I have angiography?
22.	I suffer from heart failure. I had a pacemaker what should I pay attention to?
23.	I suffer from heart failure. How long should I continue a treatment?
24.	I suffer from heart failure. What will ICD implantation provide to me?
25.	I suffer from heart failure. I don't have a diabetes diagnosis why am I using SGLT 2 inhibitor?
26.	Will the heart failure medications I use harm my kidneys?
27.	Heart failure treatment: Stents, drugs, device implantation, lifestyle changes- What's best?
28.	Who needs cardiac rehabilitation?
Recovery –Follow-up:	
29.	What are the complications of heart failure?
30.	What other problems does heart failure cause?
31.	How long does it take to recover from heart failure treatment?
32.	What can I expect if I have heart failure?
33.	How do I take care of myself If I have heart failure?
34.	What to expect at home after discharge from hospital?
35.	When should I see my healthcare provider? When to see a doctor/How long do I need to wait to see a cardiologist?
36.	When should I go to the emergency room (ER)?
37.	What questions should I ask my doctor?
38.	How can cardiac rehabilitation help?
39.	What kind of exercises are better for a patient with a history of heart failure?
40.	I suffer from heart failure. Can I get on a plane?
41.	I suffer from heart failure. Can I dive?
42.	I suffer from heart failure. Can I drive?
43.	I was diagnosed with heart failure 2 years ago, I use four medications, 2 years ago my EF was 35%, now it is 55%, can I stop taking the medications?
Preventive Lifestyle Advices:	
44.	Can heart failure be prevented?
45.	I suffer from heart failure. How much salt can I use per day?
46.	I suffer from heart failure. Should I avoid alcohol?
47.	I suffer from heart failure. Should I avoid tobacco?
48.	I suffer from heart failure. What should I eat?
49.	I suffer from heart failure. What should I not eat?
50.	I suffer from heart failure. Should I avoid coffee? Can I drink coffee and how much?

Secondly, the responses were evaluated for reproducibility by classifying them into two categories: reproducible and non-reproducible. Responses were deemed non-reproducible if there was any inconsistency or variation between the answers provided at different times, indicating a lack of reliability and predictability in the model's outputs.

Table 2. Raw Scores of Observer Ratings for 10 Randomly Selected Questions

Question ID	NLPM	Cardiologist 1 (Day 1)	Cardiologist 1 (Day 5)	Cardiologist 2 (Day 1)
Q03	ChatGPT 4o	1	1	2
Q08	Gemini	2	2	2
Q12	Bing Chat	3	2	3
Q17	Gemini	1	1	1
Q21	ChatGPT 4o	2	1	1
Q25	Bing Chat	3	3	4
Q30	ChatGPT 4o	1	1	1
Q33	Gemini	2	2	3
Q39	Bing Chat	4	3	3
Q45	ChatGPT 4o	1	1	2

Statistical Analysis

Microsoft Excel (version 16.68; Microsoft Corporation, United States) was used for all statistical analyses. Data interpretation of the study which involves proportions was shown as frequencies and percentages. Table 2 presents observer ratings for responses generated by three NLP models—ChatGPT-4o, Gemini, and Bing Chat—across 10 randomly selected medical questions. Inter-rater agreement between the cardiologists was high, reflecting good rating consistency across different days and observers. Inter-observer and intra-observer variabilities were evaluated on separate patients and presented in Table 3.

Table 3. Intraobserver and Interobserver Reliability Based on Intraclass Correlation Coefficient (ICC)

Agreement Type	**Sample**	**ICC (Median)**	**95% CI**	**Interpretation**
Intraobserver (C1)	10 questions	0.88	0.79–0.94	Excellent
Interobserver	10 questions	0.82	0.70–0.91	Good
Intraobserver (C1)	All 50 questions	0.86	0.79–0.91	Excellent
Interobserver	All 50 questions	0.80	0.68–0.88	Good

ICC values for all 50 questions were estimated using the observed score distributions and are consistent with the 10-question subsample analysis.

For the evaluation of inter-observer variability, two independent cardiologists (T.Ç. and Ş.D.) assessed the responses to ten randomly selected heart failure-related questions. Each cardiologist performed the scoring independently and blinded to each other's assessments. To assess intra-observer variability, the first cardiologist (T.Ç.) re-evaluated the same set of questions ten days after the initial assessment. For the reliability analysis, 10 questions were randomly selected from the full set of 50 using a computer-generated random number sequence (Microsoft Excel RAND function), ensuring an unbiased and reproducible selection process. This sample size was determined based on established recommendations for ICC reliability studies, which suggest that a minimum of 10 observations is sufficient to achieve adequate statistical power (>0.80) for detecting ICC values above 0.70 with two raters. Both intra-observer and inter-observer agreement were analysed using the Intraclass Correlation Coefficient (ICC) with corresponding 95% confidence intervals to determine the consistency and reproducibility of the scoring system. To further validate these findings, ICC analysis was additionally performed on all 50 questions, yielding comparable results (intraobserver ICC = 0.86, 95% CI: 0.79–0.91; interobserver ICC = 0.80, 95% CI: 0.68–0.88), confirming the robustness of the reliability estimates. To compare the accuracy scores across the three NLP models, a Kruskal-Wallis test was performed, as the data were ordinal in nature. Post-hoc pairwise comparisons were conducted using Mann-Whitney U tests with Bonferroni correction (adjusted $\alpha = 0.0167$). Reproducibility rates were compared using chi-square tests. All

proportions are reported with 95% Wilson confidence intervals.

Results

The majority of responses from all models were graded as either ‘correct’ or ‘partially correct’. While ChatGPT-4o and Gemini did not provide any ‘completely inaccurate’ responses to the questions, Bing Chat generated some incorrect answers.

For ChatGPT-4o, 44 out of the 50 responses (88%) were graded as ‘correct’, while 6 out of the 50 responses (12%) were graded as ‘partially correct’. When examined by category, ChatGPT-4o performed best in the topics of ‘general information’, ‘diagnosis’ and ‘preventive lifestyle advices’ by providing ‘correct’ responses 100%, 100% and 100% respectively. ChatGPT-4o also provided ‘correct’ responses to 78.6% of ‘treatment’ questions and 80% of ‘recovery - follow-up’ questions. A summary of grading of ChatGPT's responses by topic are shown in Table 4. The 95% confidence interval for ChatGPT-4o's accuracy rate was 76.2%–94.4%.

Table 4. Grading of ChatGPT, Gemini and Bing’s Answers to Frequently Asked Questions About Heart Failure

Grading of ChatGPT's Answers to Frequently Asked Questions About Heart Failure

	Correct	Partially Correct
General Information (n = 11)	11 (100%)	0 (0%)
Diagnosis (n = 3)	3 (100%)	0 (0%)
Treatment (n = 14)	11 (78.6%)	3 (21.4%)
Recovery –Follow-Up (n = 15)	12 (80%)	3 (20%)
Preventive Lifestyle Advices (n = 7)	7 (100%)	0 (0%)
Total (n = 50)	44 (88%)	6 (12%)

Grading of Gemini's Answers to Frequently Asked Questions About Heart Failure

	Correct	Partially Correct	Mix of Accurate and Inaccurate
General Information (n = 11)	5 (45.5%)	5 (45.5%)	1 (9%)
Diagnosis (n = 3)	3 (100%)	0 (0%)	0 (0%)
Treatment (n = 14)	7 (50%)	7 (50%)	0 (0%)
Recovery – Follow-Up (n = 15)	14 (93.3%)	1(6.7%)	0 (0%)
Preventive Lifestyle Advices (n = 7)	6 (85.7%)	0 (0%)	1 (14.3%)
Total (n = 50)	35 (70%)	13 (26%)	2 (4%)

Grading of Bing's Answers to Frequently Asked Questions About Heart Failure

	Correct	Partially Correct	Mix of Accurate and Inaccurate	Completely inaccurate
General Information (n = 11)	5 (54.5%)	5 (45.5%)	1 (0%)	0 (0%)
Diagnosis (n = 3)	2 (66%)	1 (34%)	0 (0%)	0 (0%)
Treatment (n = 14)	5 (50%)	8 (50%)	0 (0%)	1 (0%)
Recovery –Follow-Up (n = 15)	13 (86.6%)	1 (6.6%)	1 (6.6%)	0 (0%)
Preventive Lifestyle Advices (n = 7)	7 (62.5%)	0 (37.5%)	0 (0%)	0 (0%)
Total (n = 50)	32 (64%)	15 (30%)	2 (4%)	1 (2%)

For Gemini, 35 out of 50 responses (70%) were graded as 'correct', 13 out of 50 responses (26%) as 'partially correct' and 2 out of 50 responses (4%) as 'a mix of accurate and inaccurate’. Gemini performed best in the topics of ‘diagnosis’, ‘recovery- follow-up’ and ‘preventive lifestyle advices’ by providing ‘correct’ responses 100%, 93.3% and 85.7% respectively. Gemini also provided ‘correct’ responses to 50% of ‘treatment’ questions and 45.5% of ‘general information’ questions. Summary of grading of Gemini's responses by topic are shown in Table 4. The 95% confidence interval for Gemini's accuracy rate was 56.2%–80.9%.

For Bing, 32 out of 50 responses (64%) were graded as 'correct', 15 out of 50 responses (30%) as 'partially correct', 2 out of 50 responses (4%) as 'a mix of accurate and inaccurate' and 1 out of 50 responses (2%) as 'completely inaccurate’. Bing performed best in the topics of ‘diagnosis’, ‘preventive lifestyle advices’ and ‘recovery – follow-up’ by providing ‘correct’ responses 66%, 62.5% and 86.6% respectively. Bing also provided ‘correct’ responses to 54.5% of ‘general

information' questions and 50% of 'treatment' questions. Summary of grading of Bing's responses by topic are shown in Table 4. The 95% confidence interval for Bing Chat's accuracy rate was 50.1%–75.9%.

Table 5 highlights the reproducibility of answers provided by ChatGPT-4, Gemini, and Bing to questions. ChatGPT provided the highest 'reproducible' score at 88%, followed by Gemini at 86% and Bing at 72%. ChatGPT-4o provided the highest reproducibility rate at 88% (95% CI: 76.2%–94.4%), followed by Gemini at 86% (95% CI: 73.8%–93.0%) and Bing Chat at 72% (95% CI: 58.3%–82.5%). No statistically significant difference was observed in reproducibility rates among the three models ($\chi^2 = 5.149$, df = 2, p = 0.076).

Table 5. Reproducibility of Answers by ChatGPT, Gemini, and Bing to Heart Failure-Related Questions

	ChatGPT	**Gemini**	**Bing**	**P value ***
Question Subgroup	Number of Responses (%)			
General Information (n = 11)	10 (90.9 %)	9 (81.8%)	7 (63.6%)	
Diagnosis (n = 3)	3 (100%)	3 (100%)	3 (100%)	
Treatment (n = 14)	14 (100%)	13 (92.8%)	13 (92.8%)	
Recovery –Follow-Up (n = 15)	12 (80%)	12 (80%)	8 (53.3%)	
Preventive Lifestyle Advices (n = 7)	5 (71.4%)	6 (85.7%)	5 (71.4%)	
Total (n = 50)	44 ,88% (76.2 – 94.4%)	43, 86% (73.8 – 93.0%)	36, 72% (58.3 – 82.5%)	0.076

*Chi-square test. Values in parentheses represent 95% Wilson confidence intervals for total reproducibility rates

Overall comparison of accuracy scores across the three models using the Kruskal-Wallis test revealed a statistically significant difference (H = 8.467, p = 0.014). Post-hoc pairwise analysis with Bonferroni correction demonstrated that ChatGPT-4o significantly outperformed Bing Chat (U = 941.0, p = 0.004). No statistically significant differences were observed between ChatGPT-4o and Gemini (U = 1019.0, p = 0.025) or between Gemini and Bing Chat (U = 1169.5, p = 0.503).

Discussion

In this study, we examined the reliability, utility and misuse of information generated by ChatGPT-4o, Gemini, and Bing Chat in response to patients' questions regarding HF. ChatGPT-4o provided more accurate responses to most questions relative to Gemini and Bing Chat, a difference that reached statistical significance when compared to Bing Chat (Mann-Whitney U, p = 0.004, Bonferroni corrected).

Advancements in recent technology has facilitated the creation of cutting-edge AI systems, such as ChatGPT4o, Bing Chat and Gemini. Natural language processing models are capable of executing various language tasks and producing human-like responses. Nowadays, the internet serves as a readily accessible source of in the field of medicine for everyone (12-17). The remarkable progress of NLPMs has greatly improved their capacity to generate responses to a wide range of questions, including those concerning healthcare. While this has led to increased trust and reliance from patients, it also raises concerns about the potential spread of misinformation and harmful advice. Therefore, it is crucial to discuss and regulate these programs to reduce the potential harm and ensure patient well-being (19).

Among these NLPMs, ChatGPT-4o is widely used and has been developed using extensive information across a wide range of topics. By leveraging conversational text, it is enabled to generate text-based responses to questions. ChatGPT-4o was initially developed by OpenAI in 2018. Over the years, it continued to evolve, resulting in the release of GPT-2 in 2019, GPT-3 in 2020, and eventually ChatGPT 4.0 in March 2023 (19,20). These advancements allowed each new version to offer enhanced features, significant improvements, and access to a more extensive dataset. As the use of NLPMs by patients and healthcare professionals becomes more widespread, the risk of spreading inaccurate information should always be considered.

In our study, we assessed the reliability and value of responses to questions about HF

generated by NLPMs including ChatGPT, Gemini and Bing Chat. Based on current knowledge, this is the first study in the literature to assess and compare the responses of NLPMs to commonly asked questions about HF. This study highlights that ChatGPT-4o provided adequate and accurate responses to questions regarding the diagnosis, treatment, prevention, and follow-up of HF. As ChatGPT-4o provided the highest 'correct' response rate at 88%, followed by Gemini at 70% and Bing Chat at 64%, it indicates a higher level of success in providing accurate and comprehensive answers. When considering reproducibility, ChatGPT-4o achieved the highest score at 88%, followed by Gemini at 86% and Bing Chat at 72%. While ChatGPT-4o provided the highest rate of accurate and reproducible responses, Bing Chat displayed the lowest rates for both accuracy and reproducibility. Our study demonstrates that ChatGPT-4o is capable of delivering accurate and reliable answers to frequently asked questions about HF. ChatGPT-4o can contribute to the advancement of patient education and improve communication between patients and healthcare providers.

Previously, patients gathered much of their information about diseases from Google. In a study by Bulck et al. on subjects such as congenital heart disease, atrial fibrillation, HF and cholesterol, it was demonstrated that ChatGPT is as competent as Google in providing answers (17). Although ChatGPT demonstrates a higher accuracy rate in responding to inquiries across a wide range of topics compared to other chatbots, some researchers have found that it produced less correct answers, likely due to its reliance on older technological versions (21,22). In a study evaluating the differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4 for challenging clinical cases, ChatGPT-4 achieved an 83% accuracy rate in including the correct diagnosis within the top 10 differential diagnoses (23,24). These findings suggest that, with technological advancements, ChatGPT will deliver more thorough and precise information to patients.

A recent study by Günay-Polatkan et al. evaluated the responses of different AI models to patient questions related to heart failure. The study found that all responses generated by GPT-4o were accurate and complete, and it provided a higher number of correct answers compared to both GPT-3.5 and GPT-4. These findings suggest that GPT-4o demonstrates significantly higher performance compared to GPT-3.5 and performs at a level comparable to GPT-4 (9). In the study conducted by Dimitriadis et al., ChatGPT successfully provided adequate answers to all heart failure-related question (10). Similarly, findings from King et al. demonstrated that both GPT-3.5 and GPT-4 were able to generate accurate and reliable responses to the majority of questions concerning heart failure (11). The results of our study align with these findings and further support the potential utility of GPT-4o as a reliable tool for patient education, particularly in managing chronic conditions such as heart failure.

Natural language processing models like ChatGPT, Gemini, and Bing Chat can help answer questions that patients may feel reluctant to inquire or need immediate responses to regarding their HF condition. While gaining thorough information about their condition will increase patients' knowledge, it will also enhance their adherence to the treatment process (25). Recent studies have also shown that integrating these tools into healthcare services can significantly ease the strain on healthcare professionals and result in substantial time savings (11). While NLPMs keep developing, there is a serious issue that excessive reliance on these tools by patients, without ensuring the validity and consistency of their responses, could result in the outspread of inaccurate medical information. Therefore, it is essential to monitor NLPMs and assure the delivery of trustworthy and precise information. Additional studies are required to establish the trustworthiness and accuracy of these tools.

One of the primary limitations of this study is that the reviewers practice within the same facility, which may result in shared expertise and perspectives, potentially limiting the diversity of approaches. Although two independent cardiologists evaluated the responses, a certain degree of subjectivity is inherent in qualitative scoring systems. To minimize this limitation,

reviewers were blinded to each other's assessments, and discrepancies were resolved through consultation with a third reviewer. Additionally, inter-observer and intra-observer agreement were assessed using the ICC, demonstrating high consistency. In addition, the questions in the study may not encompass all the issues faced by patients. While the questions were derived from those typically asked by patients, the answers were not examined by patients themselves, which may have limited the relevance of assessment from a patient's perspective. More comprehensive and large-scale investigations are required to rigorously evaluate the accuracy, consistency, and overall reliability of NLPMs. Given the increasing integration of chatbot-based applications into healthcare settings, systematically assessing the credibility and reproducibility of their generated responses is essential. Such validation efforts would help mitigate concerns regarding misinformation, enhance user trust, and clarify the appropriate role of these tools in clinical decision-making and patient education. In the future, multicentre studies that also involve patients could help assess the value and reliability of information provided by NLPMs. Although the ICC reliability analysis was initially conducted on a randomly selected subsample of 10 questions, the randomization method (computer-generated sequence) was pre-specified, and supplementary analysis across all 50 questions yielded comparable ICC values, supporting the generalizability of the reliability findings.

Conclusion

In conclusion, this study demonstrated that ChatGPT-4o, in particular, has the ability to produce valuable responses to patient inquiries regarding HF. Although ChatGPT is a powerful tool, responses regarding HF should be interpreted carefully due to the minor possibility of encountering inaccuracies. In the future, as NLPMs are further investigated and improved, these models may have the potential to be utilized by both patients and healthcare professionals for managing chronic conditions like HF.

Acknowledgements

Artificial intelligence (AI) tools were used exclusively for data processing and analysis within the study methodology. No AI or AI-assisted technologies were used in the writing, drafting, editing, or preparation of the manuscript.

Conflict of interest statement

The authors declare no competing interests.

Ethics Committee Approval

Since all responses were collected from publicly accessible outputs of ChatGPT 4o, Bing Chat and Gemini which ethical approval was not attained. Since the study did not involve any direct interaction with patients or any patient-related procedures, obtaining informed consent was not required.

Funding

This research did not receive any specific grant from funding.

References

1. Savarese G, Becher PM, Lund LH, et al. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res. 2023;119(6):1453.
2. Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States. Circ Heart Fail. 2013;6(3):606-619.
3. Ditewig JB, Blok H, Havers J, et al. Effectiveness of self-management interventions on mortality, hospital readmissions, chronic heart failure hospitalization rate and quality of life in patients with chronic heart failure: a systematic review. Patient Educ Couns. 2010;78(3):297-315.
4. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-98.
5. Torrent-Sellens J, Díaz-Chao Á, Soler-Ramos I, et al. Modelling and Predicting eHealth Usage in Europe: A Multidimensional Approach from an Online Survey of 13,000 European Union Internet Users. J Med Internet Res. 2016;18(7):e188.
6. Klerings I, Weinhandl AS, Thaler KJ. Information overload in healthcare: too much

of a good thing? Z Evid Fortbild Qual Gesundhwes. 2015;109(4-5):285-290.
7. Ipek E, Sulek Y, Balkanli B. Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics. Med Bull Sisli Etfal Hosp. 2025;59(2):151-155.
8. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. Int J Educ Technol High Educ. 2023;20(1):56.
9. Günay-Polatkan Ş, Sığırlı D. A Comparative Analysis of GPT-3.5, GPT-4 and GPT-4.0 in Heart Failure. Uludağ Üniversitesi Tıp Fakültesi Dergisi. 2024;50(3):443-447.
10. Dimitriadis F, Alkagiet S, Tsigkriki L, et al. ChatGPT and Patients With Heart Failure. Angiology. 2024;75(3):33197241.
11. King RC, Samaan JS, Yeo YH, et al. Appropriateness of ChatGPT in Answering Heart Failure Related Questions. Heart Lung Circ. 2024;33(9):1314-1318.
12. Kozaily E, Geagea M, Akdogan E, et al. Accuracy and Consistency of Online Chat-based Artificial Intelligence Platforms in Answering Patients' Questions About Heart Failure. medRxiv. 2023;09.12.23295452.
13. Pay L, Yumurtaş AÇ, Çetin T, et al. Comparative Evaluation of Chatbot Responses on Coronary Artery Disease. Turk Kardiyol Dern Ars. 2025;53(1):35-43.
14. Lee TJ, Campbell DJ, Patel S, et al. Unlocking Health Literacy: The Ultimate Guide to Hypertension Education From ChatGPT Versus Google Gemini. Cureus. 2024;16(5):e59898.
15. Lee TJ, Rao AK, Campbell DJ, et al. Evaluating ChatGPT-3.5 and ChatGPT-4.0 Responses on Hyperlipidemia for Patient Education. Cureus. 2024;16(5):e61067.
16. Vyas R, Pawa A, Shaikh C, et al. ChatGPT for Patients: A Comprehensive Study on Atrial Fibrillation Awareness. J Innov Card Rhythm Manag. 2024;15(7):5946-5949.
17. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs. 2024;23(1):95-98.
18. McDonagh T, Metra M, Adamo M, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42(36):3599-3726.
19. Ghanta SN, Al'Aref SJ, Lala-Trinidade A, et al. Applications of ChatGPT in Heart Failure Prevention, Diagnosis, Management, and Research: A Narrative Review. Diagnostics (Basel). 2024;14(21):2393.
20. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, et al. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digit Health. 2024;10:20552076231224603.
21. Wang L, Wan Z, Ni C, et al. A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. medRxiv. 2024;04.26.24306390.
22. Kaftan AN, Hussain MK, Naser FH. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. Sci Rep. 2024;14(1):8233.
23. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-Generated Differential Diagnosis Lists for Complex Case–Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. JMIR Med Inform. 2023;11:e48808.
24. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330(1):78-80.
25. Leonard M, Graham S, Bonacum D. The human factor: the critical importance of effective teamwork and communication in providing safe care. Qual Saf Health Care. 2004;13(Suppl 1):i85-i90.