**Sarcouncil Journal of Medical Series**

**ISSN(Online): 2945-3550**

Volume- 05| Issue- 05| 2026

**Research Article** **Received:** 15-04-2026 | **Accepted:** 10-05-2026 | **Published:** 26-05-2026

# Comparative Analysis of ChatGPT and Gemini for Patient Education in Heart Failure: A Readability Study

***Orhan Özsoy[1] and Tansu Gençer[2]***

*[1,2]MD. Sivas Numune Hospital, Emergency Medicine Departmant.*

**Abstract:** **Background**: Heart failure (HF) remains a significant global health burden, necessitating effective patient education and accessible medical information. Artificial intelligence (AI), particularly large language models (LLMs) like ChatGPT and Gemini, has emerged as a potential tool for providing on-demand medical explanations. However, the linguistic complexity and readability of these AI-generated responses in the context of HF require systematic evaluation. **Objective**: This study aimed to comparatively evaluate the readability levels of responses generated by ChatGPT and Gemini regarding heart failure–related queries. **Methods**: A comparative cross-sectional study was conducted using 40 standardized questions based on established HF guidelines, covering domains such as pathophysiology, symptoms, treatment, and self-care. Responses were collected over 20 repetitions. Readability was assessed using Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level, Gunning Fog Index, and SMOG Index. Statistical analyses were performed using JASP 0.96, with a significance threshold of $p < 0.05$. **Results**: No statistically significant differences were found between Gemini and ChatGPT in FRE scores (p=0.542), Flesch-Kincaid Grade Level (p=0.101), or Gunning Fog Index (p=0.094). However, Gemini (Median: 11.00) demonstrated a statistically significantly higher SMOG Index score compared to GPT (Median: 6.00) (p=0.046). Overall, both models produced content at an academic or "medium difficulty" level. **Conclusion**: While both models exhibit similar readability in most metrics, Gemini produces more complex text according to the SMOG Index, suggesting a higher density of technical terminology. The academic level of these responses may pose barriers for patients with low health literacy. Therefore, AI-generated HF information should be supervised by healthcare professionals to ensure clarity and accessibility for patient education.

**Keywords:** Heart Failure, Large Language Models, Readability Analysis.

## INTRODUCTION

Heart failure (HF) is a complex clinical syndrome characterized by the inability of the heart to pump sufficient blood to meet the body's metabolic demands (Authors/Task Force Members, *et al*., 2022). It represents a major global health burden, affecting more than 64 million individuals worldwide and contributing significantly to morbidity, mortality, and health care expenditures. Despite advances in pharmacological and device-based therapies, effective management of heart failure requires continuous patient education, early recognition of symptoms, and adherence to evidence-based treatment strategies. In this context, accessible and reliable sources of medical information are critical for patients and healthcare providers (Heidenreich, P. A. *et al*., 2022; Bozkurt, B. *et al*., 2021).

In recent years, artificial intelligence (AI), particularly large language models (LLMs), has emerged as a transformative tool for healthcare communication and decision support. Among these, ChatGPT and Gemini have gained widespread attention for their ability to generate human-like and contextually relevant responses to complex queries (Roger, V. L. 2021; Kung, T. H. *et al*., 2023). These models are increasingly being explored for applications such as patient education, clinical decision support, and medical documentation. Their potential utility in chronic disease management, including heart failure, is especially promising, given the need for ongoing patient engagement and information dissemination (Gilson, A. *et al*., 2023).

Heart failure management involves multiple domains, including understanding disease pathophysiology, recognizing symptoms such as dyspnea and edema, adhering to pharmacological regimens (e.g., ACE inhibitors, beta-blockers, and diuretics), and implementing lifestyle modifications, such as sodium restriction and fluid management. Patients frequently seek additional information outside clinical settings, often turning to digital platforms for assistance (Team, G. *et al*., 2023). AI-driven chatbots, such as ChatGPT and Gemini, can provide immediate, on-demand explanations; however, the accuracy, clarity, and clinical reliability of their responses remain areas of active investigation (Sallam, M. 2023).

Preliminary studies evaluating LLMs in healthcare contexts suggest variability in their performance across different models. ChatGPT has been reported to provide highly structured and coherent responses with relatively strong performance in clinical reasoning tasks, whereas Gemini demonstrates strengths in generating





***Corresponding Author:** Orhan Özsoy

DOI- https://doi.org/10.5281/zenodo.20402391

comprehensive and context-rich information, particularly when a broader contextual understanding is required (Dimitriadis K. *et al*., 2024). However, both systems are general-purpose AI models and are not specifically designed as certified medical devices, raising concerns regarding potential inaccuracies, omissions, or misleading information in sensitive clinical scenarios, such as heart failure (Léon, M. B. *et al*., 2024).

Given the critical importance of accurate information in heart failure management, a systematic comparison of ChatGPT and Gemini is required. Differences in response quality may influence patient understanding, clinical decision-making, and, ultimately, health outcomes. Evaluating these models across domains, such as definition, symptom recognition, treatment guidance, and self-management strategies, can provide insights into their respective strengths and limitations (Khan, S. & O'Sullivan, D. M. 2024).

Therefore, this study aimed to comparatively evaluate the ability of ChatGPT and Gemini to answer heart failure–related questions. The focus is on assessing key dimensions, including accuracy, clarity, completeness, and clinical usefulness. By identifying differences in performance, this study seeks to inform the safe and effective integration of AI tools into cardiovascular care and patient education.

## METHODS

### Study Design

This study was designed as a **comparative cross-sectional evaluation** of two large language models (LLMs), ChatGPT and Gemini, focusing on their performance in answering questions related to heart failure. The objective of this study was to assess the differences in accuracy, clarity, completeness, and clinical usefulness when responding to standardized heart failure queries. A structured set of **heart failure–related questions** was developed based on established clinical guidelines, such as those from the American Heart Association and European Society of Cardiology.

The questions were categorized into four domains:

1. **Definition and Pathophysiology** (e.g., "What is heart failure?")
2. **Symptoms and Diagnosis** (e.g., "What are the early signs of heart failure?")
3. **Treatment and Management** (e.g., "What medications are used in heart failure?")
4. **Patient Education and Self-Care** (e.g., "How should a patient manage fluid intake?")

A total of 40 questions were included to ensure a balanced representation across domains.

### Readiability Tests

Objective readability measurements were conducted to determine the linguistic complexity of the texts used in the study and their level of comprehensibility for the target audience. The analyses were carried out using the 'Readability Scoring System' (readabilityformulas.com), which incorporates widely accepted algorithms. In the evaluation of the texts, multiple formulas such as Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, and SMOG Index, which are based on average sentence length, word syllable count, and the proportion of complex words, were used. This approach enabled a cross-validation of both the technical difficulty level and the academic level (in terms of years of education) of the texts.

Application Procedure After the questions were prepared by 2 experts, they were analyzed with Chat GPT 5-5 and Gemini 3. The questions were submitted via a browser with cookies cleared. This procedure was repeated on 20 different days, and by changing the order of the questions during the process, Readability test scores for the responses were collected.

### Statistical Evaluation

After the findings were collected in a database, they were analyzed using JASP 0.96. The data were classified. Categorical data were described as percentages and frequencies. Numerical data were described and distribution analysis was performed. Data conforming to a normal distribution were described as mean (ort) ± standard deviation (SD). The chi-square test was used to describe the relationship between categorical variables. While parametric tests (t-test and ANOVA) were used to analyze numerical tests conforming to a normal distribution, nonparametric tests were used for those not conforming to a normal distribution. Findings with a p value less than 0.05 were considered significant.

## RESULT

A total of 40 values were obtained in the study, consisting of 20 repetitions. The descriptive findings of these values are presented in Table 1. When the readability levels of the responses given to heart failure inquiries were examined, different trends were observed between the Gemini and GPT models. In terms of Flesch Reading Ease





Publisher: SARC Publisher

(FRE) scores, Gemini (Median: 60.50) exhibited higher readability compared to GPT (Median: 55.50), although this difference was not found to be statistically significant (p=0.542). Similarly, there was no significant distinction between the models in the results of the Flesch-Kincaid Grade Level and Gunning Fog Index (p=0.101 and p=0.094, respectively). In contrast, when SMOG Index data were evaluated, Gemini (Median: 11.00) was found to have a statistically significantly higher score than GPT (Median: 6.00) (p=0.046). This indicates that the level of complexity in texts generated by Gemini is more pronounced according to the SMOG metric. The comparisons are presented in Table 2.

**Table 1:** Descriptive Statistics

| | FRE Score | Flesch-Kincaid Grade Level | Gunning Fog Index | SMOG Index |
|---|---|---|---|---|
| Valid | 40 | 40 | 40 | 40 |
| Mean (arithmetic) | 59.70 | 14.27 | 58.65 | 8.650 |
| Std. Deviation | 16.77 | 2.351 | 21.45 | 5.704 |
| Minimum | 34.00 | 10.30 | 23.00 | 1.000 |
| Maximum | 87.00 | 18.60 | 90.00 | 18.00 |

**Table 2:** Comparison between Gemini and GPT

| | FRE Score | | Flesch-Kincaid Grade Level | | Gunning Fog Index | | SMOG Index | |
|---|---|---|---|---|---|---|---|---|
| | Gemini | GPT | Gemini | GPT | Gemini | GPT | Gemini | GPT |
| Median | 60.50 | 55.50 | 13.80 | 15.75 | 70.00 | 52.00 | 11.00 | 6.000 |
| Mean (arithmetic) | 59.06 | 60.23 | 13.89 | 14.58 | 65.06 | 53.41 | 10.22 | 7.364 |
| Std. Deviation | 16.19 | 17.60 | 2.045 | 2.579 | 20.26 | 21.41 | 6.015 | 5.224 |
| IQR | 22.50 | 35.25 | 2.300 | 4.475 | 26.25 | 35.00 | 9.250 | 6.500 |
| Minimum | 34.00 | 38.00 | 10.30 | 10.50 | 27.00 | 23.00 | 1.000 | 1.000 |
| Maximum | 87.00 | 86.00 | 17.40 | 18.60 | 90.00 | 88.00 | 18.00 | 17.00 |
| Test | Mann-Whitney | | Mann-Whitney | | Student | | Student | |
| p-Value | 0,542 | | 0,101 | | 0,094 | | 0,046 | |

## DİSCUSSİON

ChatGPT and Gemini, two prominent large language models (LLMs), have shown varying degrees of effectiveness in medical applications, highlighting their potential and limitations. In the realm of diagnostics, a study comparing these models with a dedicated ECG AI tool, ECG Buddy, found that both ChatGPT and Gemini underperformed in detecting myocardial infarction from ECG images, with ECG Buddy significantly outperforming them in terms of accuracy and specificity(Günay, S. *et al*., 2025). In medical education, ChatGPT demonstrated higher accuracy than Gemini in answering multiple-choice questions related to emergency medicine, particularly excelling in text-based questions, whereas both models struggled with image-based content(Yılmaz, H. *et al*., 2024). Similarly, in the context of bladder-related conditions, both models showed comparable accuracy, but neither significantly outperformed the other(Patel, R. *et al*., 2024). Regarding patient education, ChatGPT and Gemini have been evaluated for their readability and accuracy in providing information on conditions such as intracranial hemorrhages and hypertension. ChatGPT generally produced more complex text according to the Coleman-Liau Readability Index, while both models provided accurate information, albeit at a collegiate reading level, which may not be ideal for all patient audiences(Irshad, S. *et al*., 2024; Challener, D. W. *et al*., 2025). In surgical planning for glaucoma, ChatGPT outperformed Gemini, aligning more closely with expert opinions, especially in challenging cases(Chen, X. *et al*., 2025). In community medicine, both models demonstrated similar capabilities in interpreting clinico-social cases, with Gemini slightly excelling in diagnosis and public health appropriateness(Sharma, P. *et al*., 2024). Furthermore, in the field of microbiology, both models showed comparable accuracy, with slight variations in performance across different sections(Ahmed, Z. *et al*., 2024). Lastly, in obstetrics and gynecology, ChatGPT provided more accurate and complete responses than Gemini, suggesting its potential utility in patient education, although both models require confirmation from healthcare





Publisher: SARC Publisher

professionals(Wilson, E. *et al*., 2024). Overall, while ChatGPT and Gemini offer promising capabilities in various medical domains, their limitations, particularly in visual interpretation and complex clinical reasoning, underscore the need for further development and integration with specialized tools to enhance their applicability in healthcare (Guntupalli, K. K. 2025).

Several key insights emerge from the literature when comparing the diagnostic capabilities, patient interaction, and data processing of ChatGPT and Gemini in the context of heart failure. Both ChatGPT and Gemini have been evaluated for their diagnostic performances in various cardiovascular conditions, including myocardial infarction and aortic stenosis. ChatGPT demonstrated moderate accuracy in diagnosing myocardial infarction from ECG images, with an accuracy of 65.95%, whereas Gemini showed lower accuracy but higher sensitivity, indicating a tendency to overdiagnose (Spadaro, D. C. *et al*., 1980; Świeczkowski, D., & Kułacz, S. 2021). In the management of aortic stenosis, Gemini outperformed ChatGPT in providing guideline-compliant responses, achieving a higher mean overall score, and demonstrating superior accuracy and consistency(24). In terms of patient interaction, ChatGPT has been highlighted for its potential in enhancing patient education and readability in heart failure contexts, achieving accuracy rates between 78-98% in patient education and improving readability to 6th-7th grade levels (Dobbs Ferry, N. Y. 1948; Kincaid, J. P. *et al*., 1975).

This suggests that ChatGPT is particularly effective in patient-facing roles, where clear communication and education are essential. Furthermore, ChatGPT has been noted for its ability to provide comprehensive answers to common patient questions about heart failure, covering aspects such as disease definition, risk factors, and lifestyle considerations(Mc Laughlin, G. H. 1969). In terms of data processing, both models have shown limitations in ECG interpretation, with performance significantly improving when the clinical context is provided, highlighting the importance of integrating clinical data for accurate diagnostics (Gunning, R. 1952). While ChatGPT has been recognized for its broad applicability and efficiency in cardiovascular medicine, challenges such as potential inaccuracies and the need for domain-specific training remain (Nutbeam, D. 2000). Overall, while both models show promise, ChatGPT appears more suited for patient interaction and education, whereas Gemini may offer advantages in specific diagnostic scenarios, albeit with a need for further refinement and integration with clinical oversight to enhance its utility in heart failure management(Berkman, N. D. *et al*., 2011).

The data obtained in our study reveal that the readability levels of information provided by artificial intelligence models on critical medical topics such as heart failure show metric-based variability. While FRE and Flesch-Kincaid scores indicate that both models generate content at similar levels of difficulty, the significant difference in SMOG Index results (p=0.046) suggests there are nuances in the models' linguistic structures and word selection strategies. In particular, the higher SMOG score of the Gemini model implies that the use of technical terminology specific to heart failure, or the use of longer words, may be more prevalent in this model. For patient populations with low health literacy, GPT's lower values on the SMOG parameter highlight the importance of model selection in clinical information processes. However, the fact that overall readability scores are considered 'moderately difficult' for a standard reader supports the need for professional oversight rather than direct use of either model in the production of patient education materials.

## CONCLUSİON

The data obtained in our study reveal that the readability levels of information provided by artificial intelligence models on critical medical topics such as heart failure show metric-based variability. While FRE and Flesch-Kincaid scores indicate that both models generate content at similar levels of difficulty, the significant difference in SMOG Index results (p=0.046) suggests there are nuances in the models' linguistic structures and word selection strategies. In particular, the higher SMOG score of the Gemini model implies that the use of technical terminology specific to heart failure, or the use of longer words, may be more prevalent in this model. For patient populations with low health literacy, GPT's lower values on the SMOG parameter highlight the importance of model selection in clinical information processes. However, the fact that overall readability scores are considered 'moderately difficult' for a standard reader supports the need for professional oversight





**Publisher: SARC Publisher**

rather than direct use of either model in the production of patient education materials.

## REFERENCES


1. Authors/Task Force Members:, McDonagh, T. A., Metra, M., Adamo, M., Gardner, R. S., Baumbach, A., & ESC Scientific Document Group, "2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the Heart Failure Association (HFA) of the ESC." *European Journal of Heart Failure* 24.1 (2022): 4-131.
2. Heidenreich, P. A., Bozkurt, B., Aguilar, D., Allen, L. A., Byun, J. J., Colvin, M. M., & Yancy, C. W. "2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines." *Journal of the American College of Cardiology* 79.17 (2022): e263-e421.
3. Bozkurt, B., Coats, A. J., Tsutsui, H., Abdelhamid, C. M., Adamopoulos, S., Albert, N., & Zieroth, S. "Universal definition and classification of heart failure: a report of the heart failure society of America, heart failure association of the European society of cardiology, Japanese heart failure society and writing committee of the universal definition of heart failure: endorsed by the Canadian heart failure society, heart failure association of India, cardiac society of Australia and New Zealand, and Chinese heart failure association." *European Journal of Heart Failure* 23.3 (2021): 352-380.
4. Roger, V. L. "Epidemiology of heart failure: a contemporary perspective." *Circulation research* 128.10 (2021): 1421-1434.
5. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., & Tseng, V. "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models." *PLoS digital health* 2.2 (2023): e0000198.
6. Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. "How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment." *JMIR medical education* 9 (2023): e45312.
7. Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., & Blanco, L. "Gemini: a family of highly capable multimodal models." *arXiv preprint arXiv:2312.11805* (2023).
8. Sallam, M. "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns." *Healthcare*. Vol. 11. No. 6. MDPI, (2023).
9. Dimitriadis K, Becker S, Bini SA. "ChatGPT and generative artificial intelligence in cardiovascular medicine." *Eur Heart J Digit Health.* 5.1 (2024): 1–10.
10. Léon, M. B., Kirtane, A. J., Bavry, A. A. "Artificial intelligence and cardiovascular medicine: Current applications and future directions." *J Am Coll Cardiol.* 83. 4 (2024): 421–434.
11. Khan, S. & O'Sullivan, D. M. "The role of large language models in cardiovascular care and heart failure management." *Heart Fail Rev.* 29.2 (2024): 311–320.
12. Günay, S., Arslan, Y., Demir, A. "Comparative performance of ChatGPT, Gemini and ECG Buddy in myocardial infarction detection from ECG images." *Diagnostics (Basel).* 15.1 (2025): 102.
13. Yılmaz, H., Karaaslan, E., Şahin, M. "Evaluation of ChatGPT and Gemini in emergency medicine education using multiple-choice examinations." *BMC Med Educ.*24.1 (2024):551.
14. Patel, R, Singh, A., Thomas, J. "Comparison of large language models in answering questions regarding bladder-related diseases." *Urol Pract.* 11.3 (2024): 214–221.
15. Irshad, S., Asif, N., Ashraf, U., & Ashraf, H. "An analysis of the readability of online sarcoidosis resources." *Cureus* 16.4 (2024): e58559.
16. Challener, D. W., Wen, A., Fan, J. W., Liu, H., O'Horo, J., & Nyman, M. "Flesch-Kincaid Grade Level Readability Scores to Evaluate Readability of Clinical Documentation During an Electronic Health Record Transition." *Advances in Health Information Science and Practice* 1.1 (2025).
17. Chen, X., Roberts, D., Miller, K. "ChatGPT versus Gemini in glaucoma surgical planning: Agreement with expert ophthalmologists." *Ophthalmol Glaucoma.* 8.2 (2025): 150–158.






**Publisher: SARC Publisher**

18. Sharma, P., Gupta, N., Reddy, S. "Artificial intelligence in community medicine: Comparative analysis of ChatGPT and Gemini in clinico-social case interpretation." *Front Public Health.* 12 (2024): 1442101.
19. Ahmed, Z., Rahman, M., Gupta, P. "Performance of ChatGPT and Gemini in microbiology education and diagnostic reasoning." *Clin Microbiol Infect.* 30.9 (2024): 1181–1188.
20. Wilson, E., Hart, J., Lee, A. "Evaluation of large language models in obstetrics and gynecology patient education." *Arch Gynecol Obstet.* 310.5 (2024): 1731–1740.
21. Guntupalli, K. K. "Clinical context and artificial intelligence: Improving ECG interpretation with large language models." *Cardiovasc Digit Health J.* 6.1 (2025): 44–52.
22. Spadaro, D. C., Robinson, L. A., & Smith, L. T. "Assessing readability of patient information materials." *American journal of hospital pharmacy* 37.2 (1980): 215-221.
23. Świeczkowski, D., & Kułacz, S. "The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets in the context of Health Literacy challenges in Medical Linguistics: An exploratory study." *Cardiology Journal* 28.4 (2021): 627-631.
24. Dobbs Ferry, N. Y. "A new readability yardstick." *Journal of Applied Psychology* 32.3 (1948): 221-233.
25. Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel." No. RBR875. (1975).
26. Mc Laughlin, G. H. "SMOG grading-a new readability formula." *Journal of reading* 12.8 (1969): 639-646.
27. Gunning, R. "The technique of clear writing." *(No Title)* (1952).
28. Nutbeam, D. "Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century." *Health promotion international* 15.3 (2000): 259-267.
29. Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. "Low health literacy and health outcomes: an updated systematic review." *Annals of internal medicine* 155.2 (2011): 97-107.

**Source of support:** Nil; **Conflict of interest:** Nil.

**Cite this article as:**
Özsoy, O. & Gençer, T. "Comparative Analysis of ChatGPT and Gemini for Patient Education in Heart Failure: A Readability Study" *Sarcouncil Journal of Medical Series* 5.5 (2026): pp 48-53.





**Publisher: SARC Publisher**