ARCHIVES OF THE TURKISH SOCIETY OF CARDIOLOGY

Official journal of the TURKISH SOCIETY OF CARDIOLOGY

Can Large Language Models Guide Aortic Stenosis Management? A Comparative Analysis of ChatGPT and Gemini AI

Büyük Dil Modelleri Aort Darlığı Yönetimine Rehberlik Edebilir mi? ChatGPT ve Gemini AI'nın Karşılaştırmalı Analizi

ABSTRACT

Objective: Management of aortic stenosis (AS) requires integrating complex clinical, imaging, and risk stratification data. Large language models (LLMs) such as ChatGPT and Gemini AI have shown promise in healthcare, but their performance in valvular heart disease, particularly AS, has not been thoroughly assessed. This study systematically compared ChatGPT and Gemini AI in addressing guideline-based and clinical scenario questions related to AS.

Method: Forty open-ended AS-related questions were developed, comprising 20 knowledge-based and 20 clinical scenario items based on the 2021 European Society of Cardiology/European Association for Cardio-Thoracic Surgery (ESC/EACTS) guidelines. Both models were queried independently. Responses were evaluated by two blinded cardiologists using a structured 4-point scoring system. Composite scores were categorized, and comparisons were performed using Wilcoxon signed-rank and chi-square tests.

Results: Gemini AI achieved a significantly higher mean overall score than ChatGPT (3.96 ± 0.17 vs. 3.56 ± 0.87; P = 0.003). Fully guideline-compliant responses were more frequent with Gemini AI (95.0%) than with ChatGPT (72.5%), although the overall compliance distribution difference did not reach conventional significance (P = 0.067). Gemini AI performed more consistently across both question types. Inter-rater agreement was excellent for ChatGPT (κ = 0.94) and moderate for Gemini AI (κ = 0.66).

Conclusion: Gemini AI demonstrated superior accuracy, consistency, and guideline adherence compared to ChatGPT. While LLMs show potential as adjunctive tools in cardiovascular care, expert oversight remains essential, and further model refinement is needed before clinical integration, particularly in AS management.

Keywords: Aortic stenosis, artificial intelligence, clinical decision support, guideline adherence, large language models

ÖZET

Amaç: Aort darlığı (AD) yönetimi; karmaşık klinik, görüntüleme ve risk sınıflandırma verilerinin entegrasyonunu gerektirir. ChatGPT ve Gemini AI gibi büyük dil modelleri (LLM'ler) sağlık hizmetlerinde umut verici sonuçlar göstermiştir, ancak kapak hastalıklarında, özellikle de AD'deki performansları yeterince değerlendirilmemiştir. Bu çalışma, AD ile ilişkili kılavuz temelli ve klinik senaryo sorularında ChatGPT ile Gemini AI'nın sistematik olarak karşılaştırılmasını amaçlamıştır.

Yöntem: 2021 ESC/EACTS kılavuzları temel alınarak, 20 bilgi temelli ve 20 klinik senaryo sorusundan oluşan toplam 40 açık uçlu AD sorusu geliştirildi. Her iki model de bağımsız olarak sorgulandı. Yanıtlar, ikisi kardiyolog olan iki bağımsız değerlendirici tarafından körleme yöntemiyle, yapılandırılmış 4 puanlık bir sistemle puanlandı. Kompozit puanlar kategorize edildi ve karşılaştırmalar Wilcoxon işaretli sıralar testi ve ki-kare testi ile yapıldı.

Bulgular: Gemini AI, ChatGPT'ye kıyasla anlamlı derecede daha yüksek ortalama toplam puan elde etti (3.96 ± 0.17 vs 3.56 ± 0.87; P = 0.003). Kılavuzlara tamamen uyumlu yanıtlar Gemini AI tarafından daha sık verildi (%95.0 vs %72.5), ancak genel uyum dağılımı geleneksel anlamlılık düzeyine ulaşmadı (P = 0.067). Gemini AI her iki soru türünde de daha tutarlı performans sergiledi. Değerlendiriciler arası uyum ChatGPT için mükemmel (κ = 0.94), Gemini AI için ise orta düzeydeydi (κ = 0.66).

Sonuç: Gemini AI, doğruluk, tutarlılık ve kılavuz uyumu açısından ChatGPT'ye üstünlük göstermiştir. LLM'ler kardiyovasküler bakımda tamamlayıcı araçlar olarak potansiyel taşısa da, uzman denetimi vazgeçilmezdir ve özellikle AD yönetiminde klinik entegrasyon öncesi modellerin daha da geliştirilmesi gerekmektedir.

Anahtar Kelimeler: Aort darlığı, yapay zeka, klinik karar destek, kılavuz uyumu, büyük dil modelleri

ORIGINAL ARTICLE
KLİNİK ÇALIŞMA

Ali Sezgin[1]
Veysel Ozan Tanık[1]
Murat Akdoğan[1]
Yusuf Bozkurt Şahin[1]
Kürşat Akbuğa[1]
Vedat Hekimsoy[1]
Çağatay Tunca[1]
Erhan Saraçoğlu[1]
Bülent Özlek[2]

[1]*Department of Cardiology, Ankara Etlik City Hospital, Ankara, Türkiye*
[2]*Department of Cardiology, Muğla Sıtkı Koçman University, Faculty of Medicine, Muğla, Türkiye*

Corresponding author:
Bülent Özlek
✉ bulent_ozlek@hotmail.com

Received: June 05, 2025
Accepted: August 07, 2025

Cite this article as: Sezgin A, Tanık VO, Akdoğan M, et al. Can Large Language Models Guide Aortic Stenosis Management? A Comparative Analysis of ChatGPT and Gemini AI. *Turk Kardiyol Dern Ars.* 2025;53(0):000–000.

DOI: 10.5543/tkda.2025.54968

Aortic stenosis (AS) is one of the most prevalent and clinically significant valvular heart diseases, particularly affecting elderly populations in developed countries. Its burden continues to rise alongside global population aging, contributing substantially to cardiovascular morbidity and mortality. Severe AS, when left untreated, carries a poor prognosis, with survival rates as low as 50% at two years in symptomatic individuals. Early diagnosis and timely intervention—whether through surgical aortic valve replacement (SAVR) or transcatheter aortic valve implantation (TAVI)—are therefore crucial for improving clinical outcomes.[1] However, determining disease severity and selecting an optimal therapeutic strategy require integrating complex clinical, imaging, and risk stratification data.

In recent years, artificial intelligence (AI) has emerged as a promising tool to support clinical decision-making in cardiology. Among AI applications, large language models (LLMs) such as ChatGPT and Gemini AI have attracted particular attention for their ability to generate human-like responses informed by extensive biomedical knowledge. These systems can interpret clinical scenarios, extract guideline-based recommendations, and assist healthcare providers in managing complex cases. Their roles in patient education, documentation support, and diagnostic guidance are expanding rapidly across various medical disciplines.[2,3]

Despite this growing interest, the reliability and clinical utility of LLMs in specialized areas such as valvular heart disease remain uncertain. AS presents unique diagnostic and therapeutic challenges, particularly in nuanced situations such as low-flow, low-gradient states or asymptomatic patients with high-risk features. Whether AI models can comprehend and apply these subtleties in alignment with current guidelines has yet to be systematically evaluated.[4]

This study aims to evaluate and compare the performance of two widely accessible AI models—ChatGPT and Gemini AI—in their ability to answer guideline-based and scenario-driven questions related to aortic stenosis. By involving expert cardiologists in the evaluation process, we assessed the models' compliance with the 2021 European Society of Cardiology/European Association for Cardio-Thoracic Surgery (ESC/EACTS) valvular heart disease guidelines[1] and explored their potential and limitations as decision-support tools in contemporary cardiology practice.

ABBREVIATIONS

AI	Artificial intelligence
AS	Aortic stenosis
EF	Ejection fraction
ESC/EACTS	European Society of Cardiology/European Association for Cardio-Thoracic Surgery
LLMs	Large language models
SAVR	Surgical aortic valve replacement
TAVI	Transcatheter aortic valve implantation

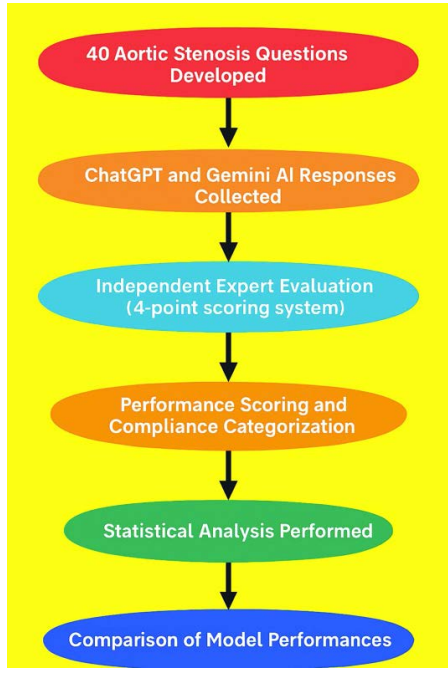

Figure 1. Flowchart summarizing the evaluation process of ChatGPT and Gemini AI models. Forty aortic stenosis-related questions were created and separately submitted to both AI models. Responses were independently assessed by blinded cardiologists using a 4-point scoring system. Compliance categorization and statistical analyses were then performed to compare each model's adherence to clinical guidelines and overall decision-making performance.

Materials and Methods

This study was conducted as a cross-sectional evaluation to assess and compare the aortic stenosis-related clinical knowledge and decision-making performance of two LLMs: ChatGPT-4 (OpenAI) and Gemini AI (Google). Specifically, we used the GPT-4-turbo model, available via ChatGPT-Plus, and the Gemini 1.5 Pro model, accessed via Google's web interface, both in May 2025. The evaluation was based on the 2021 ESC/EACTS Guidelines for the Management of Valvular Heart Disease.[1] No human or animal participants were involved in this study.

The overall study workflow is summarized in Figure 1. A total of 40 open-ended questions were developed by two experienced cardiologists, comprising 20 knowledge-based items and 20 clinical scenarios (Table 1). The questions were phrased in a standardized open-ended format and independently reviewed to ensure clinical neutrality and consistency. All prompts were submitted in a zero-shot setting—without prior examples, role prompts, or chain-of-thought instructions—to eliminate prompt engineering bias. The identical questions were presented to both models in clean, isolated sessions. These questions were specifically designed to reflect the core principles of diagnosis, risk stratification, and treatment decision-making in aortic stenosis, as outlined in guideline-recommended practices.[1] The clinical scenarios were designed to simulate realistic patient cases with varying presentations, comorbidities, and surgical risks, providing a comprehensive framework for assessing the AI models' ability to reason through complex situations. Both AI models were presented with the same set of questions under standardized conditions. Each question was submitted separately to the respective model in independent sessions, using a new prompt environment to prevent memory-based carryover effects. All prompts were presented in a uniform

Table 1. Expert-assigned scores for each of the 40 questions evaluating ChatGPT and Gemini AI. Scores are shown alongside their corresponding guideline compliance category

Question no	Question type	Question text	Overall ChatGPT score	ChatGPT category	Overall Gemini score	Gemini category
1	Knowledge-based	What are the echocardiographic criteria for diagnosing severe aortic stenosis?	4.0	Fully compliant	4.0	Fully compliant
2	Knowledge-based	What are the surgical indications for asymptomatic patients with aortic stenosis?	1.0	Incorrect / Misleading	3.0	Partially compliant
3	Knowledge-based	What is the diagnostic approach in low-flow/low-gradient aortic stenosis?	3.5	Mostly compliant	4.0	Fully compliant
4	Knowledge-based	What is the STS (Society of Thoracic Surgeons) risk score, and how does it affect treatment decisions?	4.0	Fully compliant	4.0	Fully compliant
5	Knowledge-based	What are the indications for transcatheter aortic valve implantation (TAVI) in aortic stenosis?	4.0	Fully compliant	4.0	Fully compliant
6	Knowledge-based	In patients with aortic stenosis and coronary artery disease, what is the order of intervention?	4.0	Fully compliant	4.0	Fully compliant
7	Knowledge-based	What are the most common complications after transcatheter aortic valve replacement (TAVR)?	4.0	Fully compliant	4.0	Fully compliant
8	Knowledge-based	What should be done in patients with severe aortic stenosis (AS), normal ejection fraction (EF), and low gradient?	4.0	Fully compliant	4.0	Fully compliant
9	Knowledge-based	What is the recommended approach for young patients with bicuspid aortic stenosis?	4.0	Fully compliant	4.0	Fully compliant
10	Knowledge-based	Is beta-blocker therapy recommended for patients with aortic stenosis?	3.0	Partially compliant	4.0	Fully compliant
11	Knowledge-based	What are the advantages of TAVI in geriatric patients?	4.0	Fully compliant	4.0	Fully compliant
12	Knowledge-based	How should aortic stenosis be managed in patients with reduced EF?	4.0	Fully compliant	4.0	Fully compliant
13	Knowledge-based	What is the role of angiotensin-converting enzyme (ACE) inhibitors in aortic stenosis?	4.0	Fully compliant	4.0	Fully compliant
14	Knowledge-based	What does a paradoxical low-flow state mean in echocardiography?	4.0	Fully compliant	4.0	Fully compliant
15	Knowledge-based	What is the role of computed tomography (CT) in the diagnosis of aortic stenosis?	3.0	Partially compliant	4.0	Fully compliant
16	Knowledge-based	What are the non-surgical treatment options for severe AS?	4.0	Fully compliant	4.0	Fully compliant
17	Knowledge-based	When is stress echocardiography necessary in the evaluation of aortic stenosis?	4.0	Fully compliant	4.0	Fully compliant
18	Knowledge-based	How should pregnant patients with aortic stenosis be monitored and managed?	4.0	Fully compliant	4.0	Fully compliant
19	Knowledge-based	How does the calcification process develop in aortic valve biology?	4.0	Fully compliant	4.0	Fully compliant
20	Knowledge-based	What is ventriculo-arterial coupling in aortic stenosis and its prognostic significance?	4.0	Fully compliant	4.0	Fully compliant
21	Clinical scenario	What is your treatment approach for an 82-year-old patient with EF 55%, NYHA III (New York Heart Association Class III), and STS risk of 10%?	4.0	Fully compliant	4.0	Fully compliant
22	Clinical scenario	What would you recommend for a 70-year-old patient with EF 38%, aortic valve area (AVA) 0.6 cm^2, and a mean gradient of 28 mmHg?	4.0	Fully compliant	4.0	Fully compliant
23	Clinical scenario	What would you suggest for a 58-year-old asymptomatic patient with normal EF, AVA 0.75 cm^2, and Vmax 4.3 m/s?	4.0	Fully compliant	4.0	Fully compliant
24	Clinical scenario	What is your treatment choice for a 79-year-old patient with EF 60%, chronic kidney disease (CKD) stage 3, NYHA II, and STS 6.5%?	4.0	Fully compliant	4.0	Fully compliant

Table 1 (cont). Expert-assigned scores for each of the 40 questions evaluating ChatGPT and Gemini AI. Scores are shown alongside their corresponding guideline compliance category

Question no	Question type	Question text	Overall ChatGPT score	ChatGPT category	Overall Gemini score	Gemini category
25	Clinical scenario	Would you recommend further testing for a 65-year-old patient with EF 48% and a mean gradient of 32 mmHg?	4.0	Fully compliant	4.0	Fully compliant
26	Clinical scenario	What is the management plan for an 87-year-old with severe chronic obstructive pulmonary disease (COPD) and anemia, AVA 0.7 cm^2?	3.0	Partially compliant	4.0	Fully compliant
27	Clinical scenario	What is your approach for a 55-year-old patient with a congenital bicuspid valve and AVA 0.9 cm^2?	4.0	Fully compliant	4.0	Fully compliant
28	Clinical scenario	TAVI or SAVR (surgical aortic valve replacement): What would you recommend for a 60-year-old patient with EF 30% and low-gradient severe AS?	2.0	Not compliant	4.0	Fully compliant
29	Clinical scenario	How would you manage a 72-year-old patient with active malignancy and severe AS findings?	4.0	Fully compliant	4.0	Fully compliant
30	Clinical scenario	What is the next step for an 80-year-old with AVA 0.6 cm^2, a gradient of 22 mmHg, and a low stroke volume?	3.0	Partially compliant	4.0	Fully compliant
31	Clinical scenario	Surgery or TAVI: What is the best option for a 66-year-old with multivessel disease and severe AS?	3.0	Partially compliant	4.0	Fully compliant
32	Clinical scenario	Follow-up or intervention: What would you do for a 76-year-old with a high frailty score and AVA of 0.5 cm^2?	4.0	Fully compliant	4.0	Fully compliant
33	Clinical scenario	Is surgery indicated in a 59-year-old asymptomatic patient with left ventricular (LV) hypertrophy and EF 52%?	1.0	Incorrect / Misleading	4.0	Fully compliant
34	Clinical scenario	Is a patient with prior coronary artery bypass grafting (CABG) surgery at age 83 a suitable TAVI candidate?	4.0	Fully compliant	4.0	Fully compliant
35	Clinical scenario	How should combined treatment be planned for a 64-year-old with 90% left anterior descending (LAD) stenosis and severe AS?	1.0	Incorrect / Misleading	4.0	Fully compliant
36	Clinical scenario	What confirmatory test would you perform for a 70-year-old with AVA 0.8, low-flow/low-gradient AS, and EF 45%?	4.0	Fully compliant	4.0	Mostly compliant
37	Clinical scenario	Is urgent intervention needed for a 75-year-old patient with syncope despite medical therapy?	4.0	Fully compliant	4.0	Mostly compliant
38	Clinical scenario	How does chronic atrial fibrillation affect management in a 68-year-old patient with AVA 0.7?	4.0	Fully compliant	3.5	Mostly compliant
39	Clinical scenario	What is the recommendation for an 85-year-old with cognitive impairment and high surgical risk?	4.0	Fully compliant	4.0	Partially compliant
40	Clinical scenario	How should a 73-year-old patient with EF 58%, AVA 0.8, and severe diastolic dysfunction be evaluated?	3.0	Partially compliant	4.0	Fully compliant

format as plain-text, open-ended clinical questions. No system messages or role instructions were provided prior to submission. Each model received identical questions individually in a clean session without any prior conversation history. Prompt length and phrasing were standardized to maintain consistency across models and ensure a fair evaluation. Responses were recorded without editing or refinement, and all identifying data were removed to ensure objectivity. Each response was independently evaluated by two cardiologists who were blinded to the source of the answer (ChatGPT or Gemini). A structured 4-point ordinal scoring system was used:

- 4 = Fully Guideline-Compliant – Accurate, complete, and consistent with current ESC/EACTS recommendations.
- 3 = Partially Guideline-Compliant – Generally accurate but with minor omissions or incomplete details.
- 2 = Non-Compliant – Significant deviation from guideline recommendations.
- 1 = Incorrect or Misleading – Contains factual errors or potentially unsafe suggestions.

The arithmetic mean of the two reviewers' scores was calculated for each AI response to produce a composite performance metric. These mean values, which could include decimal scores (e.g., 3.5), were then grouped into interpretive performance bands to enhance clinical interpretability:

- 4.0 → Fully compliant
- 3.5–3.9 → Mostly compliant
- 3.0–3.4 → Partially compliant
- 2.0–2.9 → Not compliant
- <2.0 → Incorrect/Misleading.

Table 2. Descriptive statistics of AI scores

Model	Mean score	Median score	Standard deviation	Minimum score	Maximum score
ChatGPT	3.56	4.0	0.87	1.0	4.0
Gemini AI	3.96	4.0	0.17	3.0	4.0

AI: Artificial intelligence.

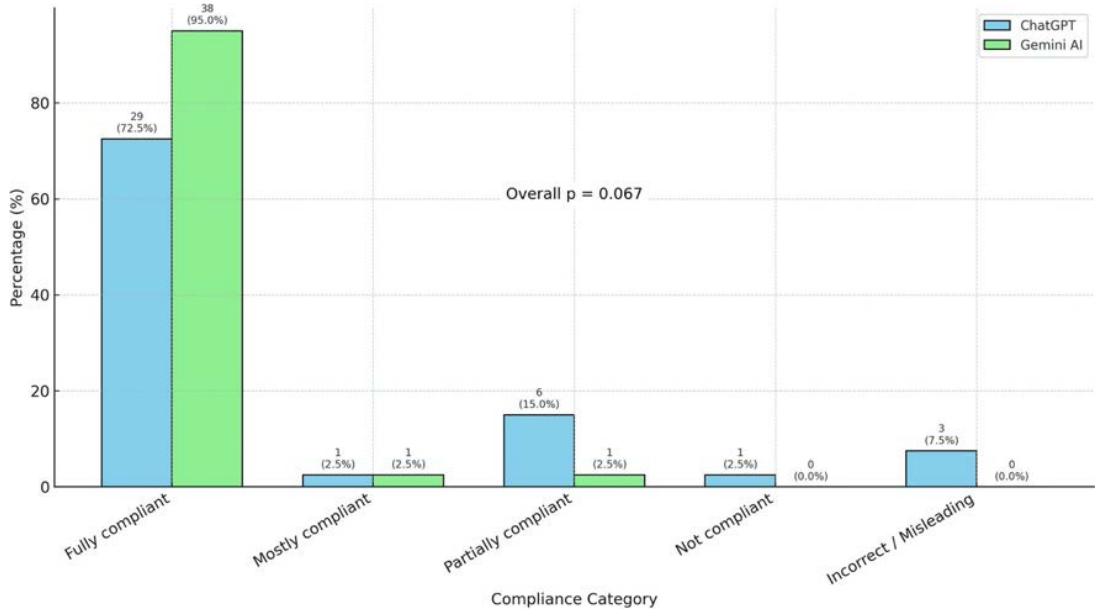

Figure 2. Distribution of responses across compliance categories for ChatGPT and Gemini AI. Fully compliant responses were significantly more frequent with Gemini AI than with ChatGPT (95.0% vs. 72.5%), while ChatGPT had higher rates of partially compliant, not compliant, and incorrect/misleading answers. Although the overall difference did not reach traditional statistical significance (overall P = 0.067, chi-square test), a numerical trend favoring Gemini AI was observed. Data are presented as both absolute counts (n) and percentages (%).

This approach aligns with scoring strategies used in recent AI evaluation studies,[2] where ordinal expert scores are averaged and mapped to descriptive categories. Such methods are widely employed to consolidate expert ratings while preserving both clinical relevance and statistical interpretability.

The study was conducted in accordance with the recommendations outlined in the Declaration of Helsinki for biomedical research involving human subjects. As it did not involve human or animal participants, patient data, or identifiable information, institutional ethics committee approval was not required. While the study focuses on evaluating the consistency of AI in the therapeutic management of AS, no artificial intelligence tools were utilized in the preparation, analysis, or writing of this manuscript.

Statistical Analysis

Descriptive statistics were calculated for each model, including the mean, median, and score distribution. The Wilcoxon signed-rank test was applied to compare the overall performance scores of ChatGPT and Gemini AI. This non-parametric test was chosen due to the ordinal nature of the scoring system (1–4) and the paired design of the data, as each model answered the same set of questions. Inter-rater agreement between the two cardiologists was measured using Cohen's kappa coefficient, with values above 0.8 interpreted as excellent agreement and values between 0.6 and 0.8 interpreted as moderate agreement. Additionally, a chi-square test was performed to compare the distribution of responses across categorical compliance levels (fully compliant, mostly compliant, partially compliant, not compliant, and incorrect/misleading) between the two AI models. All statistical analyses were two-tailed, and a p-value <0.05 was considered statistically significant. Analyses were conducted using SPSS (IBM Corp., Armonk, NY, USA) and Python (v3.10) with the matplotlib and scikit-learn libraries for figure generation and advanced modeling.

Results

Descriptive statistics of expert-assigned scores are shown in Table 2. ChatGPT had a lower average score (mean score = 3.56) and greater variability [standard deviation (SD) = 0.87] compared with Gemini AI, which achieved scores closer to the maximum (mean score = 3.96; SD = 0.17). Although both models had a median score of 4.0, the broader score range for ChatGPT suggests a higher proportion of partially compliant or incorrect responses.

Overall Compliance with Guidelines

A detailed summary of the categorical score distribution is provided in Figure 2.

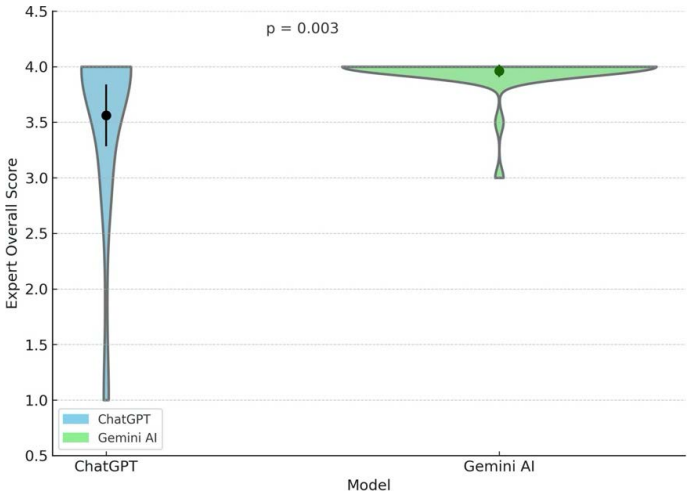

Figure 3. Violin plot showing the distribution of expert-assigned overall scores for ChatGPT and Gemini AI. Each model's distribution is displayed with mean scores (dots) and 95% confidence intervals (error bars). Gemini AI achieved a significantly higher mean overall score than ChatGPT (P = 0.003, Wilcoxon signed-rank test). Expert scores were based on a 4-point ordinal scale measuring adherence to clinical guidelines.

Among ChatGPT's 40 responses, 29 (72.5%) were classified as fully guideline-compliant, while 11 (27.5%) were categorized as partially compliant, not compliant, or incorrect. Specifically, six responses (15.0%) were partially compliant, one (2.5%) was not compliant, and three (7.5%) were considered incorrect or potentially misleading. In contrast, Gemini AI demonstrated greater consistency and a higher level of compliance. Of its 40 responses, 38 (95.0%) were fully compliant, while only two (5%) fell outside this category—one was mostly compliant and one was partially compliant. No responses from Gemini were rated as incorrect or non-compliant. Although Gemini AI demonstrated a numerically higher rate of fully compliant responses compared with ChatGPT (95.0% vs. 72.5%), the overall difference across all compliance categories did not reach statistical significance (P = 0.067, chi-square test).

Performance Across Question Types

When analyzed by question type, Gemini AI achieved full compliance in 90% of theoretical questions and 85% of clinical scenarios. ChatGPT reached full compliance in 75% of theoretical questions and 70% of clinical scenarios. ChatGPT's weaker performance in clinical questions reflected its difficulty integrating patient complexity, surgical risk assessment, and low-gradient decision-making algorithms.

Comparison of the AI Models

A Wilcoxon signed-rank test was conducted to compare the overall performance of ChatGPT and Gemini AI based on expert-assigned scores. Gemini AI achieved a significantly higher average score (3.96 ± 0.17) compared with ChatGPT (3.56 ± 0.87). The difference between the two models was statistically significant (P = 0.003), indicating that Gemini AI provided a more consistent and guideline-compliant output. The distribution of expert scores for both models is shown in Figure 3.

Inter-Rater Agreement

Cohen's kappa coefficient was used to evaluate inter-rater reliability. Agreement was excellent for ChatGPT evaluations (κ = 0.94) and moderate for Gemini AI evaluations (κ = 0.66). This apparent discrepancy likely reflects a ceiling effect: the almost unanimous high scores given to Gemini AI responses reduced kappa sensitivity despite minimal actual disagreement.

Observations on Common Errors

ChatGPT's partially or non-compliant responses were often due to vague thresholds (e.g., citing ejection fraction [EF] <50–55% instead of the guideline-defined <55%), incorrect prioritization of treatment options, or omission of specific criteria for surgical decision-making in low-flow states. In contrast, Gemini AI's rare deficiencies involved subtle oversimplifications but seldom contradicted ESC/EACTS recommendations.

Discussion

This study provides a new, structured comparison of two publicly available large language models—ChatGPT and Gemini AI—in the specific and complex clinical setting of AS management. While previous research has evaluated LLMs on general medical examinations or broad clinical scenarios,[5-7] to our knowledge, this is the first study focused exclusively on a high-stakes, guideline-driven cardiovascular condition. This targeted approach enables a detailed assessment of each model's ability to reason through real-world clinical decision-making processes.

Gemini AI demonstrated superior adherence to guidelines, achieving full compliance in 95.0% of responses compared with 72.5% for ChatGPT. Although the overall compliance rates across all categories did not reach statistical significance, the notable difference in the proportion of fully compliant responses between the models may represent a clinically meaningful advantage for Gemini AI. Notably, Gemini AI produced no responses classified as non-compliant or incorrect, whereas ChatGPT generated several partially compliant and some misleading responses. These differences were most evident in complex clinical scenarios, such as low-flow, low-gradient AS or asymptomatic high-risk patients, where knowing how to interpret hemodynamic data and stratify risk is essential. These findings suggest that while both models possess strong factual knowledge, Gemini AI may interpret clinical complexities more reliably. These results are consistent with prior studies showing that newer or more specialized LLMs outperform generalist models in clinical reasoning.[8] Our findings highlight that domain-specific complexity—such as the nuanced decision-making thresholds in AS—can reveal significant performance gaps in AI models not specifically tailored for medical tasks.

Several factors may explain the observed differences. First, Gemini AI may incorporate more current or finely tuned medical training data than ChatGPT-4. Second, Gemini's response algorithms could focus on guideline-based patterns and threshold-specific logic, thereby reducing variability and generalization errors observed in ChatGPT.[9-11] Prior research suggests that "prompt engineering" and "model alignment" strategies have a significant impact on AI performance in clinical settings.[12] Additionally, differences in how each model processes conditional decision trees—a central element in AS management—may explain the discrepancies.

Our findings reinforce the potential of LLMs as adjunctive tools in cardiovascular care, particularly in domains with stringent, evidence-based guidelines. For example, LLMs could assist in standardized case triage, preliminary risk stratification, or continuing medical education. Recent work has also demonstrated the value of LLMs in aligning with cardiology guidelines and supporting structured decision-making workflows.[13] However, the variability noted with ChatGPT underscores the ongoing need for expert oversight. In AS, where therapeutic decisions such as intervention timing critically impact survival, reliance on non-validated AI recommendations could pose safety risks. Although this study assessed performance quantitatively, it did not formally investigate how each model internally constructs its reasoning. Understanding which textual features or decision thresholds influence LLM-generated recommendations is crucial for clinician confidence. Incorporating transparent reasoning mechanisms has been shown to significantly improve user trust and facilitate model adoption in clinical contexts.[14] Future studies should consider integrating explainability frameworks—such as rationale tracing or language-based model introspection—to improve interpretability and transparency.

Given the rapid pace of development in large language models, it is important to interpret our findings as a reflection of model capabilities at a specific point in time (May 2025). Both ChatGPT and Gemini AI are evolving platforms that may undergo substantial changes in performance, reasoning strategies, and guideline adherence in future iterations. This transient nature introduces inherent challenges for reproducibility and long-term clinical reliability. Accordingly, our study should be viewed as an early-stage evaluation rather than a definitive benchmark. These dynamics have been highlighted in recent work on biomedical LLM benchmarks, where inconsistent performance across versions in zero- and few-shot settings underscores the need for ongoing, adaptive benchmarking models rather than static one-time assessments.[15] Future research should focus on developing adaptive validation frameworks that account for version changes and enable continuous performance monitoring over time. The relatively high accuracy of Gemini AI indicates that, with additional medical fine-tuning, LLMs could eventually support—or even partially automate—decision-support tools for managing valvular heart disease. However, current models are still inadequate for unsupervised clinical use, consistent with recent concerns about hallucinations, overconfidence, and factual inaccuracies.[10,16]

Recent evaluations of LLMs on medical licensing examinations have reported overall accuracies ranging from 60% to 80%.[9,10] Our findings show slightly higher compliance rates, likely due to the study's use of tightly structured, guideline-based questions rather than broad knowledge domains. Nevertheless, the drop in ChatGPT's performance for clinical scenarios echoes prior observations: LLMs often excel at knowledge recall but struggle when integration, synthesis, and nuanced judgment are required.[11,12] Building on these observations, it is crucial to emphasize the potential clinical implications of integrating AI into the management of valvular heart disease. With further refinement and external validation, LLMs could help clinicians streamline diagnostic workflows, identify high-risk patients for early intervention, and ensure adherence to evidence-based practices. However, AI outputs should be viewed as complementary rather than definitive, serving to augment, not replace clinical expertise. Collaborative models that combine AI-driven suggestions with physician oversight may ultimately provide the most balanced and safe approach to leveraging these technologies in complex cardiovascular care.

Ongoing developments in LLMs suggest that domain-specific fine-tuning ("medically aligned LLMs") and multi-modal capabilities (e.g., integration of imaging data) will be critical next steps.[4,12] Future research should evaluate LLM performance in real-time clinical simulations, across broader valvular diseases (e.g., mitral regurgitation, tricuspid valve disease), and in diverse clinical settings, including low-resource environments. Ethical considerations, such as explainability, bias minimization, and clinician-AI collaborative workflows, will also need to be addressed.[16]

Study Limitations

This study has several important limitations that warrant consideration. Although expert-based scoring provides a structured framework for evaluation, some degree of subjectivity is inevitable, even with high inter-rater agreement. The study focused exclusively on AS and did not examine model performance in other valvular pathologies such as mitral or tricuspid disease, which limits generalizability. Moreover, both ChatGPT and Gemini AI are rapidly evolving platforms; their performance may vary significantly across different versions, and the present findings reflect only the capabilities of the specific model versions available in May 2025. All interactions were conducted in English, and no assessment was made of model performance in non-English clinical settings or multicultural contexts. Another limitation relates to the artificial nature of the testing environment. The use of isolated, pre-formulated prompts does not fully replicate the dynamic, iterative decision-making processes encountered in real-world clinical practice. In particular, multimodal data inputs—such as echocardiographic images, laboratory values, or structured electronic health records—were not incorporated, potentially underestimating the complexity of actual clinical reasoning. Additionally, the impact of AI-assisted responses on clinical decision-making accuracy, workflow efficiency, or patient outcomes was not assessed. Since both models rely exclusively on textual input, their outputs are highly sensitive to prompt clarity, phrasing, and completeness, raising concerns about reproducibility and context dependence. Lastly, the study did not evaluate how clinicians with varying levels of expertise interpret or act upon AI-generated responses. Human-AI interaction dynamics, cognitive bias, and user trust are all critical factors that could affect the utility, safety, and real-world applicability of such tools. These limitations collectively underscore the importance of cautious interpretation and the continued need for rigorous validation before any clinical deployment of large language models.

Conclusion

This study represents the first structured evaluation directly comparing two LLMs—ChatGPT and Gemini AI—in the context of guideline-driven management of AS. Our findings show that while both models demonstrate substantial factual knowledge, significant variability exists in their clinical reasoning capabilities.

Gemini AI consistently provided more accurate and guideline-compliant responses, whereas ChatGPT exhibited greater variability, particularly in complex clinical scenarios that required nuanced judgment. These results underscore the critical importance of expert validation when employing AI tools in high-stakes, patient-specific decision-making. Although LLMs show promise as adjuncts in medical education and preliminary decision support, they are not yet suitable for independent clinical application, particularly in complex domains such as valvular heart disease. Until such models undergo specialized medical alignment, fine-tuning, and rigorous real-world validation, human expertise remains irreplaceable. Future developments should prioritize enhancing LLM accuracy, minimizing response variability, and improving transparency and interpretability. Ethical deployment frameworks, strong data privacy safeguards, and clinician-centered integration strategies will be essential to safely leverage AI's potential. With continued evolution and responsible implementation, AI tools may ultimately contribute to a more efficient, equitable, and evidence-based cardiovascular care system.

Ethics Committee Approval: This study did not involve human or animal participants, patient data, or identifiable information. Therefore, institutional ethics committee approval was not required.

Informed Consent: Written informed consent was not required for this study.

Conflict of Interest: The authors have no conflicts of interest to declare.

Funding: The authors declared that this study received no financial support.

Use of AI for Writing Assistance: No artificial intelligence tools were utilized in the preparation, analysis, or writing of this manuscript.

Author Contributions: Concept – A.S., V.O.T., M.A.; Design – A.S., V.O.T., M.A.; Supervision – A.S., V.O.T., M.A., Y.B.Ş., K.A., V.H., Ç.T., E.S., B.Ö.; Materials – A.S., V.O.T.; Data Collection and/or Processing - A.S., Y.B.Ş., K.A.; Analysis and/or Interpretation - A.S., V.O.T., V.H., Ç.T., E.S., B.Ö.; Literature Review – A.S., V.O.T., B.Ö.; Writing – A.S., V.O.T., M.A., Y.B.Ş., K.A., V.H., Ç.T., E.S., B.Ö.; Critical Review – A.S., V.O.T., M.A., Y.B.Ş., K.A., V.H., Ç.T., E.S., B.Ö.

Peer-review: Externally peer-reviewed.

References

1. Vahanian A, Beyersdorf F, Praz F, et al.; ESC/EACTS Scientific Document Group. 2021 ESC/EACTS Guidelines for the management of valvular heart disease. *Eur Heart J*. 2022;43(7):561-632. Erratum in: *Eur Heart J*. 2022;43(21):2022. [CrossRef]
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. [CrossRef]
3. Gilson A, Safranek CW, Huang T, et al. 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 Med Educ*. 2023;9:e45312. Erratum in: *JMIR Med Educ*. 2024;10:e57594. [CrossRef]
4. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. *Eur Heart J*. 2024;45(5):332-345. [CrossRef]
5. Geneş M, Yaşar S, Fırtına S, et al. Artificial Intelligence in Cardiac Rehabilitation: Assessing ChatGPT's Knowledge and Clinical Scenario Responses. *Turk Kardiyol Dern Ars*. 2025;53(3):173-177. [CrossRef]
6. Lang Q, Zhong C, Liang Z, et al. Six application scenarios of artificial intelligence in the precise diagnosis and treatment of liver cancer. *Artif Intell Rev*. 2021;54(7):5307-5346. [CrossRef]
7. Simon ST, Mandair D, Tiwari P, Rosenberg MA. Prediction of Drug-Induced Long QT Syndrome Using Machine Learning Applied to Harmonized Electronic Health Record Data. *J Cardiovasc Pharmacol Ther*. 2021;26(4):335-340. [CrossRef]
8. Carl N, Schramm F, Haggenmüller S, et al. Large language model use in clinical oncology. *NPJ Precis Oncol*. 2024;8(1):240. [CrossRef]
9. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7(1):258. [CrossRef]
10. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and Other Large Language Models in Medical Education - Scoping Literature Review. *Med Sci Educ*. 2024;35(1):555-567. [CrossRef]
11. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. Erratum in: *Nature*. 2023;620(7973):E19. [CrossRef]
12. Hirata K, Matsui Y, Yamada A, et al. Generative AI and large language models in nuclear medicine: current status and future prospects. *Ann Nucl Med*. 2024;38(11):853-864. Erratum in: *Ann Nucl Med*. 2025;39(4):404-405. [CrossRef]
13. Ferreira Santos J, Ladeiras-Lopes R, Leite F, Dores H. Applications of large language models in cardiovascular disease: a systematic review. *Eur Heart J Digit Health*. 2025;6(4):540-553. [CrossRef]
14. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI*. 2024;3:e53207. [CrossRef]
15. Chen Q, Hu Y, Peng X, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat Commun*. 2025;16(1):3280. [CrossRef]
16. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. [CrossRef]