ORIGINAL RESEARCH / ORİJİNAL ARAŞTIRMA

DOI: 10.4274/kvbulten.galenos.2026.08370
Bull Cardiovasc Acad 2026;4(1):20-28

Guideline-based Evaluation of Large Language Models in Heart Failure with Preserved Ejection Fraction

Korunmuş Ejeksiyon Fraksiyonlu Kalp Yetersizliğinde Büyük Dil Modellerinin Kılavuz Temelli Değerlendirilmesi

Hatice Taşkan[1], Selen Eşki[2], Özkan Eravcı[3], Doğan İliş[4], Sinem Çakal[1]

[1]University of Health Sciences Türkiye, İstanbul Haseki Training and Research Hospital, Department of Cardiology, İstanbul, Türkiye
[2]Doğubayazıt Dr. Yaşar Eryılmaz State Hospital, Clinic of Cardiology, Ağrı, Türkiye
[3]University of Health Sciences Türkiye, Gülhane Training and Research Hospital, Department of Cardiology, Ankara, Türkiye
[4]Kafkas University Faculty of Medicine, Department of Cardiology, Kars, Türkiye

Abstract

Objectives: Large language models (LLM) are increasingly used by clinicians as informal decision-support tools. However, most benchmark studies emphasize aggregate accuracy and do not clarify whether errors are evenly distributed across clinical tasks. We evaluated whether contemporary LLMs exhibit a domain-specific reliability gap in the management of heart failure with preserved ejection fraction (HFpEF).

Material and Methods: In this cross-sectional benchmark study, four contemporary LLMs (Gemini 2.5 Pro, ChatGPT-5, Microsoft Copilot, and Claude Sonnet 4.5) were queried with 80 guideline-anchored items derived from the 2022 American Heart Association/American College of Cardiology/Heart Failure Society of America heart failure guideline, the 2021 European Society of Cardiology (ESC) guideline, and the 2023 ESC focused update. Items were classified into four domains: diagnosis and pathophysiology (n=16), comorbidity management (n=11), treatment and drug management (n=12), and guideline comparison and clinical scenarios (n=41). Responses were evaluated against pre-specified gold-standard answers by three cardiologists using a four-point ordinal rubric; disagreements were resolved by consensus, with senior adjudication when required. Domain-level differences were analyzed using Fisher's exact test.

Results: Across 320 responses, comprehensive accuracy ranged from 65.0% to 87.5%. No model produced a critical error in either the diagnosis or the comorbidity domain. In contrast, critical error rates rose markedly in the treatment and drug-management domain, reaching 50.0% for Claude Sonnet 4.5, 33.3% for ChatGPT-5, and 25.0% for Microsoft Copilot, whereas Gemini 2.5 Pro produced no critical errors. When pooled across models, critical errors were significantly more frequent in the treatment domain than in all other domains combined (odds ratio: 4.68, $p<0.001$). Failures clustered around newer-generation therapies (including sodium-glucose cotransporter 2 inhibitors and semaglutide), the HFpEF-specific iron-deficiency recommendation, and the prioritization of drugs for concomitant hypertension.

Conclusion: Contemporary LLMs perform well on conceptual and diagnostic HFpEF questions but show a clinically meaningful reliability gap in guideline-concordant pharmacotherapy. These findings suggest that treatment-related LLM outputs should not be used without independent verification against primary guideline sources.

Keywords: Heart failure with preserved ejection fraction, large language models, artificial intelligence, clinical decision support, guideline concordance, pharmacotherapy

Address for Correspondence/Yazar Adresi: Hatice Taşkan, MD, University of Health Sciences Türkiye, İstanbul Haseki Training and Research Hospital, Department of Cardiology, İstanbul, Türkiye
E-mail: haticetaskan@gmail.com **ORCID ID:** orcid.org/0000-0001-8309-6076
Received/Geliş Tarihi: 16.04.2026 **Accepted/Kabul Tarihi:** 26.04.2026 **Published Date/Yayınlanma Tarihi:** 30.04.2026

Cite this article as/Atıf: Taşkan H, Eşki S, Eravcı Ö, İliş D, Çakal S. Guideline-based evaluation of large language models in heart failure with preserved ejection fraction. Bull Cardiovasc Acad. 2026;4(1):20-28

Öz

Amaç: Büyük dil modelleri (BDM), klinisyenler tarafından giderek daha sık gayri resmi karar destek aracı olarak kullanılmaktadır. Ancak mevcut çalışmalar genellikle toplam doğruluğa odaklanmakta, hataların klinik görevler arasında eşit dağılıp dağılmadığını yeterince göstermemektedir. Bu çalışma, güncel BDM'lerin korunmuş ejeksiyon fraksiyonlu kalp yetersizliği (KEF-KY) yönetiminde alana özgü bir güvenilirlik açığı sergileyip sergilemediğini değerlendirmeyi amaçladı.

Yöntem ve Gereçler: Bu kesitsel kıyaslama çalışmasında dört güncel BDM (Gemini 2.5 Pro, ChatGPT-5, Microsoft Copilot ve Claude Sonnet 4.5), 2022 Amerikan Kalp Derneği/Amerikan Kardiyoloji Koleji/Amerikan Kalp Yetmezliği Derneği, 2021 Avrupa Kardiyoloji Derneği (ESC) ve 2023 ESC güncellenmiş kalp yetersizliği kılavuzlarından türetilen 80 madde ile sorgulandı. Sorular tanı ve patofizyoloji (n=16), komorbidite yönetimi (n=11), tedavi ve ilaç yönetimi (n=12) ile kılavuz karşılaştırması ve klinik senaryolar (n=41) olmak üzere dört alanda sınıflandırıldı. Yanıtlar, önceden belirlenmiş altın standart cevaplara göre üç kardiyolog tarafından dört puanlı ordinal bir ölçekle değerlendirildi; anlaşmazlıklar konsensus ile çözüldü ve gerektiğinde kıdemli bir kardiyolog tarafından karara bağlandı. Alan düzeyindeki farklılıklar Fisher'ın kesin testi ile incelendi.

Bulgular: Toplam 320 yanıtın değerlendirilmesinde kapsamlı doğruluk oranı %65,0 ile %87,5 arasında değişti. Tanı ve komorbidite yönetimi alanlarında hiçbir model kritik hata üretmedi. Buna karşılık tedavi ve ilaç yönetimi alanında kritik hata oranları belirgin olarak arttı; bu oran Claude Sonnet 4.5 için %50,0, ChatGPT-5 için %33,3 ve Microsoft Copilot için %25,0 idi; Gemini 2.5 Pro'da kritik hata saptanmadı. Modeller birlikte değerlendirildiğinde, tedavi alanındaki kritik hataların diğer tüm alanların toplamına göre anlamlı olarak daha sık olduğu görüldü (olasılık oranı: 4,68; p<0,001). Hatalar özellikle sodyum-glukoz kotransporter 2 inhibitörleri, semaglutid, KEF-KY'ye özgü demir eksikliği yaklaşımı ve eşlik eden hipertansiyonda ilaç önceliklendirmesi konularında kümelendi.

Sonuç: Güncel BDM'ler KEF-KY ile ilişkili kavramsal ve tanısal sorularda yüksek performans gösterirken, kılavuz uyumlu farmakoterapi alanında klinik açıdan anlamlı bir güvenilirlik açığı sergilemektedir. Bu bulgular, BDM'lerin özellikle tedavi önerileri için bağımsız doğrulama yapılmadan kullanılmaması gerektiğini düşündürmektedir.

Anahtar Kelimeler: Korunmuş ejeksiyon fraksiyonlu kalp yetersizliği, büyük dil modelleri, yapay zeka, klinik karar destek, kılavuz uyumu, farmakoterapi

INTRODUCTION

Heart failure (HF) with preserved ejection fraction (HFpEF) is defined as left ventricular ejection fraction (LVEF) ≥50%, and approximately 50% of HF cases worldwide are with preserved EF (1).

Its prevalence varies geographically but is generally on the rise, with projections suggesting that HFpEF may soon surpass HF with reduced ejection fraction (HFrEF) in incidence (2). Despite its high prevalence, HFpEF remains challenging to diagnose and manage. The increasing burden underscores the importance of ongoing research into its pathophysiology, risk factors, and treatment options, including recent pharmacotherapeutic advances such as empagliflozin and dapagliflozin, which have shown benefit in clinical trials (3,4). Overall, HFpEF constitutes a significant and growing global health concern, affecting millions and necessitating targeted strategies for diagnosis, management, and prevention.

The recent evolution in HF management highlights the central role of sodium-glucose cotransporter 2 (SGLT2) inhibitors, supported by robust clinical trial evidence and incorporated into updated guidelines such as the 2022 American Heart Association (AHA)/American College of Cardiology (ACC)/Heart Failure Society of America (HFSA) and 2023 European Society of Cardiology (ESC) recommendations (5,6). The 2023 ESC focused update emphasizes early initiation of SGLT2 inhibitors, including in acute settings, and underscores their broad applicability (5). Despite their proven efficacy, the evidence for mineralocorticoid receptor antagonists and angiotensin receptor-neprilysin inhibitors remains less definitive, relying largely on subgroup analyses, which translates into variations in guideline classifications (6,7).

These rapidly evolving treatments and guidelines impose a substantial cognitive load on clinicians. As a result, residents, general practitioners, and specialist physicians, including cardiologists, are increasingly turning to large language models (LLMs) as informal decision-support tools. While LLMs perform well on general medical questions, this aggregate success can obscure the reality at the patient's bedside (8,9). A model that accurately describes HFpEF may still provide unsafe guidance for the treatment and management of a complex gray-zone patient in real-world clinical practice.

Concerns about liability, confidentiality, and the preservation of diagnostic autonomy further underscore the need for rigorous, task-specific validation before clinical deployment (10). Empirical studies reveal that while LLMs generate coherent and medically plausible outputs, their accuracy varies across different evidence-based medicine tasks, with notable issues in numerical accuracy, source verifiability, and methodological rigor (11). Furthermore, studies on phenotyping HFpEF emphasize the heterogeneity of clinical approaches, underscoring the need for

integrated, patient-specific data analysis rather than reliance on artificial intelligence (AI) outputs alone (12).

Whether LLM reliability holds across different tasks, such as diagnosis and drug selection, remains an open question. If errors are concentrated in specific areas, a hierarchical collaboration model should be adopted in which AI is trusted only for certain tasks. In this study, we evaluated four current LLMs (Gemini 2.5 Pro, ChatGPT-5, Microsoft Copilot, and Claude Sonnet 4.5) using a bank of 80 questions based on the AHA/ACC/HFSA and ESC guidelines. We divided the questions into four domains: diagnosis and pathophysiology, comorbidity management, treatment and drug administration, and guideline comparisons. Our aim was to test whether the models exhibited a "domain-specific reliability gap" in pharmacotherapy decisions while demonstrating high performance on conceptual questions.

MATERIAL AND METHODS

Study Design

This was a cross-sectional benchmark study evaluating four contemporary LLMs using a structured question bank derived from the 2022 AHA/ACC/HFSA HF guideline, the 2021 ESC guideline, and its 2023 focused update (5-7). Because the study involved no human participants or patient data, institutional review board approval was not required. The study is reported in accordance with the TRIPOD-LLM reporting guideline for studies using LLMs (13).

Question Bank Construction

A practising cardiologist prospectively drafted 80 items that were explicitly stratified across four clinical domains. The number of items per domain was chosen to reflect the relative clinical weight of each domain in contemporary HFpEF practice while preserving adequate power for domain-level comparisons. To ensure internal validity and a consistent difficulty level across domains, the questions were deliberately balanced between direct factual recall (e.g., specific diagnostic cutoffs or recommendation classes) and complex clinical integration requiring interpretation (e.g., therapeutic decision-making in vignette-style scenarios with multiple comorbidities). Furthermore, a standardized structural template, informed by the patient, intervention, comparison, outcome framework, was utilized during item drafting. Question topics were mapped directly to specific guideline recommendations, tables, and class/level of evidence statements. Each item carried a pre-specified gold-standard answer anchored strictly to the originating guideline text; these gold-standard answers were drafted by the same author and independently verified against the source guideline sections by a second cardiologist before any model was queried.

Question Categories

The diagnosis and pathophysiology domain (n=16) included items covering the LVEF threshold for HFpEF, the HFA-PEFF and H2FPEF diagnostic algorithms, N-terminal proBNP cut-offs, echocardiographic parameters of diastolic dysfunction, invasive hemodynamic evaluation, assessment of pulmonary hypertension, cardiopulmonary exercise testing, and physiology of exercise intolerance.

The comorbidity management domain (n=11) included items on atrial fibrillation, chronic kidney disease, iron deficiency, diabetes, amyloidosis, and sarcoidosis.

The treatment and drug management domain (n=12) was deliberately anchored to items requiring integration of a class of recommendation or a level of evidence from one or both guidelines specifically those concerning SGLT2 inhibitors, mineralocorticoid receptor antagonists, angiotensin receptor-neprilysin inhibitors, beta-blockers, digoxin, angiotensin-converting enzyme inhibitors/angiotensin II receptor blockers, diuretics, semaglutide, and lipid or sodium targets.

The guideline comparison and clinical scenarios domain (n=41) included items that explicitly contrasted the two guidelines' specific recommendations and included short clinical vignettes (e.g., first-line therapy for a patient with HFpEF, an LVEF of 55%, and stage 3 chronic kidney disease). This domain was the largest by design, as guideline-to-guideline synthesis represents the most common real-world use case for clinicians consulting an AI tool.

LLMs Evaluated

Four contemporary, publicly accessible LLMs were selected to represent the major commercial AI ecosystems available to clinicians at the time of the study: Gemini 2.5 Pro (Google), ChatGPT-5 (OpenAI), Microsoft Copilot (Microsoft; based on the ChatGPT family with retrieval augmentation), and Claude Sonnet 4.5 (Anthropic). All queries were performed between 10.11.2025 and 15.11.2025; the specific model build available through each interface during that window is reflected in the reported results. All models were queried through their respective default public chat interfaces without system prompts, retrieval plug-ins, or custom instructions, reflecting the exact conditions under which a typical clinician operates them.

Query Protocol

To mitigate account-specific personalization, algorithmic caching, or IP-based localization bias,a rigorous, multi-user querying protocol was used. The 80 questions were independently submitted to all four models by three separate cardiologists from different computers and IP addresses, with a 48-hour interval between each investigator's session.

Each question was submitted in a zero-shot fashion without follow-up prompts. Queries and corresponding gold-standard answers were performed in Turkish, reflecting investigators' primary working language and a substantial real-world user base. The order of question submission was held constant across models. For every question, a new chat conversation was initiated to eliminate in-context carry over. All generated responses were saved verbatim in the study database for offline evaluation.

"Materially similar" was defined qualitatively; although minor stylistic or conversational variations existed in the models' phrasing across different sessions, the core medical advice, pharmacological recommendations, and cited evidence levels remained identical across all three investigators' independent queries. Because the outputs obtained by the three cardiologists were materially similar for the same model-item pairs, responses were consolidated into a single model-question-level output for final scoring, yielding 320 evaluated responses (80 items×4 models).

Scoring Rubric and Consensus Adjudication

A joint panel-based grading approach was used. The three investigating cardiologists evaluated the AI outputs against the pre-specified gold-standard answers in joint review sessions. During the evaluation, model identities were masked (labeled model A-D). Each response was discussed, and a single final score was determined by majority vote or by unanimous consensus based on a four-point ordinal rubric:

1. (Comprehensive): The response fully captured the guideline-concordant answer, including the relevant class of recommendation and level of evidence where applicable.

2. (Adequate but incomplete): The core recommendation was correct, but key qualifiers or evidence levels were missing.

3. (Partially incorrect): The response contained at least one clinically meaningful error alongside correct elements.

4. (Completely incorrect): The response contradicted the guideline-concordant answer in a clinically meaningful or potentially harmful way.

In instances where a unanimous decision could not be reached among the three primary evaluators, an independent, senior cardiologist—blinded to the initial internal discussions—adjudicated the final score.

A composite success rate was defined as the proportion of responses scored 1 or 2. A critical error rate was defined as the proportion of responses scoring 4.

Statistical Analysis

Because the primary purpose of the study was to compare the structural distribution of errors across clinical domains rather than to establish the absolute superiority of one model over another, the analyses were descriptive in intent, and confirmatory testing was restricted to the pre-specified domain×outcome comparison. For each model and domain, we report the distribution of the four ordinal scores; the comprehensive rate (Score 1); the composite success rate (Scores 1+2); and the critical error rate (Score 4). Exact binomial (Clopper-Pearson) 95% confidence intervals (CI) were calculated for all proportions to convey uncertainty appropriate to the subgroup sizes. To quantify the heterogeneity of model performance within each domain, we calculated the inter-model range of the comprehensive rates. The pre-specified primary inferential test was Fisher's exact test, which compared the proportion of critical errors in the treatment and drug management domain with that in all other domains combined and was performed both pooled across models and separately for each model. A two-sided p-value <0.05 was considered statistically significant; given the descriptive framing, no correction for multiple per-model tests was applied, and per-model results are reported as supportive. All statistical analyses were performed using IBM SPSS Statistics (version 26.0, IBM Corp., Armonk, NY, USA).

RESULTS

Overall Performance

A total of 320 responses (80 questions×4 models) were evaluated. The overall performance metrics varied across systems, as summarized in Table 1. Comprehensive and accurate response rates (Score 1) ranged from 65.0% (95% CI 53.5-75.3) for Claude Sonnet 4.5 to 87.5% (95% CI 78.2-93.8) for Gemini 2.5 Pro. When considering the composite success rate (Scores 1+2), performance remained relatively high across all systems, ranging from 83.75% to 93.75%. However, critical error rates (Score 4) remained non-trivial, ranging from 6.25% (Gemini 2.5 Pro; 95% CI 2.1-14.0) to 13.75% (Claude Sonnet 4.5; 95% CI 7.1-23.3)

Table 1. Overall distribution of response quality across four LLMs (n=80 items per model)

Model	Score 1 (%)	Score 2 (%)	Score 3 (%)	Score 4 (%)	Success 1+2 (%)
Gemini 2.5 Pro	87.5	6.25	0.0	6.25	93.75
ChatGPT-5	83.75	5.0	0.0	11.25	88.75
Microsoft Copilot	77.5	8.75	3.75	10.0	86.25
Claude Sonnet 4.5	65.0	18.75	2.5	13.75	83.75

Score 1: Comprehensive, Score 2: Adequate but incomplete, Score 3: Partially incorrect, Score 4: Completely incorrect

As illustrated in Figure 1 , Gemini 2.5 Pro achieved the most favorable overall profile, combining the highest proportion of comprehensive responses with the lowest critical error rate. In contrast, Claude Sonnet 4.5 demonstrated the lowest comprehensive accuracy and the highest proportion of critical errors. The remaining models showed intermediate performance across these metrics (Figure 1).

Domain-specific Performance and the Reliability Gap

The distribution of errors varied markedly across the four clinical domains (Tables 2 and 3). In the diagnosis and pathophysiology domain, all four models performed consistently well; Gemini 2.5 Pro, ChatGPT-5, and Microsoft Copilot each achieved a comprehensive coverage rate of 93.75%, while Claude Sonnet 4.5 achieved 75.0%. Notably, no critical errors (Score 4) were observed in either this domain or the comorbidity management domain, where comprehensive coverage rates ranged from 72.7% to 100%.

By contrast, reliability collapsed in the treatment and drug management domain. Comprehensive coverage rates in this domain fell as low as 33.3% (95% CI 9.9-65.1) for Claude Sonnet

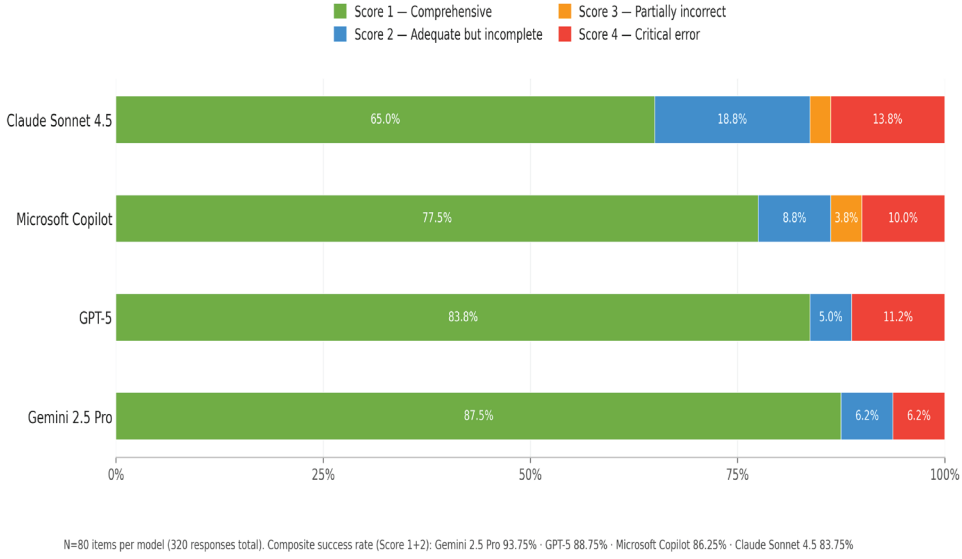

Figure 1. Distribution of response quality scores across LLMs in clinical evaluation
Distribution of LLM response quality for HFpEF (n=320). While overall accuracy rates varied between 65.0% and 87.5% across the four systems, all models maintained a composite success rate (Scores 1+2) exceeding 83%
LLM: Large language model, HFpEF: Heart failure with preserved ejection fraction

Table 2. Comprehensive (Score 1) response rate by clinical domain. Exact binomial 95% confidence intervals for the treatment and drug management domain (n=12) are shown in parentheses

Model	Diagnosis and pathophysiology (n=16)	Comorbidity management (n=11)	Treatment and drugs (n=12)	Guideline comparison and scenarios (n=41)
Gemini 2.5 Pro	93.8	90.9	100.0 (73.5-100)	80.5
ChatGPT-5	93.8	100.0	66.7 (34.9-90.1)	80.5
Microsoft Copilot	93.8	72.7	58.3 (27.7-84.8)	78.0
Claude Sonnet 4.5	75.0	100.0	33.3 (9.9-65.1)	61.0
Inter-model range	**18.8**	**27.3**	**66.7**	**19.5**

Table 3. Critical error rate (Score 4) by clinical domain. Exact binomial 95% confidence intervals for the treatment and drug management domain (n=12) are shown in parentheses

Model	Diagnosis and pathophysiology	Comorbidity management	Treatment and drugs	Guideline comparison and scenarios
Gemini 2.5 Pro	0.0	0.0	0.0 (0-26.5)	12.2
ChatGPT-5	0.0	0.0	33.3 (9.9-65.1)	12.2
Microsoft Copilot	0.0	0.0	25.0 (5.5-57.2)	12.2
Claude Sonnet 4.5	0.0	0.0	50.0 (21.1-78.9)	12.2

4.5 and 58.3% (95% CI 27.7-84.8) for Microsoft Copilot. Score 4 responses accounted for 50.0% (95% CI 21.1-78.9) of Claude Sonnet 4.5's pharmacotherapy answers, 33.3% (95% CI 9.9-65.1) for ChatGPT-5, and 25.0% (95% CI 5.5-57.2) for Microsoft Copilot. Gemini 2.5 Pro was the only model without a single critical error in this domain (0/12; 95% CI 0-26.5). The heterogeneity of model performance (the inter-model range of comprehensive rates) was 66.7% for treatment more than threefold higher than that in any other domain. When pooled across all four models, critical errors were significantly more frequent in the treatment and drug management domain than in all other domains combined (13/48 vs. 20/272; odds ratio 4.68, Fisher's exact $p<0.001$). In per-model analyses, this difference reached statistical significance for Claude Sonnet 4.5 ($p=0.001$) and ChatGPT-5 ($p=0.025$), approached significance for Microsoft Copilot ($p=0.09$), and was non-significant (Fisher's exact $p=1.00$) for Gemini 2.5 Pro, which produced no critical errors in the treatment domain.

In the guideline comparison and clinical scenarios section (n=41), accuracy rates were relatively homogeneous (61.0%-80.5%). Interestingly, critical error rates converged at 12.2% (5/41) across all four models, indicating that each system independently produced completely incorrect responses to exactly five items. However, item-level inspection showed that these were not the same five questions for each model; the error count converged, while the error content diverged.

Item-level Failure Analysis

To determine whether these errors were stochastic or systematic, a post-hoc analysis of critical errors at the item level was performed (Table 4). Of the 80 items, 11 produced a Score 4 response in at least two models ("common-failure items"), and 6 failed in at least three models. Only a single item—regarding the guideline-appropriate treatment approach for iron deficiency in HFpEF—elicited completely incorrect responses from all four models simultaneously.

Every item that failed in three or more models was directly related to a drug-class recommendation. These errors clustered around SGLT2 inhibitors (including the 2023 focused updates), glucagon like peptide-1 (GLP-1) receptor agonists (semaglutide following the STEP-HFpEF trial), intravenous iron therapy, and drug prioritization in concomitant hypertension. Conversely, no common failures were observed in conceptual questions regarding diagnosis or pathophysiology.

When the distribution of these 11 common-failure items across individual models was examined, Claude Sonnet 4.5 produced critical errors on 10 of the 11 items, ChatGPT-5 on 9, Microsoft Copilot on 7, and Gemini 2.5 Pro on 3. The substantially lower failure count of Gemini 2.5 Pro on these items mirrors its overall performance profile. One possible explanation is that this apparent advantage may reflect differences in model updating or in the recency of knowledge rather than a fundamentally different reasoning capability; however, this hypothesis was not directly tested in the present study.

DISCUSSION

In this guideline-anchored benchmark of four contemporary LLMs on HFpEF, we observed three findings with direct implications for how these systems should—and should not—be

Table 4. Items on which multiple large language models produced a critical error (Score 4), grouped by failure cluster

Failure cluster	Question focus	Domain	Models scoring 4 (count)
Iron deficiency	IV vs. oral iron in HFpEF	Scenarios	All four models (4/4)
SGLT2 inhibitors	Drug class for CKD progression in HFpEF	Treatment	ChatGPT-5, Copilot, Sonnet (3/4)
	SGLT2i class of recommendation (2023 ESC update)	Scenarios	ChatGPT-5, Copilot, Sonnet (3/4)
	Impact of CKD on SGLT2i use	Treatment	ChatGPT-5, Sonnet (2/4)
GLP-1/obesity	Semaglutide in obesity-related HFpEF (STEP-HFpEF)	Scenarios	ChatGPT-5, Copilot, Sonnet (3/4)
Diabetes prioritization	DM management priority in HFpEF (AHA/ACC/HFSA)	Treatment	GPT-5, Copilot, Sonnet (3/4)
Hypertension prioritization	Hypertension priority (AHA/ACC/HFSA) for CV benefit	Scenarios	Gemini, ChatGPT-5, Copilot (3/4)
	Hypertension priority in HFpEF (ESC)	Scenarios	Gemini, Sonnet (2/4)
Atrial fibrillation in HFpEF	Anticoagulant class in HFpEF with AF	Treatment	ChatGPT-5, Sonnet (2/4)
	Catheter ablation class in HFpEF with AF	Treatment	Copilot, Sonnet (2/4)
ARNI class discrepancy	ARNI class difference (AHA/ACC/HFSA vs. ESC)	Scenarios	ChatGPT-5, Sonnet (2/4)

Clusters are ordered by the number of converged critical errors. "Sonnet" refers to Claude Sonnet 4.5
SGLT2: Sodium-glucose cotransporter 2, GLP-1: Glucagon like peptide-1, HFpEF: Heart failure with preserved ejection fraction, ARNI: Angiotensin receptor-neprilysin inhibitor, CKD: Chronic kidney disease, AHA: American Heart Association, ACC: American College of Cardiology, HFSA: Heart Failure Society of America, ESC: European Society of Cardiology, AF: Atrial fibrillation

used in cardiology practice. First, LLMs perform uniformly well on conceptual, diagnostic, and pathophysiological questions about HFpEF, with no critical errors observed across the diagnostic and comorbidity domains. Second, in the domain of treatment and drug management, the same models exhibited a dramatic decline in reliability, with critical error rates as high as 50% on direct pharmacotherapy questions. Third, the magnitude of inter-model heterogeneity was more than three times as great in the treatment domain as in any other domain, indicating that errors are not randomly distributed but are concentrated in the single domain most likely to directly influence prescribing decisions. The pooled Fisher's exact test confirmed that this concentration is not a statistical artefact of small numbers (odds ratio: 4.68, $p<0.001$). We refer to this pattern as a domain-specific reliability gap. We define this gap as task-type–dependent performance heterogeneity, in which an LLM's accuracy fluctuates significantly with the clinical nature of the question rather than with the disease entity alone.

The post-hoc analysis of items in which multiple models converged on the same critical error provides a mechanistic explanation for this gap. Shared failures were thematically narrow: every item on which at least three of the four models failed concerned either a newer-generation therapy (e.g., SGLT2 inhibitors, semaglutide) or a specific class-of-recommendation hierarchy. Moreover, the convergence of critical error rates at exactly 12.2% across models in the scenarios domain, despite divergent item identities, suggests that the difficulty ceiling of clinical scenario synthesis is shared across systems, while the specific failure points are model-idiosyncratic a pattern consistent with model-specific gaps superimposed on a common substrate of genuinely difficult content.

These common failures cluster around treatments for which the evidence base is changing most rapidly. The efficacy of SGLT2 inhibitors and GLP-1 receptor agonists in organ-preserving strategies, even in non-diabetic patients, is supported by new evidence. Landmark studies such as EMPEROR-Preserved and DELIVER have shown that SGLT2 inhibitors are now a foundational class for the management of HFpEF, displacing several older agents from guideline-directed therapy (14,15). Similarly, the role of semaglutide in obesity-related HFpEF is supported by the STEP-HFpEF study (16). When multiple models produce completely incorrect answers on these items, the pattern is consistent with temporal misalignment between model knowledge and the current guideline state. We acknowledge that, while highly plausible, this remains a speculative hypothesis, as it was not directly tested in our study design. LLMs encode the statistical average of the evidence in their training window; when the therapeutic paradigm has recently shifted, this statistical average is no longer the correct answer.

Furthermore, the single item on which all four models produced a completely incorrect response (the treatment approach for iron deficiency in HFpEF) highlights a qualitatively different failure mechanism: inappropriate generalization. The robust evidence base for intravenous iron in HF was established predominantly in patients with HFrEF (17,18). A model that has correctly learned that "iron deficiency in HF is treated with intravenous iron" will confidently produce that answer for HFpEF, failing to recognize that the HFpEF-specific recommendation is deliberately more cautious. This pattern suggests a potentially important failure mode in which models may inappropriately extrapolate from HFrEF to HFpEF, rather than simply relying on outdated knowledge. This highlights a critical vulnerability often discussed in broader clinical AI literature: "contextual blindness", where models fail to distinguish between high-quality evidence in a related phenotype and the more nuanced and restricted recommendations required for the specific target phenotype.

Most existing guidance on the use of LLMs in medicine treats the AI knowledge base as monolithic and recommends uniform caution. However, our findings, together with recent literature on task-dependent AI performance, suggest that a more nuanced, task-stratified approach may be more appropriate than uniform caution across all clinical tasks (19,20). For HFpEF specifically, the present findings suggest that contemporary LLMs may be more reliable for conceptual and diagnostic tasks than for treatment-related decision support. In these domains, the models showed consistently high performance with no observed critical errors. In the treatment and drug-management domain, current LLMs do not appear sufficiently reliable for unsupervised decision support. With a 1-in-2 critical error rate for the weakest model and a 1-in-3 rate even for one flagship system, the consequences of acting on an LLM recommendation without independent verification are potentially serious. Taken together, these findings suggest that caution is warranted: contemporary LLMs should be used primarily as conceptual aids, and pharmacotherapy recommendations should be independently verified against primary guideline sources.

Aggregate benchmarks of LLMs on medical examinations have reported encouraging overall accuracy, but they rarely stratify results by the cognitive type of the underlying task (21). Recent empirical evaluations have begun to demonstrate that LLM errors in medicine are not uniformly distributed but tend to cluster within specific clinical tasks, such as complex evidence synthesis (22,23). Our findings add a critical layer to this emerging literature. In our previous multidimensional audit of cardiovascular decision-making, we demonstrated that LLM safety failures are highly model-dependent and prone to "dysfunctional humility"—where cautious phrasing is dangerously paired with unsafe recommendations (24). In our

prior study, we identified a shared multimodel vulnerability in the management of high-risk non-ST-segment elevation myocardial infarction, in which models inappropriately deviated from ESC Class III guidance. The present study extends this paradigm by showing that reliability varies across task types, even within a single disease phenotype, rather than remaining uniform across all forms of clinical questioning.

Future investigations should extend this framework to retrieval-augmented generation configurations, testing whether the reliability gap in the treatment domain can be closed by directly grounding the model in source documents. Additionally, this domain-stratified approach should be applied to other cardiovascular conditions with dynamic evidence bases to determine if the reliability gap is a universal property of AI in cardiology.

Study Limitations

Several limitations must be acknowledged. First, queries were performed in Turkish. Although the guidelines are international and the clinical content is language-independent, we cannot exclude a language effect on LLM performance, and prior work suggests that non-English medical queries may degrade LLM accuracy subtly. Second, the treatment and drug management domain contained only 12 items. Although the magnitude of the observed effect is large and the pooled Fisher's exact test remains significant, between-model comparisons in this subgroup should be interpreted cautiously, and the exact 95% CIs for per-model estimates are correspondingly wide. Third, although model identity was masked during scoring, the three evaluators were not blinded to one another during consensus adjudication; this reflects standard practice for rubric-based clinical adjudication but could introduce subtle group-dynamic bias. Finally, the benchmark reflects the state of these models at a single point in time. Consequently, these findings represent a temporal snapshot rather than a definitive model ranking, because results will inevitably shift as models are retrained and guidelines are updated.

CONCLUSION

Contemporary LLMs handle conceptual and diagnostic questions about HFpEF uniformly well, but exhibit a clinically meaningful reliability gap in guideline-concordant pharmacotherapy. Errors concentrate in the single domain most likely to influence prescribing, driven by temporal misalignment and inappropriate generalization. Until retrieval-grounded systems are independently validated, these data suggest that caution is warranted and that LLMs are currently best used as conceptual aids, with all pharmacotherapy recommendations subject to rigorous human verification.

*Ethics

Ethics Committee Approval: This study involved no human participants, human tissue, or identifiable patient data; it evaluated publicly accessible large language models using synthetic guideline-derived questions. Institutional review board approval and informed consent were therefore not applicable.

Informed Consent: This study involved no human participants, human tissue, or identifiable patient data; therefore, informed consent was not applicable.

Footnotes

Data Availability

The complete question bank, pre-specified gold-standard answers, and raw scored responses are available from the corresponding author upon reasonable request.

Authorship Contributions

Concept: H.T., S.E., S.Ç., Design: H.T., S.E., S.Ç., Data Collection or Processing: H.T., S.E., Ö.E., D.İ., Analysis or Interpretation: H.T., S.E., Ö.E., D.İ., S.Ç., Literature Search: H.T., S.E., Ö.E., D.İ., Writing: H.T., S.E., Ö.E., D.İ., S.Ç.

Conflict of Interest: No conflict of interest was declared by the authors.

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

REFERENCES

1. Fayol A, Wack M, Livrozet M, Carves JB, Domengé O, Vermersch E, et al. Aetiological classification and prognosis in patients with heart failure with preserved ejection fraction. ESC Heart Fail. 2022;9(1):519-530.
2. Kapelios CJ, Shahim B, Lund LH, Savarese G. Epidemiology, clinical characteristics and cause-specific outcomes in heart failure with preserved ejection fraction. Card Fail Rev. 2023;9:e14.
3. Vaduganathan M, Docherty KF, Claggett BL, Jhund PS, de Boer RA, Hernandez AF, et al. SGLT-2 inhibitors in patients with heart failure: a comprehensive meta-analysis of five randomised controlled trials. Lancet. 2022;400(10354):757-767.
4. Kittleson MM, Panjrath GS, Amancherla K, Davis LL, Deswal A, Dixon DL, et al. 2023 ACC expert consensus decision pathway on management of heart failure with preserved ejection fraction: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol. 2023;81(18):1835-1878.
5. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2023 focused update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2023;44(37):3627-3639.
6. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 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. J Am Coll Cardiol. 2022;79(17):e263-e421.

7. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42(36):3599-3726.
8. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls : a narrative review. Ann Intern Med. 2024;177(2):210-220.
9. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969.
10. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120.
11. Qi W, Pan L. Epistemic and ethical limits of large language models in evidence-based medicine: from knowledge to judgment. Front Digit Health. 2026;7:1706383.
12. Shah SJ, Kitzman DW, Borlaug BA, van Heerebeek L, Zile MR, Kass DA, et al. Phenotype-specific treatment of heart failure with preserved ejection fraction: a multiorgan roadmap. Circulation. 2016;134(1):73-90.
13. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-69.
14. Solomon SD, McMurray JJV, Claggett B, de Boer RA, DeMets D, Hernandez AF, et al. Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. N Engl J Med. 2022;387(12):1089-1098.
15. Anker SD, Butler J, Filippatos G, Ferreira JP, Bocchi E, Böhm M, et al. Empagliflozin in heart failure with a preserved ejection fraction. N Engl J Med. 2021;385(16):1451-1461.
16. Kosiborod MN, Abildstrøm SZ, Borlaug BA, Butler J, Rasmussen S, Davies M, et al. Semaglutide in patients with heart failure with preserved ejection fraction and obesity. N Engl J Med. 2023;389(12):1069-1084.
17. Ponikowski P, Kirwan BA, Anker SD, McDonagh T, Dorobantu M, Drozdz J, et al. Ferric carboxymaltose for iron deficiency at discharge after acute heart failure: a multicentre, double-blind, randomised, controlled trial. Lancet. 2020;396(10266):1895-1904.
18. Mentz RJ, Garg J, Rockhold FW, Butler J, De Pasquale CG, Ezekowitz JA, et al. Ferric carboxymaltose in heart failure with iron deficiency. N Engl J Med. 2023;389(11):975-986.
19. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-950.
20. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180.
22. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep. 2025;15(1):39426.
23. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-2622.
24. Eşki S, Taşkan H. Dysfunctional humility and safety failures in large language models: a multidimensional evaluation in cardiovascular decision-making. Health Policy Technol. 2026;15(4):101183.