Unmasking True Clinical Competence: The Importance of Adaptive and Open-Ended Evaluation for Large Language Models in Cardiology

Chieh-Ju Chao[1,2*†], Abhinav Kumar[3*], Aakash Mishra[3], Yu-Chiang Wang[1], Chieh-Mei Tsai[1], Nima Baba Ali[4], Somanshu Sharma[4], Juan M. Farina[4], Reza Arsanjani[4], Yucheng Jiang[3], Monica Lam[3†]

1. Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA
2. Institute of Human-centered Artificial Intelligence, Stanford University, Palo Alto, CA, USA
3. Department of Computer Science, Stanford University, Palo Alto, CA, USA
4. Department of Cardiovascular Diseases, Mayo Clinic, Scottsdale, AZ, USA

*Co-first author; †Co-corresponding author

Address of Correspondence:
200 1st Street SW, Room: Gonda 4-478
Rochester, MN 55095, USA
Chieh-Ju Chao, MD
Assistant Professor of Medicine, Department of Cardiovascular Medicine, Mayo Clinic
Email: chao.chiehju@mayo.edu
Twitter: @chiehjuchao1

Abstract

Background: Large language models (LLMs) achieve impressive accuracy on multiple-choice (MC) medical examinations, but this performance alone may not accurately reflect their clinical reasoning abilities. MC evaluations of LLMs risk inflating apparent competence by rewarding recall rather than genuine clinical adaptability, particularly in specialized medical domains such as cardiology.
Objectives: To rigorously evaluate the clinical reasoning, contextual adaptability, and answer-reasoning consistency of a state-of-the-art reasoning-based LLM in cardiology using MC, open-ended, and clinically modified question formats.
Methods: We assessed GPT o1-preview using 185 board-style cardiology questions from the American College of Cardiology Self-Assessment Program (ACCSAP) in MC and open-ended formats. A subset of 66 questions underwent modifications of critical clinical parameters (e.g.,

ascending aorta diameter, ejection fraction) to evaluate model adaptability to context changes. The model's answer and reasoning correctness were graded by cardiology experts. Statistical differences were analyzed using exact McNemar tests.
Results: GPT o1-preview demonstrated high baseline accuracy on MC questions (93.0% answers, 92.4% reasoning). Performance significantly decreased with open-ended questions (80.0% answers, 80.5% reasoning; $p<0.001$). For modified MC questions, accuracy decreased significantly (answers: 93.9% to 66.7%; reasoning: 93.9% to 71.2%; both $p<0.001$), as well as answer-reasoning concordance (93.9% to 66.7%, $p<0.001$).
Conclusions: Using existing MC question formats substantially overestimates the performance and clinical reasoning capabilities of GPT o1-preview. Incorporating open-ended, clinically adaptive questions and evaluating answer-reasoning concordance are essential for accurately assessing the real-world clinical decision-making competencies of LLMs in cardiology.

Keywords: Large language model; cardiology; clinical reasoning; board examination; evaluation framework; artificial intelligence

Introduction

Recent advancements in large language models (LLMs) have shown promising performance across numerous medical tasks, with several studies suggesting these models can match or exceed human experts on standardized medical board-style examinations(1–6). Such models include generic LLMs, medicine-specific LLMs(7,8), and more recently, "reasoning" models, such as GPT o1-preview, which decompose a query into logical steps before arriving at a final answer (9). Although LLM systems have been reported to achieve comparable or suprahuman performance in these questions(1–6), there is still a question of whether current evaluation methodologies truly reflect these models' capability in clinical settings. Moreover, most studies have focused on general medical license exam questions (1,10,11), with some evaluation in subspecialties like Cardiology(12,13).

Current evaluation frameworks for LLMs in medicine rely heavily on multiple-choice (MC) question formats (14). MC examinations, as being questioned for their effectiveness in evaluating human medical trainees (15–17), can also be misleading when assessing LLM performance. The prevalence of MC question answering frameworks in current LLM research is largely due to pragmatic considerations. MC answers are easily evaluated and can be scaled across a large number of questions and multiple datasets. MC questions can be compromised by cues in the answer choices or "best guess" strategies (16–19), which LLMs can exploit without necessarily understanding the underlying clinical concepts. In addition, since many board-style questions exist in widely available databases (1,11), there is a risk that some have been seen during an LLM's pre-training phase, which can inflate accuracy during evaluation.

An alternative approach is to evaluate using long-form, open-ended questions (8,20), where the model must generate a complete written rationale and final answer without relying on multiple-choice cues. However, evaluating open-ended responses presents significant challenges, particularly in specialized medical domains, where expertise is required to judge answer correctness (21). In cardiology practice, critical clinical decisions hinge on key quantitative metrics, such as left ventricular ejection fraction for guiding implantable cardioverter-defibrillator therapy and ascending aorta dimensions for surgical repair thresholds. The capability of making decisions according to these contextual clinical variations would serve as a critical indicator of clinical reasoning capability in LLM assessments. This approach, which evaluates a model's ability to interpret and adapt to shifting clinical parameters, has not been systematically incorporated into prior LLM evaluation frameworks (1–6).

In this study, we present an evaluation framework aimed at overcoming the limitations of the conventional MC question format and more accurately gauging LLM performance, including answering and reasoning accuracy in the field of cardiology. In doing so, we highlight both the potential and the current limitations of GPT o1-preview, a state-of-the-art reasoning LLM, in cardiology practice.

Methods

In this work, we evaluated GPT o1-preview, one of OpenAI's latest LLMs that has a specific emphasis on its reasoning capacity on a set of 185 subspecialty board-style cardiology questions. We use two primary question formats, including multiple-choice and open-ended, to evaluate whether performance differs in the presence (or absence) of explicit answer options. We additionally assess the model's ability to understand the clinical context by modifying a subset of 66 questions, in which we change key clinical details so that the correct answer differs from the original correct answer. By comparing performance on original vs. modified questions using expert cardiologist evaluators, we evaluate GPT o1-preview's ability to adapt to nuanced changes in clinical scenarios. **Figure 1** illustrates our study design.

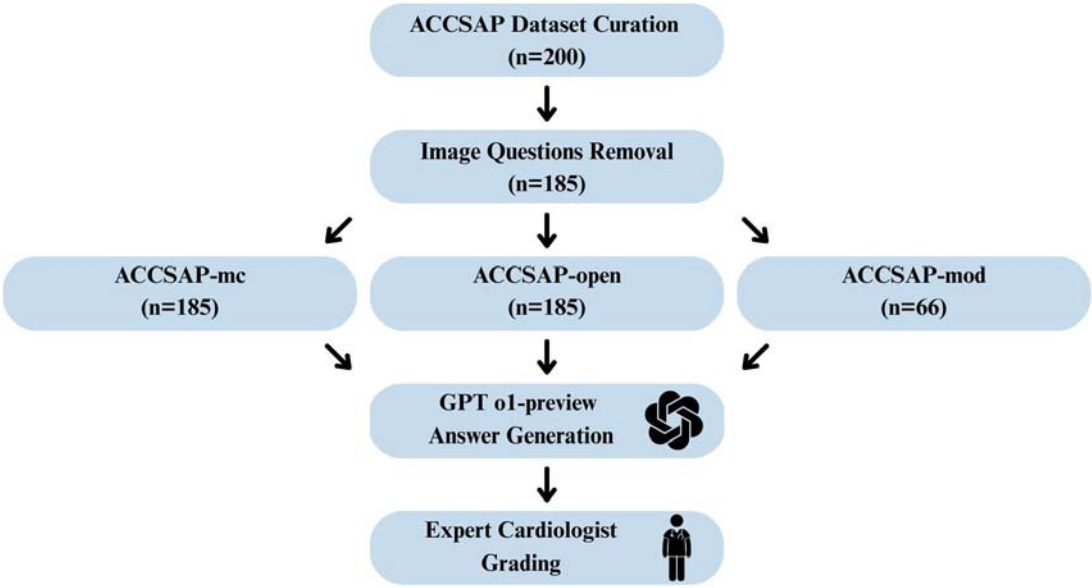

Figure 1. Study design for assessing clinical reasoning capabilities of GPT o1-preview. The question dataset was curated from the ACCSAP, excluding image-based questions. Questions were formulated into multiple-choice (ACCSAP-mc), open-ended (ACCSAP-open), and clinically modified (ACCSAP-mod) formats. Model responses underwent expert grading to quantify performance across question formats.

Baseline Question Dataset

We conducted an evaluation of GPT o1-preview on a dataset of 185 board-style cardiology questions. We leveraged the American College of Cardiology's Self-Assessment Program (ACCSAP), a board-preparation resource designed for the cardiology subspecialty board exam. This question bank, which is behind a paywall, contains a total of 640 questions covering various cardiology topics and is less likely to be leaked for LLM training due to its restricted access.

To create a representative benchmark, we initially used ACCSAP's random selection function to obtain 200 questions, ensuring balanced coverage across all question categories. As our study focuses solely on evaluating the model's ability to understand clinical text, we followed prior research methodologies and excluded questions that contained figures in the question stem(12), leaving 185 questions. This allows for a more precise assessment of the model's textual comprehension without the confounding factor of medical image interpretation.

To extend our evaluation framework beyond the provided multiple-choice questions, we introduced two question formats of the original dataset:

1) Multiple choice questions with a "None" option (ACCSAP-mc, n=185) – We added an additional "None" option to the provided answer choices to mitigate common process-of-

elimination strategies used by reasoning models and simulate a more realistic clinical decision-making scenario.

2) Open-ended questions (ACCSAP-open, n=185) – We removed answer choices and revised the question stem to prompt models to generate the best possible answer in an open-ended format, encouraging more nuanced reasoning rather than pattern recognition.

Clinical Context Question Modification

Building on the baseline ACCSAP dataset, we aimed to test GPT o1-preview's ability to adapt to nuanced changes in clinical parameters rather than simple recall for memorized solutions. All 185 questions were reviewed for modifiable content, including specific symptoms and numerical measurements that are critical for informed clinical decision-making.

A subset of 66 questions (ACCSAP-mod) was selected for contextual modification to evaluate LLM's flexibility in clinical reasoning(22). Specifically, each modification involved altering one or more key clinical parameters, such as decreasing the ascending aorta diameter from 35 mm to 55 mm or adjusting the left ventricular ejection fraction (LVEF) from 15% to 55%, so that the correct answer to the modified question would differ from the original question's correct answer. Modifications were made according to current cardiology guidelines to ensure each revised question had a legitimate new best answer. **Table 1** provided a few examples of how the questions were modified (not the questions used in our test).

able 1. Examples of Question Modification

riginal Question*	**Modified Question**	**Change in the answer**
63-year-old man with a edical history of coronary artery sease, for which he underwent ercutaneous coronary tervention 3 years earlier. A cent echocardiogram showed ft ventricular ejection fraction VEF) 26%, unchanged from 6 onths prior. He is optimized on uideline-directed medical erapy (GDMT). n electrocardiogram shows ormal sinus rhythm and left undle branch block (LBBB) with	A 63-year-old man with a medical history of coronary artery disease, for which he underwent percutaneous coronary intervention 3 years earlier. A recent echocardiogram showed left ventricular ejection fraction (LVEF) 26%, unchanged from 6 months prior. He is optimized on guideline-directed medical therapy (GDMT). An electrocardiogram shows normal sinus rhythm and first-degree AV block. What's the next step?	**Original**: Implant a biventricular implantable cardioverter-defibrillator. **Modified**: Implant a single-chamber implantable cardioverter-defibrillator.

able 1. Examples of Question Modification		
RS width 160 msec. What's the ext step?		
58-year-old man with a history bicuspid aortic valve (BAV). n echocardiogram 1 year earlier owed normal left ventricular V) function, and a BAV with a ean gradient across the aortic lve (AoV) was 16 mm Hg. The cending aortic diameter was 4.6 n. Repeated echocardiogram day shows similar LV function, d the mean gradient across the oV is 18 mm Hg. The ascending rtic diameter is 5 cm. hat's the best next step?	A 58-year-old man with a history of bicuspid aortic valve (BAV). An echocardiogram 1 year earlier showed normal left ventricular (LV) function, and a BAV with a mean gradient across the aortic valve (AoV) was 16 mm Hg. The ascending aortic diameter was 4.6 cm. Repeated echocardiogram today shows similar LV function, and the mean gradient across the AoV is 18 mm Hg. The ascending aortic diameter is 4.6 cm. What's the best next step?	**Original**: Refer to surgical intervention **Modified**: Repeat echocardiogram in 1 year.
65-year-old woman with onischemic dilated rdiomyopathy (left ventricular ection fraction [EF] 35%) is ere for follow-up. Her edications are carvedilol 6.25 g twice daily and losartan 50 g daily. ysical examination showed rmal blood pressure 137/84 and art rate 60. Laboratory results owed normal creatinine and ectrolytes. What's the next best ep in managing heart failure?	A 65-year-old woman with nonischemic dilated cardiomyopathy (left ventricular ejection fraction [EF] 35%) is here for follow-up. Her medications are carvedilol 25 mg twice daily and losartan 50 mg daily. Physical examination showed normal blood pressure 137/84 and heart rate 35. Laboratory results showed normal creatinine and electrolytes. What's the next best step in managing heart failure?	**Original**: Substitute sacubitril/valsartan for losart **Modified**: Reduce carvedilol.
68-year-old man presents to the inic for a postoperative assessment. He had a 26 mm oprosthetic mitral valve (MV) placement for severe mitral gurgitation 6 months earlier. hat's the best medication gimen for the patient?	A 68-year-old man presents to the clinic for a postoperative reassessment. He had a 26 mm bioprosthetic mitral valve (MV) replacement for severe mitral regurgitation 2 months earlier. What's the best medication regimen for the patient?	**Original**: Aspirin 81 mg daily. **Modified**: Warfarin with target INR 2.5.

able 1. Examples of Question Modification		
72-year-old man with a history aortic stenosis (AS), ypertension, and hyperlipidemia. is medications are amlodipine) mg daily and atorvastatin 40 g daily. ital signs are within normal nits. Transthoracic chocardiogram (TTE) reveals ft ventricular (LV) ejection action 65% and mild LV ypertrophy. The aortic valve is alcified with peak velocity 4.3 /sec, mean gradient 42 mm Hg, nd calculated aortic valve area AVA) 0.8 cm2. What is the best ext step?	A 72-year-old man with a history of aortic stenosis (AS), hypertension, and hyperlipidemia. His medications are amlodipine 10 mg daily and atorvastatin 40 mg daily. Vital signs are within normal limits. Transthoracic echocardiogram (TTE) reveals left ventricular (LV) ejection fraction 45% and mild LV hypertrophy. The aortic valve is calcified with peak velocity 3.2 m/sec, mean gradient 28 mm Hg, and calculated aortic valve area (AVA) 0.8 cm2. What is the best next step?	**Original**: Exercise treadmill testing. **Modified**: Dobutamine stress echocardiography.

Abbreviated content to demonstrate the clinical context of the original ACC questions.

In the multiple□choice (MC) format, these modifications are designed to admit one correct answer for each question. Similarly, in the open□ended format, the question stem was reworded to incorporate the new clinical values, so that the LLM needed to derive a different final answer and reasoning. This approach distinguished truly context□aware reasoning from mere pattern matching or recall of prior training data.

For analysis, we compared the model's performance on the original vs. modified versions of the same question in both multiple-choice and open-ended formats.

Model Selection

We selected GPT o1□preview(23) for this study for its advanced reasoning architecture; it goes beyond simple pattern recognition to systematically work through multi□step reasoning. Compared to earlier GPT models, GPT o1□preview uses an internal chain□of□thought approach, allowing it to "think aloud" over relevant details and integrate guidelines before arriving at a final conclusion. This design more closely mirrors the reasoning process of human experts, making o1□preview suitable for complex, context□dependent questions such as those

found in advanced board-style exams. The evaluation of GPT o1-preview is performed using Microsoft Azure services, ensuring that our test data is not leaked to OpenAI for training purposes.

Model Inference and Data Collection

For each question and each format, o1-preview was prompted with the question and asked to "answer a general practitioner's question" by generating a final answer to the question and providing a detailed reasoning to justify the answer. In the multiple-choice question format, GPT o1-preview selected one of the provided lettered answers and then produced a rationale in free-text format. In the open-ended format, it produced a written answer and its reasoning in a free text format. See **Supplementary Section A.1** and **Supplementary Section A.2** for the specific prompts used in generating answers and reasoning.

Evaluation

Human expert grading was first performed by YCW and CMT and reviewed by a board-certified cardiologist (CJC). Raters examined GPT-o1-preview's final answer and accompanying reasoning beside the reference solution and assigned two independent binary labels: answer correct/incorrect and reasoning correct/incorrect. An answer was deemed correct if it exactly matched, or represented an alternative yet clinically acceptable formulation of, the ground-truth response.

The current gold standard for evaluating open-ended and reasoning-based questions relies on human clinical experts. However, this approach is often limited by availability and is not easily scalable. Therefore, we also leveraged LLM-based evaluators to assess o1-preview's performance and compared their agreement with human evaluators(20). See **Supplementary Section B** for implementation details. For both metrics, we evaluated the agreement between GPT-4o automated grading and the expert cardiologist grading using Cohen's kappa (See **Supplementary Section C**).

Statistical Analysis

We assessed question-answering performance using answer accuracy, the number of questions with the correct answer generated, reasoning accuracy, and the number of questions with the correct reasoning generated. To assess the model's internal answer–reasoning concordance, we constructed a 2 × 2 contingency table for every question set, cross-classifying answer correctness against reasoning correctness. We defined the full-concordance rate (FCR) as the proportion of questions in which both answer and reasoning were correct (A+/R+) and the full-discordance rate (FDR) as the proportion in which both were incorrect (A–/R–).

Paired comparisons were performed with exact McNemar tests: (i) baseline multiple-choice vs. baseline open-ended; (ii) baseline vs. modified multiple-choice; and (iii) baseline vs. modified open-ended. All analyses were two-sided, with statistical significance set at $p < 0.05$, and were conducted in Python 3.11 using the *statsmodels* package (v0.14).

Results

Characteristics of Question Sets

A total of 185 original questions were evaluated. **Table 2** summarizes the distribution across 10 categories, with the largest subset addressing Heart Failure and Cardiomyopathies (24.3%, n=45). Each question was presented in both open-ended and multiple-choice formats for GPT o1-preview baseline evaluation. A subset of 66 questions was modified to evaluate contextual adaptation. **Table 3** describes the distribution of the modified questions by category.

Table 2: Distribution of ACCSAP Questions

Category	Number of Questions	Percentage (%)
Heart Failure & Cardiomyopathies	45	24.3
Coronary Artery Disease	26	14.1
Valvular Heart Disease	24	13.0
Arrhythmias	21	11.1
Vascular Disease	20	10.8
Congenital Heart Disease	14	7.6
Hypertension	13	7.0
Pericardial Disease	13	7.0
Miscellaneous Topics in Cardiovascular Medicine	7	3.8
Systemic Disorders Affecting Circulatory System	2	1.1
Total	185	100.0

Table 3: Distribution of Modified ACCSAP Question Subset

Category	Number of Questions	Percentage (%)
Heart Failure & Cardiomyopathies	15	22.7
Coronary Artery Disease	9	13.6
Valvular Heart Disease	9	13.6
Vascular Disease	9	13.6
Arrhythmias	7	10.6
Hypertension	6	9.1
Pericardial Disease	6	9.1
Congenital Heart Disease	4	6.1
Systemic Disorders Affecting Circulatory System	1	1.5
Total	66	100.0

Baseline Question Performance

For multiple□choice questions, GPT o1□preview selected the correct answer in 93.0% of cases (n□=□172), and expert cardiologists rated 92.4% of the generated reasonings as correct (n□=□171). 0.5% (n=1) of multiple-choice reasonings were labeled as “alternative, clinically acceptable” solutions. Notably, the Arrhythmias, Systemic Disorders Affecting the Circulatory System, and Valvular categories all achieved 100% accuracy for both answers and reasoning. Pericardial Disease category exhibited the lowest performance, with answer and reasoning accuracies of 76.9% (n□=□10) and 69.2% (n□=□9), respectively.

For open-ended questions, expert evaluators independently rated 80.0% (n=145) of the answers as correct and 80.5% (n=146) of the reasonings as correct. 7.6% (n=14) of open-ended answers and 7.6% (n=14) of open-ended reasonings were rated as “alternative, clinically acceptable” solutions. Questions in the Systemic Disorders Affecting Circulatory System category had the highest open-ended answer and reasoning accuracy of 100% (n=2). Consistent with the MC format, questions in the Pericardial Disease category were the most challenging, with an answer accuracy of 69.2% (n=9) and a reasoning accuracy of 61.5% (n=8).

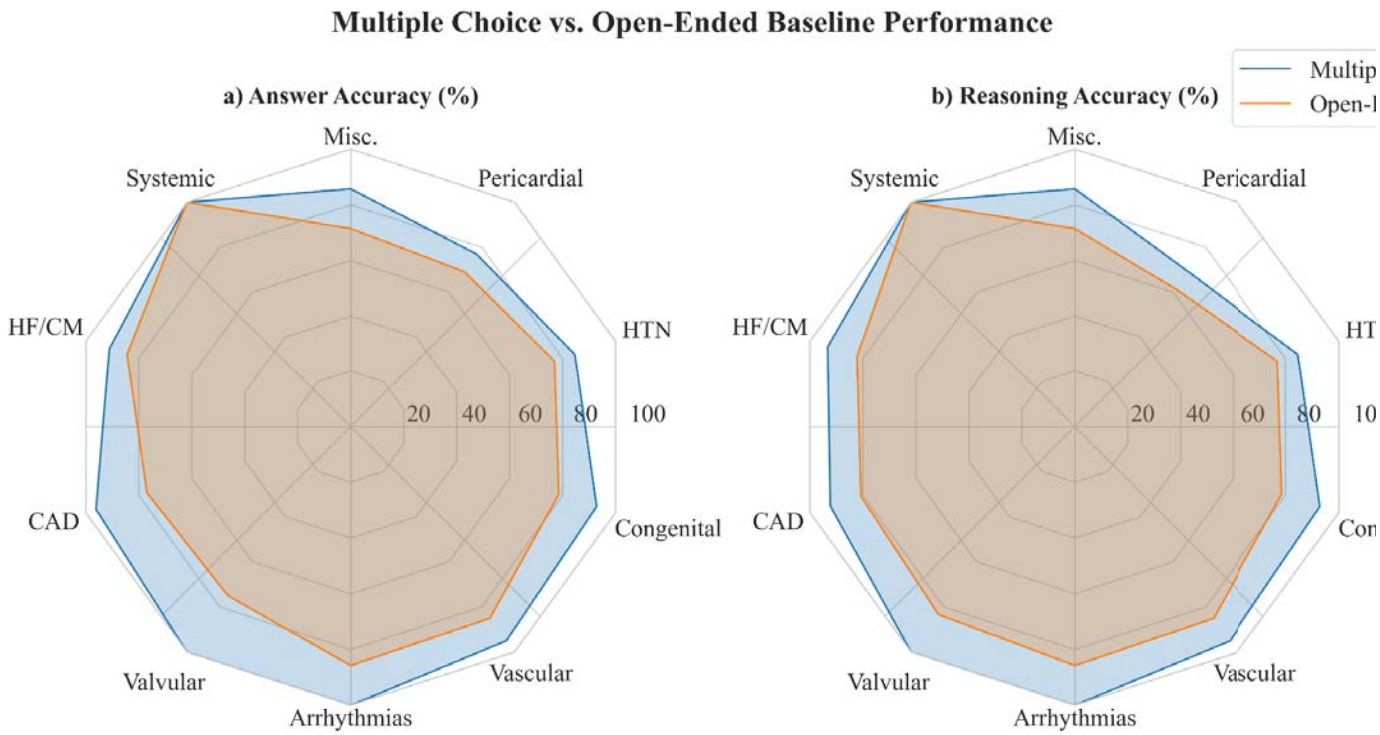

Figure 2. Comparison of performance between multiple□choice and open□ended formats across baseline cardiology questions. **2a** shows the percentage of correct answers, and **2b** shows the percentage of correct reasoning across each question category.

Overall, multiple□choice performance significantly exceeded that of the open□ended format, with answer accuracy of 93.0% versus 80.0% ($p \square < \square 0.001$) and reasoning accuracy of 92.4% versus 80.5% ($p \square < \square 0.001$). As shown in **Figure 2**, the magnitude of this gap between multiple-choice and open-ended performance varied across categories. Valvular Heart Disease had the largest difference, with answer accuracy dropping from 100.0% in multiple-choice to 75.0% in open-ended format ($p=0.03$).

<u>Modified Question Performance</u>

Baseline MC performance on the 66□question subset was strong, with overall answer and reasoning accuracies of 93.9%. After the MC questions were modified, GPT□o1□preview's answer accuracy decreased to 66.7% ($p \square < \square 0.001$), and reasoning accuracy dropped to 71.2% ($p \square = \square 0.0015$), as shown in **Figure□3**. Note, no multiple-choice reasonings were labeled as alternatively acceptable. Although no single subcategory's difference reached significance, the aggregate differences in both answer correctness and reasoning correctness were statistically significant, suggesting that GPT□o1□preview was consistently less accurate when the multiple□choice questions contained modified clinical parameters.

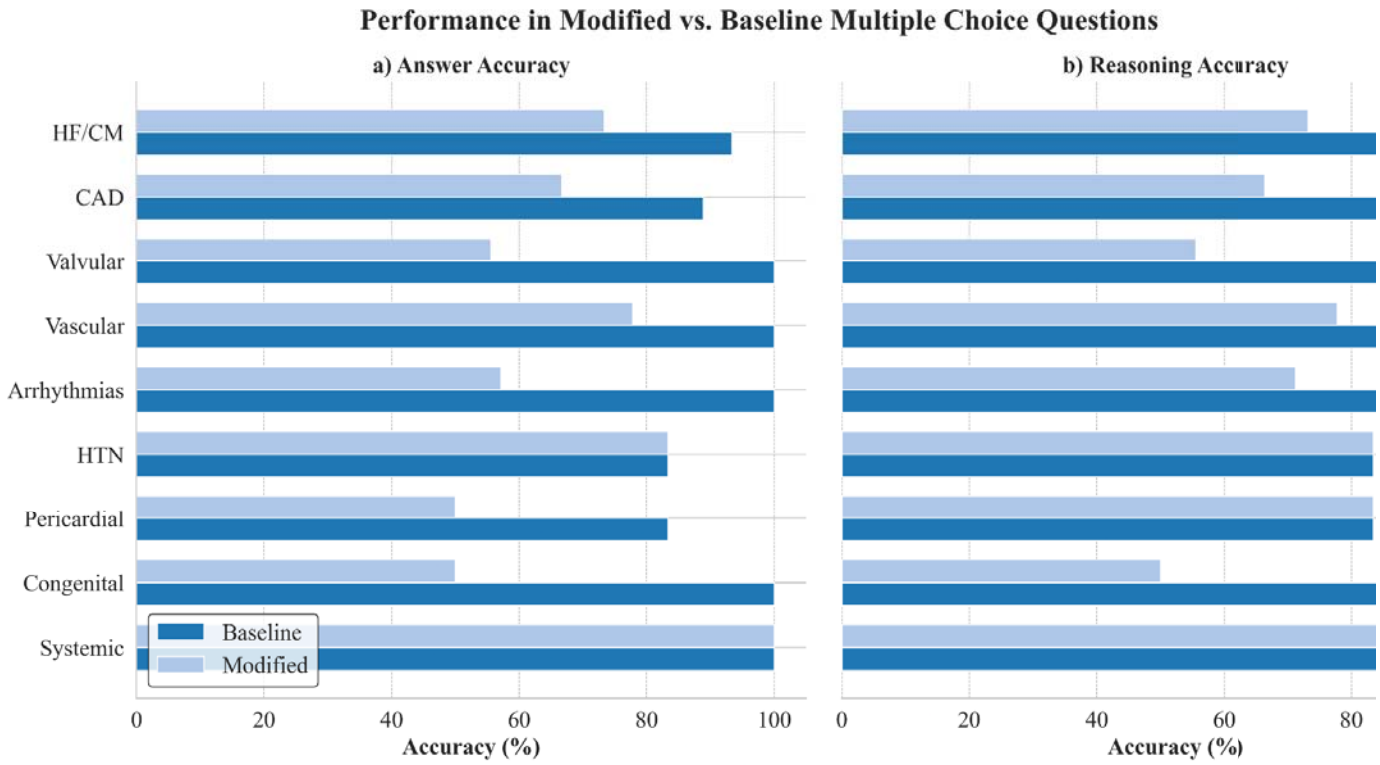

Figure 3. Comparison of performance between baseline and modified **multiple□choice** questions for each cardiology category. **3a** illustrates the percentage of correct answers, and **3b** shows the percentage of correct reasoning.

<u>LLM's Answer-Reasoning Consistency</u>

Question Set	Counts				Rates (%)	
	A+/R+	A+/R–	A–/R+	A–/R–	FCR	FDR
A. Baseline MC (N=185)	170	2	1	12	91.9	6.5
B. Baseline OE (N=185)	146	2	3	34	78.9	18.4
C. Modified MC (N=66)	44	0	3	19	66.7	28.8
D. Modified OE (N=66)	48	3	0	15	72.7	22.7

Table 4. Joint answer-reasoning performance (A/R) across baseline and modified question sets. MC = multiple□choice; OE = open□ended; A+/R+ answer and reasoning both correct; A+/R– = answer correct, reasoning incorrect; A–/R+ = answer incorrect, reasoning correct; A–/R– = both incorrect. FCR = full-concordance rate (proportion of A+/R+); FDR = full-discordance rate (proportion of A–/R–). All ratings by expert cardiologists.

Table 4 summarizes GPT□o1□preview's joint answer–reasoning performance. In baseline multiple□choice questions, fully concordant outputs were observed in 170 questions (FCR = 91.9 %), whereas 6.5 % were fully discordant. When the identical items were presented in open□ended form, FCR fell to 78.9 % (n = 146) and FDR rose to 18.4 % ($p < 0.001$), indicating lower answer–reasoning consistency once MC answer cues were removed.

Among the 66 questions available in both baseline and modified versions, baseline multiple□choice performance remained highly concordant (FCR = 93.9 %) but declined to 66.7 % after clinical modifications ($p < 0.001$). Baseline vs. modified open□ended items showed no significant difference in concordance (77.3 % vs. 72.7 %; $p = 1.0$). These findings indicate that GPT□o1□preview's performance is more vulnerable to context shifts in an MC framework than in an open□ended format.

Discussion

The main contributions of this work include: 1) Introducing an improved evaluation framework that goes beyond conventional MC formats by incorporating open-ended questions and revised clinical contexts, enabling a more comprehensive assessment of LLMs' clinical reasoning and answering abilities in realistic cardiology practice scenarios, and 2) Our results revealed a significant drop in LLM performance when evaluated with this framework, highlighting the limitations of current models in handling complex, clinically relevant reasoning tasks in a specialized medical specialty. Our findings challenge the conventional approach of evaluating LLMs with MC questions that may have been used in training and suggest the need for more rigorous LLM evaluation methods in the medical field that critically assess real-world clinical reasoning.

Cardiology is a specialized field with rapidly evolving evidence. Clinical decisions often depend on precise assessments like echocardiography, requiring thorough knowledge of both the evidence and clinical context for optimal outcomes(24–26). Previous studies in this field have predominantly assessed LLM accuracy using MC questions without systematically examining the quality or consistency of the corresponding reasoning processes (12,27–30). Consistent with prior studies using MC questions (12,27–30), GPT o1-preview demonstrated high accuracy on ACCSAP-mc questions (93.0%), superseding GPT-4's 71.0% on a different set of MC format ACCSAP questions (12). However, when evaluation extended beyond conventional MC format questions, o1-preview did not demonstrate the same level of capability across question formats and contextual adaptations (**Figures 2 and 3**). O1-preview's performance significantly deteriorated when questions were reformulated into open-ended formats (ACCSAP-open, 80.0%) or even more when minimal but critical clinical context modifications were introduced (ACCSAP-mod, 66.7%). This aligns with previous findings that LLMs show reduced

performance when multiple-choice questions undergo alterations in option ordering or response format (31).

Although the o1-preview model is claimed to possess generalized reasoning capabilities(6,32), our findings indicate these models cannot fully replicate expert clinical decision-making in cardiology. In open-ended scenarios or contextually adapted questions, o1-preview often failed to apply logical reasoning or adapt reasoning to current clinical guidelines, similar to the inflexible reasoning failure mode described in recent studies (22,33,34). We observed the most significant performance drop in the VHD subcategory (from 100.0% to 75.0%), in which clinical decisions can be distinct with semantically minor changes in a single echocardiography measurement (e.g., aortic valve area 0.8 cm^2 vs. 2.5 cm^2). This finding suggests a limitation in detecting subtle but clinically significant contextual changes.

In some ACCSAP-mod questions, o1-preview provided reasoning corresponding to the original unmodified question despite changes in critical clinical context, implying memorization of previously encountered question-answer pairs during training rather than genuine clinical understanding. When further analyzing o1-preview's answer–reasoning concordance, fully concordant outputs dropped from 93.9% to 66.7 % after minor clinical modifications in a multiple□choice setting, suggesting the model's dependence on fixed answer cues rather than flexible, context□sensitive reasoning. O1's performance was similarly low on modified MC and OE questions, indicating that the model struggled in handling "new" information in the modified clinical context, regardless of the question format. The above observations suggest that o1's performance, to some extent, still relies on pattern recognition from seen examples instead of comprehensive understanding and reasoning based on clinical context. Prior research in human medical education also challenges the assumption that knowledge memorization equates to clinical understanding(15–17,35). Our results, particularly the significant performance drop between baseline and modified questions, further highlight the limitations of using existing questions in assessing LLMs' clinical capabilities. Thus, caution is warranted when interpreting and promoting results from assessments of LLM performance based on existing question sets.

As LLMs advance into complex reasoning, their clinical evaluation frameworks must also evolve. Human expert evaluation is subject to availability constraints, potential subjectivity, and limited scalability. To address these challenges, we explored the use of automated evaluation by assessing GPT-4o's capability to judge the reasoning generated by o1-preview. While our automated evaluation using GPT-4o showed substantial agreement with expert cardiologists in assessing MC reasoning (Cohen's kappa = 0.78), agreement was considerably lower for open-ended questions (Cohen's kappa = 0.37 for answer correctness for baseline questions, 0.36 for modified questions). Therefore, despite being more resource-intensive, employing human experts remains the gold standard for evaluating models' performance in open-ended answering and reasoning tasks in the medical field(36,37).

Several limitations should be considered when interpreting our results. Although we did not include the entire ACCSAP dataset, our benchmark was randomly selected to ensure balanced representation of the original dataset while maintaining feasibility for evaluation. Excluding questions with images didn't allow assessment of medical image interpretation; however, this design eliminates the confounding factor and is consistent with prior studies. Not all original questions were suitable for context modification due to their content. As a result, the revised questions did not cover all categories. Additionally, transitioning questions from MC to an open-ended format may introduce evaluation biases with subjective judgments. Only o1-preview was evaluated as the representative state-of-the-art reasoning model, knowing its superior performance compared to other recent models.

Conclusion

Although LLMs achieve high accuracy on MC cardiology board questions, our study demonstrates that using established MC question benchmarks can substantially overestimate their performance, and lack assessment of their true clinical reasoning capabilities, particularly when adapting to diverse clinical contexts. Therefore, future assessment benchmarks for LLMs in cardiology should incorporate open-ended, clinically adaptive frameworks to more accurately reflect real-world clinical decision-making.

Data Availability

The ACCSAP questions are behind a paywall but can be accessed with payment. To prevent unauthorized distribution of the revised question set, the open-ended and revised content questions are not made publicly available. However, they can be obtained upon reasonable request from the corresponding author.

Abbreviations

MC: multiple choice
LLM: large language model
HTN: hypertension
CAD: coronary artery disease
HF/CM: heart failure & cardiomyopathies

Supplementary Material

Section A

A.1 Multiple Choice Question Answering Prompt (GPT o1-preview)

You are a cardiologist. Your task is to select the most correct answer to a general practitioner's multiple choice question and provide the most specific and concrete reasoning. Your answer should be one of the provided multiple choice answers.
Key Instructions:

1. Answer the question with only one best multiple choice option based on the scenario provided. Avoid choosing multiple recommendations or answers, even if more than one seems relevant.
2. Provide an explanation of reasoning for your answer under the section **Reasoning:** to justify your choice.
3. Provide your final answer under the section **Answer:**. This section should only contain the letter of the multiple choice answer you selected!

General Tips:
Ensure your answer is a single letter (A, B, C, D, or E), corresponding to the multiple choice option that you choose. Be specific and detailed in your reasoning. If multiple options seem equally valid, choose the one supported by the strongest evidence or most aligned with clinical guidelines. Ensure your reasoning supports why this option is the single best choice and how it addresses the clinical situation effectively.

A.2 Open-Ended Question Answering Prompt (GPT o1-preview)

You are a cardiologist. Your task is to provide the most specific and concrete reasoning and answer to a general practitioner's question. Be as specific as possible in your reasoning and answer to the question by providing exact treatment medication, treatment plan, or any other detailed information that is asked about in the question.
Key Instructions:

1. Answer the question with only one best, most appropriate, and concrete option based on the scenario provided. Avoid giving multiple recommendations or answers, even if more than one seems relevant.
2. Provide reasoning and an explanation for your answer under the section **Reasoning:** to justify your choice.
3. Provide your final answer under the section **Answer:**. This section should contain a single, concise, and definitive answer.

General Tips:
Be specific and detailed in your reasoning. Tailor your response to the type of question asked, whether it is about a recommended test, treatment plan, medication, or any other action. If multiple options seem equally valid, choose the one supported by the strongest evidence or most aligned with clinical guidelines. Ensure your reasoning supports why this option is the single best choice and how it addresses the clinical situation effectively.

Section B

To create an automated evaluation pipeline, we used two instances of GPT-4o as evaluators – one for answer evaluation and one for reasoning evaluation. For the open-ended experiments, the answer evaluation agent was provided with the question, the predicted answer, and each answer option in the original multiple-choice question individually and was asked to determine if the predicted answer is entailed by the answer option. If and only if the predicted answer was entailed by the ground truth answer, it was marked correct. For the multiple-choice experiments, the answer correctness of the selected letter (A, B, C, etc.) was determined by string matching to the ground-truth letter. For both open-ended and multiple-choice experiments, the reasoning evaluation agent was provided with the question, the predicted reasoning, and the ground truth reasoning and asked to determine if the predicted reasoning was correct or not. The final output of GPT-4o was a binary "correct/incorrect" for each evaluation. See **Supplementary Section B.1** and **Supplementary Section B.2** for specific prompts used for the evaluation.

B.1 Open-ended Answer Evaluation Prompt (GPT-4o)

> You are a cardiologist. You will be given a predicted answer and a sample answer to a cardiolog question. Your role is to determine if the predicted answer is entailed by the sample answer and why

B.2 Multiple Choice & Open-ended Reasoning Evaluation Prompt (GPT-4o)

> You are a cardiologist. You will be provided a cardiology question, the correct answer, a predicted reasoning, and a gold standard ground truth reasoning. Your role is to determine if the predicted reasoning is correct or not based on the gold standard reasoning and provide a justification as to why
> **Key Instructions:**
>
> 1. Determine if the predicted reasoning is correct based on the gold standard reasoning.
> 2. Provide a justification for your decision. This should be a detailed explanation of why the predicted reasoning is correct or incorrect based on the gold standard reasoning.
>
> **General Tips:**
> Ensure the "Is Correct" field is only "True" or "False." Be specific and detailed in your justification. Ensure your justification supports why the predicted reasoning is correct or incorrect based on the gold standard reasoning.

Section C: Agreement between human-expert and GPT-4o evaluators

Across the baseline and modified question experiments, Cohen's kappa for answer correctness in the baseline open□ended set was 0.37 and 0.54 for reasoning correctness. For baseline multiple□choice questions, the kappa for reasoning correctness was higher at 0.78, representing

substantial agreement with the human expert. In the modified open-ended experiment, however, the kappa for answer correctness was only 0.36, suggesting that GPT-4o did not align with expert cardiologists when grading open-ended questions.

Acknowledgement

We acknowledge Moaz Kamel for his contribution to the initial phase of ACCSAP question data collection.

Financial Disclosure

Chieh-Ju Chao, MD, is supported by research funding from the CV Prospective award from the Mayo Clinic Department of Cardiovascular Medicine and the AI/ML Enablement award from the Center for Digital Health at the Mayo Clinic. The Stanford team is supported by the Verdant Foundation and Microsoft Azure Cloud Credits.

References

1. Nori H., King N., McKinney SM., Carignan D., Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. ArXiv 2023. Doi: 10.48550/arxiv.2303.13375.
2. 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):210–20. Doi: 10.7326/m23-2772.
3. OpenAI., :, Hurst A., et al. GPT-4o System Card. ArXiv 2024. Doi: 10.48550/arxiv.2410.21276.
4. Katz U., Cohen E., Shachar E., et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. NEJM AI 2024;1(5). Doi: 10.1056/aidbp2300192.
5. Schubert MC., Wick W., Venkataramani V. Performance of Large Language Models on a Neurology Board–Style Examination. JAMA Netw Open 2023;6(12):e2346721. Doi: 10.1001/jamanetworkopen.2023.46721.
6. Brodeur PG., Buckley TA., Kanjee Z., et al. Superhuman performance of a large language model on the reasoning tasks of a physician. ArXiv 2024. Doi: 10.48550/arxiv.2412.10849.
7. Saab K., Tu T., Weng W-H., et al. Capabilities of Gemini Models in Medicine. ArXiv 2024. Doi: 10.48550/arxiv.2404.18416.
8. Singhal K., Tu T., Gottweis J., et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31(3):943–50. Doi: 10.1038/s41591-024-03423-7.
9. Zhong T., Liu Z., Pan Y., et al. Evaluation of OpenAI o1: Opportunities and Challenges of AGI. ArXiv 2024. Doi: 10.48550/arxiv.2409.18486.
10. Singhal K., Azizi S., Tu T., et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. Doi: 10.1038/s41586-023-06291-2.

11. Nori H., Lee YT., Zhang S., et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. ArXiv 2023. Doi: 10.48550/arxiv.2311.16452.
12. Shahid A., Shetty NS., Patel N., Gaonkar M., Arora G., Arora P. Evaluating Cardiology Certification Using the ACCSAP Question Bank: Large Language Models vs Physicians. Mayo Clin Proc 2025;100(1):160–3. Doi: 10.1016/j.mayocp.2024.07.030.
13. Lee PC., Sharma SK., Motaganahalli S., Huang A. Evaluating the Clinical Decision-Making Ability of Large Language Models Using MKSAP-19 Cardiology Questions. JACC: Adv 2023;2(9):100658. Doi: 10.1016/j.jacadv.2023.100658.
14. Gendler M., Nadkarni GN., Sudri K., et al. Large Language Models in Cardiology: A Systematic Review. MedRxiv 2024:2024.09.01.24312887. Doi: 10.1101/2024.09.01.24312887.
15. Mbakwe AB., Lourentzou I., Celi LA., Mechanic OJ., Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Heal 2023;2(2):e0000205. Doi: 10.1371/journal.pdig.0000205.
16. Hift RJ. Should essays and other “open-ended”-type questions retain a place in written summative assessment in clinical medicine? BMC Méd Educ 2014;14(1):249. Doi: 10.1186/s12909-014-0249-2.
17. Tangianu F., Mazzone A., Berti F., et al. Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? Ital J Med 2018;12(2):88–96. Doi: 10.4081/itjm.2018.980.
18. Wijk EV van., Janse RJ., Ruijter BN., et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. PLOS ONE 2023;18(7):e0288558. Doi: 10.1371/journal.pone.0288558.
19. Parekh P., Bahadoor V. The Utility of Multiple-Choice Assessment in Current Medical Education: A Critical Review. Cureus 2024;16(5):e59778. Doi: 10.7759/cureus.59778.
20. Hosseini P., Sin JM., Ren B., et al. A Benchmark for Long-Form Medical Question Answering. ArXiv 2024. Doi: 10.48550/arxiv.2411.09834.
21. Brush JE., Sherbino J., Norman GR. Diagnostic reasoning in cardiovascular medicine. BMJ 2022;376:e064389. Doi: 10.1136/bmj-2021-064389.
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. ArXiv 2025. Doi: 10.48550/arxiv.2502.04381.
23. OpenAI., :, Jaech A., et al. OpenAI o1 System Card. ArXiv 2024. Doi: 10.48550/arxiv.2412.16720.
24. Otto CM., Nishimura RA., Bonow RO., et al. 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2020;(J Am Coll Cardiol 63 2014). Doi: 10.1016/j.jacc.2020.11.018.
25. Stout KK., Daniels CJ., Aboulhosn JA., et al. 2018 AHA/ACC Guideline for the Management of Adults With Congenital Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll

Cardiol 2019;73(12):e81 e192. Doi: 10.1016/j.jacc.2018.08.1029.
26. Ommen SR., Mital S., Burke MA., et al. 2020 AHA/ACC Guideline for the Diagnosis and Treatment of Patients With Hypertrophic Cardiomyopathy. Circulation 2020;142(25):e558–631. Doi: 10.1161/cir.0000000000000937.
27. Builoff V., Shanbhag A., Miller RJH., et al. Evaluating AI proficiency in nuclear cardiology: Large language models take on the board preparation exam. J Nucl Cardiol 2025;45:102089. Doi: 10.1016/j.nuclcard.2024.102089.
28. Huwiler J., Oechslin L., Biaggi P., Tanner FC., Wyss CA. Experimental assessment of the performance of artificial intelligence in solving multiple-choice board exams in cardiology. Swiss Méd Wkly 2024;154(10):3547. Doi: 10.57187/s.3547.
29. Milutinovic S., Petrovic M., Begosh-Mayne D., et al. Evaluating Performance of ChatGPT on MKSAP Cardiology Board Review Questions. Int J Cardiol 2024;417:132576. Doi: 10.1016/j.ijcard.2024.132576.
30. Altamimi I., Alhumimidi A., Alshehri S., et al. The scientific knowledge of three large language models in cardiology: multiple-choice questions examination-based performance. Ann Med Surg 2024;86(6):3261–6. Doi: 10.1097/ms9.0000000000002120.
31. Gupta V., Pantoja D., Ross C., Williams A., Ung M. Changing Answer Order Can Decrease MMLU Accuracy. ArXiv 2024. Doi: 10.48550/arxiv.2406.19470.
32. Wu S., Peng Z., Du X., et al. A Comparative Study on Reasoning Patterns of OpenAI's o1 Model. ArXiv 2024. Doi: 10.48550/arxiv.2410.13639.
33. Hager P., Jungmann F., Holland R., et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613–22. Doi: 10.1038/s41591-024-03097-1.
34. Reese JT., Danis D., Caufield JH., et al. On the limitations of large language models in clinical diagnosis. MedRxiv 2024:2023.07.13.23292613. Doi: 10.1101/2023.07.13.23292613.
35. Groves M., O'Rourke P., Alexander H. The clinical reasoning characteristics of diagnostic experts. Méd Teach 2003;25(3):308–13. Doi: 10.1080/0142159031000100427.
36. Bavaresco A., Bernardi R., Bertolazzi L., et al. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. ArXiv 2024. Doi: 10.48550/arxiv.2406.18403.
37. Szymanski A., Ziems N., Eicher-Miller HA., Li TJ-J., Jiang M., Metoyer RA. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. Proc 30th Int Conf Intell User Interfaces 2025:952–66. Doi: 10.1145/3708359.3712091.